Convex Optimization Methods for Graphs and Statistical Modeling

by

Venkat Chandrasekaran

B.A. in Mathematics, B.S. in Electrical and Computer Engineering
Rice University (2005)

S.M. in Electrical Engineering and Computer Science
Massachusetts Institute of Technology (2007)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June 2011

Signature of Author: _____
Department of Electrical Engineering and Computer Science
April 29, 2011

Certified by: _____
Alan S. Willsky
Edwin Sibley Webster Professor of Electrical Engineering and Computer Science

Certified by: _____
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: _____
Jeslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Convex Optimization Methods for Graphs and Statistical Modeling

by Venkat Chandrasekaran

Submitted to the Department of Electrical Engineering
and Computer Science on April 29, 2011
in Partial Fulfillment of the Requirements for the Degree
of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

An outstanding challenge in many problems throughout science and engineering is to succinctly characterize the relationships among a large number of interacting entities. Models based on graphs form one major thrust in this thesis, as graphs often provide a concise representation of the interactions among a large set of variables. A second major emphasis of this thesis are classes of structured models that satisfy certain algebraic constraints. The common theme underlying these approaches is the development of computational methods based on convex optimization, which are in turn useful in a broad array of problems in signal processing and machine learning. The specific contributions are as follows:

- We propose a convex optimization method for decomposing the sum of a sparse matrix and a low-rank matrix into the individual components. Based on new rank-sparsity uncertainty principles, we give conditions under which the convex program exactly recovers the underlying components.

- Building on the previous point, we describe a convex optimization approach to latent variable Gaussian graphical model selection. We provide theoretical guarantees of the statistical consistency of this convex program in the high-dimensional scaling regime in which the number of latent/observed variables grows with the number of samples of the observed variables. The algebraic varieties of sparse and low-rank matrices play a prominent role in this analysis.

- We present a general convex optimization formulation for linear inverse problems, in which we have limited measurements in the form of linear functionals of a signal or model of interest. When these underlying models have algebraic structure, the

resulting convex programs can be solved exactly or approximately via semidefinite programming. We provide sharp estimates (based on computing certain Gaussian statistics related to the underlying model geometry) of the number of generic linear measurements required for exact and robust recovery in a variety of settings.

- We present convex graph invariants, which are invariants of a graph that are convex functions of the underlying adjacency matrix. Graph invariants characterize structural properties of a graph that do not depend on the labeling of the nodes; convex graph invariants constitute an important subclass, and they provide a systematic and unified computational framework based on convex optimization for solving a number of interesting graph problems.

We emphasize a unified view of the underlying convex geometry common to these different frameworks. We describe applications of both these methods to problems in financial modeling and network analysis, and conclude with a discussion of directions for future research.

Thesis Supervisors: Alan S. Willsky and Pablo A. Parrilo
Title: Professors of Electrical Engineering and Computer Science

Acknowledgments

I have been truly lucky to have two great advisors in Alan Willsky and Pablo Parrilo. I am extremely grateful to them for giving me complete freedom in pursuing my interests while also providing research guidance, for sharing with me their tremendous intellectual enthusiasm and curiosity, for reading our paper drafts promptly, and for the countless conversations about research, jobs and life. Thanks, Pablo and Alan, for everything.

I am very grateful to Devavrat Shah for his many words of encouragement and advice, and for serving on my thesis committee. My thanks also go to Sanjoy Mitter for his support at various stages during my time at MIT.

I enjoyed working with many wonderful researchers these last few years. This thesis is the result of some of these collaborations. In addition to Pablo, Alan, and Devavrat, I would like to thank Jin Choi, Ben Recht, Jason Johnson, Parikshit Shah, Dmitry Malioutov, James Saunderson, Ying Liu, Anima Anandkumar, Jinwoo Shin, David Gamarnik, Misha Chertkov, Sujay Sanghavi, Nathan Srebro, and Prahladh Harsha.

I am grateful to Rachel Cohen, Debbie Deng, Jennifer Donovan, Janet Fischer, Lisa Gaumond, and Brian Jones for their assistance in ensuring that my progress through MIT was smooth.

I could not have asked for a better set of people with whom to interact than those in LIDS: Jason Johnson, Jason Williams, Dmitry Malioutov, Pat Kreidl, Ayres Fan, Emily Fox, Jin Choi, Kush Varshney, Ying Liu, James Saunderson, Matt Johnson, Vincent Tan, Anima Anandkumar, Peter Jones, Lav Varshney, Parikshit Shah, Amir Ali Ahmadi, Noah Stein, Ozan Candogan, Mesrob Ohannessian, Srikanth Jagabathula, Jinwoo Shin, Hari Narayanan, Shashi Borade, and many others. Outside of LIDS, I'm very lucky to be able to count Shrini Kudekar, Rene Pfitzner, Urs Niesen, Evgeny Logvinov, Ryan Giles, and Matt Barnett as my friends.

Finally I would like to thank my family for their love, support, and encouragement.

Funding: This research was supported in part by the following grants – MURI AFOSR grant FA9550-06-1-0324, MURI AFOSR grant FA9550-06-1-0303, NSF FRG 0757207, AFOSR grant FA9550-08-1-0180, and MURI ARO grant W911NF-06-1-0076.

Contents

Chapter 1

Introduction

An outstanding challenge in many applications throughout science and engineering is to succinctly characterize the relationships among a large number of interacting entities. In a statistical model selection setting we wish to learn a "simple" statistical model to approximate the behavior observed in a collection of random variables. Modern data analysis tasks in geophysics, economics, and image processing often involve learning statistical models over collections of random variables that may number in the hundreds of thousands, or even a few million. In a computational biology setting a typical question involving gene regulatory networks is to discover the interaction patterns among a collection of genes in order to better understand how a gene influences or is influenced by other genes. Similar problems also arise in the analysis of biological, social, or chemical reaction networks in which one seeks to better understand a complicated network by decomposing it into simpler networks. Models based on graphs offer a fruitful framework to solve such problems, as graphs often provide a concise representation of the interactions among a large set of variables.

In this thesis we explore a set of research directions at the intersection of *graphs* and *statistics*. An important instance of a framework that lies in this intersection is that of *graphical models*, in which a statistical model is defined with respect to a graph. Another example is one in which we have *statistical models over the space of graphs*, so that a graph itself is viewed as a sample drawn from a probability distribution defined over some set of graphs. Natural questions that arise in standard statistical settings such as deconvolution can then be posed in a deterministic framework in this graph setting as well.

A common theme underlying our investigations is the development of tractable computational tools based on *convex optimization*, which possess numerous favorable properties. Due to their powerful modeling capabilities, convex optimization methods

can provide tractable formulations for solving difficult combinatorial problems exactly or approximately. Further convex programs may often be solved effectively using general-purpose off-the-shelf software. Finally one can also give conditions for the success of these convex relaxations based on standard optimality results from convex analysis.

■ 1.1 Main Contributions

In this section we outline the main contributions of this thesis. Details about related previous work are given in the relevant chapters. The research and results of Chapters 3, 4, 5, and 6 correspond to the papers [37], [33], [36], and [34] respectively.

Rank-Sparsity Uncertainty Principles and Matrix Decomposition

Suppose we are given a matrix that is formed by adding an unknown sparse matrix to an unknown low-rank matrix. The goal is to decompose the given matrix into its sparse and low-rank components. Such a problem is intractable to solve in general, and arises in a number of applications such as model selection in statistics, system identification in control, optical system decomposition, and matrix rigidity in computer science. Indeed sparse-plus-low-rank matrix decomposition is the main challenge in latent-variable Gaussian graphical model selection, which is discussed next (and in greater detail in Chapter 4). In Chapter 3, we propose a convex optimization formulation to splitting the specified matrix into its components, by minimizing a linear combination of the ℓ_1 norm and the nuclear norm (the sum of the singular values of a matrix) of the components. We develop a notion of *rank-sparsity incoherence*, expressed as an uncertainty principle between the sparsity pattern of a matrix and its row and column spaces, and use it to characterize both fundamental identifiability as well as (deterministic) sufficient conditions for exact recovery. The analysis is geometric in nature with the tangent spaces to the algebraic varieties of sparse and low-rank matrices playing a prominent role.

Latent Variable Gaussian Graphical Model Selection

Graphical models are widely used in many applications throughout machine learning, computational biology, statistical signal processing, and statistical physics as they offer a compact representation for the statistical structure among a large collection of random variables. Graphical models in which the underlying graph is sparse typically tend to be better suited for efficiently performing tasks such as inference and estimation.

In the setting of Gaussian graphical models where the random variables are jointly Gaussian, sparsity in the graph structure corresponds to sparsity in the inverse of the covariance matrix of the random variables, also called the concentration matrix. Thus Gaussian graphical model selection is the problem of learning a model described by a sparse concentration matrix to best approximate the observed statistics in a collection of random variables [119]. However a significant difficulty arises if we do not have sample observations of some of the relevant variables, because a whole set of *extra correlations* are induced among the observed variables due to marginalization over the unobserved, hidden variables. Is it possible to discover the number of hidden components, and to learn a statistical model over the entire collection of variables? If only we realized that much of the seemingly complicated correlation structure among the observed variables can be explained as the effect of marginalization over a few hidden variables, we would be able to learn a "simple" statistical model among the observed variables and a few additional hidden variables.

In the Gaussian setting this problem reduces to one of approximating a given matrix by the sum of a sparse matrix and a low-rank matrix: the low-rank matrix corresponds to the correlations induced by marginalization over latent variables (it is low-rank as the number of hidden variables is usually much smaller than the number of observed variables), and the sparse matrix corresponds to the conditional graphical model structure among the observed variables conditioned on the latent variables. From a statistical viewpoint this approach to modeling can be seen as a blend of dimensionality reduction (to identify latent variables) and graphical modeling (to capture remaining statistical structure not attributable to the latent variables). In Chapter 4, we propose a *tractable convex programming estimator* for latent variable Gaussian graphical model selection based on regularized maximum-likelihood; motivated by the results in Chapter 3 the regularizer uses the ℓ_1 norm for the sparse component, and the nuclear norm for the low-rank component. In addition to being computationally efficient to evaluate, this estimator enjoys favorable *statistical consistency* properties. Indeed we show that consistent model selection is possible under suitable identifiability conditions even if the number of observed/latent variables is on the same order as the number of samples of the observed variables. The rank-sparsity uncertainty principles of Chapter 3 described above are fundamental to our analysis. Previous approaches to latent variable graphical modeling using variants of the Expectation-Maximization (EM) algorithm do not share these favorable properties, as they optimize non-convex functions (hence converging

only to local optima) and have no high-dimensional consistency guarantees.

Convex Optimization for Inverse Problems

Many of the questions from the previous two sections can be viewed as instances of *inverse problems* in which we wish to recover simple and structured models given limited information. In Chapter 5 we study a general class of *linear inverse problems* in which the goal is to recover a model given a small number of linear measurements. Such problems are generally ill-posed as the number of measurements available is typically smaller than the dimension of the model. However in many practical applications of interest, models are often constrained structurally so that they only have a few degrees of freedom relative to their ambient dimension. Exploiting such structure is the key to making linear inverse problems well-posed. The class of simple models that we consider in Chapter 5 are those formed as the sum of a few atoms from some elementary atomic set; examples include well-studied cases such as sparse vectors (e.g., signal processing, statistics) and low-rank matrices (e.g., control, statistics), as well as several others such as sums of a few permutations matrices (e.g., ranked elections, multiobject tracking), low-rank tensors (e.g., vision, neuroscience), orthogonal matrices (e.g., machine learning), and atomic measures (e.g., system identification). We describe a general framework to convert such notions of simplicity into convex penalty functions, which give rise to convex optimization solutions to linear inverse problems. These convex programs can be solved via semidefinite programming under suitable conditions, and they significantly generalize previous approaches based on ℓ_1 norm and nuclear norm minimization for recovering sparse and low-rank models. Our results give general conditions and bounds on the number generic measurements under which exact or robust recovery of the underlying model is possible via convex optimization. Thus this work extends the catalog of simple models (beyond sparse vectors, i.e., compressed sensing, and low-rank matrices) that can be recovered from limited linear information via tractable convex programming.

Convex Graph Invariants

Investigating graphs from the viewpoint of statistics provides a very fruitful research agenda, as many questions from classical statistics can be posed in a deterministic setting in which data are represented as graphs. As an example suppose that we have a composite graph formed as the combination of two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ overlaid on

the same set of nodes. We are only given the composite graph without any additional information about the relative labeling of the nodes, which may reveal the structure of the individual components. Can we *deconvolve* the composite graph into the individual components? As discussed in Chapter 6 such a problem is of interest in network analysis in social and biological networks in which one seeks to decompose a complex network into simpler components to better understand the behavior of the composite network. Other problems motivated by statistics include *hypothesis testing* between families of graphs, and *generating/sampling* graphs with certain desired structural properties (see Chapter 6 for details).

An important goal towards solving these and many other graph problems is to characterize the underlying structural properties of a graph. *Graph invariants* play an important role in describing such abstract structural features, as they do not depend on the labeling of the nodes of the graph. Examples of commonly used graph invariants include the spectrum of a graph (i.e., eigenvalues of the adjacency matrix), or the degree sequence. In Chapter 6 we introduce and investigate *convex graph invariants*, which are graph invariants that are convex functions of the adjacency matrix of a graph. Examples of such functions of a graph include the maximum degree, the MAXCUT value (and its semidefinite relaxation), the second smallest eigenvalue of the Laplacian, and spectral invariants such as the sum of the k largest eigenvalues of the adjacency matrix. Convex graph invariants provide a systematic and unified computational framework based on convex optimization for solving a number of interesting graph problems such as those described above.

Chapter 2

Background

In this chapter we emphasize the main themes common to the rest of this thesis. Our exposition is brief as we only provide the basic relevant technical background, and we refer the reader to the texts [124] (on convex analysis) and [79] (on algebraic geometry) for more details. The individual chapters also give more background pertaining to the corresponding chapter.

2.1 Basics of Convex Analysis

A set $\mathcal{C} \subseteq \mathbb{R}^p$ is a *convex* set if for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and any scalar $\lambda \in [0, 1]$, we have that $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in \mathcal{C}$. A convex set $\mathcal{C}$ is also a *cone* if it is closed under positive linear combinations. Such convex cones are fundamental objects of study in convex analysis, and play an important role in all the main chapters of this thesis.

The *polar* $\mathcal{C}^*$ of a cone $\mathcal{C}$ is the cone

$$\mathcal{C}^* = \{\mathbf{x} \in \mathbb{R}^p \ : \ \langle \mathbf{x}, \mathbf{z} \rangle \leq 0 \ \forall \mathbf{z} \in \mathcal{C}\}.$$

Given a closed convex set $\mathcal{C} \in \mathbb{R}^p$ and some nonzero $\mathbf{x} \in \mathbb{R}^p$ we define the *tangent cone* at $\mathbf{x}$ with respect to $\mathcal{C}$ as

$$T_{\mathcal{C}}(\mathbf{x}) = \text{cone}\{\mathbf{z} - \mathbf{x} \ : \ \mathbf{z} \in \mathcal{C}\}. \tag{2.1}$$

Here cone$(\cdot)$ refers to the conic hull of a set obtained by taking nonnegative linear combinations of elements of the set. The cone $T_{\mathcal{C}}(\mathbf{x})$ is the set of *directions* to points in $\mathcal{C}$ from the point $\mathbf{x}$. The *normal cone* $N_{\mathcal{C}}(\mathbf{x})$ at $\mathbf{x}$ with respect to the convex set $\mathcal{C}$ is defined to be the polar cone of the tangent cone $T_{\mathcal{C}}(\mathbf{x})$, i.e., the normal cone consists of vectors that form an obtuse angle with every vector in the tangent cone $T_{\mathcal{C}}(\mathbf{x})$.

A real-valued function f defined on a convex set $\mathcal{C}$ is said to be a *convex* function

if for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and any scalar $\lambda \in [0, 1]$, we have that

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}).$$

Following standard notation in convex analysis, we denote the *subdifferential* of a convex function f at a point $\hat{\mathbf{x}}$ in its domain by $\partial f(\hat{\mathbf{x}})$. The subdifferential $\partial f(\hat{\mathbf{x}})$ consists of all $\mathbf{y}$ such that

$$f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \langle \mathbf{y}, \mathbf{x} - \hat{\mathbf{x}} \rangle, \quad \forall \mathbf{x}.$$

■ 2.2 Representation of Convex Sets

Convex programs denote those optimization problems in which we seek to minimize a convex function over a convex constraint set [24]. For example linear programming and semidefinite programming form two prominent subclasses in which linear functions are minimized over constraint sets given by affine spaces intersecting the nonnegative orthant (in linear programming) and the positive-semidefinite cone (in semidefinite programming) [11]. Roughly speaking convex programs are tractable to solve computationally if the convex objective function can be computed efficiently, and membership in the convex constraint sets can be certified efficiently. Hence, the tractable *representation* of convex sets is an important point that must be addressed in order to develop practically feasible computational solutions to convex optimization problems.

Any closed convex set has two dual representations. Specifically, an element $\mathbf{x}$ belonging to a convex set $\mathcal{C}$ is an *extreme point* if it cannot be expressed as the midpoint of the line segment between some two points in $\mathcal{C}$. With this definition the first representation of a convex set is as the convex hull of all its extreme points. With respect to this representation, certifying membership in a convex set means that we must *produce* a representation of a point as the convex combination of (a subset of) extreme points. A second representation of a convex set is as the intersection of (possibly infinitely many) halfspaces. Here certifying membership of a point in a convex set means that we need to *verify* that this point satisfies the constraints defining the convex set. Using the tools of convex duality one can transform between these two alternate representations of a convex set (see [124] for more details).

In this section we provide several examples of convex sets and their representations, with the objective of highlighting the main ideas that lead to tractable representations. In particular the concept of *lift-and-project* plays a central role in many examples of efficient representations of convex sets. The lift-and-project concept is simple – we wish

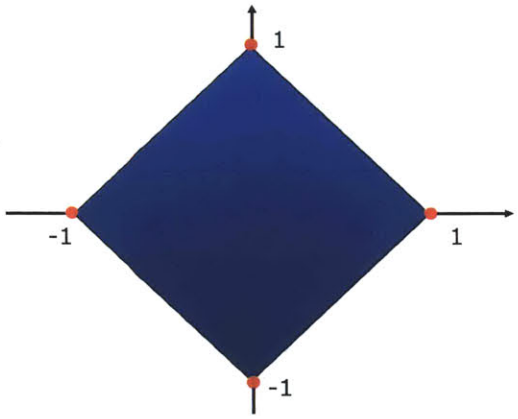

Figure 2.1. The cross-polytope in two dimensions.

to express a convex set $\mathcal{C} \in \mathbb{R}^p$ as the projection of a convex set $\mathcal{C}' \in \mathbb{R}^{p'}$ in some higher-dimensional space (i.e., $p' > p$). Such methods are useful if p' is not too much larger than p and if $\mathcal{C}'$ has an efficient representation in the higher-dimensional space $\mathbb{R}^{p'}$. Lift-and-project provides a very powerful representation tool, as will be seen in the examples to follow.

■ 2.2.1 Cross-polytope

The cross-polytope (see Figure 2.1) is the unit ball of the ℓ_1-norm:

$$B_{\ell_1}^p = \left\{ \mathbf{x} \in \mathbb{R}^p \mid \sum_i |\mathbf{x}_i| \leq 1 \right\}.$$

The ℓ_1-norm has been the focus of much attention recently due to its sparsity-inducing properties [29, 53, 54].

In a statistical model selection setting sparsity corresponds to models that consist of few nonzero parameters. Specialized to a linear regression or feature selection context, penalty functions based on the ℓ_1-norm lead to parameter vectors that are sparse, i.e., responses are expressed as the linear combination of a small number of features [135]. Specialized to a covariance selection context, ℓ_1-norm penalty functions lead to distributions defined by sparse covariance and concentration matrices [15, 16, 61]. Sparsity has also played a central role in signal processing as a variety of applications exploit the expression of signals as the sum of few elements from a dictionary, e.g.,

approximating natural images as the weighted sum of a few wavelet basis functions. The benefits of such *sparse approximations* are clear for tasks such as compression, but extend also to tasks such as signal denoising and classification.

How do we represent the p-dimensional cross-polytope $B_{\ell_1}^p$? While the cross-polytope has $2p$ vertices, a direct specification in terms of halfspaces involves 2^p inequalities:

$$B_{\ell_1}^p = \left\{ \mathbf{x} \in \mathbb{R}^p \mid \sum_i \mathbf{z}_i \mathbf{x}_i \leq 1, \ \forall \mathbf{z} \in \{-1, +1\}^p \right\}.$$

However we can obtain a tractable inequality representation by lifting to $\mathbb{R}^{2p}$ and then projecting onto the first p coordinates:

$$B_{\ell_1}^p = \left\{ \mathbf{x} \in \mathbb{R}^p \mid \exists \mathbf{z} \in \mathbb{R}^p \text{ s.t. } -\mathbf{z}_i \leq \mathbf{x}_i \leq \mathbf{z}_i \ \forall i, \ \sum_i \mathbf{z}_i \leq 1, \ \mathbf{z}_i \geq 0 \ \forall i \right\}.$$

Note that in $\mathbb{R}^{2p}$ with the additional variables $\mathbf{z}$, we have only $3p + 1$ inequalities.

Next suppose $\mathbf{x} \in B_{\ell_1}^p$ is a point on the boundary of the cross-polytope, i.e., $\|\mathbf{x}\|_{\ell_1} = 1$. Letting $\Omega \subseteq \{1, \ldots, p\}$ denote the indices at which $\mathbf{x}$ is nonzero, the normal cone at $\mathbf{x}$ with respect to $B_{\ell_1}^p$ is given as:

$$N_{B_{\ell_1}^p}(\mathbf{x}) = \{ \mathbf{z} \mid \mathbf{z}_i = t\text{sgn}(\mathbf{x}_i) \text{ for } i \in \Omega, |\mathbf{z}_i| \leq t \text{ for } i \in \Omega^c \text{ for some } t \geq 0 \}.$$

Here $\text{sgn}(\cdot)$ is the sign function.

■ 2.2.2 Nuclear-norm ball

The nuclear norm of a matrix (see Figure 2.2 for the unit ball) is the sum of its singular values:

$$\|X\|_* = \sum_i \sigma_i(X).$$

Analogous to the case of the ℓ_1-norm, the nuclear norm has received much attention recently because it induces low-rank structure in matrices in a number of settings [30, 121].

In a statistics context low-rank covariance matrices are used in factor analysis, and they represent the property that the corresponding random variables lie on or near a low-dimensional subspace. In a control setting low-rank system matrices correspond to systems with a low-dimensional state space, i.e., systems with small model order. In optical system modeling low-rank matrices represent so-called *coherent* systems, which correspond to low-pass optical filters.

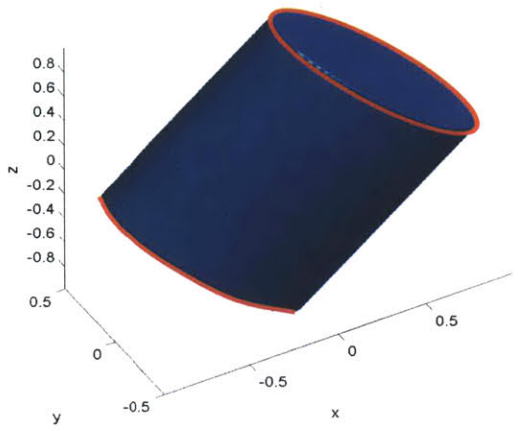

Figure 2.2. The nuclear-norm ball of 2×2 symmetric matrices. Here x, y denote the diagonal entries, and z the off-diagonal entry.

Unlike with the ℓ_1-norm the nuclear-norm of a matrix has no closed-form representation, but can instead be expressed variationally. Specifically, the spectral or operator norm $\|\cdot\|$ of a matrix (the largest singular value) is the dual norm of the nuclear-norm $\|\cdot\|_*$ [82]:

$$\|X\|_* = \max\{\mathrm{Tr}(X'Y) |\ \|Y\| \leq 1\}.$$

Further, the spectral norm admits a simple semidefinite characterization:

$$\|Y\| = \min_t \ t \qquad \text{s.t.} \begin{pmatrix} tI_n & Y \\ Y' & tI_n \end{pmatrix} \succeq 0.$$

We then obtain the following SDP characterization of the nuclear-norm:

$$\begin{aligned} \|X\|_* = \min_{W_1, W_2} \quad & \tfrac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)) \\ \text{s.t.} \quad & \begin{pmatrix} W_1 & X \\ X' & W_2 \end{pmatrix} \succeq 0. \end{aligned}$$

This semidefinite characterization can in turn be used to specify the unit ball of the nuclear-norm:

$$B^{p\times p}_{\|\cdot\|_*} = \left\{ X \in \mathbb{R}^{p\times p} \mid \|X\|_* \leq 1 \right\}.$$

Suppose $X \in B^{p\times p}_{\|\cdot\|_*}$ is a boundary point of the nuclear-norm ball, i.e., $\|X\|_* = 1$. Let $X = U\Sigma V'$ be a singular value decomposition of X, such that $U, V \in \mathbb{R}^{p\times \mathrm{rank}(X)}$

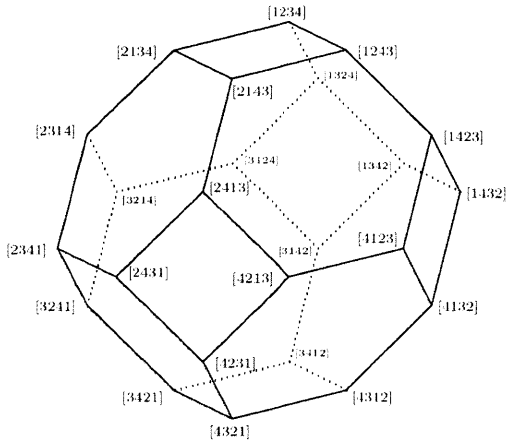

Figure 2.3. The permutahedron generated by the vector $[1, 2, 3, 4]'$.

and $\Sigma \in \mathbb{R}^{\text{rank}(X)\times\text{rank}(X)}$. Further let $W \subset \mathbb{R}^{p\times p}$ denote the subspace of matrices given by the span of matrices with either the same row space or the same column space as X:

$$W = \left\{UM' + NV' \mid M, N \in \mathbb{R}^{p\times\text{rank}(X)}\right\}.$$

Then we have the following description of the normal cone at X with respect to $B^{p\times p}_{\|\cdot\|_*}$:

$$N_{B^{p\times p}_{\|\cdot\|_*}}(X) = \left\{tUV^T + W \in \mathbb{R}^{p\times p} \mid W^TU = 0, WV = 0, \|W\|_* \leq t, t \geq 0\right\}.$$

Here $\mathcal{P}$ denotes the projection operator. Notice the parallels with the normal cone with respect to the cross-polytope.

■ 2.2.3 Permutahedron

The permutahedron (see Figure 2.3) generated by a vector $\mathbf{x} \in \mathbb{R}^p$ is the convex hull of all permutations of the vector $\mathbf{x}$:

$$P^p(\mathbf{x}) = \text{conv}\{\Pi\mathbf{x} \mid \forall \text{ permutation matrices } \Pi\}.$$

The set of permutations of the vector $[1, \ldots, p]'$ represents the set of all *rankings* of p objects. Consequently the permutahedron, and the related Birkhoff polytope (the convex hull of permutation matrices), lead to useful convex relaxation approaches in ranking and tracking problems (see Chapter 5).

The permutahedron $P^p(\mathbf{x})$ of a vector composed of distinct entries consists of $p!$ extreme points and a direct halfspace representation requires $2^p - 2$ inequalities (one for each proper subset of $\{1, \ldots, p\}$). However the permutahedron still has a tractable representation via lifting. Before describing this lifted specification, we require some notation. For any vector $\mathbf{y}$ let $\bar{\mathbf{y}}$ denote the vector obtained by sorting the entries of $\mathbf{y}$ in descending order. A vector $\mathbf{y} \in \mathbb{R}^p$ is said to be *majorized* by a vector $\mathbf{x} \in \mathbb{R}^p$ if the following conditions hold:

$$\sum_{i=1}^{k} \bar{\mathbf{y}}_i \leq \sum_{i=1}^{k} \bar{\mathbf{x}}_i, \ \forall k = 1, \ldots, p-1, \quad \text{and} \quad \sum_i \mathbf{y}_i = \sum_i \mathbf{x}_i. \tag{2.2}$$

The *majorization principle* states that the permutahedron $P^p(\mathbf{x})$ is exactly the set of vectors majorized by $\mathbf{x}$ [11]:

$$P^p(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^p \mid \mathbf{y} \text{ majorized by } \mathbf{x}\}.$$

Consequently a tractable description of the permutahedron can be obtained if the majorization inequalities of (2.2) can be expressed tractably. Since $\sum_{i=1}^{k} \bar{\mathbf{x}}_i$ is a fixed quantity, we require a tractable expression for sets of the form

$$Q_k(c) = \left\{ \mathbf{y} \in \mathbb{R}^p \mid \sum_{i=1}^{k} \bar{\mathbf{y}}_i \leq c \right\}.$$

Letting $\mathbf{e} \in \mathbb{R}^p$ denote the all-ones vector, we have that [11]

$$Q_k(c) = \left\{ \mathbf{y} \in \mathbb{R}^p \mid \exists \mathbf{z} \in \mathbb{R}^p, s \in \mathbb{R} \text{ s.t. } c - ks - \mathbf{e}'\mathbf{z} \geq 0, \ \mathbf{z} \geq 0, \ \mathbf{z} - \mathbf{y} + s\mathbf{e} \geq 0 \right\}.$$

Here the last two inequalities are to be interpreted elementwise. Consequently we have a tractable description of the permutahedron by lifting to $\mathbb{R}^{p^2+p-1}$ and using $2p^2 - 2p - 1$ inequalities and one equation.

It turns out that a more efficient representation of the permutahedron can be specified by lifting to a space of dimension $\mathcal{O}(p \log(p))$ and using only $\mathcal{O}(p \log(p))$ inequalities [71]. This representation is based on the structure of certain sorting networks, and is in some sense the most efficient possible representation of the permutahedron (see [71] for more details).

■ 2.2.4 Schur-Horn orbitope

Let $\mathbf{S}^p$ denote the space of $p \times p$ symmetric matrices, and let $\lambda(N)$ denote the sorted (in descending order) eigenvalues of a symmetric matrix N. Given a symmetric matrix

$M \in \mathbf{S}^p$ the Schur-Horn orbitope specified by M is defined as the convex hull of all matrices with the same spectrum as that of M:

$$SH^p(M) = \text{conv}\{UMU' \mid U \in \mathbb{R}^{p\times p} \text{ orthogonal}\}.$$

The Schur-Horn orbitope is the spectral analog of the permutahedron, and the projection of $SH^p(M)$ onto the set of diagonal matrices is exactly the permutahedron $P^p(\lambda(M))$.

A spectral majorization principle can be used to give a tractable representation of the Schur-Horn orbitope [11]. Specifically we have that

$$SH^p(M) = \left\{N \in \text{Sym}^{p\times p} \mid \lambda(N) \text{ majorized by } \lambda(M)\right\}.$$

Again we have the following tractable representation of sets constraining the sum of the top k eigenvalues of a matrix [11]:

$$\begin{aligned} R_k(c) &= \{N \in \text{Sym}^{p\times p} \mid \sum_{i=1}^{k} \lambda_i(N) \leq c\} \\ &= \{N \in \text{Sym}^{p\times p} \mid \exists Z \succeq 0, s \in \mathbb{R} \text{ s.t. } c - ks - \text{Tr}(Z) \geq 0,\ Z - N + sI_p \succeq 0\}. \end{aligned}$$

Here I_p represents the $p \times p$ identity matrix.

■ 2.3 Semidefinite Relaxations using Theta Bodies

In many cases of interest convex sets may not be tractable to represent, and it is of interest to develop tractable approximations. Here we describe a method to obtain a hierarchy of (increasingly complex) representations for convex sets given as the convex hulls of sets with algebraic structure. Specifically we focus on the setting in which our convex bodies arise as the convex hulls of algebraic varieties, which play a prominent role in this thesis. A *real algebraic variety* $\mathcal{A} \subseteq \mathbb{R}^p$ is the set of real solutions of a system of polynomial equations:

$$\mathcal{A} = \{\mathbf{x} : g_j(\mathbf{x}) = 0,\ \forall j\},$$

where $\{g_j\}$ is a finite collection of polynomials in p variables.

A basic question is to derive tractable representations of the convex hull $\text{conv}(\mathcal{A})$ of a variety $\mathcal{A}$. All our discussion here is based on results described in [77] for semidefinite relaxations of convex hulls of algebraic varieties using theta bodies. We only give a

brief review of the relevant constructions, and refer the reader to the vast literature on this subject for more details (see [77, 114] and the references therein).

To begin with we note that a *sum-of-squares* (SOS) polynomial in $\mathbb{R}[\mathbf{x}]$ (the ring of polynomials in the variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$) is a polynomial that can be written as the (finite) sum of squares of other polynomials in $\mathbb{R}[\mathbf{x}]$. Verifying the nonnegativity of a multivariate polynomial is intractable in general, and therefore SOS polynomials play an important role in real algebraic geometry as an SOS polynomial is easily seen to be nonnegative everywhere. Further checking whether a polynomial is an SOS polynomial can be accomplished efficiently via semidefinite programming [114].

Turning our attention to the description of the convex hull of an algebraic variety, we will assume for the sake of simplicity that the convex hull is closed. Let $I \subseteq \mathbb{R}[\mathbf{x}]$ be a polynomial ideal [79], and let $V_{\mathbb{R}}(I) \in \mathbb{R}^p$ be its real algebraic variety:

$$V_{\mathbb{R}}(I) = \{\mathbf{x} : f(\mathbf{x}) = 0, \ \forall f \in I\}.$$

One can then show that the convex hull $\text{conv}(V_{\mathbb{R}}(I))$ is given as:

$$\begin{aligned} \text{conv}(V_{\mathbb{R}}(I)) &= \{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear and nonnegative on } V_{\mathbb{R}}(I)\} \\ &= \{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear s.t. } f = h + g, \ \forall \ h \text{ nonnegative}, \ \forall \ g \in I\} \\ &= \{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear s.t. } f \text{ nonnegative modulo } I\}. \end{aligned}$$

A linear polynomial here is one that has a maximum degree of one, and the meaning of "modulo an ideal" is clear. As nonnegativity modulo an ideal may be intractable to check, we can consider a relaxation to a polynomial being SOS modulo an ideal, i.e., a polynomial that can be written as $\sum_{i=1}^{q} h_i^2 + g$ for g in the ideal. Since it is tractable to check via semidefinite programmming whether bounded-degree polynomials are SOS, the k-th theta body of an ideal I is defined as follows in [77]:

$$\text{TH}_k(I) = \{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear s.t. } f \text{ is } k\text{-sos modulo } I\}.$$

Here k-sos refers to an SOS polynomial in which the components in the SOS decomposition have degree at most k. The k-th theta body $\text{TH}_k(I)$ is a convex relaxation of $\text{conv}(V_{\mathbb{R}}(I))$, and one can verify that

$$\text{conv}(V_{\mathbb{R}}(I)) \subseteq \cdots \subseteq \text{TH}_{k+1}(I) \subseteq \text{TH}_k(V_{\mathbb{R}}(I)).$$

By the arguments given above (see also [77]) these theta bodies can be described using semidefinite programs of size polynomial in k. Hence by considering theta bodies

$\text{TH}_k(I)$ with increasingly larger k, one can obtain a hierarchy of tighter semidefinite relaxations of $\text{conv}(V_{\mathbb{R}}(I))$. We also note that in many cases of interest such semidefinite relaxations preserve low-dimensional faces of the convex hull of a variety, although these properties are not known in general.

Example The cut polytope is defined as the convex hull of all symmetric rank-one signed matrices:

$$CP^p = \text{conv}\{\mathbf{z}\mathbf{z}^T : \mathbf{z} \in \{-1, +1\}^p\}.$$

It is well-known that the cut polytope is intractable to characterize [47], and therefore we need to use tractable relaxations instead. The following popular relaxation is used in semidefinite approximations of the MAXCUT problem:

$$CP - SDP_1^p = \{M : M \text{ symmetric}, \ M \succeq 0, \ M_{ii} = 1, \forall i = 1, \cdots, p\}.$$

This is the well-studied elliptope [47], and can be interpreted as the second theta body relaxation of the cut polytope CP^p [77].

Chapter 3

Rank-Sparsity Uncertainty Principles and Matrix Decomposition

3.1 Introduction

Complex systems and models arise in a variety of problems in science and engineering. In many applications such complex systems and models are often composed of multiple simpler systems and models. Therefore, in order to better understand the behavior and properties of a complex system a natural approach is to decompose the system into its simpler components. In this chapter we consider matrix representations of systems and statistical models in which our matrices are formed by adding together *sparse* and *low-rank* matrices. We study the problem of recovering the sparse and low-rank components given no prior knowledge about the sparsity pattern of the sparse matrix, or the rank of the low-rank matrix. We propose a tractable convex program to recover these components, and provide sufficient conditions under which our procedure recovers the sparse and low-rank matrices *exactly*.

Such a decomposition problem arises in a number of settings, with the sparse and low-rank matrices having different interpretations depending on the application. In a statistical model selection setting, the sparse matrix can correspond to a Gaussian graphical model [93] and the low-rank matrix can summarize the effect of latent, unobserved variables (see Chapter 4 for a detailed investigation). In computational complexity, the notion of *matrix rigidity* [138] captures the smallest number of entries of a matrix that must be changed in order to reduce the rank of the matrix below a specified level (the changes can be of arbitrary magnitude). Bounds on the rigidity of a matrix have several implications in complexity theory [99]. Similarly, in a system identification setting the low-rank matrix represents a system with a small model order while

the sparse matrix represents a system with a sparse impulse response. Decomposing a system into such simpler components can be used to provide a simpler, more efficient description.

■ 3.1.1 Our results

Formally the decomposition problem in which we are interested can be defined as follows:

Problem Given $C = A^\star + B^\star$ where $A^\star$ is an unknown sparse matrix and $B^\star$ is an unknown low-rank matrix, recover $A^\star$ and $B^\star$ from C using no additional information on the sparsity pattern and/or the rank of the components.

In the absence of any further assumptions, this decomposition problem is fundamentally ill-posed. Indeed, there are a number of scenarios in which a unique splitting of C into "low-rank" and "sparse" parts may not exist; for example, the low-rank matrix may itself be very sparse leading to identifiability issues. In order to characterize when a unique decomposition is possible we develop a notion of *rank-sparsity incoherence*, an uncertainty principle between the sparsity pattern of a matrix and its row/column spaces. This condition is based on quantities involving the tangent spaces to the algebraic variety of sparse matrices and the algebraic variety of low-rank matrices [79]. Another point of ambiguity in the problem statement is that one could subtract a nonzero entry from $A^\star$ and add it to $B^\star$; the sparsity level of $A^\star$ is strictly improved while the rank of $B^\star$ is increased by at most 1. Therefore it is in general unclear what the "true" sparse and low-rank components are. We discuss this point in greater detail in Section 3.4.2 following the statement of the main theorem. In particular we describe how our identifiability and recovery results for the decomposition problem are to be interpreted.

Two natural identifiability problems may arise. The first one occurs if the low-rank matrix itself is very sparse. In order to avoid such a problem we impose certain conditions on the row/column spaces of the low-rank matrix. Specifically, for a matrix M let $T(M)$ be the tangent space at M with respect to the variety of all matrices with rank less than or equal to $\text{rank}(M)$. Operationally, $T(M)$ is the span of all matrices with row-space contained in the row-space of M or with column-space contained in the column-space of M; see (3.7) for a formal characterization. Let $\xi(M)$ be defined as

follows:

$$\xi(M) \triangleq \max_{N \in T(M),\ \|N\| \leq 1} \|N\|_\infty. \tag{3.1}$$

Here $\|\cdot\|$ is the spectral norm (i.e., the largest singular value), and $\|\cdot\|_\infty$ denotes the largest entry in magnitude. Thus $\xi(M)$ being small implies that (appropriately scaled) elements of the tangent space $T(M)$ are "diffuse", i.e., these elements are not too sparse; as a result M cannot be very sparse. As shown in Proposition 3.4.3 (see Section 3.4.3) a low-rank matrix M with row/column spaces that are not closely aligned with the coordinate axes has small $\xi(M)$.

The other identifiability problem may arise if the sparse matrix has all its support concentrated in one column; the entries in this column could negate the entries of the corresponding low-rank matrix, thus leaving the rank and the column space of the low-rank matrix unchanged. To avoid such a situation, we impose conditions on the sparsity pattern of the sparse matrix so that its support is not too concentrated in any row/column. For a matrix M let $\Omega(M)$ be the tangent space at M with respect to the variety of all matrices with number of nonzero entries less than or equal to $|\text{support}(M)|$. The space $\Omega(M)$ is simply the set of all matrices that have support contained within the support of M; see (3.5). Let $\mu(M)$ be defined as follows:

$$\mu(M) \triangleq \max_{N \in \Omega(M),\ \|N\|_\infty \leq 1} \|N\|. \tag{3.2}$$

The quantity $\mu(M)$ being small for a matrix implies that the *spectrum* of any element of the tangent space $\Omega(M)$ is "diffuse", i.e., the singular values of these elements are not too large. We show in Proposition 3.4.2 (see Section 3.4.3) that a sparse matrix M with "bounded degree" (a small number of nonzeros per row/column) has small $\mu(M)$.

For a given matrix M, it is impossible for both quantities $\xi(M)$ and $\mu(M)$ to be simultaneously small. Indeed, we prove that for any matrix $M \neq 0$ we must have that $\xi(M)\mu(M) \geq 1$ (see Theorem 3.3.1 in Section 3.3.3). Thus, this *uncertainty principle* asserts that there is no nonzero matrix M with all elements in $T(M)$ being diffuse *and* all elements in $\Omega(M)$ having diffuse spectra. As we describe later, the quantities ξ and μ are also used to characterize fundamental identifiability in the decomposition problem.

In general solving the decomposition problem is intractable; this is due to the fact that it is intractable in general to compute the rigidity of a matrix (see Section 3.2.2), which can be viewed as a special case of the sparse-plus-low-rank decomposition problem. Hence, we consider tractable approaches employing recently well-studied convex

relaxations. We formulate a convex optimization problem for decomposition using a combination of the ℓ_1 norm and the nuclear norm. For any matrix M the ℓ_1 norm is given by

$$\|M\|_1 = \sum_{i,j} |M_{i,j}|,$$

and the nuclear norm, which is the sum of the singular values, is given by

$$\|M\|_* = \sum_k \sigma_k(M),$$

where $\{\sigma_k(M)\}$ are the singular values of M. The ℓ_1 norm has been used as an effective surrogate for the number of nonzero entries of a vector, and a number of results provide conditions under which this heuristic recovers sparse solutions to ill-posed inverse problems [29, 53, 54]. More recently, the nuclear norm has been shown to be an effective surrogate for the rank of a matrix [64]. This relaxation is a generalization of the previously studied trace-heuristic that was used to recover low-rank positive semidefinite matrices [108]. Indeed, several papers demonstrate that the nuclear norm heuristic recovers low-rank matrices in various rank minimization problems [30, 121]. Based on these results, we propose the following optimization formulation to recover $A^\star$ and $B^\star$ given $C = A^\star + B^\star$:

$$\begin{aligned} (\hat{A}, \hat{B}) = \arg\min_{A,B} \ & \gamma\|A\|_1 + \|B\|_* \\ \text{s.t.} \ & A + B = C. \end{aligned} \tag{3.3}$$

Here γ is a parameter that provides a trade-off between the low-rank and sparse components. This optimization problem is convex, and can in fact be rewritten as a semidefinite program (SDP) [139] (see Appendix A.1).

We prove that $(\hat{A}, \hat{B}) = (A^\star, B^\star)$ is the unique optimum of (3.3) for a range of γ if $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$ (see Theorem 3.4.1 in Section 3.4.2). Thus, the conditions for *exact* recovery of the sparse and low-rank components via the convex program (3.3) involve the tangent-space-based quantities defined in (3.1) and (3.2). Essentially these conditions specify that each element of $\Omega(A^\star)$ must have a diffuse spectrum, *and* every element of $T(B^\star)$ must be diffuse. In a sense that will be made precise later, the condition $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$ required for the convex program (3.3) to provide exact recovery is slightly tighter than that required for fundamental identifiability in the decomposition problem. An important feature of our result is that it provides a simple *deterministic* condition for exact recovery. In addition, note that the conditions only depend on the

row/column spaces of the low-rank matrix $B^\star$ and the support of the sparse matrix $A^\star$, and not the magnitudes of the nonzero singular values of $B^\star$ or the nonzero entries of $A^\star$. The reason for this is that the magnitudes of the nonzero entries of $A^\star$ and the nonzero singular values of $B^\star$ play no role in the subgradient conditions with respect to the ℓ_1 norm and the nuclear norm.

In the sequel we discuss concrete classes of sparse and low-rank matrices that have small μ and ξ respectively. We also show that when the sparse and low-rank matrices $A^\star$ and $B^\star$ are drawn from certain natural random ensembles, then the sufficient conditions of Theorem 3.4.1 are satisfied with high probability; consequently, (3.3) provides exact recovery with high probability for such matrices.

■ 3.1.2 Previous work using incoherence

The concept of incoherence was studied in the context of recovering sparse representations of vectors from a so-called "overcomplete dictionary" [52]. More concretely consider a situation in which one is given a vector formed by a sparse linear combination of a few elements from a combined time-frequency dictionary, i.e., a vector formed by adding a few sinusoids and a few "spikes"; the goal is to recover the spikes and sinusoids that compose the vector from the infinitely many possible solutions. Based on a notion of time-frequency incoherence, the ℓ_1 heuristic was shown to succeed in recovering sparse solutions [51]. Incoherence is also a concept that is used in recent work under the title of *compressed sensing*, which aims to recover "low-dimensional" objects such as sparse vectors [29, 54] and low-rank matrices [30, 121] given incomplete observations. Our work is closer in spirit to that in [52], and can be viewed as a method to recover the "simplest explanation" of a matrix given an "overcomplete dictionary" of sparse and low-rank matrix atoms.

■ 3.1.3 Outline

In Section 3.2 we elaborate on the applications mentioned previously, and discuss the implications of our results for each of these applications. Section 3.3 formally describes conditions for fundamental identifiability in the decomposition problem based on the quantities ξ and μ defined in (3.1) and (3.2). We also provide a proof of the rank-sparsity uncertainty principle of Theorem 3.3.1. We prove Theorem 3.4.1 in Section 3.4, and also provide concrete classes of sparse and low-rank matrices that satisfy the sufficient conditions of Theorem 3.4.1. Section 3.5 describes the results of simulations of our

approach applied to synthetic matrix decomposition problems. We conclude with a discussion in Section 3.6. Appendix A provides additional details and proofs.

■ 3.2 Applications

In this section we describe several applications that involve decomposing a matrix into sparse and low-rank components.

■ 3.2.1 Graphical modeling with latent variables

We begin with a problem in statistical model selection. In many applications large covariance matrices are approximated as low-rank matrices based on the assumption that a small number of *latent* factors explain most of the observed statistics (e.g., principal component analysis). Another well-studied class of models are those described by graphical models [93] in which the *inverse* of the covariance matrix (also called the precision or concentration or information matrix) is assumed to be *sparse* (typically this sparsity is with respect to some graph). Consequently, a natural sparse-plus-low-rank decomposition problem arises in latent-variable graphical model selection, which we discuss in more detail in Chapter 4.

■ 3.2.2 Matrix rigidity

The *rigidity* of a matrix M, denoted by $R_M(k)$, is the smallest number of entries that need to be changed in order to reduce the rank of M below k. Obtaining bounds on rigidity has a number of implications in complexity theory [99], such as the trade-offs between size and depth in arithmetic circuits. However, computing the rigidity of a matrix is intractable in general [38, 101]. For any $M \in \mathbb{R}^{n\times n}$ one can check that $R_M(k) \leq (n-k)^2$ (this follows directly from a Schur complement argument). Generically every $M \in \mathbb{R}^{n\times n}$ is very rigid, i.e., $R_M(k) = (n-k)^2$ [138], although special classes of matrices may be less rigid. We show that the SDP (3.3) can be used to compute rigidity for certain matrices with sufficiently small rigidity (see Section 3.4.4 for more details). Indeed, this convex program (3.3) also provides a certificate of the sparse and low-rank components that form such low-rigidity matrices; that is, the SDP (3.3) not only enables us to compute the rigidity for certain matrices but additionally provides the changes required in order to realize a matrix of lower rank.

■ 3.2.3 Composite system identification

A decomposition problem can also be posed in the system identification setting. Linear time-invariant (LTI) systems can be represented by Hankel matrices, where the matrix represents the input-output relationship of the system [131]. Thus, a sparse Hankel matrix corresponds to an LTI system with a sparse impulse response. A low-rank Hankel matrix corresponds to a system with small model order, and provides a minimal realization for a system [65]. Given an LTI system H as follows

$$H = H_s + H_{lr},$$

where H_s is sparse and H_{lr} is low-rank, obtaining a simple description of H requires decomposing it into its simpler sparse and low-rank components. One can obtain these components by solving our rank-sparsity decomposition problem. Note that in practice one can impose in (3.3) the additional constraint that the sparse and low-rank matrices have Hankel structure.

■ 3.2.4 Partially coherent decomposition in optical systems

We outline an optics application that is described in greater detail in [63]. Optical imaging systems are commonly modeled using the Hopkins integral [75], which gives the output intensity at a point as a function of the input transmission via a quadratic form. In many applications the operator in this quadratic form can be well-approximated by a (finite) positive semi-definite matrix. Optical systems described by a low-pass filter are called *coherent* imaging systems, and the corresponding system matrices have *small rank*. For systems that are not perfectly coherent various methods have been proposed to find an *optimal coherent decomposition* [115], and these essentially identify the best approximation of the system matrix by a matrix of lower rank. At the other end are *incoherent* optical systems that allow some high frequencies, and are characterized by system matrices that are *diagonal*. As most real-world imaging systems are some combination of coherent and incoherent, it was suggested in [63] that optical systems are better described by a sum of coherent and incoherent systems rather than by the best coherent (i.e., low-rank) approximation as in [115]. Thus, decomposing an imaging system into coherent and incoherent components involves splitting the optical system matrix into low-rank and diagonal components. Identifying these simpler components has important applications in tasks such as optical microlithography [75, 115].

3.3 Rank-Sparsity Incoherence

Throughout this chapter, we restrict ourselves to square $n \times n$ matrices to avoid cluttered notation. All our analysis extends to rectangular $n_1 \times n_2$ matrices, if we simply replace n by $\max(n_1, n_2)$.

3.3.1 Identifiability issues

As described in the introduction, the matrix decomposition problem can be fundamentally ill-posed. We describe two situations in which identifiability issues arise. These examples suggest the kinds of additional conditions that are required in order to ensure that there exists a unique decomposition into sparse and low-rank matrices.

First, let $A^\star$ be any sparse matrix and let $B^\star = e_i e_j^T$, where e_i represents the i-th standard basis vector. In this case, the low-rank matrix $B^\star$ is also very sparse, and a valid sparse-plus-low-rank decomposition might be $\hat{A} = A^\star + e_i e_j^T$ and $\hat{B} = 0$. Thus, we need conditions that ensure that the low-rank matrix is not too sparse. One way to accomplish this is to require that the quantity $\xi(B^\star)$ be small. As will be discussed in Section 3.4.3), if the row and column spaces of $B^\star$ are "incoherent" with respect to the standard basis, i.e., the row/column spaces are not aligned closely with any of the coordinate axes, then $\xi(B^\star)$ is small.

Next, consider the scenario in which $B^\star$ is any low-rank matrix and $A^\star = -ve_1^T$ with v being the first column of $B^\star$. Thus, $C = A^\star + B^\star$ has zeros in the first column, $\mathrm{rank}(C) \leq \mathrm{rank}(B^\star)$, and C has the same column space as $B^\star$. Therefore, a reasonable sparse-plus-low-rank decomposition in this case might be $\hat{B} = B^\star + A^\star$ and $\hat{A} = 0$. Here $\mathrm{rank}(\hat{B}) = \mathrm{rank}(B^\star)$. Requiring that a sparse matrix $A^\star$ have small $\mu(A^\star)$ avoids such identifiability issues. Indeed we show in Section 3.4.3 that sparse matrices with "bounded degree" (i.e., few nonzero entries per row/column) have small μ.

3.3.2 Tangent-space identifiability

We begin by describing the sets of sparse and low-rank matrices. These sets can be considered either as differentiable manifolds (away from their singularities) or as algebraic varieties; we emphasize the latter viewpoint here. Recall that an algebraic variety is the solution set of a system of polynomial equations. The set of sparse matrices and the set of low-rank matrices can be naturally viewed as algebraic varieties. Here we describe these varieties, and discuss some of their properties. Of particular interest in

this chapter are geometric properties of these varieties such as the tangent space at a (smooth) point.

Let $\mathcal{S}(k)$ denote the set of matrices with at most k nonzeros:

$$\mathcal{S}(k) \triangleq \{M \in \mathbb{R}^{n\times n} \mid |\text{support}(M)| \leq k\}. \tag{3.4}$$

The set $\mathcal{S}(k)$ is an algebraic variety, and can in fact be viewed as a union of $\binom{n^2}{k}$ subspaces in $\mathbb{R}^{n\times n}$. This variety has dimension k, and it is smooth everywhere except at those matrices that have support size strictly smaller than k. For any matrix $M \in \mathbb{R}^{n\times n}$, consider the variety $\mathcal{S}(|\text{support}(M)|)$; M is a smooth point of this variety, and the tangent space at M is given by

$$\Omega(M) = \{N \in \mathbb{R}^{n\times n} \mid \text{support}(N) \subseteq \text{support}(M)\}. \tag{3.5}$$

In words the tangent space $\Omega(M)$ at a smooth point M is given by the set of all matrices that have support contained within the support of M. We view $\Omega(M)$ as a subspace in $\mathbb{R}^{n\times n}$.

Next let $\mathcal{L}(r)$ denote the algebraic variety of matrices with rank at most r:

$$\mathcal{L}(r) \triangleq \{M \in \mathbb{R}^{n\times n} \mid \text{rank}(M) \leq r\}. \tag{3.6}$$

It is easily seen that $\mathcal{L}(r)$ is an algebraic variety because it can be defined through the vanishing of all $(r+1)\times(r+1)$ minors. This variety has dimension equal to $r(2n-r)$, and it is smooth everywhere except at those matrices that have rank strictly smaller than r. Consider a rank-r matrix M with SVD $M = UDV^T$, where $U, V \in \mathbb{R}^{n\times r}$ and $D \in \mathbb{R}^{r\times r}$. The matrix M is a smooth point of the variety $\mathcal{L}(\text{rank}(M))$, and the tangent space at M with respect to this variety is given by

$$T(M) = \{UY_1^T + Y_2V^T \mid Y_1, Y_2 \in \mathbb{R}^{n\times r}\}. \tag{3.7}$$

In words the tangent space $T(M)$ at a smooth point M is the span of all matrices that have either the same row-space as M or the same column-space as M. As with $\Omega(M)$ we view $T(M)$ as a subspace in $\mathbb{R}^{n\times n}$.

Before analyzing whether $(A^\star, B^\star)$ can be recovered in general (for example, using the SDP (3.3)), we ask a simpler question. Suppose that we had prior information about the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$, in addition to being given $C = A^\star + B^\star$. Can we then *uniquely* recover $(A^\star, B^\star)$ from C? Assuming such prior knowledge of

the tangent spaces is unrealistic in practice; however, we obtain useful insight into the kinds of conditions required on sparse and low-rank matrices for exact decomposition. A necessary and sufficient condition for unique identifiability of $(A^\star, B^\star)$ with respect to the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ is that these spaces intersect transversally:

$$\Omega(A^\star) \cap T(B^\star) = \{0\}.$$

That is, the subspaces $\Omega(A^\star)$ and $T(B^\star)$ have a trivial intersection. The sufficiency of this condition for unique decomposition is easily seen. For the necessity part, suppose for the sake of a contradiction that a nonzero matrix M belongs to $\Omega(A^\star) \cap T(B^\star)$; one can add and subtract M from $A^\star$ and $B^\star$ respectively while still having a valid decomposition, which violates the uniqueness requirement. Therefore tangent space transversality is equivalent to a "linearized" identifiability condition around $(A^\star, B^\star)$. Note that tangent space transversality is also a *sufficient condition* for local identifiability around $(A^\star, B^\star)$ with respect to the sparse and low-rank matrix varieties, based on the inverse function theorem. The transversality condition does not, however, imply *global* identifiability with respect to the sparse and low-rank matrix varieties. The following proposition, proved in Appendix A.2, provides a simple condition in terms of the quantities $\mu(A^\star)$ and $\xi(B^\star)$ for the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ to intersect transversally.

Proposition 3.3.1. *Given any two matrices $A^\star$ and $B^\star$, we have that*

$$\mu(A^\star)\xi(B^\star) < 1 \quad \Rightarrow \quad \Omega(A^\star) \cap T(B^\star) = \{0\},$$

where $\xi(B^\star)$ and $\mu(A^\star)$ are defined in (3.1) and (3.2), and the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ are defined in (3.5) and (3.7).

Thus, both $\mu(A^\star)$ and $\xi(B^\star)$ being small implies that the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ intersect transversally; consequently, we can exactly recover $(A^\star, B^\star)$ given $\Omega(A^\star)$ and $T(B^\star)$. As we shall see, the condition required in Theorem 3.4.1 (see Section 3.4.2) for exact recovery using the convex program (3.3) will be simply a mild tightening of the condition required above for unique decomposition given the tangent spaces.

■ 3.3.3 Rank-sparsity uncertainty principle

Another important consequence of Proposition 3.3.1 is that we have an elementary proof of the following rank-sparsity uncertainty principle.

Theorem 3.3.1. *For any matrix $M \neq 0$, we have that*

$$\xi(M)\mu(M) \geq 1,$$

where $\xi(M)$ and $\mu(M)$ are as defined in (3.1) and (3.2) respectively.

Proof: Given any $M \neq 0$ it is clear that $M \in \Omega(M) \cap T(M)$, i.e., M is an element of both tangent spaces. However $\mu(M)\xi(M) < 1$ would imply from Proposition 3.3.1 that $\Omega(M) \cap T(M) = \{0\}$, which is a contradiction. Consequently, we must have that $\mu(M)\xi(M) \geq 1$. □

Hence, for *any* matrix $M \neq 0$ both $\mu(M)$ and $\xi(M)$ cannot be simultaneously small. Note that Proposition 3.3.1 is an assertion involving μ and ξ for (in general) *different* matrices, while Theorem 3.3.1 is a statement about μ and ξ for the *same* matrix. Essentially the uncertainty principle asserts that no matrix can be too sparse while having "diffuse" row and column spaces. An extreme example is the matrix $e_i e_j^T$, which has the property that $\mu(e_i e_j^T)\xi(e_i e_j^T) = 1$.

■ 3.4 Exact Decomposition Using Semidefinite Programming

We begin this section by studying the optimality conditions of the convex program (3.3), after which we provide a proof of Theorem 3.4.1 with simple conditions that guarantee exact decomposition. Next we discuss concrete classes of sparse and low-rank matrices that satisfy the conditions of Theorem 3.4.1, and can thus be uniquely decomposed using (3.3).

■ 3.4.1 Optimality conditions

The orthogonal projection onto the space $\Omega(A^\star)$ is denoted $P_{\Omega(A^\star)}$, which simply sets to zero those entries with support not inside support$(A^\star)$. The subspace orthogonal to $\Omega(A^\star)$ is denoted $\Omega(A^\star)^c$, and it consists of matrices with complementary support, i.e., supported on support$(A^\star)^c$. The projection onto $\Omega(A^\star)^c$ is denoted $P_{\Omega(A^\star)^c}$.

Similarly the orthogonal projection onto the space $T(B^\star)$ is denoted $P_{T(B^\star)}$. Letting $B^\star = U\Sigma V^T$ be the SVD of $B^\star$, we have the following explicit relation for $P_{T(B^\star)}$:

$$P_{T(B^\star)}(M) = P_U M + M P_V - P_U M P_V. \tag{3.8}$$

Here $P_U = UU^T$ and $P_V = VV^T$. The space orthogonal to $T(B^\star)$ is denoted $T(B^\star)^\perp$, and the corresponding projection is denoted $P_{T(B^\star)^\perp}(M)$. The space $T(B^\star)^\perp$ consists of

matrices with row-space orthogonal to the row-space of $B^\star$ *and* column-space orthogonal to the column-space of $B^\star$. We have that

$$P_{T(B^\star)^\perp}(M) = (I_{n\times n} - P_U)M(I_{n\times n} - P_V), \tag{3.9}$$

where $I_{n\times n}$ is the $n \times n$ identity matrix.

Following standard notation in convex analysis [124], we denote the *subdifferential* of a convex function f at a point $\hat{x}$ in its domain by $\partial f(\hat{x})$. The subdifferential $\partial f(\hat{x})$ consists of all y such that

$$f(x) \geq f(\hat{x}) + \langle y, x - \hat{x}\rangle, \quad \forall x.$$

From the optimality conditions for a convex program [13], we have that $(A^\star, B^\star)$ is an optimum of (3.3) if and only if there exists a dual $Q \in \mathbb{R}^{n\times n}$ such that

$$Q \in \gamma\partial\|A^\star\|_1 \text{ and } Q \in \partial\|B^\star\|_*. \tag{3.10}$$

From the characterization of the subdifferential of the ℓ_1 norm, we have that $Q \in \gamma\partial\|A^\star\|_1$ if and only if

$$P_{\Omega(A^\star)}(Q) = \gamma\,\mathrm{sign}(A^\star), \quad \|P_{\Omega(A^\star)^c}(Q)\|_\infty \leq \gamma. \tag{3.11}$$

Here $\mathrm{sign}(A^\star_{i,j})$ equals $+1$ if $A^\star_{i,j} > 0$, -1 if $A^\star_{i,j} < 0$, and 0 if $A^\star_{i,j} = 0$. We also have that $Q \in \partial\|B^\star\|_*$ if and only if [142]

$$P_{T(B^\star)}(Q) = UV', \quad \|P_{T(B^\star)^\perp}(Q)\| \leq 1. \tag{3.12}$$

Note that these are necessary and sufficient conditions for $(A^\star, B^\star)$ to be *an* optimum of (3.3). The following proposition provides sufficient conditions for $(A^\star, B^\star)$ to be the *unique* optimum of (3.3), and it involves a slight tightening of the conditions (3.10), (3.11), and (3.12).

Proposition 3.4.1. *Suppose that $C = A^\star + B^\star$. Then $(\hat{A}, \hat{B}) = (A^\star, B^\star)$ is the* unique *optimizer of (3.3) if the following conditions are satisfied:*

1. $\Omega(A^\star) \cap T(B^\star) = \{0\}$.
2. *There exists a dual $Q \in \mathbb{R}^{n\times n}$ such that*
   (a) $P_{T(B^\star)}(Q) = UV'$

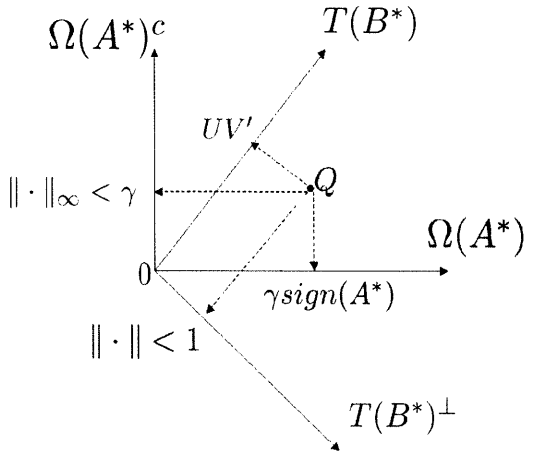

Figure 3.1. Geometric representation of optimality conditions: Existence of a dual Q. The arrows denote orthogonal projections – every projection must satisfy a condition (according to Proposition 3.4.1), which is described next to each arrow.

(b) $P_{\Omega(A^\star)}(Q) = \gamma \text{sign}(A^\star)$

(c) $\|P_{T(B^\star)^\perp}(Q)\| < 1$

(d) $\|P_{\Omega(A^\star)^c}(Q)\|_\infty < \gamma$

The proof of the proposition can be found in Appendix A.2. Figure 3.1 provides a visual representation of these conditions. In particular, we see that the spaces $\Omega(A^\star)$ and $T(B^\star)$ intersect transversely (part (1) of Proposition 3.4.1). One can also intuitively see that guaranteeing the existence of a dual Q with the requisite conditions (part (2) of Proposition 3.4.1) is perhaps easier if the intersection between $\Omega(A^\star)$ and $T(B^\star)$ is more transverse. Note that condition (1) of this proposition essentially requires identifiability with respect to the tangent spaces, as discussed in Section 3.3.2.

■ 3.4.2 Sufficient conditions based on $\mu(A^\star)$ and $\xi(B^\star)$

Next we provide simple sufficient conditions on $A^\star$ and $B^\star$ that guarantee the existence of an appropriate dual Q (as required by Proposition 3.4.1). Given matrices $A^\star$ and $B^\star$ with $\mu(A^\star)\xi(B^\star) < 1$, we have from Proposition 3.3.1 that $\Omega(A^\star) \cap T(B^\star) = \{0\}$, i.e., condition (1) of Proposition 3.4.1 is satisfied. We prove that if a slightly stronger condition holds, there exists a dual Q that satisfies the requirements of condition (2) of Proposition 3.4.1.

Theorem 3.4.1. *Given* $C = A^\star + B^\star$ *with*

$$\mu(A^\star)\xi(B^\star) < \frac{1}{6}$$

the unique optimum $(\hat{A}, \hat{B})$ *of (3.3) is* $(A^\star, B^\star)$ *for the following range of* γ*:*

$$\gamma \in \left(\frac{\xi(B^\star)}{1 - 4\mu(A^\star)\xi(B^\star)}, \frac{1 - 3\mu(A^\star)\xi(B^\star)}{\mu(A^\star)} \right).$$

Specifically $\gamma = \frac{(3\xi(B^\star))^p}{(2\mu(A^\star))^{1-p}}$ *for any choice of* $p \in [0,1]$ *is always inside the above range, and thus guarantees exact recovery of* $(A^\star, B^\star)$*. For example* $\gamma = \sqrt{\frac{3\xi(B^\star)}{2\mu(A^\star)}}$ *always guarantees exact recovery of* $(A^\star, B^\star)$*.*

Recall from the discussion in Section 3.3.2 and from Proposition 3.3.1 that $\mu(A^\star)\xi(B^\star) < 1$ is sufficient to ensure that the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ have a transverse intersection, which implies that $(A^\star, B^\star)$ are *locally* identifiable and can be recovered given $C = A^\star + B^\star$ along with side information about the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$. Theorem 3.4.1 asserts that if $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$, i.e., if the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ are *sufficiently transverse*, then the SDP (3.3) succeeds in recovering $(A^\star, B^\star)$ *without* any information about the tangent spaces.

The proof of this theorem can be found in Appendix A.2. The main idea behind the proof is that we only consider candidates for the dual Q that lie in the direct sum $\Omega(A^\star) \oplus T(B^\star)$ of the tangent spaces. Since $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$, we have from Proposition 3.3.1 that the tangent spaces $\Omega(A^\star)$ and $T(B^\star)$ have a transverse intersection, i.e., $\Omega(A^\star) \cap T(B^\star) = \{0\}$. Therefore, there exists a *unique* element $\hat{Q} \in \Omega(A^\star) \oplus T(B^\star)$ that satisfies $P_{T(B^\star)}(\hat{Q}) = UV'$ and $P_{\Omega(A^\star)}(\hat{Q}) = \gamma \text{sign}(A^\star)$. The proof proceeds by showing that if $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$ then the projections of this $\hat{Q}$ onto the orthogonal spaces $\Omega(A^\star)^c$ and $T(B^\star)^\perp$ are small, thus satisfying condition (2) of Proposition 3.4.1.

Remarks We discuss here the manner in which our results are to be interpreted. Given a matrix $C = A^\star + B^\star$ with $A^\star$ sparse and $B^\star$ low-rank, there are a number of alternative decompositions of C into "sparse" and "low-rank" components. For example, one could subtract one of the nonzero entries from the matrix $A^\star$ and add it to $B^\star$; thus, the sparsity level of $A^\star$ is strictly improved, while the rank of the modified $B^\star$ increases by at most 1. In fact one could construct many such alternative decompositions. Therefore, it may *a priori* be unclear which of these many decompositions is the "correct" one. To clarify this issue consider a matrix $C = A^\star + B^\star$ that is composed of the sum

of a sparse $A^\star$ with small $\mu(A^\star)$ and a low-rank $B^\star$ with small $\xi(B^\star)$. Recall that a sparse matrix having a small μ implies that the the sparsity pattern of the matrix is "diffuse," i.e., no row/column contains too many non-zeros (see Proposition 3.4.2 in Section 3.4.3 for a precise characterization). Similarly, a low-rank matrix with small ξ has "diffuse" row/column spaces, i.e., the row/column spaces are not aligned with any of the coordinate axes and as a result do not contain sparse vectors (see Proposition 3.4.3 in Section 3.4.3 for a precise characterization). Now let $C = A + B$ be an alternative decomposition with some of the entries of $A^\star$ moved to $B^\star$. Although the new A has a smaller support contained strictly within the support of $A^\star$ (and consequently, a smaller $\mu(A)$), the new low-rank matrix B has *sparse* vectors in its row and column spaces. Consequently we have that $\xi(B) \gg \xi(B^\star)$. Thus, while (A, B) is also a sparse-plus-low-rank decomposition, it is *not* a *diffuse* sparse-plus-low-rank decomposition, in that *both* the sparse matrix A and the low-rank matrix B do not *simultaneously* have diffuse supports and row/column spaces respectively. Also the opposite situation of removing a rank-1 term from the SVD of the low-rank matrix $B^\star$ and moving it to $A^\star$ to form a new decomposition (A, B) (now with B having strictly smaller rank than $B^\star$) faces a similar problem. In this case B has strictly smaller rank than $B^\star$, and also by construction a smaller $\xi(B)$. However the original low-rank matrix $B^\star$ has a small $\xi(B^\star)$ and thus has diffuse row/column spaces; therefore the rank-1 term that is added to $A^\star$ will *not* be sparse, and consequently the new matrix A will have $\mu(A) \gg \mu(A^\star)$. Hence the key point is that these alternate decompositions (A, B) do not satisfy the property that $\mu(A)\xi(B) < \frac{1}{6}$. Thus, our result is to be interpreted as follows: Given a matrix $C = A^\star + B^\star$ formed by adding a sparse matrix $A^\star$ with diffuse support and a low-rank matrix $B^\star$ with diffuse row/column spaces, the convex program that is studied in this chapter will recover this diffuse decomposition over the many possible alternative decompositions into sparse and low-rank components as none of these have the property of *both* components being *simultaneously* diffuse. Indeed in applications such as graphical model selection (see Section 3.2.1) it is precisely such a "diffuse" decomposition that one seeks to recover.

A related question is given a decomposition $C = A^\star + B^\star$ with $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$, do there exist small, local perturbations of $A^\star$ and $B^\star$ that give rise to alternate decompositions (A, B) with $\mu(A)\xi(B) < \frac{1}{6}$? Suppose $B^\star$ is slightly perturbed *along* the variety of rank-constrained matrices to some B. This ensures that the tangent space varies smoothly from $T(B^\star)$ to $T(B)$, and consequently that $\xi(B) \approx \xi(B^\star)$. However,

compensating for this by changing $A^\star$ to $A^\star + (B^\star - B)$ moves $A^\star$ outside the variety of sparse matrices. This is because $B^\star - B$ is *not sparse*. Thus the dimension of the tangent space $\Omega(A^\star + B^\star - B)$ is much greater than that of the tangent space $\Omega(A^\star)$, as a result of which $\mu(A^\star + B^\star - B) \gg \mu(A^\star)$; therefore we have that $\xi(B)\mu(A^\star + B^\star - B) \gg \frac{1}{6}$. The same reasoning holds in the opposite scenario. Consider perturbing $A^\star$ slightly along the variety of sparse matrices to some A. While this ensures that $\mu(A) \approx \mu(A^\star)$, changing $B^\star$ to $B^\star + (A^\star - A)$ moves $B^\star$ outside the variety of rank-constrained matrices. Therefore the dimension of the tangent space $T(B^\star + A^\star - A)$ is much greater than that of $T(B^\star)$, and also $T(B^\star + A^\star - A)$ contains sparse matrices, resulting in $\xi(B^\star + A^\star - A) \gg \xi(B^\star)$; consequently we have that $\mu(A)\xi(B^\star + A^\star - A) \gg \frac{1}{6}$.

■ 3.4.3 Sparse and low-rank matrices with $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$

We discuss concrete classes of sparse and low-rank matrices that satisfy the sufficient condition of Theorem 3.4.1 for exact decomposition. We begin by showing that sparse matrices with "bounded degree", i.e., bounded number of nonzeros per row/column, have small μ.

Proposition 3.4.2. *Let $A \in \mathbb{R}^{n\times n}$ be any matrix with at most $\deg_{\max}(A)$ nonzero entries per row/column, and with at least $\deg_{\min}(A)$ nonzero entries per row/column. With $\mu(A)$ as defined in (3.2), we have that*

$$\deg_{\min}(A) \le \mu(A) \le \deg_{\max}(A).$$

See Appendix A.2 for the proof. Note that if $A \in \mathbb{R}^{n\times n}$ has full support, i.e., $\Omega(A) = \mathbb{R}^{n\times n}$, then $\mu(A) = n$. Therefore, a constraint on the number of zeros per row/column provides a useful bound on μ. We emphasize here that simply bounding the number of nonzero entries in A does not suffice; the *sparsity pattern* also plays a role in determining the value of μ.

Next we consider low-rank matrices that have small ξ. Specifically, we show that matrices with row and column spaces that are incoherent with respect to the standard basis have small ξ. We measure the incoherence of a subspace $S \subseteq \mathbb{R}^n$ as follows:

$$\beta(S) \triangleq \max_i \|P_S e_i\|_2, \tag{3.13}$$

where e_i is the i'th standard basis vector, P_S denotes the projection onto the subspace S, and $\|\cdot\|_2$ denotes the vector ℓ_2 norm. This definition of incoherence also played an

important role in the results in [30]. A small value of $\beta(S)$ implies that the subspace S is not closely aligned with any of the coordinate axes. In general for any k-dimensional subspace S, we have that

$$\sqrt{\frac{k}{n}} \leq \beta(S) \leq 1,$$

where the lower bound is achieved, for example, by a subspace that spans any k columns of an $n \times n$ orthonormal Hadamard matrix, while the upper bound is achieved by any subspace that contains a standard basis vector. Based on the definition of $\beta(S)$, we define the incoherence of the row/column spaces of a matrix $B \in \mathbb{R}^{n \times n}$ as

$$\text{inc}(B) \triangleq \max\{\beta(\text{row-space}(B)),\ \beta(\text{column-space}(B))\}. \tag{3.14}$$

If the SVD of $B = U\Sigma V^T$ then row-space(B) = span(V) and column-space(B) = span(U). We show in Appendix A.2 that matrices with incoherent row/column spaces have small ξ; the proof technique for the lower bound here was suggested by Ben Recht [120].

Proposition 3.4.3. *Let $B \in \mathbb{R}^{n \times n}$ be any matrix with* inc(B) *defined as in (3.14), and $\xi(B)$ defined as in (3.1). We have that*

$$\text{inc}(B) \leq \xi(B) \leq 2\ \text{inc}(B).$$

If $B \in \mathbb{R}^{n \times n}$ is a full-rank matrix or a matrix such as $e_1 e_1^T$, then $\xi(B) = 1$. Therefore, a bound on the incoherence of the row/column spaces of B is important in order to bound ξ. Using Propositions 3.4.2 and 3.4.3 along with Theorem 3.4.1 we have the following corollary, which states that sparse bounded-degree matrices and low-rank matrices with incoherent row/column spaces can be uniquely decomposed.

Corollary 3.4.1. *Let $C = A^\star + B^\star$ with* $\text{deg}_{\max}(A^\star)$ *being the maximum number of nonzero entries per row/column of $A^\star$ and* inc$(B^\star)$ *being the maximum incoherence of the row/column spaces of $B^\star$ (as defined by (3.14)). If we have that*

$$\text{deg}_{\max}(A^\star)\ \text{inc}(B^\star) < \frac{1}{12},$$

then the unique optimum of the convex program (3.3) is $(\hat{A}, \hat{B}) = (A^\star, B^\star)$ for a range of values of γ:

$$\gamma \in \left(\frac{2\ \text{inc}(B^\star)}{1 - 8\ \text{deg}_{\max}(A^\star)\ \text{inc}(B^\star)},\ \frac{1 - 6\ \text{deg}_{\max}(A^\star)\ \text{inc}(B^\star)}{\text{deg}_{\max}(A^\star)} \right). \tag{3.15}$$

Specifically $\gamma = \frac{(6\ \mathrm{inc}(B^\star))^p}{(2\ \deg_{\max}(A^\star))^{1-p}}$ *for any choice of* $p \in [0,1]$ *is always inside the above range, and thus guarantees exact recovery of* $(A^\star, B^\star)$.

We emphasize that this is a result with *deterministic* sufficient conditions on exact decomposability.

■ 3.4.4 Decomposing random sparse and low-rank matrices

Next we show that sparse and low-rank matrices drawn from certain natural random ensembles satisfy the sufficient conditions of Corollary 3.4.1 with high probability. We first consider random sparse matrices with a fixed number of nonzero entries.

Random sparsity model The matrix $A^\star$ is such that $\mathrm{support}(A^\star)$ is chosen uniformly at random from the collection of all support sets of size m. There is no assumption made about the values of $A^\star$ at locations specified by $\mathrm{support}(A^\star)$.

Lemma 3.4.1. *Suppose that* $A^\star \in \mathbb{R}^{n\times n}$ *is drawn according to the random sparsity model with* m *nonzero entries. Let* $\deg_{\max}(A^\star)$ *be the maximum number of nonzero entries in each row/column of* $A^\star$. *We have that*

$$\deg_{\max}(A^\star) \leq \frac{m}{n}\log n,$$

with probability greater than $1 - \mathcal{O}(n^{-\alpha})$ *for* $m = \mathcal{O}(\alpha n)$.

The proof of this lemma follows from a standard balls and bins argument, and can be found in several references (see for example [19]).

Next we consider low-rank matrices in which the singular vectors are chosen uniformly at random from the set of all partial isometries. Such a model was considered in recent work on the matrix completion problem [30], which aims to recover a low-rank matrix given observations of a subset of entries of the matrix.

Random orthogonal model [30] A rank-k matrix $B^\star \in \mathbb{R}^{n\times n}$ with SVD $B^\star = U\Sigma V'$ is constructed as follows: The singular vectors $U, V \in \mathbb{R}^{n\times k}$ are drawn *uniformly* at random from the collection of rank-k partial isometries in $\mathbb{R}^{n\times k}$. The choices of U and V need not be mutually independent. No restriction is placed on the singular values.

As shown in [30], low-rank matrices drawn from such a model have incoherent row/column spaces.

Lemma 3.4.2. *Suppose that a rank-k matrix $B^\star \in \mathbb{R}^{n\times n}$ is drawn according to the random orthogonal model. Then we have that that* $\mathrm{inc}(B^\star)$ *(defined by (3.14)) is bounded as*

$$\mathrm{inc}(B^\star) \lesssim \sqrt{\frac{\max(k, \log n)}{n}},$$

with probability greater than $1 - \mathcal{O}(n^{-3}\log n)$.

Applying these two results in conjunction with Corollary 3.4.1, we have that sparse and low-rank matrices drawn from the random sparsity model and the random orthogonal model can be uniquely decomposed with high probability.

Corollary 3.4.2. *Suppose that a rank-k matrix $B^\star \in \mathbb{R}^{n\times n}$ is drawn from the random orthogonal model, and that $A^\star \in \mathbb{R}^{n\times n}$ is drawn from the random sparsity model with m nonzero entries. Given $C = A^\star + B^\star$, there exists a range of values for γ (given by (3.15)) so that $(\hat{A}, \hat{B}) = (A^\star, B^\star)$ is the unique optimum of the SDP (3.3) with high probability (given by the bounds in Lemma 3.4.1 and Lemma 3.4.2) provided*

$$m \lesssim \frac{n^{1.5}}{\log n\sqrt{\max(k, \log n)}}.$$

In particular, $\gamma \sim \left(\frac{\max(k,\log n)}{m \log n}\right)^{\frac{1}{3}}$ guarantees exact recovery of $(A^\star, B^\star)$.

Thus, for matrices $B^\star$ with rank k smaller than n the SDP (3.3) yields exact recovery with high probability even when the size of the support of $A^\star$ is super-linear in n.

Implications for the matrix rigidity problem Corollary 3.4.2 has implications for the matrix rigidity problem discussed in Section 3.2. Recall that $R_M(k)$ is the smallest number of entries of M that need to be changed to reduce the rank of M below k (the changes can be of arbitrary magnitude). A generic matrix $M \in \mathbb{R}^{n\times n}$ has rigidity $R_M(k) = (n-k)^2$ [138]. However, special structured classes of matrices can have low rigidity. Consider a matrix M formed by adding a sparse matrix drawn from the random sparsity model with support size $\mathcal{O}(\frac{n}{\log n})$, and a low-rank matrix drawn from the random orthogonal model with rank ϵn for some fixed $\epsilon > 0$. Such a matrix has rigidity $R_M(\epsilon n) = \mathcal{O}(\frac{n}{\log n})$, and one can recover the sparse and low-rank components that compose M with high probability by solving the SDP (3.3). To see this, note that

$$\frac{n}{\log n} \lesssim \frac{n^{1.5}}{\log n\sqrt{\max(\epsilon n, \log n)}} = \frac{n^{1.5}}{\log n\sqrt{\epsilon n}},$$

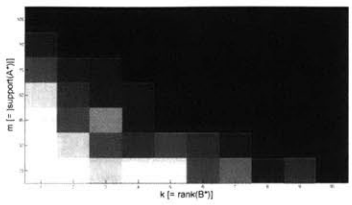

Figure 3.2. For each value of m, k, we generate 25×25 random m-sparse $A^\star$ and random rank-k $B^\star$ and attempt to recover $(A^\star, B^\star)$ from $C = A^\star + B^\star$ using (3.3). For each value of m, k we repeated this procedure 10 times. The figure shows the probability of success in recovering $(A^\star, B^\star)$ using (3.3) for various values of m and k. White represents a probability of success of 1, while black represents a probability of success of 0.

which satisfies the sufficient condition of Corollary 3.4.2 for exact recovery. Therefore, while the rigidity of a matrix is intractable to compute in general [38, 101], for such low-rigidity matrices M one can compute the rigidity $R_M(\epsilon n)$; in fact the SDP (3.3) provides a certificate of the sparse and low-rank matrices that form the low rigidity matrix M.

■ 3.5 Simulation Results

We confirm the theoretical predictions in this chapter with some simple experimental results. We also present a heuristic to choose the trade-off parameter γ. All our simulations were performed using YALMIP [98] and the SDPT3 software [136] for solving SDPs.

In the first experiment we generate random 25×25 matrices according to the random sparsity and random orthogonal models described in Section 3.4.4. To generate a random rank-k matrix $B^\star$ according to the random orthogonal model, we generate $X, Y \in \mathbb{R}^{25 \times k}$ with i.i.d. Gaussian entries and set $B^\star = XY^T$. To generate an m-sparse matrix $A^\star$ according to the random sparsity model, we choose a support set of size m uniformly at random and the values within this support are i.i.d. Gaussian. The goal is to recover $(A^\star, B^\star)$ from $C = A^\star + B^\star$ using the SDP (3.3). Let tol_γ be defined as:

$$\text{tol}_\gamma = \frac{\|\hat{A} - A^\star\|_F}{\|A^\star\|_F} + \frac{\|\hat{B} - B^\star\|_F}{\|B^\star\|_F}, \tag{3.16}$$

where $(\hat{A}, \hat{B})$ is the solution of (3.3), and $\|\cdot\|_F$ is the Frobenius norm. We declare success in recovering $(A^\star, B^\star)$ if $\text{tol}_\gamma < 10^{-3}$. (We discuss the issue of choosing γ in the

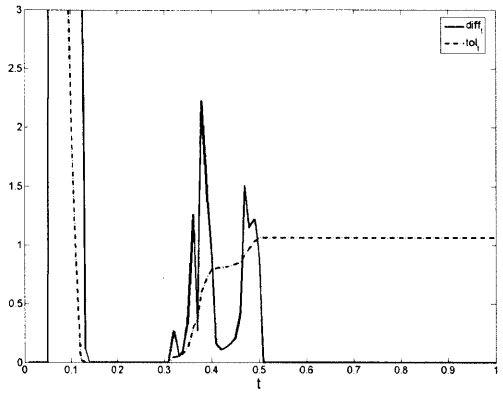

Figure 3.3. Comparison between tol_t and diff_t for a randomly generated example with $n = 25, m = 25, k = 2$.

next experiment.) Figure 3.2 shows the success rate in recovering $(A^\star, B^\star)$ for various values of m and k (averaged over 10 experiments for each m, k). Thus we see that one can recover sufficiently sparse $A^\star$ and sufficiently low-rank $B^\star$ from $C = A^\star + B^\star$ using (3.3).

Next we consider the problem of choosing the trade-off parameter γ. Based on Theorem 3.4.1 we know that exact recovery is possible for a *range* of γ. Therefore, one can simply check the stability of the solution $(\hat{A}, \hat{B})$ as γ is varied without knowing the appropriate range for γ in advance. To formalize this scheme we consider the following SDP for $t \in [0, 1]$, which is a slightly modified version of (3.3):

$$\begin{aligned} (\hat{A}_t, \hat{B}_t) = \arg\min_{A,B} \quad & t\|A\|_1 + (1-t)\|B\|_* \\ \text{s.t.} \quad & A + B = C. \end{aligned} \tag{3.17}$$

There is a one-to-one correspondence between (3.3) and (3.17) given by $t = \frac{\gamma}{1+\gamma}$. The benefit in looking at (3.17) is that the range of valid parameters is compact, i.e., $t \in [0, 1]$, as opposed to the situation in (3.3) where $\gamma \in [0, \infty)$. We compute the difference between solutions for some t and $t - \epsilon$ as follows:

$$\text{diff}_t = (\|\hat{A}_{t-\epsilon} - \hat{A}_t\|_F) + (\|\hat{B}_{t-\epsilon} - \hat{B}_t\|_F), \tag{3.18}$$

where $\epsilon > 0$ is some small fixed constant, say $\epsilon = 0.01$. We generate a random $A^\star \in \mathbb{R}^{25\times 25}$ that is 25-sparse and a random $B^\star \in \mathbb{R}^{25\times 25}$ with rank $= 2$ as described above.

Given $C = A^\star + B^\star$, we solve (3.17) for various values of t. Figure 3.3 shows two curves – one is tol_t (which is defined analogous to tol_γ in (3.16)) and the other is diff_t. Clearly we do not have access to tol_t in practice. However, we see that diff_t is near zero in exactly three regions. For sufficiently small t the optimal solution to (3.17) is $(\hat{A}_t, \hat{B}_t) = (A^\star + B^\star, 0)$, while for sufficiently large t the optimal solution is $(\hat{A}_t, \hat{B}_t) = (0, A^\star + B^\star)$. As seen in the figure, diff_t stabilizes for small and large t. The third "middle" range of stability is where we typically have $(\hat{A}_t, \hat{B}_t) = (A^\star, B^\star)$. Notice that outside of these three regions diff_t is not close to 0 and in fact changes rapidly. Therefore if a reasonable guess for t (or γ) is not available, one could solve (3.17) for a range of t and choose a solution corresponding to the "middle" range in which diff_t is stable and near zero. A related method to check for stability is to compute the sensitivity of the cost of the optimal solution with respect to γ, which can be obtained from the dual solution.

■ 3.6 Discussion

We have studied the problem of exactly decomposing a given matrix $C = A^\star + B^\star$ into its sparse and low-rank components $A^\star$ and $B^\star$. This problem arises in a number of applications in model selection, system identification, complexity theory, and optics. We characterized fundamental identifiability in the decomposition problem based on a notion of rank-sparsity incoherence, which relates the sparsity pattern of a matrix and its row/column spaces via an uncertainty principle. As the general decomposition problem is intractable to solve, we propose a natural SDP relaxation (3.3) to solve the problem, and provide sufficient conditions on sparse and low-rank matrices so that the SDP exactly recovers such matrices. Our sufficient conditions are deterministic in nature; they essentially require that the sparse matrix must have support that is not too concentrated in any row/column, while the low-rank matrix must have row/column spaces that are not closely aligned with the coordinate axes. Our analysis centers around studying the tangent spaces with respect to the algebraic varieties of sparse and low-rank matrices. Indeed the sufficient conditions for identifiability and for exact recovery using the SDP can also be viewed as requiring that certain tangent spaces have a transverse intersection. The implications of our results for the matrix rigidity problem are also demonstrated. An interesting problem for further research is the development of special-purpose algorithms that take advantage of structure in (3.3) to provide a

more efficient solution than a general-purpose SDP solver.

Chapter 4

Latent Variable Graphical Model Selection via Convex Optimization

4.1 Introduction

Statistical model selection in the high-dimensional regime arises in a number of applications. In many data analysis problems in geophysics, radiology, genetics, climate studies, and image processing, the number of samples available is comparable to or even smaller than the number of variables. However, it is well-known that empirical statistics such as sample covariance matrices are not well-behaved when both the number of samples and the number of variables are large and comparable to each other (see [103]). Model selection in such a setting is therefore both challenging and of great interest. In order for model selection to be well-posed given limited information, a key assumption that is often made is that the underlying model to be estimated only has *a few degrees of freedom.* Common assumptions are that the data are generated according to a graphical model, or a stationary time-series model, or a simple factor model with a few latent variables. Sometimes geometric assumptions are also made in which the data are viewed as samples drawn according to a distribution supported on a low-dimensional manifold.

A model selection problem that has received considerable attention recently is the estimation of covariance matrices in the high-dimensional setting. As the sample covariance matrix is poorly behaved in such a regime [85,103], some form of *regularization* of the sample covariance is adopted based on assumptions about the true underlying covariance matrix. For example approaches based on banding the sample covariance matrix [15] have been proposed for problems in which the variables have a natural ordering (e.g., times series), while "permutation-invariant" methods that use thresholding

are useful when there is no natural variable ordering [16, 61]. These approaches provide consistency guarantees under various sparsity assumptions on the true covariance matrix. Other techniques that have been studied include methods based on shrinkage [94, 145] and factor analysis [62]. A number of papers have studied covariance estimation in the context of *Gaussian graphical model selection.* In a Gaussian graphical model the *inverse* of the covariance matrix, also called the concentration matrix, is assumed to be sparse, and the sparsity pattern reveals the conditional independence relations satisfied by the variables. The model selection method usually studied in such a setting is ℓ_1-regularized maximum-likelihood, with the ℓ_1 penalty applied to the entries of the inverse covariance matrix to induce sparsity. The consistency properties of such an estimator have been studied [92,119,126], and under suitable conditions [92,119] this estimator is also "sparsistent", i.e., the estimated concentration matrix has the same sparsity pattern as the true model from which the samples are generated. An alternative approach to ℓ_1-regularized maximum-likelihood is to estimate the sparsity pattern of the concentration matrix by performing regression separately on each variable [107]; while such a method consistently estimates the sparsity pattern, it does not directly provide estimates of the covariance or concentration matrix.

In many applications throughout science and engineering, a challenge is that one may not have access to observations of all the relevant phenomena, i.e., some of the relevant variables may be hidden or unobserved. Such a scenario arises in data analysis tasks in psychology, computational biology, and economics. In general latent variables pose a significant difficulty for model selection because one may not know the number of relevant latent variables, nor the relationship between these variables and the observed variables. Typical algorithmic methods that try to get around this difficulty usually fix the number of latent variables as well as the some structural relationship between latent and observed variables (e.g., the graphical model structure between latent and observed variables), and use the EM algorithm to fit parameters [44]. This approach suffers from the problem that one optimizes non-convex functions, and thus one may get stuck in sub-optimal local minima. An alternative method that has been suggested is based on a greedy, local, combinatorial heuristic that assigns latent variables to groups of observed variables, based on some form of clustering of the observed variables [60]; however, this approach has no consistency guarantees.

In this chapter we study the problem of latent-variable graphical model selection in the setting where all the variables, both observed and hidden, are jointly Gaussian.

More concretely let the covariance matrix of a finite collection of jointly Gaussian random variables $X_O \cup X_H$ be denoted by $\Sigma_{(O\ H)}$, where X_O are the observed variables and X_H are the unobserved, hidden variables. The marginal statistics corresponding to the observed variables X_O are given by the marginal covariance matrix Σ_O, which is simply a submatrix of the full covariance matrix $\Sigma_{(O\ H)}$. However suppose that we parameterize our model by the concentration matrix $K_{(O\ H)} = \Sigma_{(O\ H)}^{-1}$, which as discussed above reveals the connection to graphical models. In such a parametrization, the *marginal concentration matrix* Σ_O^{-1} corresponding to the observed variables X_O is given by the Schur complement [82] with respect to the block K_H:

$$\tilde{K}_O = \Sigma_O^{-1} = K_O - K_{O,H} K_H^{-1} K_{H,O}.$$

Thus if we only observe the variables X_O, we only have access to Σ_O (or $\tilde{K}_O$). The two terms that compose $\tilde{K}_O$ above have interesting properties. The matrix K_O specifies the concentration matrix of the *conditional statistics* of the observed variables given the latent variables. If these conditional statistics are given by a sparse graphical model then K_O is *sparse*. On the other hand the matrix $K_{O,H} K_H^{-1} K_{H,O}$ serves as a *summary* of the effect of marginalization over the hidden variables H. This matrix has small rank if the number of latent, unobserved variables H is small relative to the number of observed variables O (the rank is equal to $|H|$). Therefore the marginal concentration matrix $\tilde{K}_O$ of the observed variables X_O is generally *not sparse* due to the additional low-rank term $K_{O,H} K_H^{-1} K_{H,O}$. Hence standard graphical model selection techniques applied directly to the observed variables X_O are not useful.

A modeling paradigm that infers the effect of the latent variables X_H would be more suitable in order to provide a simple explanation of the underlying statistical structure. Hence we *decompose* $\tilde{K}_O$ into the sparse and low-rank components, which reveals the conditional graphical model structure in the observed variables as well as the *number* of and effect due to the unobserved latent variables. Such a method can be viewed as a blend of principal component analysis and graphical modeling. In standard graphical modeling one would directly approximate a concentration matrix by a sparse matrix in order to learn a sparse graphical model. On the other hand in principal component analysis the goal is to explain the statistical structure underlying a set of observations using a small number of latent variables (i.e., approximate a covariance matrix as a low-rank matrix). In our framework based on decomposing a concentration matrix, we learn a graphical model among the observed variables *conditioned* on a few (additional)

latent variables. Notice that in our setting these latent variables are *not* principal components, as the conditional statistics (conditioned on these latent variables) are given by a graphical model. Therefore we refer to these latent variables informally as *hidden components.*

Our first contribution in Section 4.3 is to address the fundamental question of *identifiability* of such latent-variable graphical models given the marginal statistics of only the observed variables. The critical point is that we need to tease apart the correlations induced due to marginalization over the latent variables from the conditional graphical model structure among the observed variables. As the identifiability problem is one of *uniquely* decomposing the sum of a sparse matrix and a low-rank matrix into the individual components, we recall the conditions derived in Chapter 3 that relate unique identifiability to properties of the tangent spaces to the algebraic varieties of sparse and low-rank matrices. Specifically let $\Omega(K_O)$ denote the tangent space at K_O to the algebraic variety of sparse matrices, and let $T(K_{O,H}K_H^{-1}K_{H,O})$ denote the tangent space at $K_{O,H}K_H^{-1}K_{H,O}$ to the algebraic variety of low-rank matrices. Then the *statistical* question of identifiability of K_O and $K_{O,H}K_H^{-1}K_{H,O}$ given $\tilde{K}_O$ is determined by the *geometric* notion of *transversality* of the tangent spaces $\Omega(K_O)$ and $T(K_{O,H}K_H^{-1}K_{H,O})$. The study of the transversality of these tangent spaces leads us to natural conditions for identifiability. In particular we show that latent-variable models in which (1) the sparse matrix K_O has a small number of nonzeros per row/column, and (2) the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$ has row/column spaces that are not closely aligned with the coordinate axes, are identifiable. These two conditions have natural statistical interpretations. The first condition ensures that there are no densely-connected subgraphs in the conditional graphical model structure among the observed variables X_O given the hidden components, i.e., that these conditional statistics are indeed specified by a sparse graphical model. Such statistical relationships may otherwise be mistakenly attributed to the effect of marginalization over some latent variable. The second condition ensures that the effect of marginalization over the latent variables is "spread out" over many observed variables; thus, the effect of marginalization over a latent variable is not confused with the conditional graphical model structure among the observed variables. In fact the first condition is often assumed in some papers on standard graphical model selection without latent variables (see for example [119]). We note here that question of parameter identifiability was recently studied for models with discrete-valued latent variables (i.e., mixture models, hidden Markov models) [2]. However, this work is not

applicable to our setting in which both the latent and observed variables are assumed to be jointly Gaussian.

As our next contribution we propose a *regularized maximum-likelihood decomposition* framework to approximate a given sample covariance matrix by a model in which the concentration matrix decomposes into a sparse matrix and a low-rank matrix. Motivated by the combined ℓ_1 norm and nuclear norm heuristic proposed in Chapter 3 for sparse/low-rank matrix decomposition, we propose the following penalized likelihood method given a sample covariance matrix Σ_O^n formed from n samples of the observed variables:

$$\begin{aligned}(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \quad & -\ell(S-L;\Sigma_O^n) \quad + \quad \lambda_n \left(\gamma\|S\|_1 + \mathrm{Tr}(L)\right) \\ \text{s.t.} \quad & S - L \succ 0, \quad L \succeq 0.\end{aligned} \tag{4.1}$$

Here ℓ represents the Gaussian log-likelihood function and is given by $\ell(K;\Sigma) = \log\det(K) - \mathrm{Tr}(K\Sigma)$ for $K \succ 0$, where Tr is the trace of a matrix and det is the determinant. The matrix $\hat{S}_n$ provides an estimate of K_O, which represents the conditional concentration matrix of the observed variables; the matrix $\hat{L}_n$ provides an estimate of $K_{O,H}K_H^{-1}K_{H,O}$, which represents the effect of marginalization over the latent variables. Notice that the regularization function is a combination of the ℓ_1 norm applied to S and the nuclear norm applied to L (the nuclear norm reduces to the trace over the cone of symmetric, positive-semidefinite matrices), with γ providing a tradeoff between the two terms. This variational formulation is a *convex optimization* problem. In particular it is a regularized max-det problem and can be solved in polynomial time using standard off-the-shelf solvers.

Our main result in Section 4.4 is a proof of the consistency of the estimator (4.1) in the high-dimensional regime in which both the number of observed variables and the number of hidden components are allowed to grow with the number of samples (of the observed variables). We show that for a suitable choice of the regularization parameter λ_n, there exists a range of values of γ for which the estimates $(\hat{S}_n, \hat{L}_n)$ have the same sparsity (and sign) pattern and rank as $(K_O, K_{O,H}(K_H)^{-1}K_{H,O})$ with high probability (see Theorem 4.4.1). The key technical requirement is an identifiability condition for the two components of the marginal concentration matrix $\tilde{K}_O$ with respect to the Fisher information (see Section 4.3.2). We make connections between our condition and the irrepresentability conditions required for support/graphical-model recovery using ℓ_1 regularization [119, 148]. Our results provide numerous scaling regimes under which

consistency holds in latent-variable graphical model selection. For example we show that under suitable identifiability conditions consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables (see Section 4.4.3).

Related previous work The problem of decomposing the sum of a sparse matrix and a low-rank matrix, with no additional noise, into the individual components was initially studied in [37]; the results of that paper are described in Chapter 3. In subsequent work Candès et al. [31] also studied this noise-free sparse-plus-low-rank decomposition problem, and provided guarantees for exact recovery using the convex program proposed in [37]. The problem setup considered in the present chapter is quite different and is more challenging because we are only given access to an inexact sample covariance matrix, and we are interested in recovering components that preserve both the sparsity pattern and the rank of the components in the true underlying model. In addition to proving such a consistency result for the estimator (4.1), we also provide a statistical interpretation of our identifiability conditions and describe natural classes of latent-variable Gaussian graphical models that satisfy these conditions. As such our work is closer in spirit to the many recent papers on covariance selection, but with the important difference that some of the variables are not observed.

Outline Section 4.2 gives some background on graphical models as well as the algebraic varieties of sparse and low-rank matrices. It also provides a formal statement of the problem. Section 4.3 discusses conditions under which latent-variable models are identifiable, and Section 4.4 states the main results of this chapter. We provide experimental demonstration of the consistency of our estimator on synthetic data in Section 4.5. Section 4.6 concludes the chapter with a brief discussion. Appendix B include additional details and proofs of all of our technical results.

■ 4.2 Background and Problem Statement

We briefly discuss concepts from graphical modeling and give a formal statement of the latent-variable model selection problem. We also describe various properties of the algebraic varieties of sparse matrices and of low-rank matrices. Although some of these have been introduced previously, we emphasize again that the following matrix norms are employed throughout this chapter:

- $\|M\|_2$: denotes the spectral norm, which is the largest singular value of M.

- $\|M\|_\infty$: denotes the largest entry in magnitude of M.
- $\|M\|_F$: denotes the Frobenius norm, which is the square-root of the sums of the squares of the entries of M.
- $\|M\|_*$: denotes the nuclear norm, which is the sum of the singular values of M. This reduces to the trace for positive-semidefinite matrices.
- $\|M\|_1$: denotes the sum of the absolute values of the entries of M.

A number of *matrix operator* norms are also used. For example, let $\mathcal{Z} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be a linear operator acting on matrices. Then the induced operator norm $\|\mathcal{Z}\|_{q\to q}$ is defined as:

$$\|\mathcal{Z}\|_{q\to q} \triangleq \max_{N\in\mathbb{R}^{p\times p},\ \|N\|_q\leq 1} \|\mathcal{Z}(N)\|_q. \tag{4.2}$$

Therefore, $\|\mathcal{Z}\|_{F\to F}$ denotes the spectral norm of the matrix operator $\mathcal{Z}$. The only vector norm used is the Euclidean norm, which is denoted by $\|\cdot\|$.

■ 4.2.1 Gaussian graphical models with latent variables

A graphical model [93] is a statistical model defined with respect to a graph $(V, \mathcal{E})$ in which the nodes index a collection of random variables $\{X_v\}_{v\in V}$, and the edges represent the conditional independence relations (Markov structure) among the variables. The absence of an edge between nodes $i, j \in V$ implies that the variables X_i, X_j are independent conditioned on all the other variables. A *Gaussian graphical model* (also commonly referred to as a Gauss-Markov random field) is one in which all the variables are jointly Gaussian [132]. In such models the sparsity pattern of the inverse of the covariance matrix, or the *concentration* matrix, directly corresponds to the graphical model structure. Specifically, consider a Gaussian graphical model in which the covariance matrix is given by $\Sigma \succ 0$ and the concentration matrix is given by $K = \Sigma^{-1}$. Then an edge $\{i, j\} \in \mathcal{E}$ is present in the underlying graphical model if and only if $K_{i,j} \neq 0$.

Our focus in this chapter is on Gaussian models in which some of the variables may not be observed. Suppose O represents the set of nodes corresponding to observed variables X_O, and H the set of nodes corresponding to unobserved, hidden variables X_H with $O \cup H = V$ and $O \cap H = \emptyset$. The joint covariance is denoted by $\Sigma_{(O\ H)}$, and joint concentration matrix by $K_{(O\ H)} = \Sigma^{-1}_{(O\ H)}$. The submatrix Σ_O represents the marginal covariance of the observed variables X_O, and the corresponding marginal concentration

matrix is given by the Schur complement with respect to the block K_H:

$$\tilde{K}_O = \Sigma_O^{-1} = K_O - K_{O,H}K_H^{-1}K_{H,O}. \tag{4.3}$$

The submatrix K_O specifies the concentration matrix of the conditional statistics of the observed variables conditioned on the hidden components. If these conditional statistics are given by a sparse graphical model then K_O is sparse. On the other hand the marginal concentration matrix $\tilde{K}_O$ of the marginal distribution of X_O is *not* sparse in general due to the extra correlations induced from marginalization over the latent variables X_H, i.e., due to the presence of the additional term $K_{O,H}K_H^{-1}K_{H,O}$. Hence, standard graphical model selection techniques in which the goal is to approximate a sample covariance by a sparse graphical model are not well-suited for problems in which some of the variables are hidden. However, the matrix $K_{O,H}K_H^{-1}K_{H,O}$ is a low-rank matrix if the number of hidden variables is much smaller than the number of observed variables (i.e., $|H| \ll |O|$). Therefore, a more appropriate model selection method is to approximate the sample covariance by a model in which the concentration matrix decomposes into the sum of a sparse matrix and a low-rank matrix. The objective here is to learn a sparse graphical model among the observed variables *conditioned* on some latent variables, as such a model explicitly accounts for the extra correlations induced due to unobserved, hidden components.

■ 4.2.2 Problem statement

In order to analyze latent-variable model selection methods, we need to define an appropriate notion of model selection consistency for latent-variable graphical models. Notice that given the two components K_O and $K_{O,H}K_H^{-1}K_{H,O}$ of the concentration matrix of the marginal distribution (4.3), there are *infinitely* many configurations of the latent variables (i.e., matrices $K_H \succ 0, K_{O,H} = K_{H,O}^T$) that give rise to the *same* low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$. Specifically for any non-singular matrix $B \in \mathbb{R}^{|H|\times|H|}$, one can apply the transformations $K_H \to BK_HB^T, K_{O,H} \to K_{O,H}B^T$ and still preserve the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$. In *all* of these models the marginal statistics of the observed variables X_O remain the same upon marginalization over the latent variables X_H. The key *invariant* is the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$, which *summarizes* the effect of marginalization over the latent variables. These observations give rise to the following notion of consistency:

Definition 4.2.1. *A pair of (symmetric) matrices (S, L) with $S, L \in \mathbb{R}^{|O|\times|O|}$ is an*

algebraically consistent *estimate of a latent-variable Gaussian graphical model given by the concentration matrix* $K_{(O\ H)}$ *if the following conditions hold:*

1. *The sign-pattern of* S *is the same as that of* K_O*:*

$$\text{sign}(S_{i,j}) = \text{sign}((K_O)_{i,j}), \quad \forall (i,j).$$

Here we assume that $\text{sign}(0) = 0$.

2. *The rank of* L *is the same as the rank of* $K_{O,H}K_H^{-1}K_{H,O}$*:*

$$\text{rank}(L) = \text{rank}(K_{O,H}K_H^{-1}K_{H,O}).$$

3. *The concentration matrix* $S - L$ *can be realized as the marginal concentration matrix of an appropriate latent-variable model:*

$$S - L \succ 0, \qquad L \succeq 0.$$

The first condition ensures that S provides the correct structural estimate of the conditional graphical model (given by K_O) of the observed variables conditioned on the hidden components. This property is the same as the "sparsistency" property studied in standard graphical model selection [92, 119]. The second condition ensures that the number of hidden components is correctly estimated. Finally, the third condition ensures that the pair of matrices (S, L) leads to a realizable latent-variable model. In particular this condition implies that there exists a valid latent-variable model on $|O \cup H|$ variables in which (a) the conditional graphical model structure among the observed variables is given by S, (b) the number of latent variables $|H|$ is equal to the rank of L, and (c) the extra correlations induced due to marginalization over the latent variables is equal to L. Any method for matrix factorization (see for e.g., [143]) can be used to factorize the low-rank matrix L, depending on the properties that one desires in the factors (e.g., sparsity).

We also study parametric consistency in the usual sense, i.e., we show that one can produce estimates (S, L) that converge in various norms to the matrices $(K_O, K_{O,H}K_H^{-1}K_{H,O})$ as the number of samples available goes to infinity. Notice that proving (S, L) is close to $(K_O, K_{O,H}K_H^{-1}K_{H,O})$ in some norm does not in general imply that the support/sign-pattern and rank of (S, L) are the same as those of $(K_O, K_{O,H}K_H^{-1}K_{H,O})$. Therefore parametric consistency is different from algebraic consistency, which requires that (S, L) have the same support/sign-pattern and rank as $(K_O, K_{O,H}K_H^{-1}K_{H,O})$.

Goal Let $K^*_{(O\ H)}$ denote the concentration matrix of a Gaussian model. Suppose that we have n samples $\{X_O^i\}_{i=1}^n$ of the observed variables X_O. We would like to produce estimates $(\hat{S}_n, \hat{L}_n)$ that, with high-probability, are both algebraically consistent and consistent in the parametric sense (in some norm).

■ 4.2.3 Likelihood function and Fisher information

Given n samples $\{X^i\}_{i=1}^n$ of a finite collection of jointly Gaussian zero-mean random variables with concentration matrix K^*, we define the sample covariance as follows:

$$\Sigma^n \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T. \tag{4.4}$$

It is then easily seen that the log-likelihood function is given by:

$$\ell(K; \Sigma^n) = \log\det(K) - \mathrm{Tr}(K\Sigma^n), \tag{4.5}$$

where $\ell(K; \Sigma_n)$ is a function of K. Notice that this function is strictly concave for $K \succ 0$. Now consider the latent-variable modeling problem in which we wish to model a collection of random variables X_O (with sample covariance Σ_O^n) by adding some extra variables X_H. With respect to the parametrization (S, L) (with S representing the conditional statistics of X_O given X_H, and L summarizing the effect of marginalization over the additional variables X_H), the likelihood function is given by:

$$\bar{\ell}(S, L; \Sigma_O^n) = \ell(S\ -\ L; \Sigma_O^n).$$

The function $\bar{\ell}$ is *jointly concave* with respect to the parameters (S, L) whenever $S - L \succ 0$, and it is this function that we use in our variational formulation (4.1) to learn a latent-variable model.

In the analysis of a convex program involving the likelihood function, the Fisher information plays an important role as it is the negative of the Hessian of the likelihood function and thus controls the curvature. As the first term in the likelihood function is linear, we need only study higher-order derivatives of the log-determinant function in order to compute the Hessian. Letting $\mathcal{I}$ denote the Fisher information matrix, we have that [24]

$$\mathcal{I}(K^*) \triangleq -\nabla_K^2 \log\det(K)|_{K=K^*} = (K^*)^{-1} \otimes (K^*)^{-1},$$

for $K^* \succ 0$. If K^* is a $p \times p$ concentration matrix, then the Fisher information matrix $\mathcal{I}(K^*)$ has dimensions $p^2 \times p^2$. Next consider the latent-variable situation with the

variables indexed by O being observed and the the variables indexed by H being hidden. The concentration matrix $\tilde{K}^*_O = (\Sigma^*_O)^{-1}$ of the marginal distribution of the observed variables O is given by the Schur complement (4.3), and the corresponding Fisher information matrix is given by

$$\mathcal{I}(\tilde{K}^*_O) = (\tilde{K}^*_O)^{-1} \otimes (\tilde{K}^*_O)^{-1} = \Sigma^*_O \otimes \Sigma^*_O.$$

Notice that this is precisely the $|O|^2 \times |O|^2$ submatrix of the full Fisher information matrix $\mathcal{I}(K^*_{(O\ H)}) = \Sigma^*_{(O\ H)} \otimes \Sigma^*_{(O\ H)}$ with respect to all the parameters $K^*_{(O\ H)} = (\Sigma^*_{(O\ H)})^{-1}$ (corresponding to the situation in which *all* the variables $X_{O \cup H}$ are observed). The matrix $\mathcal{I}(K^*_{(O\ H)})$ has dimensions $|O \cup H|^2 \times |O \cup H|^2$, while $\mathcal{I}(\tilde{K}^*_O)$ is an $|O|^2 \times |O|^2$ matrix. To summarize, we have for all $i, j, k, l \in O$ that:

$$\mathcal{I}(\tilde{K}^*_O)_{(i,j),(k,l)} = [\Sigma^*_{(O\ H)} \otimes \Sigma^*_{(O\ H)}]_{(i,j),(k,l)} = \mathcal{I}(K^*_{(O\ H)})_{(i,j),(k,l)}.$$

In Section 4.3.2 we impose various conditions on the Fisher information matrix $\mathcal{I}(\tilde{K}^*_O)$ under which our regularized maximum-likelihood formulation provides consistent estimates with high probability.

■ 4.2.4 Curvature of rank variety

Recall from Chapter 3 that $\mathcal{S}(k)$ denotes the algebraic variety of matrices with at most k nonzero entries, and that $\mathcal{L}(r)$ denotes the algebraic variety of matrices with rank at most r. The sparse matrix variety $\mathcal{S}(k)$ has the property that it has *zero* curvature at any smooth point. Consequently the tangent space at a smooth point M is the *same* as the tangent space at any point in a neighborhood of M. This property is implicitly used in the analysis of ℓ_1 regularized methods for recovering sparse models. The situation is more complicated for the low-rank matrix variety, because the curvature at any smooth point is nonzero. Therefore we need to study how the tangent space changes from one point to a neighboring point by analyzing how this variety curves locally. Indeed the amount of curvature at a point is directly related to the "angle" between the tangent space at that point and the tangent space at a neighboring point. For any subspace T of matrices, let $\mathcal{P}_T$ denote the projection onto T. Given two subspaces T_1, T_2 of the same dimension, we measure the "twisting" between these subspaces by considering the following quantity.

$$\rho(T_1, T_2) \triangleq \|\mathcal{P}_{T_1} - \mathcal{P}_{T_2}\|_{2 \to 2} = \max_{\|N\|_2 \leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2. \tag{4.6}$$

In Appendix B.1 we briefly review relevant results from matrix perturbation theory; the key tool used to derive these results is the resolvent of a matrix [87]. Based on these tools we prove the following two results in Appendix B.2, which bound the twisting between the tangent spaces at nearby points. The first result provides a bound on the quantity ρ between the tangent spaces at a point and at its neighbor.

Proposition 4.2.1. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest non-zero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\rho(T(M + \Delta), T(M)) \leq \frac{2}{\sigma} \|\Delta\|_2.$$

The next result bounds the error between a point and its neighbor in the normal direction.

Proposition 4.2.2. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest non-zero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\| \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\|\mathcal{P}_{T(M)^{\perp}}(\Delta)\|_2 \leq \frac{\|\Delta\|_2^2}{\sigma}.$$

These results suggest that the closer the smallest singular value is to zero, the more curved the variety is locally. Therefore we control the twisting between tangent spaces at nearby points by bounding the smallest singular value away from zero.

■ 4.3 Identifiability

In the absence of additional conditions, the latent-variable model selection problem is ill-posed. In this section we discuss a set of conditions on latent-variable models that ensure that these models are identifiable given marginal statistics for a subset of the variables. Recall that the identifiability conditions of Chapter 3 are directly applicable here, and we rephrase these in the context of latent-variable graphical models.

Structure between latent and observed variables

Suppose that the low-rank matrix that summarizes the effect of the hidden components is itself sparse. This leads to identifiability issues in the sparse-plus-low-rank decomposition problem. Statistically the additional correlations induced due to marginalization

over the latent variables could be mistaken for the conditional graphical model structure of the observed variables. In order to avoid such identifiability problems the effect of the latent variables must be "diffuse" across the observed variables. To address this point the quantity $\xi(T(M))$ was introduced in Chapter 3 (see also [37]) to measure the incoherence of the row/column spaces of M with respect to the standard basis.

Curvature and change in ξ: As noted previously an important technical point is that the algebraic variety of low-rank matrices is locally curved at any smooth point. Consequently the quantity ξ changes as we move along the low-rank matrix variety smoothly. The quantity $\rho(T_1, T_2)$ introduced in (4.6) also allows us to bound the variation in ξ as follows.

Lemma 4.3.1. *Let T_1, T_2 be two matrix subspaces of the same dimension with the property that $\rho(T_1, T_2) < 1$, where ρ is defined in* (4.6)*. Then we have that*

$$\xi(T_2) \leq \frac{1}{1 - \rho(T_1, T_2)} \left[\xi(T_1) + \rho(T_1, T_2)\right].$$

This lemma is proved in Appendix B.2.

Structure among observed variables

An identifiability problem also arises if the conditional graphical model among the observed variables contains a densely connected subgraph. These statistical relationships might be mistaken as correlations induced by marginalization over latent variables. Therefore we need to ensure that the conditional graphical model among the observed variables is sparse. We impose the condition that this conditional graphical model must have small "degree", i.e., no observed variable is directly connected to too many other observed variables conditioned on the hidden components. Notice that bounding the degree is a more refined condition than simply bounding the total number of non-zeros as the *sparsity pattern* also plays a role. As described in Chapter 3 (see also [37]), the quantity $\mu(\Omega(M))$ provides an appropriate measure of the sparsity pattern of a matrix for the purposes of unique identifiability.

■ 4.3.1 Transversality of tangent spaces

From Chapter 3 we recall that the transversality of the tangent spaces at the sparse and low-rank components with respect to the respective algebraic varieties governs their identifiability. In order to quantify the level of transversality between the tangent spaces

Ω and T we study the *minimum gain* with respect to some norm of the addition operator restricted to the cartesian product $\mathcal{Y} = \Omega \times T$. More concretely let $\mathcal{A} : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ represent the addition operator, i.e., the operator that adds two matrices. Then given any matrix norm $\|\cdot\|$ on $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, the minimum gain of $\mathcal{A}$ restricted to $\mathcal{Y}$ is defined as follows:

$$\epsilon(\Omega, T, \|\cdot\|) \triangleq \min_{(S,L)\in\Omega\times T,\ \|(S,L)\|=1} \|\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}\mathcal{A}\mathcal{P}_{\mathcal{Y}}(S, L)\|,$$

where $\mathcal{P}_{\mathcal{Y}}$ denotes the projection onto the space $\mathcal{Y}$, and $\mathcal{A}^{\dagger}$ denotes the adjoint of the addition operator (with respect to the standard Euclidean inner-product). The tangent spaces Ω and T have a *transverse* intersection if and only if $\epsilon(\Omega, T, \|\cdot\|) > 0$. The "level" of transversality is measured by the magnitude of $\epsilon(\Omega, T, \|\cdot\|)$. Note that if the norm $\|\cdot\|$ used is the Frobenius norm, then $\epsilon(\Omega, T, \|\cdot\|_F)$ is the square of the *minimum singular value* of the addition operator $\mathcal{A}$ restricted to $\Omega \times T$.

A natural norm with which to measure transversality is the dual norm of the regularization function in (4.1), as the subdifferential of the regularization function is specified in terms of its dual. The reasons for this will become clearer as we proceed through this chapter. Recall that the regularization function used in the variational formulation (4.1) is given by:

$$f_{\gamma}(S, L) = \gamma\|S\|_1 + \|L\|_*,$$

where the nuclear norm $\|\cdot\|_*$ reduces to the trace function over the cone of positive-semidefinite matrices. This function is a norm for all $\gamma > 0$. The dual norm of f_{γ} is given by

$$g_{\gamma}(S, L) = \max\left\{\frac{\|S\|_{\infty}}{\gamma}, \|L\|_2\right\}.$$

The following simple lemma records a useful property of the g_{γ} norm that is used several times throughout this chapter.

Lemma 4.3.2. *Let Ω and T be tangent spaces at any points with respect to the algebraic varieties of sparse and low-rank matrices. Then for any matrix M, we have that $\|\mathcal{P}_{\Omega}(M)\|_{\infty} \leq \|M\|_{\infty}$ and that $\|\mathcal{P}_{T}(M)\|_2 \leq 2\|M\|_2$. Further we also have that $\|\mathcal{P}_{\Omega^{\perp}}(M)\|_{\infty} \leq \|M\|_{\infty}$ and that $\|\mathcal{P}_{T^{\perp}}(M)\|_2 \leq \|M\|_2$. Thus for any matrices M, N and for $\mathcal{Y} = \Omega\times T$, one can check that $g_{\gamma}(\mathcal{P}_{\mathcal{Y}}(M, N)) \leq 2g_{\gamma}(M, N)$ and that $g_{\gamma}(\mathcal{P}_{\mathcal{Y}^{\perp}}(M, N)) \leq g_{\gamma}(M, N)$.*

Next we define the quantity $\chi(\Omega, T, \gamma)$ as follows in order to study the transversality of the spaces Ω and T with respect to the g_γ norm:

$$\chi(\Omega, T, \gamma) \triangleq \max\left\{\frac{\xi(T)}{\gamma}, 2\mu(\Omega)\gamma\right\} \tag{4.7}$$

Here μ and ξ are defined in Chapter 3. We then have the following result (proved in Appendix B.3):

Lemma 4.3.3. *Let $S \in \Omega, L \in T$ be matrices such that $\|S\|_\infty = \gamma$ and let $\|L\|_2 = 1$. Then we have that $g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S, L)) \in [1 - \chi(\Omega, T, \gamma), 1 + \chi(\Omega, T, \gamma)]$, where $\mathcal{Y} = \Omega \times T$ and $\chi(\Omega, T, \gamma)$ is defined in (4.7). In particular we have that $1 - \chi(\Omega, T, \gamma) \leq \epsilon(\Omega, T, g_\gamma)$.*

The quantity $\chi(\Omega, T, \gamma)$ being small implies that the addition operator is essentially isometric when restricted to $\mathcal{Y} = \Omega \times T$. Stated differently the magnitude of $\chi(\Omega, T, \gamma)$ is a measure of the level of transversality of the spaces Ω and T. If $\mu(\Omega)\xi(T) < \frac{1}{2}$ then $\gamma \in (\xi(T), \frac{1}{2\mu(\Omega)})$ ensures that $\chi(\Omega, T, \gamma) < 1$, which in turn implies that the tangent spaces Ω and T have a transverse intersection.

Observation: Thus we have that the smaller the quantities $\mu(\Omega)$ and $\xi(T)$, the more transverse the intersection of the spaces Ω and T.

■ 4.3.2 Conditions on Fisher information

The main focus of Section 4.4 is to analyze the regularized maximum-likelihood convex program (4.1) by studying its optimality conditions. The log-likelihood function is well-approximated in a neighborhood by a quadratic form given by the Fisher information (which measures the curvature, as discussed in Section 4.2.3). Let $\mathcal{I}^* = \mathcal{I}(\tilde{K}_O^*)$ denote the Fisher information evaluated at the true marginal concentration matrix $\tilde{K}_O^* = K_O^* - K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$, where $K_{(O\ H)}^*$ represents the concentration matrix of the full model (see equation (4.3)). The appropriate measure of transversality between the tangent spaces[1] $\Omega = \Omega(K_O^*)$ and $T = T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ is then in a space in which the inner-product is given by $\mathcal{I}^*$. Specifically, we need to analyze the minimum gain of the operator $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}$ restricted to the space $\mathcal{Y} = \Omega \times T$. Therefore we impose several conditions on the Fisher information $\mathcal{I}^*$. We define quantities that control the gains of $\mathcal{I}^*$ restricted to Ω and T separately; these ensure that elements of Ω and

[1] We implicitly assume that these tangent spaces are subspaces of the space of *symmetric* matrices.

elements of T are individually identifiable under the map $\mathcal{I}^*$. In addition we define quantities that, in conjunction with bounds on $\mu(\Omega)$ and $\xi(T)$, allow us to control the gain of $\mathcal{I}^*$ restricted to the direct-sum $\Omega \oplus T$.

$\mathcal{I}^*$ restricted to Ω: The minimum gain of the operator $\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega$ restricted to Ω is given by

$$\alpha_\Omega \triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The maximum effect of elements in Ω in the orthogonal direction $\Omega^\perp$ is given by

$$\delta_\Omega \triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The operator $\mathcal{I}^*$ is injective on Ω if $\alpha_\Omega > 0$. The ratio $\frac{\delta_\Omega}{\alpha_\Omega} \leq 1 - \nu$ implies the irrepresentability condition imposed in [119], which gives a sufficient condition for consistent recovery of graphical model structure using ℓ_1-regularized maximum-likelihood. Notice that this condition is a generalization of the usual Lasso irrepresentability conditions [148], which are typically imposed on the covariance matrix. Finally we also consider the following quantity, which controls the behavior of $\mathcal{I}^*$ restricted to Ω in the spectral norm:

$$\beta_\Omega \triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}^*(M)\|_2.$$

$\mathcal{I}^*$ restricted to T: Analogous to the case of Ω one could control the gains of the operators $\mathcal{P}_{T^\perp} \mathcal{I}^* \mathcal{P}_T$ and $\mathcal{P}_T \mathcal{I}^* \mathcal{P}_T$. However as discussed previously one complication is that the tangent spaces at nearby smooth points on the rank variety are in general different, and the amount of twisting between these spaces is governed by the local curvature. Therefore we control the gains of the operators $\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}$ and $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ for all tangent spaces T' that are "close to" the nominal T (at the true underlying low-rank matrix), measured by $\rho(T, T')$ (4.6) being small. The minimum gain of the operator $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ restricted to T' (close to T) is given by

$$\alpha_T \triangleq \min_{\rho(T',T) \leq \frac{\xi(T)}{2}} \min_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Similarly the maximum effect of elements in T' in the orthogonal direction $T'^\perp$ (for T' close to T) is given by

$$\delta_T \triangleq \max_{\rho(T',T) \leq \frac{\xi(T)}{2}} \max_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Implicit in the definition of α_T and δ_T is the fact that the outer minimum and maximum are only taken over spaces T' that are tangent spaces to the rank-variety. The operator $\mathcal{I}^*$ is injective on all tangent spaces T' such that $\rho(T', T) \leq \frac{\xi(T)}{2}$ if $\alpha_T > 0$. An irrepresentability condition (analogous to those developed for the sparse case) for tangent spaces near T to the rank variety would be that $\frac{\delta_T}{\alpha_T} \leq 1 - \nu$. Finally we also control the behavior of $\mathcal{I}^*$ restricted to T' close to T in the ℓ_∞ norm:

$$\beta_T \triangleq \max_{\rho(T',T) \leq \frac{\xi(T)}{2}} \max_{M \in T, \|M\|_\infty = 1} \|\mathcal{I}^*(M)\|_\infty.$$

The two sets of quantities $(\alpha_\Omega, \delta_\Omega)$ and (α_T, δ_T) essentially control how $\mathcal{I}^*$ behaves when restricted to the spaces Ω and T *separately* (in the natural norms). The quantities β_Ω and β_T are useful in order to control the gains of the operator $\mathcal{I}^*$ restricted to the *direct sum* $\Omega \oplus T$. Notice that although the magnitudes of elements in Ω are measured most naturally in the ℓ_∞ norm, the quantity β_Ω is specified with respect to the spectral norm. Similarly elements of the tangent spaces T' to the rank variety are most naturally measured in the spectral norm, but β_T provides control in the ℓ_∞ norm. These quantities, combined with $\mu(\Omega)$ and $\xi(T)$, provide the "coupling" necessary to control the behavior of $\mathcal{I}^*$ restricted to elements in the direct sum $\Omega \oplus T$. In order to keep track of fewer quantities, we summarize the six quantities as follows:

$$\begin{aligned} \alpha &\triangleq \min(\alpha_\Omega, \alpha_T) \\ \delta &\triangleq \max(\delta_\Omega, \delta_T) \\ \beta &\triangleq \max(\beta_\Omega, \beta_T). \end{aligned}$$

Main assumption There exists a $\nu \in (0, \frac{1}{2}]$ such that:

$$\frac{\delta}{\alpha} \leq 1 - 2\nu.$$

This assumption is to be viewed as a generalization of the irrepresentability conditions imposed on the covariance matrix [148] or the Fisher information matrix [119] in order to provide consistency guarantees for sparse model selection using the ℓ_1 norm. With this assumption we have the following proposition, proved in Appendix B.3, about the gains of the operator $\mathcal{I}^*$ restricted to $\Omega \oplus T$. This proposition plays a fundamental role in the analysis of the performance of the regularized maximum-likelihood procedure (4.1).

Proposition 4.3.1. *Let Ω and T be the tangent spaces defined in this section, and let $\mathcal{I}^*$ be the Fisher information evaluated at the true marginal concentration matrix. Further let α, δ, β be as defined above. Suppose that*

$$\mu(\Omega)\xi(T) \leq \frac{1}{6}\left(\frac{\nu\alpha}{\beta(2-\nu)}\right)^2,$$

and that γ is in the following range:

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right].$$

Then we have the following two conclusions for $\mathcal{Y} = \Omega \times T'$ with $\rho(T', T) \leq \frac{\xi(T)}{2}$:

1. *The minimum gain of $\mathcal{I}^*$ restricted to $\Omega \oplus T'$ is bounded below:*

$$\min_{(S,L)\in\mathcal{Y},\ \|S\|_\infty=\gamma,\ \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \quad \geq \quad \frac{\alpha}{2}.$$

   Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \geq \frac{\alpha}{2} g_\gamma(S,L).$$

2. *The effect of elements in $\mathcal{Y} = \Omega \times T'$ on the orthogonal complement $\mathcal{Y}^\perp = \Omega^\perp \times T'^\perp$ is bounded above:*

$$\left\| \mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\left(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\right)^{-1} \right\|_{g_\gamma \to g_\gamma} \leq 1-\nu.$$

   Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \leq (1-\nu) g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)).$$

The last quantity we consider is the spectral norm of the marginal covariance matrix $\Sigma_O^* = (\tilde{K}_O^*)^{-1}$:

$$\psi \triangleq \|\Sigma_O^*\|_2 = \|(\tilde{K}_O^*)^{-1}\|_2. \tag{4.8}$$

A bound on ψ is useful in the probabilistic component of our analysis, in order to derive convergence rates of the sample covariance matrix to the true covariance matrix. We also observe that

$$\|\mathcal{I}^*\|_{2\to 2} = \|(\tilde{K}_O^*)^{-1} \otimes (\tilde{K}_O^*)^{-1}\|_{2\to 2} = \psi^2.$$

■ 4.4 Regularized Maximum-Likelihood Convex Program and Consistency

■ 4.4.1 Setup

Let $K^*_{(O\ H)}$ denote the full concentration matrix of a collection of zero-mean jointly-Gaussian observed and latent variables, let $p = |O|$ denote the number of observed variables, and let $h = |H|$ denote the number of latent variables. We are given n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O. We consider the high-dimensional setting in which (p, h, n) are all allowed to grow simultaneously. The quantities $\alpha, \delta, \beta, \nu, \psi$ defined in the previous section are accounted for in our analysis, although we suppress the dependence on these quantities in the statement of our main result. We explicitly keep track of the quantities $\mu(\Omega(K^*_O))$ and $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ as these control the complexity of the latent-variable model given by $K^*_{(O\ H)}$. In particular μ controls the sparsity of the conditional graphical model among the observed variables, while ξ controls the incoherence or "diffusivity" of the extra correlations induced due to marginalization over the hidden variables. Based on the tradeoff between these two quantities, we obtain a number of classes of latent-variable graphical models (and corresponding scalings of (p, h, n)) that can be consistently recovered using the regularized maximum-likelihood convex program (4.1) (see Section 4.4.3 for details). Specifically we show that consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables. We present our main result next demonstrating the consistency of the estimator (4.1), and then discuss classes of latent-variable graphical models and various scaling regimes in which our estimator is consistent.

■ 4.4.2 Main results

Given n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O, the sample covariance is defined as:

$$\Sigma^n_O = \frac{1}{n}\sum_{i=1}^{n} X^i_O (X^i_O)^T.$$

As discussed in Section 4.2.2 the goal is to produce an estimate given by a pair of matrices (S, L) of the latent-variable model represented by $K^*_{(O\ H)}$. We study the consistency properties of the following regularized maximum-likelihood convex program:

$$\begin{aligned}(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L}\ & \mathrm{Tr}[(S-L)\,\Sigma^n_O] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \mathrm{Tr}(L)] \\ \text{s.t.}\quad & S - L \succ 0, \quad L \succeq 0.\end{aligned} \tag{4.9}$$

Here λ_n is a regularization parameter, and γ is a tradeoff parameter between the rank and sparsity terms. Notice from Proposition 4.3.1 that the choice of γ depends on the values of $\mu(\Omega(K_O^*))$ and $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$; essentially these quantities correspond to the degree of the conditional graphical model structure of the observed variables and the incoherence of the low-rank matrix summarizing the effect of the latent variables (see Section 4.3). While these quantities may not be known *a priori*, we discuss a method to choose γ numerically in our experimental results (see Section 4.5). The following theorem shows that the estimates $(\hat{S}_n, \hat{L}_n)$ provided by the convex program (4.9) are consistent for a suitable choice of λ_n. In addition to the appropriate identifiability conditions (as specified by Proposition 4.3.1), we also impose lower bounds on the minimum nonzero entry of the sparse conditional graphical model matrix K_O^* and on the minimum nonzero singular value of the low-rank matrix $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ summarizing the effect of the hidden variables. We suppress the dependence on $\alpha, \beta, \delta, \nu, \psi$, and emphasize the dependence on $\mu(\Omega(K_O^*))$ and $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$ because these control the complexity of the underlying latent-variable graphical model as discussed above.

Theorem 4.4.1. *Let $K_{(O\ H)}^*$ denote the concentration matrix of a Gaussian model. We have n samples $\{X_O^i\}_{i=1}^n$ of the p observed variables denoted by O. Let $\Omega = \Omega(K_O^*)$ and $T = T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent spaces at K_O^* and at $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ with respect to the sparse and low-rank matrix varieties respectively.*

***Assumptions**: Suppose that the following conditions hold:*

1. *The quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 4.3.1 for identifiability, and γ is chosen in the range specified by Proposition 4.3.1.*

2. *The number of samples n available is such that*

$$n \gtrsim \frac{p}{\xi(T)^4}.$$

3. *The regularization parameter λ_n is chosen as*

$$\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

4. *The minimum nonzero singular value σ of $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is bounded as*

$$\sigma \gtrsim \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}.$$

5. *The minimum magnitude nonzero entry of* θ *of* K_O^* *is bounded as*

$$\theta \gtrsim \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}.$$

Conclusions: *Then with probability greater than* $1 - 2\exp\{-p\}$ *we have:*

1. *Algebraic consistency: The estimate* $(\hat{S}_n, \hat{L}_n)$ *given by the convex program* (4.9) *is algebraically consistent, i.e., the support and sign pattern of* $\hat{S}_n$ *is the same as that of* K_O^**, and the rank of* $\hat{L}_n$ *is the same as that of* $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$.

2. *Parametric consistency: The estimate* $(\hat{S}_n, \hat{L}_n)$ *given by the convex program* (4.9) *is parametrically consistent:*

$$g_\gamma(\hat{S}_n - K_O^*, \hat{L}_n - K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

The proof of this theorem is given in Appendix B.4. The theorem essentially states that if the minimum nonzero singular value of the low-rank piece $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ and minimum nonzero entry of the sparse piece K_O^* are bounded away from zero, then the convex program (4.9) provides estimates that are both algebraically consistent and parametrically consistent (in the ℓ_∞ and spectral norms). In Section 4.4.4 we also show that these results easily lead to parametric consistency rates for the corresponding estimate $(\hat{S}_n - \hat{L}_n)^{-1}$ of the marginal covariance Σ_O^* of the observed variables.

Remarks Notice that the condition on the minimum singular value of $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is more stringent than on the minimum nonzero entry of K_O^*. One role played by these conditions is to ensure that the estimates $(\hat{S}_n, \hat{L}_n)$ do not have smaller support size/rank than $(K_O^*, K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$. However the minimum singular value bound plays the additional role of bounding the curvature of the low-rank matrix variety around the point $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$, which is the reason for this condition being more stringent. Notice also that the number of hidden variables h does not explicitly appear in the sample complexity bound in Theorem 4.4.1, which only depends on $p, \mu(\Omega(K_O^*)), \xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$. However the dependence on h is implicit in the dependence on $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$, and we discuss this point in greater detail in the following section.

Finally we remark that algebraic and parametric consistency hold under the assumptions of Theorem 4.4.1 for a *range* of values of γ:

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right].$$

In particular the assumptions on the sample complexity, the minimum nonzero singular value of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, and the minimum magnitude nonzero entry of K^*_O are governed by the lower end of this range for γ. These assumptions can be weakened if we only require consistency for a smaller range of values of γ. The following corollary conveys this point with a specific example:

Corollary 4.4.1. *Consider the same setup and notation as in Theorem 4.4.1. Suppose that the quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 4.3.1 for identifiability. Suppose that we make the following assumptions:*

1. *Let γ be chosen to be equal to $\frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}$ (the upper end of the range specified in Proposition 4.3.1), i.e., $\gamma \asymp \frac{1}{\mu(\Omega)}$.*
2. $n \gtrsim \mu(\Omega)^4\, p$.
3. $\sigma \gtrsim \frac{\mu(\Omega)^2}{\xi(T)}\sqrt{\frac{p}{n}}$.
4. $\theta \gtrsim \sqrt{\frac{p}{n}}$.
5. $\lambda_n \asymp \mu(\Omega)\sqrt{\frac{p}{n}}$.

Then with probability greater than $1 - 2\exp\{-p\}$ we have estimates $(\hat{S}_n, \hat{L}_n)$ that are algebraically consistent, and parametrically consistent with the error bounded as

$$g_\gamma(\hat{S}_n - K^*_O, \hat{L}_n - K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}) \lesssim \mu(\Omega)\sqrt{\frac{p}{n}}.$$

The proof of this corollary is analogous to that of Theorem 4.4.1. We emphasize that in practice it is often beneficial to have consistent estimates for a range of values of γ (as in Theorem 4.4.1). Specifically the stability of the sparsity pattern and rank of the estimates $(\hat{S}_n, \hat{L}_n)$ for a range of tradeoff parameters is useful in order to choose a suitable value of γ, as prior information about the quantities $\mu(\Omega(K^*_O))$ and $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ is not typically available (see Section 4.5).

■ 4.4.3 Scaling regimes

Next we consider classes of latent-variable models that satisfy the conditions of Theorem 4.4.1. Recall that n denotes the number of samples, p denotes the number of observed variables, and h denotes the number of latent variables. We assume that the

parameters $\alpha, \beta, \delta, \nu, \psi$ defined in Section 4.3.2 remain constant, and do not scale with the other parameters such as (p, h, n) or $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ or $\mu(\Omega(K^*_O))$. In particular we focus on the tradeoff between $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ and $\mu(\Omega(K^*_O))$ (the quantities that control the complexity of a latent-variable graphical model), and the resulting scaling regimes for consistent estimation. Let $d = \deg(K^*_O)$ denote the degree of the conditional graphical model among the observed variables, and let $i = \text{inc}(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O})$ denote the incoherence of the correlations induced due to marginalization over the latent variables (we suppress the dependence on n). These quantities are defined in Chapter 3, and we have from the propositions therein that

$$\mu(\Omega(K^*_O)) \leq d, \quad \xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O})) \leq 2i.$$

Since $\alpha, \beta, \delta, \nu, \psi$ do not scale with the other parameters, we also have from Proposition 4.3.1 that the product of μ and ξ must be bounded by a constant. Thus, we study latent-variable models in which

$$d\, i = \mathcal{O}(1).$$

As we describe next, there are non-trivial classes of latent-variable graphical models in which this condition holds.

Bounded degree and incoherence: The first class of latent-variable models that we consider are those in which the conditional graphical model among the observed variables (given by K^*_O) has constant degree d. Recall from Chapter 3 that the incoherence i of the effect of the latent variables (given by $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$) can be as small as $\sqrt{\frac{h}{p}}$. Consequently latent-variable models in which

$$d = \mathcal{O}(1), \quad h = \mathcal{O}(p),$$

can be estimated consistently from $n = \mathcal{O}(p)$ samples as long as the low-rank matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ is almost maximally incoherent, i.e., $i = \mathcal{O}(\sqrt{\frac{h}{p}})$ so the effect of marginalization over the latent variables is diffuse across almost all the observed variables. Thus consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables.

Polylogarithmic degree models The next class of models that we study are those in which the degree d of the conditional graphical model of the observed variables grows poly-logarithmically with p. Consequently, the incoherence i of the matrix

$K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ must decay as the inverse of poly-log(p). Using the fact that maximally incoherent low-rank matrices $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ can have incoherence as small as $\sqrt{\frac{h}{p}}$, latent-variable models in which

$$d = \mathcal{O}(\log(p)^q), \quad h = \mathcal{O}\left(\frac{p}{\log(p)^{2q}}\right),$$

can be consistently estimated as long as $n = \mathcal{O}_P(p \text{ poly-log}(p))$.

■ 4.4.4 Rates for covariance matrix estimation

The main result Theorem 4.4.1 gives the number of samples required for consistent estimation of the sparse and low-rank parts that compose the marginal concentration matrix $\tilde{K}^*_O$. Here we prove a corollary that gives rates for covariance matrix estimation, i.e., the quality of the estimate $(\hat{S}_n - \hat{L}_n)^{-1}$ with respect to the "true" marginal covariance matrix Σ^*_O.

Corollary 4.4.2. *Under the same conditions as in Theorem 4.4.1, we have with probability greater than* $1 - 2\exp\{-p\}$ *that*

$$g_\gamma(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^*_O]) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

Specifically this implies that $\|(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^*_O\|_2 \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.

Proof: The proof of this lemma follows directly from duality. Based on the analysis in Appendix B.4 (in particular using the optimality conditions of the modified convex program (B.14)), we have that

$$g_\gamma(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^n_O]) \leq \lambda_n.$$

We also have from the bound on the number of samples n that (see Appendix B.4.7)

$$g_\gamma(\mathcal{A}^\dagger[\Sigma^*_O - \Sigma^n_O]) \lesssim \lambda_n$$

Based on the choice of λ_n in Theorem 4.4.1, we then have the desired bound. □

■ 4.4.5 Proof strategy for Theorem 4.4.1

Standard results from convex analysis [124] state that $(\hat{S}_n, \hat{L}_n)$ is a minimum of the convex program (4.9) if the zero matrix belongs to the subdifferential of the objective

function evaluated at $(\hat{S}_n, \hat{L}_n)$ (in addition to $(\hat{S}_n, \hat{L}_n)$ satisfying the constraints). The subdifferential of the ℓ_1 norm at a matrix M is given by

$$N \in \partial \|M\|_1 \quad \Leftrightarrow \quad \mathcal{P}_{\Omega(M)}(N) = \operatorname{sign}(M), \ \|\mathcal{P}_{\Omega(M)^\perp}(N)\|_\infty \leq 1.$$

For a symmetric positive semidefinite matrix M with SVD $M = UDU^T$, the subdifferential of the trace function restricted to the cone of positive semidefinite matrices (i.e., the nuclear norm over this set) is given by:

$$N \in \partial[\operatorname{Tr}(M) + \mathbb{I}_{M \succeq 0}] \quad \Leftrightarrow \quad \mathcal{P}_{T(M)}(N) = UU^T, \ \mathcal{P}_{T(M)^\perp}(N) \preceq I,$$

where $\mathbb{I}_{M \succeq 0}$ denotes the characteristic function of the set of positive semidefinite matrices (i.e., the convex function that evaluates to 0 over this set and ∞ outside). The key point is that elements of the subdifferential decompose with respect to the tangent spaces $\Omega(M)$ and $T(M)$. This decomposition property plays a critical role in our analysis. In particular it states that the optimality conditions consist of two parts, one part corresponding to the tangent spaces Ω and T and another corresponding to the normal spaces $\Omega^\perp$ and $T^\perp$.

Consider the optimization problem (4.9) with the additional (non-convex) constraints that the variable S belongs to the algebraic variety of sparse matrices and that the variables L belongs to the algebraic variety of low-rank matrices. While this new optimization problem is non-convex, it has a very interesting property. At a globally optimal solution (and indeed at any locally optimal solution) $(\tilde{S}, \tilde{L})$ such that $\tilde{S}$ and $\tilde{L}$ are smooth points of the algebraic varieties of sparse and low-rank matrices, the first-order optimality conditions state that the Lagrange multipliers corresponding to the additional variety constraints must lie in the *normal spaces* $\Omega(\tilde{S})^\perp$ and $T(\tilde{L})^\perp$. This fundamental observation, combined with the decomposition property of the subdifferentials of the ℓ_1 and nuclear norms, suggests the following high-level proof strategy.

1. Let $(\tilde{S}, \tilde{L})$ be the globally optimal solution of the optimization problem (4.9) with the additional constraints that (S, L) belong to the algebraic varieties of sparse/low-rank matrices; specifically constrain S to lie in $\mathcal{S}(|\text{support}(K_O^*)|)$ and constrain L to lie in $\mathcal{L}(\text{rank}(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$. Show first that $(\tilde{S}, \tilde{L})$ are smooth points of these varieties.

2. The first part of the subgradient optimality conditions of the original convex program (4.9) corresponding to components *on* the tangent spaces $\Omega(\tilde{S})$ and $T(\tilde{L})$

is satisfied. This conclusion can be reached because the additional Lagrange multipliers due to the variety constraints lie in the normal spaces $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$.

3. Finally show that the second part of the subgradient optimality conditions of (4.9) corresponding to components in the normal spaces $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$ is also satisfied.

Combining these steps together we show that $(\tilde{S}, \tilde{L})$ satisfy the optimality conditions of the *original convex program* (4.9). Consequently $(\tilde{S}, \tilde{L})$ is also the optimum of the convex program (4.9). As this estimate is also the solution to the problem with the variety constraints, the algebraic consistency of $(\tilde{S}, \tilde{L})$ can be directly concluded. We emphasize here that the variety-constrained optimization problem is used solely as an analysis tool in order to prove consistency of the estimates provided by the convex program (4.9). These steps describe our broad strategy, and we refer the reader to Appendix B.4 for details. The key technical complication is that the tangent spaces at $\tilde{L}$ and $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ are in general different. We bound the twisting between these tangent spaces by using the fact that the minimum non-zero singular value of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ is bounded away from zero (as assumed in Theorem 4.4.1 and using Proposition 4.2.1).

■ 4.5 Simulation Results

In this section we give experimental demonstration of the consistency of our estimator (4.9) on synthetic examples, and its effectiveness in modeling real-world stock return data. Our choices of λ_n and γ are guided by Theorem 4.4.1. Specifically, we choose λ_n to be proportional to $\sqrt{\frac{p}{n}}$. For γ we observe that the support/sign-pattern and the rank of the solution $(\hat{S}_n, \hat{L}_n)$ are the same for a *range* of values of γ. Therefore one could solve the convex program (4.9) for several values of γ, and choose a solution in a suitable range in which the sign-pattern and rank of the solution are stable. In practical problems with real-world data these parameters may be chosen via cross-validation. For small problem instances we solve the convex program (4.9) using a combination of YALMIP [98] and SDPT3 [136], which are standard off-the-shelf packages for solving convex programs. For larger problem instances we use the special purpose solver LogdetPPA [141] developed for log-determinant semidefinite programs.

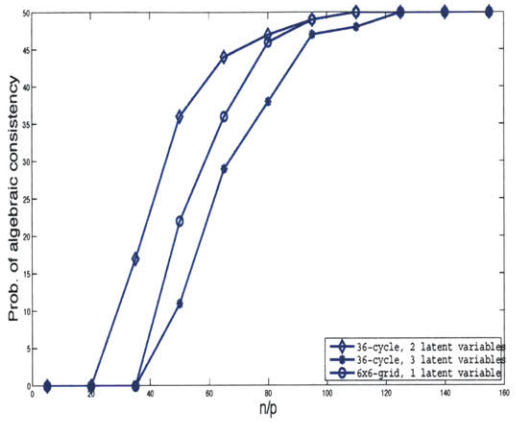

Figure 4.1. Synthetic experiments: Plot showing probability of consistent estimation of the number of latent variables, and the conditional graphical model structure of the observed variables. the three models studied are (a) 36-node conditional graphical model given by a cycle with $h = 2$ latent variables, (b) 36-node conditional graphical model given by a cycle with $h = 3$ latent variables, and (c) 36-node conditional graphical model given by a 6×6 grid with $h = 1$ latent variable. For each plotted point, the probability of consistent estimation is obtained over 50 random trials.

■ 4.5.1 Synthetic data

In the first set of experiments we consider a setting in which we have access to samples of the observed variables of a latent-variable graphical model. We consider several latent-variable Gaussian graphical models. The first model consists of $p = 36$ observed variables and $h = 2$ hidden variables. The conditional graphical model structure of the observed variables is a cycle with the edge partial correlation coefficients equal to 0.25; thus, this conditional model is specified by a sparse graphical model with degree 2. The second model is the same as the first one, but with $h = 3$ latent variables. The third model consists of $h = 1$ latent variable, and the conditional graphical model structure of the observed variables is given by a 6×6 nearest-neighbor grid (i.e., $p = 36$ and degree 4) with the partial correlation coefficients of the edges equal to 0.15. In all three of these models each latent variable is connected to a random subset of 80% of the observed variables (and the partial correlation coefficients corresponding to these edges are also random). Therefore the effect of the latent variables is "spread out" over most of the observed variables, i.e., the low-rank matrix summarizing the effect of the latent variables is incoherent.

Figure 4.2. Stock returns: The figure on the left shows the sparsity pattern (black denotes an edge, and white denotes no edge) of the concentration matrix of the conditional graphical model (135 edges) of the stock returns, conditioned on 5 latent variables, in a latent-variable graphical model (number of parameters equals 639). This model is learned using (4.9), and the KL divergence with respect to a Gaussian distribution specified by the sample covariance is 17.7. The figure on the left shows the concentration matrix of the graphical model (646 edges) of the stock returns, learned using standard sparse graphical model selection based on solving an ℓ_1-regularized maximum-likelihood program (number of parameters equals 730). The KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is 44.4.

For each model we generate n samples of the observed variables, and use the resulting sample covariance matrix Σ_O^n as input to our convex program (4.9). Figure 4.1 shows the probability of recovery of the support/sign-pattern of the conditional graphical model structure in the observed variables and the number of latent variables (i.e., probability of obtaining algebraically consistent estimates) as a function of n. This probability is evaluated over 50 experiments for each value of n.

In all of these cases standard graphical model selection applied directly to the observed variables is not useful as the marginal concentration matrix of the observed variables is not well-approximated by a sparse matrix. Both these sets of experiments agree with our theoretical results that the convex program (4.9) is an algebraically consistent estimator of a latent-variable model given (sufficiently many) samples of only the observed variables.

■ 4.5.2 Stock return data

In the next experiment we model the statistical structure of monthly stock returns of 84 companies in the S&P 100 index from 1990 to 2007; we disregard 16 companies that were listed after 1990. The number of samples n is equal to 216. We compute the

sample covariance based on these returns and use this as input to (4.9).

The model learned using (4.9) for suitable values of λ_n, γ consists of $h = 5$ latent variables, and the conditional graphical model structure of the stock returns conditioned on these hidden components consists of 135 edges. Therefore the number of parameters in the model is $84 + 135 + (5 \times 84) = 639$. The resulting KL divergence between the distribution specified by this model and a Gaussian distribution specified by the sample covariance is 17.7. Figure 4.2 (left) shows the *conditional* graphical model structure. The strongest edges in this conditional graphical model, as measured by partial correlation, are between Baker Hughes - Schlumberger, A.T.&T. - Verizon, Merrill Lynch - Morgan Stanley, Halliburton - Baker Hughes, Intel - Texas Instruments, Apple - Dell, and Microsoft - Dell. It is of interest to note that in the Standard Industrial Classification[2] system for grouping these companies, several of these pairs are in different classes.

We compare these results to those obtained using a sparse graphical model learned using ℓ_1-regularized maximum-likelihood (see for example [119]), without introducing any latent variables. Figure 4.2 (right) shows this graphical model structure. The number of edges in this model is 646 (the total number of parameters is equal to $646 + 84 = 730$), and the resulting KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is 44.4. Indeed to obtain a comparable KL divergence to that of the latent-variable model described above, one would require a graphical model with over 3000 edges.

These results suggest that a latent-variable graphical model is better suited than a standard sparse graphical model for modeling the statistical structure among stock returns. This is likely due to the presence of global, long-range correlations in stock return data that are better modeled via latent variables.

■ 4.6 Discussion

We have studied the problem of modeling the statistical structure of a collection of random variables as a sparse graphical model conditioned on a few additional hidden components. As a first contribution we described conditions under which such latent-variable graphical models are identifiable given samples of only the observed variables.

[2]See the United States Securities and Exchange Commission website at http://www.sec.gov/info/edgar/siccodes.htm

We also proposed a convex program based on regularized maximum-likelihood for latent-variable graphical model selection; the regularization function is a combination of the ℓ_1 norm and the nuclear norm. Given samples of the observed variables of a latent-variable Gaussian model we proved that this convex program provides consistent estimates of the number of hidden components as well as the conditional graphical model structure among the observed variables conditioned on the hidden components. Our analysis holds in the high-dimensional regime in which the number of observed/latent variables are allowed to grow with the number of samples of the observed variables. In particular we discuss certain scaling regimes in which consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables. These theoretical predictions are verified via a set of experiments on synthetic data.

Chapter 5

Convex Geometry of Linear Inverse Problems

5.1 Introduction

Deducing the state or structure of a system from partial, noisy measurements is a fundamental task throughout the sciences and engineering. A commonly encountered difficulty that arises in such inverse problems is the very limited availability of data relative to the ambient dimension of the signal to be estimated. However many interesting signals or models in practice contain few degrees of freedom relative to their ambient dimension. For instance a small number of genes may constitute a signature for disease, very few parameters may be required to specify the correlation structure in a time series, or a sparse collection of geometric constraints might completely specify a molecular configuration. Such low-dimensional structure plays an important role in making inverse problems well-posed. In this chapter we propose a unified approach to transform notions of simplicity into convex penalty functions, thus obtaining convex optimization formulations for inverse problems.

We describe a model as simple if it can be written as a linear combination of a few elements from an atomic set. Concretely let $\mathbf{x} \in \mathbb{R}^p$ be formed as follows:

$$\mathbf{x} = \sum_{i=1}^{k} c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A}, c_i \geq 0, \tag{5.1}$$

where $\mathcal{A}$ is a set of atoms that constitute simple building blocks of general signals. Here we assume that $\mathbf{x}$ is *simple* so that k is relatively small. For example $\mathcal{A}$ could be the finite set of unit-norm one-sparse vectors in which case $\mathbf{x}$ is a sparse vector, or $\mathcal{A}$ could be the infinite set of unit-norm rank-one matrices in which case $\mathbf{x}$ is a low-rank matrix. These two cases arise in many applications, and have received a

tremendous amount of attention recently as several authors have shown that sparse vectors and low-rank matrices can be recovered from highly incomplete information [29, 30, 53, 54, 121]. However a number of other structured mathematical objects also fit the notion of simplicity described in (5.1). The set $\mathcal{A}$ could be the collection of unit-norm rank-one tensors, in which case $\mathbf{x}$ is a low-rank tensor and we are faced with the familiar challenge of low-rank tensor decomposition. Such problems arise in numerous applications in computer vision and image processing [1], and in neuroscience [9]. Alternatively $\mathcal{A}$ could be the set of permutation matrices; sums of a few permutation matrices are objects of interest in ranking [84] and multi-object tracking. As yet another example, $\mathcal{A}$ could consist of measures supported at a single point so that $\mathbf{x}$ is an atomic measure supported at just a few points. This notion of simplicity arises in problems in system identification and statistics.

In each of these examples as well as several others, a fundamental problem of interest is to recover $\mathbf{x}$ given limited *linear* measurements. For instance the question of recovering a sparse function over the group of permutations (i.e., the sum of a few permutation matrices) given linear measurements in the form of partial Fourier information was investigated in the context of ranked election problems [84]. Similar linear inverse problems arise with atomic measures in system identification, with orthogonal matrices in machine learning, and with simple models formed from several other atomic sets (see Section 5.2.2 for more examples). Hence we seek tractable computational tools to solve such problems. When $\mathcal{A}$ is the collection of one-sparse vectors, a method of choice is to use the ℓ_1 norm to induce sparse solutions. This method, as mentioned previously, has seen a surge interest in the last few years as it provides a tractable convex optimization formulation to exactly recover sparse vectors under various conditions [29, 53, 54]. Also as discussed before, the nuclear norm has been proposed more recently as an effective convex surrogate for solving rank minimization problems subject to various affine constraints [30, 121].

Motivated by the success of these methods we propose a general convex optimization framework in Section 5.2 in order to recover objects with structure of the form (5.1) from limited linear measurements. The guiding question behind our framework is: how do we take a concept of simplicity such as sparsity and derive the ℓ_1 norm as a convex heuristic? In other words what is the natural procedure to go from the set of one-sparse vectors $\mathcal{A}$ to the ℓ_1 norm? We observe that the convex hull of (unit-Euclidean-norm) one-sparse vectors is the unit ball of the ℓ_1 norm, or the cross-polytope. Similarly the

convex hull of the (unit-Euclidean-norm) rank-one matrices is the nuclear norm ball; see Chapter 2 for illustrations. These constructions suggest a natural generalization to other settings. Under suitable conditions the convex hull $\text{conv}(\mathcal{A})$ defines the unit ball of a norm, which is called the *atomic norm* induced by the atomic set $\mathcal{A}$. We can then minimize the atomic norm subject to measurement constraints, which results in a convex programming heuristic for recovering simple models given linear measurements. As an example suppose we wish to recover the sum of a few permutation matrices given linear measurements. The convex hull of the set of permutation matrices is the *Birkhoff polytope* of doubly stochastic matrices [149], and our proposal is to solve a convex program that minimizes the norm induced by this polytope. Similarly if we wish to recover an orthogonal matrix from linear measurements we would solve a *spectral norm* minimization problem, as the spectral norm ball is the convex hull of all orthogonal matrices. As discussed in Section 5.2.5 the atomic norm minimization problem is the best convex heuristic for recovering simple models with respect to a given atomic set.

We give general conditions for exact and robust recovery using the atomic norm heuristic. In Section 5.3 we provide concrete bounds on the number of generic linear measurements required for the atomic norm heuristic to succeed. This analysis is based on computing certain *Gaussian widths* of tangent cones with respect to the unit balls of the atomic norm [76]. Arguments based on Gaussian width have been fruitfully applied to obtain bounds on the number of Gaussian measurements for the special case of recovering sparse vectors via ℓ_1 norm minimization [127,134], but computing Gaussian widths of general cones is not easy. Therefore it is important to exploit the special structure in atomic norms, while still obtaining sufficiently general results that are broadly applicable. An important theme in this chapter is the connection between Gaussian widths and various notions of *symmetry*. Specifically by exploiting symmetry structure in certain atomic norms as well as convex duality properties, we give bounds on the number of measurements required for recovery using very general atomic norm heuristics. For example we provide precise estimates of the number of generic measurements required for exact recovery of an orthogonal matrix via spectral norm minimization, and the number of generic measurements required for exact recovery of a permutation matrix by minimizing the norm induced by the Birkhoff polytope. While these results correspond to the recovery of individual atoms from random measurements, our techniques are more generally applicable to the recovery of models formed as sums of a few atoms as well. We also give tighter bounds than those previously obtained on the number of

Underlying model	Convex heuristic	# Gaussian measurements
s-sparse vector in $\mathbb{R}^p$	ℓ_1 norm	$2s(\log(p/s-1)+1)$
$m \times m$ rank-r matrix	nuclear norm	$3r(2m-r)$
sign-vector $\{-1,+1\}^p$	ℓ_∞ norm	$p/2$
$m \times m$ permutation matrix	norm induced by Birkhoff polytope	$9\,m\log(m)$
$m \times m$ orthogonal matrix	spectral norm	$(3m^2-m)/4$

Table 5.1. A summary of the recovery bounds obtained using Gaussian width arguments.

measurements required to robustly recover sparse vectors and low-rank matrices via ℓ_1 norm and nuclear norm minimization. In all of the cases we investigate, we find that the number of measurements required to reconstruct an object is proportional to its intrinsic dimension rather than the ambient dimension, thus confirming prior folklore. See Table 5.1 for a summary of these results.

Although our conditions for recovery and bounds on the number of measurements hold generally, we note that it may not be possible to obtain a computable representation for the convex hull $\text{conv}(\mathcal{A})$ of an arbitrary set of atoms $\mathcal{A}$. This leads us to another important theme of this chapter, which we discuss in Section 5.4, on the connection between algebraic structure in $\mathcal{A}$ and the semidefinite representability of the convex hull $\text{conv}(\mathcal{A})$. In particular when $\mathcal{A}$ is an algebraic variety the convex hull $\text{conv}(\mathcal{A})$ can be approximated as (the projection of) a set defined by linear matrix inequalities. Thus the resulting atomic norm minimization heuristic can be solved via semidefinite programming. A second issue that arises in practice is that even with algebraic structure in $\mathcal{A}$ the semidefinite representation of $\text{conv}(\mathcal{A})$ may not be computable in polynomial time, which makes the atomic norm minimization problem intractable to solve. A prominent example here is the tensor nuclear norm ball, which is obtained by taking the convex hull of the rank-one tensors. In order to address this problem we study a hierarchy of semidefinite relaxations using *theta bodies* [77] (described in Chapter 2), which approximate the original (intractable) atomic norm minimization problem. A third point we highlight is that while these semidefinite relaxations are more tractable to solve, we require more measurements for exact recovery of the underlying model than if we solve the original intractable atomic norm minimization problem. Hence we have a tradeoff between the complexity of the recovery algorithm and the number of measurements required for recovery. We illustrate this tradeoff with the cut polytope,

which is intractable to compute, and its relaxations.

Outline Section 5.2 describes the construction of the atomic norm, gives several examples of applications in which these norms may be useful to recover simple models, and provides general conditions for recovery by minimizing the atomic norm. In Section 5.3 we investigate the number of generic measurements for exact or robust recovery using atomic norm minimization, and give estimates in a number of settings by analyzing the Gaussian width of certain tangent cones. We address the problem of semidefinite representability and tractable relaxations of the atomic norm in Section 5.4. Section 5.5 describes some algorithmic issues as well as a few simulation results, and we conclude with a discussion in Section 5.6.

■ 5.2 Atomic Norms and Convex Geometry

In this section we describe the construction of an atomic norm from a collection of simple atoms. In addition we give several examples of atomic norms, and discuss their properties in the context of solving ill-posed linear inverse problems. We denote the Euclidean norm by $\|\cdot\|$.

■ 5.2.1 Definition

Let $\mathcal{A}$ be a collection of atoms that is a compact subset of $\mathbb{R}^p$. We will assume throughout this chapter that no element $\mathbf{a} \in \mathcal{A}$ lies in the convex hull of the other elements $\text{conv}(\mathcal{A}\backslash\mathbf{a})$, i.e., the elements of $\mathcal{A}$ are the extreme points of $\text{conv}(\mathcal{A})$. Let $\|\mathbf{x}\|_{\mathcal{A}}$ denote the gauge of $\mathcal{A}$ [124]:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t > 0 \ : \ x \in t \ \text{conv}(\mathcal{A})\}. \tag{5.2}$$

Note that the gauge is always a convex, extended-real valued function for any set $\mathcal{A}$. By convention this function evaluates to $+\infty$ if $\mathbf{x}$ does not lie in the affine hull of $\text{conv}(\mathcal{A})$. We will assume without loss of generality that the centroid of $\text{conv}(\mathcal{A})$ is at the origin, as this can be achieved by appropriate recentering. With this assumption the gauge function can be rewritten as:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\left\{\sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}} \ : \ \mathbf{x} = \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\mathbf{a}, \ \ c_{\mathbf{a}} \geq 0 \ \forall \mathbf{a} \in \mathcal{A}\right\},$$

with the sum being replaced by an integral when $\mathcal{A}$ is uncountable. If $\mathcal{A}$ is centrally symmetric about the origin (i.e., $\mathbf{a} \in \mathcal{A}$ if and only if $-\mathbf{a} \in \mathcal{A}$) we have that $\|\cdot\|_{\mathcal{A}}$ is

a norm, which we call the *atomic norm* induced by $\mathcal{A}$. The support function of $\mathcal{A}$ is given as:

$$\|\mathbf{x}\|_{\mathcal{A}}^* = \sup\{\langle \mathbf{x}, \mathbf{a}\rangle \ : \ \mathbf{a} \in \mathcal{A}\}. \tag{5.3}$$

If $\|\cdot\|_{\mathcal{A}}$ is a norm the support function $\|\cdot\|_{\mathcal{A}}^*$ is the dual norm of this atomic norm. From this definition we see that the unit ball of $\|\cdot\|_{\mathcal{A}}$ is equal to $\text{conv}(\mathcal{A})$. In many examples of interest the set $\mathcal{A}$ is not centrally symmetric, so that the gauge function does not define a norm. However our analysis is based on the underlying convex geometry of $\text{conv}(\mathcal{A})$, and our results are applicable even if $\|\cdot\|_{\mathcal{A}}$ does not define a norm. Therefore, with an abuse of terminology we generally refer to $\|\cdot\|_{\mathcal{A}}$ as the atomic norm of the set $\mathcal{A}$ even if $\|\cdot\|_{\mathcal{A}}$ is not a norm. We note that the duality characterization between (5.2) and (5.3) when $\|\cdot\|_{\mathcal{A}}$ is a norm is in fact applicable even in infinite-dimensional Banach spaces by Bonsall's atomic decomposition theorem [21], but our focus is on the finite-dimensional case in this work. We investigate in greater detail the issues of representability and efficient approximation of these atomic norms in Section 5.4.

Equipped with a convex penalty function given a set of atoms, we propose a convex optimization method to recover a "simple" model give limited linear measurements. Specifically suppose that $\mathbf{x}^\star$ is formed according to (5.1) from a set of atoms $\mathcal{A}$. Further suppose that we have a known linear map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$, and we have linear information about $\mathbf{x}^\star$ as follows:

$$\mathbf{y} = \Phi\mathbf{x}^\star. \tag{5.4}$$

The goal is to reconstruct $\mathbf{x}^\star$ given $\mathbf{y}$. We consider the following convex formulation to accomplish this task:

$$\begin{aligned} \hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \quad & \|\mathbf{x}\|_{\mathcal{A}} \\ \text{s.t.} \quad & \mathbf{y} = \Phi\mathbf{x}. \end{aligned} \tag{5.5}$$

When $\mathcal{A}$ is the set of one-sparse atoms this problem reduces to standard ℓ_1 norm minimization. Similarly when $\mathcal{A}$ is the set of rank-one matrices this problem reduces to nuclear norm minimization. More generally if the atomic norm $\|\cdot\|_{\mathcal{A}}$ is tractable to evaluate, then (5.5) potentially offers an efficient convex programming formulation for reconstructing $\mathbf{x}^\star$ from the limited information $\mathbf{y}$. The *dual problem* of (5.5) is given as follows:

$$\begin{aligned} \max_{\mathbf{z}} \quad & \mathbf{y}^T\mathbf{z} \\ \text{s.t.} \quad & \|\Phi^\dagger\mathbf{z}\|_{\mathcal{A}}^* \leq 1. \end{aligned} \tag{5.6}$$

Here $\Phi^\dagger$ denotes the adjoint (or transpose) of the linear measurement map Φ.

The convex formulation (5.5) can be suitably modified in case we only have access to inaccurate, noisy information. Specifically suppose that we have noisy measurements $\mathbf{y} = \Phi\mathbf{x}^\star + \omega$ where ω represents the noise term. A natural convex formulation is one in which the constraint $\mathbf{y} = \Phi\mathbf{x}$ of (5.5) is replaced by the relaxed constraint $\|\mathbf{y} - \Phi\mathbf{x}\| \leq \delta$, where δ is an upper bound on the size of the noise ω:

$$\begin{aligned} \hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \quad & \|\mathbf{x}\|_{\mathcal{A}} \\ \text{s.t.} \quad & \|\mathbf{y} - \Phi\mathbf{x}\| \leq \delta. \end{aligned} \tag{5.7}$$

We say that we have *exact recovery* in the noise-free case if $\hat{\mathbf{x}} = \mathbf{x}^\star$ in (5.5), and *robust recovery* in the noisy case if the error $\|\hat{\mathbf{x}} - \mathbf{x}^\star\|$ is small in (5.7). In Section 5.2.4 and Section 5.3 we give conditions under which the atomic norm heuristics (5.5) and (5.7) recover $\mathbf{x}^\star$ exactly or approximately. Atomic norms have found fruitful applications in problems in approximation theory of various function classes [8, 46, 86, 116]. However this prior body of work was concerned with infinite-dimensional Banach spaces, and none of these references consider nor provide recovery guarantees that are applicable in our setting.

■ 5.2.2 Examples

Next we provide several examples of atomic norms that can be viewed as special cases of the construction above. These norms are obtained by convexifying atomic sets that are of interest in various applications.

Sparse vectors. The problem of recovering sparse vectors from limited measurements has received a great deal of attention, with applications in many problem domains. In this case the atomic set $\mathcal{A} \subset \mathbb{R}^p$ can be viewed as the set of unit-norm one-sparse vectors $\{\pm\mathbf{e}_i\}_{i=1}^p$, and k-sparse vectors in $\mathbb{R}^p$ can be constructed using a linear combination of k elements of the atomic set. In this case it is easily seen that the convex hull $\text{conv}(\mathcal{A})$ is given by the *cross-polytope* (i.e., the unit ball of the ℓ_1 norm; see Chapter 2), and the atomic norm $\|\cdot\|_{\mathcal{A}}$ corresponds to the ℓ_1 norm in $\mathbb{R}^p$.

Low-rank matrices. Recovering low-rank matrices from limited information is also a problem that has received considerable attention as it finds applications in problems in statistics, control, and machine learning. The atomic set $\mathcal{A}$ here can be viewed as the set of rank-one matrices of unit-Euclidean-norm. The convex hull $\text{conv}(\mathcal{A})$ is the *nuclear norm ball* of matrices in which the sum of the singular values is less than or equal to one (see Chapter 2).

Permutation matrices. A problem of interest in a ranking context [84] or an object tracking context is that of recovering permutation matrices from partial information. Suppose that a small number k of rankings of m candidates is preferred by a population. Such preferences can be modeled as the sum of a few $m \times m$ permutation matrices, with each permutation corresponding to a particular ranking. By conducting surveys of the population one can obtain partial linear information of these preferred rankings. The set $\mathcal{A}$ here is the collection of permutation matrices (consisting of $m!$ elements), and the convex hull $\text{conv}(\mathcal{A})$ is the *Birkhoff polytope* or the set of doubly stochastic matrices [149]. The centroid of the Birkhoff polytope is the matrix $\mathbf{1}\mathbf{1}^T/m$, so it needs to be recentered appropriately. We mention here recent work by Jagabathula and Shah [84] on recovering a sparse function over the symmetric group (i.e., the sum of a few permutation matrices) given partial Fourier information; although the algorithm proposed in [84] is tractable it is not based on convex optimization.

Binary vectors. In integer programming one is often interested in recovering vectors in which the entries take on values of ± 1. Suppose that there exists such a sign-vector, and we wish to recover this vector given linear measurements. This corresponds to a version of the multi-knapsack problem [102]. In this case $\mathcal{A}$ is the set of all sign-vectors, and the convex hull $\text{conv}(\mathcal{A})$ is the *hypercube* or the unit ball of the ℓ_∞ norm. The image of this hypercube under a linear map is also referred to as a zonotope [149].

Vectors from lists. Suppose there is an unknown vector $\mathbf{x} \in \mathbb{R}^p$, and that we are given the entries of this vector without any information about the locations of these entries. For example if $\mathbf{x} = [3\ 1\ 2\ 2\ 4]'$, then we are only given the list of numbers $\{1, 2, 2, 3, 4\}$ without their positions in $\mathbf{x}$. Further suppose that we have access to a few linear measurements of $\mathbf{x}$. Can we recover $\mathbf{x}$ by solving a convex program? Such a problem is of interest in recovering partial rankings of elements of a set. An extreme case is one in which we only have two preferences for rankings, i.e., a vector in $\{1, 2\}^p$ composed only of one's and two's, which reduces to a special case of the problem above of recovering binary vectors (in which the number of entries of each sign is fixed). For this problem the set $\mathcal{A}$ is the set of all permutations of $\mathbf{x}$ (which we know since we have the list of numbers that compose $\mathbf{x}$), and the convex hull $\text{conv}(\mathcal{A})$ is the *permutahedron* [129, 149] (see Chapter 2). As with the Birkhoff polytope, the permutahedron also needs to be recentered about the point $\mathbf{1}^T\mathbf{x}/p$.

Matrices constrained by eigenvalues. This problem is in a sense the non-commutative analog of the one above. Suppose that we are given the eigenvalues λ of

a symmetric matrix, but no information about the eigenvectors. Can we recover such a matrix given some additional linear measurements? In this case the set $\mathcal{A}$ is the set of all symmetric matrices with eigenvalues λ, and the convex hull $\text{conv}(\mathcal{A})$ is given by the *Schur-Horn orbitope* [129] (see Chapter 2).

Orthogonal matrices. In many applications matrix variables are constrained to be orthogonal, which is a non-convex constraint and may lead to computational difficulties. We consider one such simple setting in which we wish to recover an orthogonal matrix given limited information in the form of linear measurements. In this example the set $\mathcal{A}$ is the set of $m \times m$ orthogonal matrices, and $\text{conv}(\mathcal{A})$ is the *spectral norm ball.*

Measures. Recovering a measure given its moments is another question of interest that arises in system identification and statistics. Suppose one is given access to a linear combination of moments of an atomically supported measure. How can we reconstruct the support of the measure? The set $\mathcal{A}$ here is the moment curve, and its convex hull $\text{conv}(\mathcal{A})$ goes by several names including the *Caratheodory orbitope* [129]. Discretized versions of this problem correspond to the set $\mathcal{A}$ being a finite number of points on the moment curve; the convex hull $\text{conv}(\mathcal{A})$ is then a *cyclic polytope* [149].

Cut matrices. In some problems one may wish to recover low-rank matrices in which the entries are constrained to take on values of ± 1. Such matrices can be used to model basic user preferences, and are of interest in problems such as collaborative filtering [133]. The set of atoms $\mathcal{A}$ could be the set of rank-one signed matrices, i.e., matrices of the form $\mathbf{z}\mathbf{z}^T$ with the entries of $\mathbf{z}$ being ± 1. The convex hull $\text{conv}(\mathcal{A})$ of such matrices is the *cut polytope* [47]. An interesting issue that arises here is that the cut polytope is in general intractable to characterize. However there exist several well-known tractable semidefinite relaxations to this polytope [47, 72], and one can employ these in constructing efficient convex programs for recovering cut matrices. We discuss this point in greater detail in Section 5.4.3.

Low-rank tensors. Low-rank tensor decompositions play an important role in numerous applications throughout signal processing and machine learning [91]. Developing computational tools to recover low-rank tensors is therefore of great interest. In principle we could solve a tensor nuclear norm minimization problem, in which the tensor nuclear norm ball is obtained by taking the convex hull of rank-one tensors. A computational challenge here is that the tensor nuclear norm is in general intractable to compute; in order to address this problem we discuss further convex relaxations to the tensor nuclear norm using theta bodies in Section 5.4. A number of additional

technical issues also arise with low-rank tensors including the non-existence in general of a singular value decomposition analogous to that for matrices [90], and the difference between the rank of a tensor and its border rank [45].

Nonorthogonal factor analysis. Suppose that a data matrix admits a factorization $X = AB$. The matrix nuclear norm heuristic will find a factorization into *orthogonal* factors in which the columns of A and rows of B are mutually orthogonal. However if *a priori* information is available about the factors, precision and recall could be improved by enforcing such priors. These priors may sacrifice orthogonality, but the factors might better conform with assumptions about how the data are generated. For instance in some applications one might know in advance that the factors should only take on a discrete set of values [133]. In this case, we might try to fit a sum of rank-one matrices that are bounded in ℓ_∞ norm rather than in ℓ_2 norm. Another prior that commonly arises in practice is that the factors are non-negative (i.e., in non-negative matrix factorization). These and other priors on the basic rank-one summands induce different norms on low-rank models than the standard nuclear norm [64], and may be better suited to specific applications.

■ 5.2.3 Background on tangent and normal cones

In order to properly state our results, we recall some basic concepts from convex analysis. A convex set $\mathcal{C}$ is a *cone* if it is closed under positive linear combinations. The polar $\mathcal{C}^*$ of a cone $\mathcal{C}$ is the cone

$$\mathcal{C}^* = \{x \in \mathbb{R}^p \ : \ \langle x, z \rangle \leq 0 \ \forall z \in \mathcal{C}\}.$$

Given some nonzero $\mathbf{x} \in \mathbb{R}^p$ we define the *tangent cone* at $\mathbf{x}$ with respect to the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}}\text{conv}(\mathcal{A})$ as

$$T_{\mathcal{A}}(\mathbf{x}) = \text{cone}\{\mathbf{z} - \mathbf{x} \ : \ \|\mathbf{z}\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}}\}. \tag{5.8}$$

The cone $T_{\mathcal{A}}(\mathbf{x})$ is equal to the set of *descent directions* of the atomic norm $\|\cdot\|_{\mathcal{A}}$ at the point $\mathbf{x}$, i.e., the set of all directions $\mathbf{d}$ such that the directional derivative is negative. This notation is slightly overloaded relative to the notation in Chapter 2.

The *normal cone* $N_{\mathcal{A}}(\mathbf{x})$ at $\mathbf{x}$ with respect to the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}}\text{conv}(\mathcal{A})$ is defined to be the set of all directions $\mathbf{s}$ that form obtuse angles with every descent direction of the atomic norm $\|\cdot\|_{\mathcal{A}}$ at the point $\mathbf{x}$:

$$N_{\mathcal{A}}(\mathbf{x}) = \{\mathbf{s} : \langle \mathbf{s}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \ \forall \mathbf{z} \text{ s.t. } \|\mathbf{z}\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}}\}. \tag{5.9}$$

The normal cone is equal to the set of all hyperplanes given by normal vectors $\mathbf{s}$ that support the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}}\text{conv}(\mathcal{A})$ at $\mathbf{x}$. Observe that the polar cone of the tangent cone $T_{\mathcal{A}}(\mathbf{x})$ is the normal cone $N_{\mathcal{A}}(\mathbf{x})$ and vice-versa. Moreover we have the following basic characterization

$$N_{\mathcal{A}}(\mathbf{x}) = \text{cone}(\partial\|\mathbf{x}\|_{\mathcal{A}}),$$

which states that the normal cone $N_{\mathcal{A}}(\mathbf{x})$ is the conic hull of the subdifferential of the atomic norm at $\mathbf{x}$.

■ 5.2.4 Recovery condition

The following result gives a characterization of the favorable underlying geometry required for exact recovery. Let $\text{null}(\Phi)$ denote the nullspace of the operator Φ.

Proposition 5.2.1. *We have that $\hat{\mathbf{x}} = \mathbf{x}^\star$ is the unique optimal solution of (5.5) if and only if $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \{0\}$.*

Proof. Eliminating the equality constraints in (5.5) we have the equivalent optimization problem

$$\min_{\mathbf{d}} \ \|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} \quad \text{s.t. } \mathbf{d} \in \text{null}(\Phi).$$

Suppose $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = 0$. Since $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} \leq \|\mathbf{x}^\star\|$ implies $\mathbf{d} \in T_{\mathcal{A}}(\mathbf{x}^\star)$, we have that $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} > \|\mathbf{x}^\star\|_{\mathcal{A}}$ for all $\mathbf{d} \in \text{null}(\Phi) \setminus \{0\}$. Conversely $\mathbf{x}^\star$ is the unique optimal solution of (5.5) if $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} > \|\mathbf{x}^\star\|_{\mathcal{A}}$ for all $\mathbf{d} \in \text{null}(\Phi) \setminus \{0\}$, which implies that $\mathbf{d} \notin T_{\mathcal{A}}(\mathbf{x}^\star)$. □

Proposition 5.2.1 asserts that the atomic norm heuristic succeeds if the nullspace of the sampling operator does not intersect the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ at $\mathbf{x}^\star$. In Section 5.3 we provide a characterization of tangent cones that determines the number of Gaussian measurements required to guarantee such an empty intersection.

A tightening of this empty intersection condition can also be used to address the noisy approximation problem. The following proposition characterizes when $\mathbf{x}^\star$ can be *well-approximated* using the convex program (5.7).

Proposition 5.2.2. *Suppose that we are given n noisy measurements $\mathbf{y} = \Phi\mathbf{x}^\star + \omega$ where $\|\omega\| \leq \delta$, and $\Phi : \mathbb{R}^p \to \mathbb{R}^n$. Let $\hat{\mathbf{x}}$ denote an optimal solution of (5.7). Further suppose for all $\mathbf{z} \in T_{\mathcal{A}}(\mathbf{x}^\star)$ that we have $\|\Phi\mathbf{z}\| \geq \epsilon\|\mathbf{z}\|$. Then $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$.*

Proof. The set of descent directions at $\mathbf{x}^\star$ with respect to the atomic norm ball is given by the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. The error vector $\hat{\mathbf{x}} - \mathbf{x}^\star$ lies in $T_{\mathcal{A}}(\mathbf{x}^\star)$ because $\hat{\mathbf{x}}$ is a minimal atomic norm solution, and hence $\|\hat{\mathbf{x}}\|_{\mathcal{A}} \leq \|\mathbf{x}^\star\|_{\mathcal{A}}$. It follows by the triangle inequality that

$$\|\Phi(\hat{\mathbf{x}} - \mathbf{x}^\star)\| \leq \|\Phi\hat{\mathbf{x}} - \mathbf{y}\| + \|\Phi\mathbf{x}^\star - \mathbf{y}\| \leq 2\delta. \tag{5.10}$$

By assumption we have that

$$\|\Phi(\hat{\mathbf{x}} - \mathbf{x}^\star)\| \geq \epsilon\|\hat{\mathbf{x}} - \mathbf{x}^\star\|, \tag{5.11}$$

which allows us to conclude that $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$. □

Therefore, we need only concern ourselves with estimating the minimum value of $\frac{\|\Phi\mathbf{z}\|}{\|\mathbf{z}\|}$ for non-zero $\mathbf{z} \in T_{\mathcal{A}}(\mathbf{x}^\star)$. We denote this quantity as the minimum gain of the measurement operator Φ restricted to the cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. In particular if this minimum gain is bounded away from zero, then the atomic norm heuristic also provides robust recovery when we have access to noisy linear measurements of $\mathbf{x}^\star$.

■ 5.2.5 Why atomic norm?

The atomic norm induced by a set $\mathcal{A}$ possesses a number of favorable properties that are useful for recovering "simple" models from limited linear measurements. The key point to note from Section 5.2.4 is that the smaller the tangent cone at a point $\mathbf{x}^\star$ with respect to $\text{conv}(\mathcal{A})$, the easier it is to satisfy the empty-intersection condition of Proposition 5.2.1.

Based on this observation it is desirable that points in $\text{conv}(\mathcal{A})$ with smaller tangent cones correspond to simpler models, while points in $\text{conv}(\mathcal{A})$ with larger tangent cones generally correspond to more complicated models. The construction of $\text{conv}(\mathcal{A})$ by taking the convex hull of $\mathcal{A}$ ensures that this is the case. The extreme points of $\text{conv}(\mathcal{A})$ correspond to the simplest models, i.e., those models formed from a single element of $\mathcal{A}$. Further the low-dimensional faces of $\text{conv}(\mathcal{A})$ consist of those elements that are obtained by taking linear combinations of a few basic atoms from $\mathcal{A}$. These are precisely the properties desired as points lying in these low-dimensional faces of $\text{conv}(\mathcal{A})$ have smaller tangent cones than those lying on larger faces.

We also note that the atomic norm is in some sense the best possible convex heuristic for recovering simple models. Specifically the unit ball of any convex penalty heuristic must satisfy a key property: the tangent cone at any $\mathbf{a} \in \mathcal{A}$ with respect to this unit ball

must contain the vectors $\mathbf{a}' - \mathbf{a}$ for all $\mathbf{a}' \in \mathcal{A}$. The best convex penalty function is one in which the tangent cones at $\mathbf{a} \in \mathcal{A}$ to the unit ball are the smallest possible, while still satisfying this requirement. This is because, as described above, smaller tangent cones are more likely to satisfy the empty intersection condition required for exact recovery. It is clear that the smallest such convex set is precisely $\text{conv}(\mathcal{A})$, hence implying that the atomic norm is the best convex heuristic for recovering simple models.

Our reasons for proposing the atomic norm as a useful convex heuristic are quite different from previous justifications of the ℓ_1 norm and the nuclear norm. In particular let $f : \mathbb{R}^p \to \mathbb{R}$ denote the cardinality function that counts the number of nonzero entries of a vector. Then the ℓ_1 norm is the *convex envelope* of f restricted to the unit ball of the ℓ_∞ norm, i.e., the best convex underestimator of f restricted to vectors in the ℓ_∞-norm ball. This view of the ℓ_1 norm in relation to the function f is often given as a justification for its effectiveness in recovering sparse vectors. However if we consider the convex envelope of f restricted to the Euclidean norm ball, then we obtain a very different convex function than the ℓ_1 norm! With more general atomic sets, it may not be clear *a priori* what the bounding set should be in deriving the convex envelope. In contrast the viewpoint adopted in this chapter leads to a natural, unambiguous construction of the ℓ_1 norm and other general atomic norms. Further as explained above it is the favorable *facial structure* of the atomic norm ball that makes the atomic norm a suitable convex heuristic to recover simple models, and this connection is transparent in the definition of the atomic norm.

■ 5.3 Recovery from Generic Measurements

We consider the question of using the convex program (5.5) to recover "simple" models formed according to (5.1) from a *generic* measurement operator or map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$. Specifically, we wish to compute estimates on the number of measurements n so that we have exact recovery using (5.5) for *most* operators comprising of n measurements. That is, the measure of n-measurement operators for which recovery fails using (5.5) must be exponentially small. In order to conduct such an analysis we study random *Gaussian* maps Φ, in which the entries are independent and identically distributed Gaussians. These measurement operators have the property that the nullspace $\text{null}(\Phi)$ is uniformly distributed among the set of all $(p-n)$-dimensional subspaces in $\mathbb{R}^p$. In particular we analyze when such operators satisfy the conditions of Proposition 5.2.1

and Proposition 5.2.2 for exact recovery.

■ 5.3.1 Recovery conditions based on Gaussian width

Proposition 5.2.1 requires that the nullspace of the measurement operator Φ must miss the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. Gordon [76] gave a solution to the problem of characterizing the probability that a random subspace (of some fixed dimension) distributed uniformly misses a cone. We begin by defining the Gaussian width of a set, which plays a key role in Gordon's analysis.

Definition 5.3.1. *The* Gaussian width *of a set $S \subset \mathbb{R}^p$ is defined as:*

$$w(S) := \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{z}\in S} \mathbf{g}^T\mathbf{z}\right],$$

where $\mathbf{g} \sim \mathcal{N}(0, I)$ is a vector of independent zero-mean unit-variance Gaussians.

Gordon characterized the likelihood that a random subspace misses a cone $\mathcal{C}$ purely in terms of the dimension of the subspace and the Gaussian width $w(\mathcal{C} \cap \mathbb{S}^{p-1})$, where $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ is the unit sphere. Before describing Gordon's result formally, we introduce some notation. Let λ_k denote the expected length of a k-dimensional Gaussian random vector. By elementary integration, we have that $\lambda_k = \sqrt{2}\Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2})$. Further by induction one can show that λ_k is tightly bounded as $\frac{k}{\sqrt{k+1}} \leq \lambda_k \leq \sqrt{k}$.

The main idea underlying Gordon's theorem is a bound on the minimum gain of an operator restricted to a set. Specifically, recall that $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \{0\}$ is the condition required for recovery by Proposition 5.2.1. Thus if we have that the minimum gain of Φ restricted to vectors in the set $T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ is bounded away from zero, then it is clear that $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \emptyset$. We refer to such minimum gains restricted to a subset of the sphere as *restricted minimum singular values*, and the following theorem of Gordon gives a bound these quantities [76]:

Theorem 5.3.1 (Gordon's Minimum Restricted Singular Values Theorem)**.** *Let Ω be a closed subset of $\mathbb{S}^{p-1}$. Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ be a random map with i.i.d. zero-mean Gaussian entries having variance one. Then provided that $\lambda_k \geq w(\Omega) + \epsilon$, we have*

$$\mathbb{P}\left[\min_{\mathbf{z}\in\Omega} \|\Phi\mathbf{z}\|_2 \geq \epsilon\right] \geq 1 - \frac{5}{2}\exp\left(-\frac{1}{18}(\lambda_k - w(\Omega) - \epsilon)^2\right). \tag{5.12}$$

This theorem is not explicitly stated as such in [76] but the proof follows directly as a result of Gordon's arguments. Theorem 5.3.1 allows us to characterize exact recovery in

the noise-free case using the convex program (5.5), and robust recovery in the noisy case using the convex program (5.7). Specifically, we consider the number of measurements required for exact or robust recovery when the measurement map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ consists of i.i.d. zero-mean Gaussian entries having variance $1/n$. The normalization of the variance ensures that the columns of Φ are approximately unit-norm, and is necessary in order to properly define a signal-to-noise ratio. The following corollary summarizes the main results of interest in our setting:

Corollary 5.3.1. *Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ be a random map with i.i.d. zero-mean Gaussian entries having variance $1/n$. Further let $\Omega = T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ denote the spherical part of the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$.*

1. *Suppose that we have measurements $\mathbf{y} = \Phi\mathbf{x}^\star$, and we solve the convex program (5.5). Then $\mathbf{x}^\star$ is the unique optimum of (5.5) with high probability provided that*

$$n \geq w(\Omega)^2 + \mathcal{O}(1).$$

2. *Suppose that we have noisy measurements $\mathbf{y} = \Phi\mathbf{x}^\star + \omega$, with the noise ω bounded as $\|\omega\| \leq \delta$, and that we solve the convex program (5.7). Letting $\hat{\mathbf{x}}$ denote the optimal solution of (5.7), we have that $\|\mathbf{x}^\star - \hat{\mathbf{x}}\| \leq \frac{2\delta}{\epsilon}$ with high probability provided*

$$n \geq \frac{w(\Omega)^2}{(1-\epsilon)^2} + \mathcal{O}(1).$$

Proof. The two results are simple consequences of Theorem 5.3.1:

1. The first part follows by setting $\epsilon = 0$ in Theorem 5.3.1.

2. For $\epsilon \in (0, 1)$ we have from Theorem 5.3.1 that

$$\|\Phi(\mathbf{z})\| = \|\mathbf{z}\| \left\| \Phi\left(\frac{\mathbf{z}}{\|\mathbf{z}\|}\right) \right\| \geq \frac{\epsilon}{\sqrt{n}} \|\mathbf{z}\| \tag{5.13}$$

   for all $\mathbf{z} \in \mathcal{T}_{\mathcal{A}}(\mathbf{x}^\star)$ with high probability. Therefore we can apply Proposition 5.2.2 to conclude that $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$ with high probability, provided that $n \geq \frac{w(\Omega)^2}{(1-\epsilon)^2} + \mathcal{O}(1)$.

□

Gordon's theorem thus provides a simple characterization of the number of measurements required for reconstruction with the atomic norm. Indeed the Gaussian width of $\Omega = T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ is the only quantity that we need to compute in order to obtain bounds for both exact and robust recovery. Unfortunately it is in general not easy to compute Gaussian widths. Rudelson and Vershynin [127] have worked out Gaussian widths for the special case of tangent cones at sparse vectors on the boundary of the ℓ_1 ball, and derived results for sparse vector recovery using ℓ_1 minimization that improve upon previous results. In the next section we give various well-known properties of the Gaussian width that are useful in some computations. In Section 5.3.3 we discuss a new approach to width computations that gives near-optimal recovery bounds in a variety of settings.

■ 5.3.2 Properties of Gaussian width

In this section we record several elementary properties of the Gaussian width that are useful for computation. We begin by making some basic observations, which are easily derived.

First we note that the width is monotonic. If $S_1 \subseteq S_2 \subseteq \mathbb{R}^p$, then it is clear from the definition of the Gaussian width that

$$w(S_1) \leq w(S_2).$$

Second we note that if we have a set $S \subseteq \mathbb{R}^p$, then the Gaussian width of S is equal to the Gaussian width of the convex hull of S:

$$w(S) = w(\text{conv}(S)).$$

This result follows from the basic fact in convex analysis that the maximum of a convex function over a convex set is achieved at an extreme point of the convex set. Third if $V \subset \mathbb{R}^p$ is a subspace in $\mathbb{R}^p$, then we have that

$$w(V \cap \mathbb{S}^{p-1}) = \sqrt{\dim(V)},$$

which follows from standard results on random Gaussians. This result also agrees with the intuition that a random Gaussian map Φ misses a k-dimensional subspace with high probability as long as $\dim(\text{null}(\Phi)) \geq k+1$. Finally, if a cone $S \subset \mathbb{R}^p$ is such that $S = S_1 \oplus S_2$, where $S_1 \subset \mathbb{R}^p$ is a k-dimensional cone, $S_2 \subset \mathbb{R}^p$ is a $(p-k)$-dimensional

cone that is orthogonal to S_1, and $\oplus$ denotes the direct sum operation, then the width can be decomposed as follows:

$$w(S \cap \mathbb{S}^{p-1})^2 \leq w(S_1 \cap \mathbb{S}^{p-1})^2 + w(S_2 \cap \mathbb{S}^{p-1})^2.$$

These observations are useful in a variety of situations. For example a width computation that frequently arises is one in which $S = S_1 \oplus S_2$ as described above, with S_1 being a k-dimensional subspace. It follows that the width of $S \cap \mathbb{S}^{p-1}$ is bounded as

$$w(S \cap \mathbb{S}^{p-1})^2 \leq k + w(S_2 \cap \mathbb{S}^{p-1})^2.$$

These basic operations involving Gaussian widths were used by Rudelson and Vershynin [127] to compute the Gaussian widths of tangent cones at sparse vectors with respect to the ℓ_1 norm ball.

Another tool for computing Gaussian widths is based on Dudley's inequality [57,96], which bounds the width of a set in terms of the covering number of the set at all scales.

Definition 5.3.2. *Let S be an arbitrary compact subset of $\mathbb{R}^p$. The* covering number *of S in the Euclidean norm at resolution ϵ is the smallest number, $\mathfrak{N}(S, \epsilon)$, such that $\mathfrak{N}(S, \epsilon)$ Euclidean balls of radius ϵ cover S.*

Theorem 5.3.2 (Dudley's Inequality)**.** *Let S be an arbitrary compact subset of $\mathbb{R}^p$, and let $\mathbf{g}$ be a random vector with i.i.d. zero-mean, unit-variance Gaussian entries. Then*

$$w(S) \leq 24 \int_0^\infty \sqrt{\log(\mathfrak{N}(S, \epsilon))} d\epsilon. \tag{5.14}$$

We note here that a weak converse to Dudley's inequality can be obtained via Sudakov's Minoration [96] by using the covering number for just a single scale. Specifically, we have the following *lower bound* on the Gaussian width of a compact subset $S \subset \mathbb{R}^p$ for any $\epsilon > 0$:

$$w(S) \geq c\epsilon \sqrt{\log(\mathfrak{N}(S, \epsilon))}.$$

Here $c > 0$ is some universal constant.

Although Dudley's inequality can be applied quite generally, estimating covering numbers is difficult in most instances. There are a few simple characterizations available for spheres and Sobolev spaces, and some tractable arguments based on Maurey's empirical method [96]. However it is not evident how to compute these numbers for general convex cones. Also, in order to apply Dudley's inequality we need to estimate

the covering number at all scales. Further Dudley's inequality can be quite loose in its estimates, and it often introduces extraneous polylogarithmic factors. In the next section we describe a new mechanism for estimating Gaussian widths, which provides near-optimal guarantees for recovery of sparse vectors and low-rank matrices, as well as for several of the recovery problems discussed in Section 5.3.4.

■ 5.3.3 New results on Gaussian width

We discuss a new dual framework for computing Gaussian widths. In particular we express the Gaussian width of a cone in terms of the dual of the cone. To be fully general let $\mathcal{C}$ be a non-empty convex cone in $\mathbb{R}^p$, and let $\mathcal{C}^*$ denote the polar of $\mathcal{C}$. We can then upper bound the Gaussian width of any cone $\mathcal{C}$ in terms of the polar cone $\mathcal{C}^*$:

Proposition 5.3.1. *Let $\mathcal{C}$ be any non-empty convex cone in $\mathbb{R}^p$, and let $\mathbf{g} \sim \mathcal{N}(0, I)$ be a random Gaussian vector. Then we have the following bound:*

$$w(\mathcal{C} \cap \mathbb{S}^{p-1}) \leq \mathbb{E}_{\mathbf{g}} \left[\operatorname{dist}(\mathbf{g}, \mathcal{C}^*) \right],$$

where dist *here denotes the Euclidean distance between a point and a set.*

The proof is given in Appendix C.1, and it follows from an appeal to convex duality. Proposition 5.3.1 is more or less a restatement of the fact that the support function of a convex cone is equal to the distance to its polar cone. As it is the square of the Gaussian width that is of interest to us (see Corollary 5.3.1), it is often useful to apply Jensen's inequality to make the following approximation:

$$\mathbb{E}_{\mathbf{g}}[\operatorname{dist}(\mathbf{g}, \mathcal{C}^*)]^2 \leq \mathbb{E}_{\mathbf{g}}[\operatorname{dist}(\mathbf{g}, \mathcal{C}^*)^2]. \tag{5.15}$$

The inspiration for our characterization in Proposition 5.3.1 of the width of a cone in terms of the expected distance to its dual came from the work of Stojnic [134], who used linear programming duality to construct Gaussian-width-based estimates for analyzing recovery in sparse reconstruction problems. Specifically, Stojnic's relatively simple approach recovered well-known phase transitions in sparse signal recovery [55], and also generalized to block sparse signals and other forms of structured sparsity.

This new dual characterization yields a number of useful bounds on the Gaussian width, which we describe here. In the following section we use these bounds to derive new recovery results. The first result is a bound on the Gaussian width of a cone in terms of the Gaussian width of its polar.

Lemma 5.3.1. *Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a non-empty closed, convex cone. Then we have that*

$$w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 + w(\mathcal{C}^* \cap \mathbb{S}^{p-1})^2 \leq p.$$

Proof. Combining Proposition 5.3.1 and (5.15), we have that

$$w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 \leq \mathbb{E}_{\mathbf{g}} \left[\operatorname{dist}(\mathbf{g}, \mathcal{C}^*)^2 \right],$$

where as before $\mathbf{g} \sim \mathcal{N}(0, I)$. For any $\mathbf{z} \in \mathbb{R}^p$ we let $\Pi_{\mathcal{C}}(\mathbf{z}) = \arg\inf_{\mathbf{u} \in \mathcal{C}} \|\mathbf{z} - \mathbf{u}\|$ denote the projection of $\mathbf{z}$ onto $\mathcal{C}$. From standard results in convex analysis [124], we note that one can decompose any $\mathbf{z} \in \mathbb{R}^p$ into orthogonal components as follows:

$$\mathbf{z} = \Pi_{\mathcal{C}}(\mathbf{z}) + \Pi_{\mathcal{C}^*}(\mathbf{z}), \quad \langle \Pi_{\mathcal{C}}(\mathbf{z}), \Pi_{\mathcal{C}^*}(\mathbf{z}) \rangle = 0.$$

Therefore we have the following sequence of bounds:

$$\begin{aligned} w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 &\leq \mathbb{E}_{\mathbf{g}} \left[\operatorname{dist}(\mathbf{g}, \mathcal{C}^*)^2 \right] \\ &= \mathbb{E}_{\mathbf{g}} \left[\|\Pi_{\mathcal{C}}(\mathbf{g})\|^2 \right] \\ &= \mathbb{E}_{\mathbf{g}} \left[\|\mathbf{g}\|^2 - \|\Pi_{\mathcal{C}^*}(\mathbf{g})\|^2 \right] \\ &= p - \mathbb{E}_{\mathbf{g}} \left[\|\Pi_{\mathcal{C}^*}(\mathbf{g})\|^2 \right] \\ &= p - \mathbb{E}_{\mathbf{g}} \left[\operatorname{dist}(\mathbf{g}, \mathcal{C})^2 \right] \\ &\leq p - w(\mathcal{C}^* \cap \mathbb{S}^{p-1})^2. \end{aligned}$$

□

In many recovery problems one is interested in computing the width of a self-dual cone. For such cones the following corollary to Lemma 5.3.1 gives a simple solution:

Corollary 5.3.2. *Let $\mathcal{C} \subset \mathbb{R}^p$ be a self-dual cone, i.e., $\mathcal{C} = -\mathcal{C}^*$. Then we have that*

$$w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 \leq \frac{p}{2}.$$

Proof. The proof follows directly from Lemma 5.3.1 as $w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 = w(\mathcal{C}^* \cap \mathbb{S}^{p-1})^2$. □

Our next bound for the width of a cone $\mathcal{C}$ is based on the volume of its polar $\mathcal{C}^* \cap \mathbb{S}^{p-1}$. The *volume* of a measurable subset of the sphere is the fraction of the sphere $\mathbb{S}^{p-1}$ covered by the subset. Thus it is a quantity between zero and one.

Theorem 5.3.3 (Gaussian width from volume of the polar). *Let $\mathcal{C} \subseteq \mathbb{R}^p$ be any closed, convex, solid cone, and suppose that its polar $\mathcal{C}^*$ is such that $\mathcal{C}^* \cap \mathbb{S}^{p-1}$ has a volume of $\Theta \in [0,1]$. Then for $p \geq 9$ we have that*

$$w(\mathcal{C} \cap \mathbb{S}^{p-1}) \leq 3\sqrt{\log\left(\frac{4}{\Theta}\right)}.$$

The proof of this theorem is given in Appendix C.2. The main property that we appeal to in the proof is *Gaussian isoperimetry*. In particular there is a formal sense in which a spherical cap[1] is the "extremal case" among all subsets of the sphere with a given volume Θ. Other than this observation the proof mainly involves a sequence of integral calculations.

Note that if we are given a specification of a cone $\mathcal{C} \subset \mathbb{R}^p$ in terms of a membership oracle, it is possible to efficiently obtain good numerical estimates of the volume of $\mathcal{C} \cap \mathbb{S}^{p-1}$ [58]. Moreover, simple symmetry arguments often give relatively accurate estimates of these volumes. Such estimates can then be plugged into Theorem 5.3.3 to yield bounds on the width.

■ 5.3.4 New recovery bounds

We use the bounds derived in the last section to obtain new recovery results. First using the dual characterization of the Gaussian width in Proposition 5.3.1, we are able to obtain sharp bounds on the number of measurements required for recovering sparse vectors and low-rank matrices from random Gaussian measurements using convex optimization (i.e., ℓ_1-norm and nuclear norm minimization).

Proposition 5.3.2. *Let $\mathbf{x}^\star \in \mathbb{R}^p$ be an s-sparse vector. Letting $\mathcal{A}$ denote the set of unit-Euclidean-norm one-sparse vectors, we have that*

$$w(T_{\mathcal{A}}(\mathbf{x}^\star))^2 \leq \begin{cases} 2s\left(\log\left(\frac{p-s}{s}\right)+1\right)\right) & s < \frac{1}{1+e}p \\ 2s(\log(p-s)+1) & \text{otherwise.} \end{cases}$$

Thus, when $s < 0.26p$, $2s(\log(p/s-1)+1)$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via ℓ_1 norm minimization with high probability. Moreover, $2s(\log(p-s)+1)$ measurements suffice for any value of s.

[1] A *spherical cap* is a subset of the sphere obtained by intersecting the sphere $\mathbb{S}^{p-1}$ with a halfspace.

Proposition 5.3.3. *Let $\mathbf{x}^\star$ be an $m_1 \times m_2$ rank-r matrix with $m_1 \leq m_2$. Letting $\mathcal{A}$ denote the set of unit-Euclidean-norm rank-one matrices, we have that*

$$w(T_{\mathcal{A}}(\mathbf{x}^\star))^2 \leq 3r(m_1 + m_2 - r).$$

Thus $3r(m_1+m_2-r)$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via nuclear norm minimization with high probability.

The proofs of these propositions are given in Appendix C.3. The number of measurements required by these bounds is on the same order as previously known results [28, 53], but with improved constants. We also note that we have robust recovery at these thresholds. Further these results do not require explicit recourse to any type of restricted isometry property [28], and the proofs are simple and based on elementary integrals.

Next we obtain a set of recovery results by appealing to Corollary 5.3.2 on the width of a self-dual cone. These examples correspond to the recovery of individual atoms (i.e., the extreme points of the set $\operatorname{conv}(\mathcal{A})$), although the same machinery is applicable in principle to estimate the number of measurements required to recover models formed as sums of a few atoms (i.e., points lying on low-dimensional faces of $\operatorname{conv}(\mathcal{A})$). We first obtain a well-known result on the number of measurements required for recovering sign-vectors via ℓ_∞ norm minimization.

Proposition 5.3.4. *Let $\mathbf{x}^\star \in \{-1, +1\}^p$ be a sign-vector in $\mathbb{R}^p$, and let $\mathcal{A}$ be the set of all such sign-vectors. Then we have that*

$$w(T_{\mathcal{A}}(\mathbf{x}^\star))^2 \leq \frac{p}{2}.$$

Thus $\frac{p}{2}$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via ℓ_∞-norm minimization with high probability.

Proof. The tangent cone at any signed vector $\mathbf{x}^\star$ with respect to the ℓ_∞ ball is a rotation of the nonnegative orthant. Thus we only need to compute the Gaussian width of an orthant in $\mathbb{R}^p$. As the orthant is self-dual, we have the required bound from Corollary 5.3.2. □

This result agrees with previously computed bounds in [56, 102], which relied on a more complicated combinatorial argument. Next we compute the number of measurements required to recover orthogonal matrices via spectral-norm minimization (see

Section 5.2.2). Let $\mathbb{O}(m)$ denote the group of $m \times m$ orthogonal matrices, viewed as a subgroup of the set of nonsingular matrices in $\mathbb{R}^{m \times m}$.

Proposition 5.3.5. *Let $\mathbf{x}^\star \in \mathbb{R}^{m \times m}$ be an orthogonal matrix, and let $\mathcal{A}$ be the set of all orthogonal matrices. Then we have that*

$$w(T_{\mathcal{A}}(\mathbf{x}^\star))^2 \leq \frac{3m^2 - m}{4}.$$

Thus $\frac{3m^2-m}{4}$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via spectral-norm minimization with high probability.

Proof. Due to the symmetry of the orthogonal group, it suffices to consider the tangent cone at the identity matrix I with respect to the spectral norm ball. Recall that the spectral norm ball is the convex hull of the orthogonal matrices. Therefore the *tangent space* at the identity matrix with respect to the orthogonal group $\mathbb{O}(m)$ is a subset of the tangent cone $T_{\mathcal{A}}(I)$. It is well-known that this tangent space is the set of all $m \times m$ skew-symmetric matrices. Thus we only need to compute the component S of $T_{\mathcal{A}}(I)$ that lies in the subspace of symmetric matrices:

$$\begin{aligned} S &= \operatorname{cone}\{M - I : \|M\|_{\mathcal{A}} \leq 1,\ M \text{ symmetric}\} \\ &= \operatorname{cone}\{UDU^T - UU^T : \|D\|_{\mathcal{A}} \leq 1,\ D \text{ diagonal},\ U \in \mathbb{O}(m)\} \\ &= \operatorname{cone}\{U(D - I)U^T : \|D\|_{\mathcal{A}} \leq 1,\ D \text{ diagonal},\ U \in \mathbb{O}(m)\} \\ &= -\mathrm{PSD}_m. \end{aligned}$$

Here PSD_m denotes the set of $m \times m$ symmetric positive-semidefinite matrices. As this cone is self-dual, we can apply Corollary 5.3.2 in conjunction with the observations in Section 5.3.2 to conclude that

$$w(T_{\mathcal{A}}(I))^2 \leq \binom{m}{2} + \frac{1}{2}\binom{m+1}{2} = \frac{3m^2 - m}{4}.$$

□

We note that the number of degrees of freedom in an $m \times m$ orthogonal matrix (i.e., the dimension of the manifold of orthogonal matrices) is $\frac{m(m-1)}{2}$. Proposition 5.3.4 and Proposition 5.3.5 point to the importance of obtaining recovery bounds with sharp constants. Larger constants in either result would imply that the number of measurements required exceeds the ambient dimension of the underlying $\mathbf{x}^\star$. In these and many

other cases of interest Gaussian width arguments not only give order-optimal recovery results, but also provide precise constants that result in sharp recovery thresholds.

Finally we give a third set of recovery results that appeal to the Gaussian width bound of Theorem 5.3.3. The following measurement bound applies to cases when $\text{conv}(\mathcal{A})$ is a *symmetric polytope* (roughly speaking, all the vertices are "equivalent"), and is a simple corollary of Theorem 5.3.3.

Corollary 5.3.3. *Suppose that the set $\mathcal{A}$ is a finite collection of m points, with the convex hull $\text{conv}(\mathcal{A})$ being a vertex-transitive polytope [149] whose vertices are the points in $\mathcal{A}$. Using the convex program (5.5) we have that $9\log(m)$ random Gaussian measurements suffice, with high probability, for exact recovery of a point in $\mathcal{A}$, i.e., a vertex of $\text{conv}(\mathcal{A})$.*

Proof. We recall the basic fact from convex analysis that the normal cones at the vertices of a convex polytope in $\mathbb{R}^p$ provide a partitioning of $\mathbb{R}^p$. As $\text{conv}(\mathcal{A})$ is a vertex-transitive polytope, the normal cone at a vertex covers $\frac{1}{m}$ fraction of $\mathbb{R}^p$. Applying Theorem 5.3.3, we have the desired result. □

Clearly we require the number of vertices to be bounded as $m \leq \exp\{\frac{p}{9}\}$, so that the estimate of the number of measurements is not vacuously true. This result has useful consequences in settings in which $\text{conv}(\mathcal{A})$ is a *combinatorial polytope*, as such polytopes are often vertex-transitive. We have the following example on the number of measurements required to recover permutation matrices:

Proposition 5.3.6. *Let $\mathbf{x}^\star \in \mathbb{R}^{m\times m}$ be a permutation matrix, and let $\mathcal{A}$ be the set of all $m \times m$ permutation matrices. Then $9\,m\log(m)$ random Gaussian measurements suffice, with high probability, to recover $\mathbf{x}^\star$ by solving the optimization problem (5.5), which minimizes the norm induced by the Birkhoff polytope of doubly stochastic matrices.*

Proof. This result follows from Corollary 5.3.3 by noting that there are $m!$ permutation matrices of size $m \times m$. □

■ 5.4 Representability and Algebraic Geometry of Atomic Norms

■ 5.4.1 Role of algebraic structure

All of our discussion thus far has focussed on arbitrary atomic sets $\mathcal{A}$. As seen in Section 5.2 the geometry of the convex hull $\text{conv}(\mathcal{A})$ completely determines conditions

under which exact recovery is possible using the convex program (5.5). In this section we address the question of *computationally representing* the convex hull $\text{conv}(\mathcal{A})$ (or equivalently of computing the atomic norm $\|\cdot\|_{\mathcal{A}}$). These issues are critical in order to be able to solve the convex optimization problem (5.5). Although the convex hull $\text{conv}(\mathcal{A})$ is a well-defined object, in general we may not even be able to computationally represent it (for example, if $\mathcal{A}$ is a fractal). In order to obtain exact or approximate representations (analogous to the cases of the ℓ_1 norm and the nuclear norm) it is important to impose some structure on the atomic set $\mathcal{A}$. We focus on cases in which the set $\mathcal{A}$ has algebraic structure. Specifically let the ring of multivariate polynomials in p variables be denoted by $\mathbb{R}[\mathbf{x}] = \mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_p]$. We then consider real algebraic varieties [18]:

Definition 5.4.1. *A* real algebraic variety $S \subseteq \mathbb{R}^p$ *is the set of real solutions of a system of polynomial equations:*

$$S = \{\mathbf{x} : g_j(\mathbf{x}) = 0, \ \forall j\},$$

where $\{g_j\}$ *is a finite collection of polynomials in* $\mathbb{R}[\mathbf{x}]$.

Indeed all of the atomic sets $\mathcal{A}$ considered in this chapter are examples of algebraic varieties. Algebraic varieties have the remarkable property that (the closure of) their convex hull can be arbitrarily well-approximated in a constructive manner as (the projection of) a set defined by linear matrix inequality constraints. A potential complication may arise, however, if these semidefinite representations are intractable to compute in polynomial time. In such cases it is possible to approximate the convex hulls via a hierarchy of tractable semidefinite relaxations. We describe these results in more detail in Section 5.4.2. Therefore the atomic norm minimization problems such as (5.7) arising in such situations can be solved exactly or approximately via semidefinite programming.

Algebraic structure also plays a second important role in atomic norm minimization problems. If an atomic norm $\|\cdot\|_{\mathcal{A}}$ is intractable to compute, we may approximate it via a more tractable norm $\|\cdot\|_{app}$. However not every approximation of the atomic norm is equally good for solving inverse problems. As illustrated in Figure 5.1 we can construct approximations of the ℓ_1 ball that are tight in a *metric* sense, with $(1-\epsilon)\|\cdot\|_{app} \leq \|\cdot\|_{\ell_1} \leq (1+\epsilon)\|\cdot\|_{app}$, but where the tangent cones at sparse vectors in the new norm are halfspaces. In such a case, the number of measurements required to recover

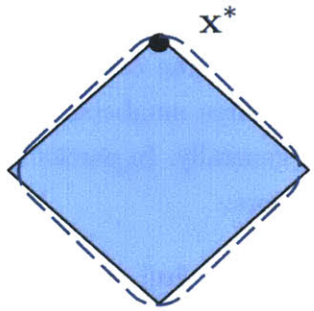

Figure 5.1. The convex body given by the dotted line is a good metric approximation to the ℓ_1 ball. However as its "corners" are "smoothed out", the tangent cone at $\mathbf{x}^\star$ goes from being a proper cone (with respect to the ℓ_1 ball) to a halfspace (with respect to the approximation).

the sparse vector ends up being on the same order as the ambient dimension. (Note that the ℓ_1-norm is in fact tractable to compute; we simply use it here for illustrative purposes.) The key property that we seek in approximations to an atomic norm $\|\cdot\|_{\mathcal{A}}$ is that they *preserve algebraic structure* such as the vertices/extreme points and more generally the low-dimensional faces of the $\text{conv}(\mathcal{A})$. As discussed in Section 5.2.5 points on such low-dimensional faces correspond to simple models, and algebraic-structure preserving approximations ensure that the tangent cones at simple models with respect to the approximations are not too much larger than the corresponding tangent cones with respect to the original atomic norms.

■ 5.4.2 Semidefinite relaxations using Theta bodies – an example

In this section we give an example of a family of semidefinite relaxations to the atomic norm minimization problem; the hierarchy of relaxations is obtained using the Theta-bodies construction of [77] (see Chapter 2 for a brief summary), and is applicable whenever the atomic set has algebraic structure. To begin with if we approximate the atomic norm $\|\cdot\|_{\mathcal{A}}$ by another atomic norm $\|\cdot\|_{\tilde{\mathcal{A}}}$ defined using a *larger* collection of atoms $\mathcal{A} \subseteq \tilde{\mathcal{A}}$, it is clear that

$$\|\cdot\|_{\tilde{\mathcal{A}}} \leq \|\cdot\|_{\mathcal{A}}.$$

Consequently outer approximations of the atomic set give rise to approximate norms that provide lower bounds on the optimal value of the problem (5.5).

In order to provide such lower bounds on the optimal value of (5.5), we discuss semidefinite relaxations of the convex hull $\text{conv}(\mathcal{A})$ based on Theta bodies. Specifically

we discuss an example application of these relaxations to the problem of approximating the tensor nuclear norm. We focus on the case of tensors of order three that lie in $\mathbb{R}^{m\times m\times m}$, i.e., tensors indexed by three numbers, for notational simplicity, although our discussion is applicable more generally. In particular the atomic set $\mathcal{A}$ is the set of unit-Euclidean-norm rank-one tensors:

$$\begin{aligned}\mathcal{A} &= \{\mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w} : \mathbf{u},\mathbf{v},\mathbf{w}\in\mathbb{R}^m,\ \|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1\} \\ &= \{N\in\mathbb{R}^{m^3} : N = \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w},\ \mathbf{u},\mathbf{v},\mathbf{w}\in\mathbb{R}^m,\ \|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1\},\end{aligned}$$

where $\mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}$ is the tensor product of three vectors. Note that the second description is written as the projection onto $\mathbb{R}^{m^3}$ of a variety defined in $\mathbb{R}^{m^3+3m}$. The nuclear norm is then given by (5.2), and is intractable to compute in general. Now let $I_{\mathcal{A}}$ denote a polynomial ideal of polynomial maps from $\mathbb{R}^{m^3+m}$ to $\mathbb{R}$:

$$I_{\mathcal{A}} = \{g : g = \sum_{i,j,k=1}^{m} g_{ijk}(N_{ijk} - \mathbf{u}_i\mathbf{v}_j\mathbf{w}_k) + g_u(\mathbf{u}^T\mathbf{u}-1) + g_v(\mathbf{v}^T\mathbf{v}-1) + g_w(\mathbf{w}^T\mathbf{w}-1), \forall g_{ijk}, g_u, g_v, g_w\}.$$

Here $g_u, g_v, g_w, \{g_{ijk}\}_{i,j,k}$ are polynomials in the variables $N, \mathbf{u}, \mathbf{v}, \mathbf{w}$. Following the program described above for constructing approximations, a family of semidefinite relaxations to the tensor nuclear norm ball can be prescribed in this manner via the theta bodies $\mathrm{TH}_k(I_{\mathcal{A}})$.

■ 5.4.3 Tradeoff between relaxation and number of measurements

As discussed in Section 5.2.5 the atomic norm is the best convex heuristic for solving ill-posed linear inverse problems of the type considered in this chapter. However we may wish to approximate the atomic norm in cases when it is intractable to compute exactly, and the discussion in the preceding section provides one approach to constructing a family of relaxations. As one might expect the tradeoff for using such approximations, i.e., a *weaker* convex heuristic than the atomic norm, is an increase in the number of measurements required for exact or robust recovery. The reason for this is that the approximate norms have *larger* tangent cones at their extreme points, which makes it harder to satisfy the empty intersection condition of Proposition 5.2.1. We highlight this tradeoff here with an illustrative example involving the cut polytope.

The cut polytope is defined as the convex hull of all cut matrices:

$$\mathcal{P} = \mathrm{conv}\{\mathbf{z}\mathbf{z}^T : \mathbf{z}\in\{-1,+1\}^m\}.$$

As described in Section 5.2.2 low-rank matrices that are composed of ± 1's as entries are of interest in collaborative filtering [133], and the norm induced by the cut polytope is a potential convex heuristic for recovering such matrices from limited measurements. However it is well-known that the cut polytope is intractable to characterize [47], and therefore we need to use tractable relaxations instead. We consider the following two relaxations of the cut polytope. The first is the popular relaxation that is used in semidefinite approximations of the MAXCUT problem:

$$\mathcal{P}_1 = \{M : M \text{ symmetric}, \ M \succeq 0, \ M_{ii} = 1, \forall i = 1, \cdots, p\}.$$

This is the well-studied elliptope [47], and can also be interpreted as the second theta body relaxation (see Chapter 2) of the cut polytope $\mathcal{P}$ [77]. We also investigate the performance of a second, weaker relaxation:

$$\mathcal{P}_2 = \{M : M \text{ symmetric}, \ M_{ii} = 1, \forall i, \ |M_{ij}| \leq \pm 1, \forall i \neq j\}.$$

This polytope is simply the convex hull of symmetric matrices with ± 1's in the off-diagonal entries, and 1's on the diagonal. We note that $\mathcal{P}_2$ is an extremely weak relaxation of $\mathcal{P}$, but we use it here only for illustrative purposes. It is easily seen that

$$\mathcal{P} \subset \mathcal{P}_1 \subset \mathcal{P}_2,$$

with all the inclusions being strict. Figure 5.2 gives a toy sketch that highlights all the main geometric aspects of these relaxations. In particular $\mathcal{P}_1$ has many more extreme points that $\mathcal{P}$, although the set of vertices of $\mathcal{P}_1$, i.e., points that have full-dimensional normal cones, are precisely the cut matrices (which are the vertices of $\mathcal{P}$) [47]. The convex polytope $\mathcal{P}_2$ contains many more vertices compared to $\mathcal{P}$ as shown in Figure 5.2. As expected the tangent cones at vertices of $\mathcal{P}$ become increasingly larger as we use successively weaker relaxations. The following result summarizes the number of random measurements required for recovering a cut matrix, i.e., a rank-one sign matrix, using the norms induced by each of these convex bodies.

Proposition 5.4.1. *Suppose $\mathbf{x}^\star \in \mathbb{R}^{m \times m}$ is a rank-one sign matrix, i.e., a cut matrix, and we are given n random Gaussian measurements of $\mathbf{x}^\star$. We wish to recover $\mathbf{x}^\star$ by solving a convex program based on the norms induced by each of $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2$. We have exact recovery of $\mathbf{x}^\star$ in each of these cases with high probability under the following conditions on the number of measurements:*

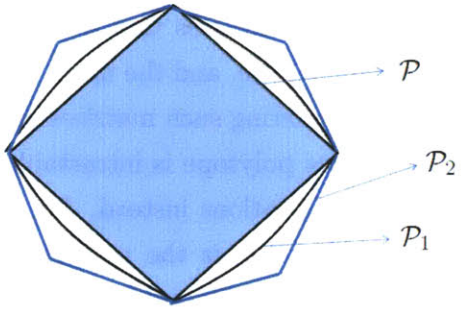

Figure 5.2. A toy sketch illustrating the cut polytope $\mathcal{P}$, and the two approximations $\mathcal{P}_1$ and $\mathcal{P}_2$. Note that $\mathcal{P}_1$ is a sketch of the standard semidefinite relaxation that has the *same* vertices as $\mathcal{P}$. On the other hand $\mathcal{P}_2$ is a polyhedral approximation to $\mathcal{P}$ that has many more vertices as shown in this sketch.

1. *Using* $\mathcal{P}$: $n = \mathcal{O}(m)$.
2. *Using* $\mathcal{P}_1$: $n = \mathcal{O}(m)$.
3. *Using* $\mathcal{P}_2$: $n = \frac{m^2-m}{4}$.

Proof. For the first part, we note that $\mathcal{P}$ is a symmetric polytope with 2^{m-1} vertices. Therefore we can apply Corollary 5.3.3 to conclude that $n = \mathcal{O}(m)$ measurements suffices for exact recovery.

For the second part we note that the tangent cone at $\mathbf{x}^\star$ with respect to the nuclear norm ball of $m \times m$ matrices contains within it the tangent cone at $\mathbf{x}^\star$ with respect to the polytope $\mathcal{P}_1$. Hence we appeal to Proposition 5.3.3 to conclude that $n = \mathcal{O}(m)$ measurements suffices for exact recovery.

Finally, we note that $\mathcal{P}_2$ is essentially the hypercube in $\binom{m}{2}$ dimensions. Appealing to Proposition 5.3.4, we conclude that $n = \frac{m^2-m}{4}$ measurements suffices for exact recovery. □

It is not too hard to show that these bounds are order-optimal, and that they cannot be improved. Thus we have a rigorous demonstration in this particular instance of the fact that the number of measurements required for exact recovery increases as the relaxations get weaker (and as the tangent cones get larger). The principle underlying this illustration holds more generally, namely that there exists a tradeoff between the complexity of the convex heuristic and the number of measurements required for exact or robust recovery. It would be of interest to quantify this tradeoff in other settings,

for example, in problems in which we use increasingly tighter relaxations of the atomic norm via theta bodies.

We also note that the tractable relaxation based on $\mathcal{P}_1$ is only off by a constant factor with respect to the optimal heuristic based on the cut polytope $\mathcal{P}$. This suggests the potential for tractable heuristics to approximate hard atomic norms with provable approximation ratios, akin to methods developed in the literature on approximation algorithms for hard combinatorial optimization problems.

■ 5.4.4 Terracini's lemma and lower bounds on recovery

Algebraic structure in the atomic set $\mathcal{A}$ provides yet another interesting insight, namely for giving *lower bounds* on the number of measurements required for exact recovery. The recovery condition of Proposition 5.2.1 states that the nullspace null(Φ) of the measurement operator $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ must miss the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ at the point of interest $\mathbf{x}^\star$. Suppose that this tangent cone contains a q-dimensional subspace. It is then clear from straightforward linear algebra arguments that the number of measurements n must exceed q. Indeed this bound must hold for *any* linear measurement scheme. Thus the dimension of the subspace contained inside the tangent cone provides a simple lower bound on the number of linear measurements.

In this section we discuss a method to obtain estimates of the dimension of a subspace component of the tangent cone. We focus again on the setting in which $\mathcal{A}$ is an algebraic variety. Indeed in all of the examples of Section 5.2.2, the atomic set $\mathcal{A}$ is an algebraic variety. In such cases simple models $\mathbf{x}^\star$ formed according to (5.1) can be viewed as elements of *secant varieties*.

Definition 5.4.2. *Let $\mathcal{A} \in \mathbb{R}^p$ be an algebraic variety. Then the k'th* secant variety *$\mathcal{A}^k$ is defined as the union of all affine spaces passing through any $k+1$ points of $\mathcal{A}$.*

Algebraic geometry has a long history of investigations of secant varieties, as well as tangent spaces to these secant varieties [79]. In particular a question of interest is to characterize the dimensions of secant varieties and tangent spaces. In our context, estimates of these dimensions are useful in giving lower bounds on the number of measurements required for recovery. Specifically we have the following result, which states that certain linear spaces must lie in the tangent cone at $\mathbf{x}^\star$ with respect to conv($\mathcal{A}$):

Proposition 5.4.2. *Let $\mathcal{A} \subset \mathbb{R}^p$ be a smooth variety, and let $\mathcal{T}(\mathbf{u}, \mathcal{A})$ denote the tangent space at any $\mathbf{u} \in \mathcal{A}$ with respect to $\mathcal{A}$. Suppose $\mathbf{x} = \sum_{i=1}^k c_i \mathbf{a}_i$, $\forall \mathbf{a}_i \in \mathcal{A}, c_i \geq 0$,*

such that

$$\|\mathbf{x}\|_{\mathcal{A}} = \sum_{i=1}^{k} c_i.$$

Then the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ *contains the following linear space:*

$$\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A}) \subset T_{\mathcal{A}}(\mathbf{x}^\star),$$

where $\oplus$ *denotes the direct sum of subspaces.*

Proof. We note that if we perturb $\mathbf{a}_1$ slightly to *any* neighboring $\mathbf{a}_1'$ so that $\mathbf{a}_1' \in \mathcal{A}$, then the resulting $\mathbf{x}' = c_1\mathbf{a}_1' + \sum_{i=2}^{k} c_2\mathbf{a}_i$ is such that $\|\mathbf{x}'\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}}$. The proposition follows directly from this observation. □

By Terracini's lemma [79] from algebraic geometry the subspace $\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A})$ is in fact the estimate for the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ at $\mathbf{x}$ with respect to the $(k-1)$'th secant variety $\mathcal{A}^{k-1}$:

Proposition 5.4.3 (Terracini's Lemma)**.** *Let* $\mathcal{A} \subset \mathbb{R}^p$ *be a smooth affine variety, and let* $\mathcal{T}(\mathbf{u}, \mathcal{A})$ *denote the tangent space at any* $\mathbf{u} \in \mathcal{A}$ *with respect to* $\mathcal{A}$*. Suppose* $\mathbf{x} \in \mathcal{A}^{k-1}$ *is a generic point such that* $\mathbf{x} = \sum_{i=1}^{k} c_i\mathbf{a}_i$, $\forall \mathbf{a}_i \in \mathcal{A}, c_i \geq 0$*. Then the tangent space* $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ *at* $\mathbf{x}$ *with respect to the secant variety* $\mathcal{A}^{k-1}$ *is given by* $\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A})$*. Moreover the dimension of* $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ *is at most (and is expected to be)* $\min\{p, (k+1)\dim(\mathcal{A}) + k\}$.

Combining these results we have that estimates of the dimension of the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ lead directly to lower bounds on the number of measurements required for recovery. The intuition here is clear as the number of measurements required must be bounded below by the number of "degrees of freedom," which is captured by the dimension of the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$. However Terracini's lemma provides us with general estimates of the dimension of $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ for generic points $\mathbf{x}$. Therefore we can directly obtain lower bounds on the number of measurements, purely by considering the dimension of the variety $\mathcal{A}$ and the number of elements from $\mathcal{A}$ used to construct $\mathbf{x}$ (i.e., the order of the secant variety in which $\mathbf{x}$ lies). As an example the dimension of the base variety of normalized order-three tensors in $\mathbb{R}^{m \times m \times m}$ is $3(m-1)$. Consequently if we were to in principle solve the tensor nuclear norm minimization problem, we should expect to require at least $\mathcal{O}(km)$ measurements to recover a rank-k tensor.

5.5 Computational Experiments

5.5.1 Algorithmic considerations

While a variety of atomic norms can be represented or approximated by linear matrix inequalities, these representations do not necessarily translate into practical implementations. Semidefinite programming can be technically solved in polynomial time, but general interior point solvers typically only scale to problems with a few hundred variables. For larger scale problems, it is often preferable to exploit structure in the atomic set $\mathcal{A}$ to develop fast, first-order algorithms.

A starting point for first-order algorithm design lies in determining the structure of the proximity operator (or Moreau envelope) associated with the atomic norm,

$$\Pi_{\mathcal{A}}(\mathbf{x};\mu) := \arg\min_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \mu\|\mathbf{z}\|_{\mathcal{A}}\,. \tag{5.16}$$

Here μ is some positive parameter. Proximity operators have already been harnessed for fast algorithms involving the ℓ_1 norm [39, 40, 66, 78, 144] and the nuclear norm [26, 100, 137] where these maps can be quickly computed in closed form. For the ℓ_1 norm, the ith component of $\Pi_{\mathcal{A}}(\mathbf{x};\mu)$ is given by

$$\Pi_{\mathcal{A}}(\mathbf{x};\mu)_i = \begin{cases} \mathbf{x}_i + \mu & \mathbf{x}_i < -\mu \\ 0 & -\mu \le \mathbf{x}_i \le \mu \\ \mathbf{x}_i - \mu & \mathbf{x}_i > \mu \end{cases}. \tag{5.17}$$

This is the so-called *soft thresholding* operator. For the nuclear norm, $\Pi_{\mathcal{A}}$ soft thresholds the singular values. In either case, the only structure necessary for the cited algorithms to converge is the convexity of the norm. Indeed, essentially any algorithm developed for ℓ_1 or nuclear norm minimization can in principle be adapted for atomic norm minimization. One simply needs to apply the operator $\Pi_{\mathcal{A}}$ wherever a shrinkage operation was previously applied.

For a concrete example, suppose f is a smooth function, and consider the optimization problem

$$\min_{\mathbf{x}} \; f(\mathbf{x}) + \mu\|\mathbf{x}\|_{\mathcal{A}}\,. \tag{5.18}$$

The classical projected gradient method for this problem alternates between taking steps along the gradient of f and then applying the proximity operator associated with the atomic norm. Explicitly, the algorithm consists of the iterative procedure

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{A}}(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k); \alpha_k \lambda) \tag{5.19}$$

where $\{\alpha_k\}$ is a sequence of positive stepsizes. Under very mild assumptions, this iteration can be shown to converge to a stationary point of (5.18) [68]. When f is convex, the returned stationary point is a globally optimal solution. Recently, Nesterov has described a particular variant of this algorithm that is guaranteed to converge at a rate no worse than $O(k^{-1})$, where k is the iteration counter [112]. Moreover, he proposes simple enhancements of the standard iteration to achieve an $O(k^{-2})$ convergence rate for convex f and a linear rate of convergence for strongly convex f.

If we apply the projected gradient method to the regularized inverse problem

$$\min_{\mathbf{x}} \ \|\Phi\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{x}\|_{\mathcal{A}} \tag{5.20}$$

then the algorithm reduces to the straightforward iteration

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{A}}(\mathbf{x}_k + \alpha_k\Phi^\dagger(\mathbf{y} - \Phi\mathbf{x}_k); \alpha_k\lambda)\,. \tag{5.21}$$

Here (5.20) is equivalent to (5.7) for an appropriately chosen $\lambda > 0$ and is useful for estimation from noisy measurements.

The basic (noiseless) atomic norm minimization problem (5.5) can be solved by minimizing a sequence of instances of (5.20) with monotonically decreasing values of λ. Each subsequent minimization is initialized from the point returned by the previous step. Such an approach corresponds to the classic Method of Multipliers [12] and has proven effective for solving problems regularized by the ℓ_1 norm and for total variation denoising [27, 146].

This discussion demonstrates that when the proximity operator associated with some atomic set $\mathcal{A}$ can be easily computed, then efficient first-order algorithms are immediate. For novel atomic norm applications, one can thus focus on algorithms and techniques to compute proximity operators associated. We note that, from a computational perspective, it may be easier to compute the proximity operator via dual atomic norm. Associated to each proximity operator is the dual operator

$$\Lambda_{\mathcal{A}}(\mathbf{x}; \mu) = \arg\min_{\mathbf{y}} \tfrac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 \text{ s.t. } \|\mathbf{y}\|_{\mathcal{A}}^* \leq \mu \tag{5.22}$$

By an appropriate change of variables, $\Lambda_{\mathcal{A}}$ is nothing more than the projection of $\mu^{-1}\mathbf{x}$ onto the unit ball in the dual atomic norm:

$$\Lambda_{\mathcal{A}}(\mathbf{x}; \mu) = \arg\min_{\mathbf{y}} \tfrac{1}{2}\|\mathbf{y} - \mu^{-1}\mathbf{x}\|^2 \text{ s.t. } \|\mathbf{y}\|_{\mathcal{A}}^* \leq 1 \tag{5.23}$$

From convex programming duality, we have $\mathbf{x} = \Pi_{\mathcal{A}}(\mathbf{x};\mu) + \Lambda_{\mathcal{A}}(\mathbf{x};\mu)$. This can be seen by observing

$$\min_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \mu\|\mathbf{z}\|_{\mathcal{A}} = \min_{\mathbf{z}} \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \leq \mu} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \langle \mathbf{y}, \mathbf{z}\rangle \tag{5.24}$$

$$= \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \leq \mu} \min_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \langle \mathbf{y}, \mathbf{z}\rangle \tag{5.25}$$

$$= \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \leq \mu} -\tfrac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2 + \tfrac{1}{2}\|\mathbf{x}\|^2 \tag{5.26}$$

In particular, $\Pi_{\mathcal{A}}(\mathbf{x};\mu)$ and $\Lambda_{\mathcal{A}}(\mathbf{x};\mu)$ form a complementary primal-dual pair for this optimization problem. Hence, we only need to able to efficiently compute the Euclidean projection onto the dual norm ball to compute the proximity operator associated with the atomic norm.

Finally, though the proximity operator provides an elegant framework for algorithm generation, there are many other possible algorithmic approaches that may be employed to take advantage of the particular structure of an atomic set $\mathcal{A}$. For instance, we can rewrite (5.22) as

$$\Lambda_{\mathcal{A}}(\mathbf{x};\mu) = \arg\min_{\mathbf{y}} \tfrac{1}{2}\|\mathbf{y}-\mu^{-1}\mathbf{x}\|^2 \;\text{ s.t. }\; \langle \mathbf{y}, \mathbf{a}\rangle \leq 1 \;\; \forall \mathbf{a} \in \mathcal{A} \tag{5.27}$$

Suppose we have access to a procedure that, given $\mathbf{z} \in \mathbb{R}^n$, can decide whether $\langle \mathbf{z}, \mathbf{a}\rangle \leq 1$ for all $\mathbf{a} \in \mathcal{A}$, or can find a violated constraint where $\langle \mathbf{z}, \hat{\mathbf{a}}\rangle > 1$. In this case, we can apply a cutting plane method or ellipsoid method to solve (5.22) or (5.6) [111, 117]. Similarly, if it is simpler to compute a subgradient of the atomic norm than it is to compute a proximity operator, then the standard subgradient method [13, 111] can be applied to solve problems of the form (5.20). Each computational scheme will have different advantages and drawbacks for specific atomic sets, and relative effectiveness needs to be evaluated on a case-by-case basis.

■ 5.5.2 Simulation results

We describe the results of numerical experiments in recovering orthogonal matrices, permutation matrices, and rank-one sign matrices (i.e., cut matrices) from random linear measurements by solving convex optimization problems. All the atomic norm minimization problems in these experiments are solved using a combination of the SDPT3 package [136] and the YALMIP parser [98].

Orthogonal matrices. We consider the recovery of 20×20 orthogonal matrices from random Gaussian measurements via *spectral norm minimization*. Specifically we

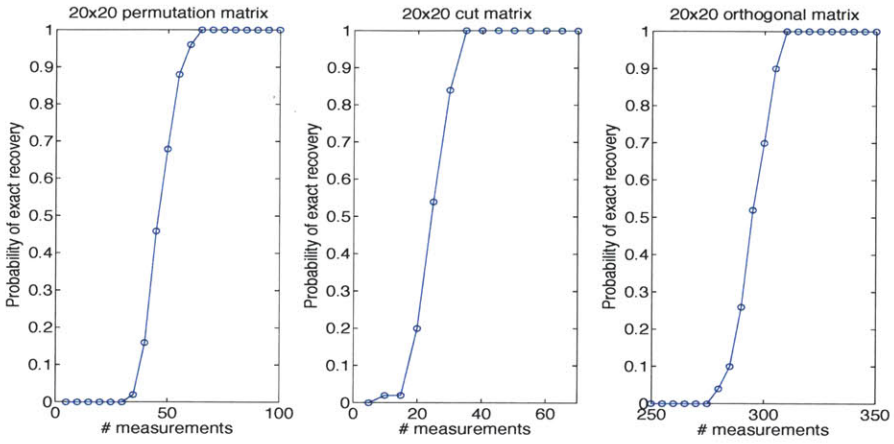

Figure 5.3. Plots of the number of measurements available versus the probability of exact recovery (computed over 50 trials) for various models.

solve the convex program (5.5), with the atomic norm being the spectral norm. Figure 5.3 gives a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

Permutation matrices. We consider the recovery of 20×20 permutation matrices from random Gaussian measurements. We solve the convex program (5.5), with the atomic norm being the norm induced by the Birkhoff polytope of 20×20 doubly stochastic matrices. Figure 5.3 gives a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

Cut matrices. We consider the recovery of 20×20 cut matrices from random Gaussian measurements. As the cut polytope is intractable to characterize, we solve the convex program (5.5) with the atomic norm being approximated by the norm induced by the semidefinite relaxation $\mathcal{P}_1$ described in Section 5.4.3. Figure 5.3 gives a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

In each of these experiments we see agreement between the observed phase transitions, and the theoretical predictions (Propositions 5.3.5, 5.3.6, and 5.4.1) of the number of measurements required for exact recovery. In particular note that the phase transition in Figure 5.3 for the number of measurements required for recovering an orthogonal matrix is very close to the prediction $n \approx \frac{3m^2-m}{4} = 295$ of Proposition 5.3.5. We refer the reader to [55, 102, 121] for similar phase transition plots for recovering sparse

vectors, low-rank matrices, and signed vectors from random measurements via convex optimization.

■ 5.6 Discussion

This chapter has illustrated that for a fixed set of base atoms, the atomic norm is the best choice of a convex regularizer for solving ill-posed inverse problems with the prescribed priors. With this in mind, our results in Section 5.3 and Section 5.4 outline methods for computing hard limits on the number of measurements required for recovery from *any* convex heuristic. Using the calculus of Gaussian widths, such bounds can be computed in a relatively straightforward fashion, especially if one can appeal to notions of convex duality and symmetry. This computational machinery of widths and dimension counting is surprisingly powerful: near-optimal bounds on estimating sparse vectors and low-rank matrices from partial information follow from elementary integration. Thus we expect that our new bounds concerning symmetric, vertex-transitive polytopes are also nearly tight. Moreover, algebraic reasoning allowed us to explore the inherent trade-offs between computational efficiency and measurement demands. More complicated algorithms for atomic norm regularization might extract structure from less information, but approximation algorithms are often sufficient for near optimal reconstructions.

This chapter serves as a foundation for many new exciting directions in inverse problems, and we close our discussion with a description of several natural possibilities for future work:

Width calculations for more atomic sets. The calculus of Gaussian widths described in Section 5.3 provides the building blocks for computing the Gaussian widths for the application examples discussed in Section 5.2. We have not yet exhaustively estimated the widths in all of these examples, and a thorough cataloging of the measurement demands associated with different prior information would provide a more complete understanding of the fundamental limits of solving underdetermined inverse problems. Moreover, our list of examples is by no means exhaustive. The framework developed in this chapter provides a compact and efficient methodology for constructing regularizers from very general prior information, and new regularizers can be easily created by translating grounded expert knowledge into new atomic norms.

Atomic norm decompositions. While the techniques of Section 5.3 and Section 5.4 provide bounds on the estimation of points in low-dimensional secant varieties of atomic sets, they do not provide a procedure for actually constructing decompositions. That is, we have provided bounds on the number of measurements required to recover points $\mathbf{x}$ of the form

$$\mathbf{x} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a}$$

when the coefficient sequence $\{c_{\mathbf{a}}\}$ is sparse, but we do not provide any methods for actually recovering c itself. These decompositions are useful, for instance, in actually computing the rank-one binary vectors optimized in semidefinite relaxations of combinatorial algorithms [3, 72, 110], or in the computation of tensor decompositions from incomplete data [91]. Is it possible to use algebraic structure to generate deterministic or randomized algorithms for reconstructing the atoms that underlie a vector $\mathbf{x}$, especially when approximate norms are used?

Large-scale algorithms. Finally, we think that the most fruitful extensions of this work lie in a thorough exploration of the empirical performance and efficacy of atomic norms on large-scale inverse problems. The proposed algorithms in Section 5.5 require only the knowledge of the proximity operator of an atomic norm, or a Euclidean projection operator onto the dual norm ball. Using these design principles and the geometry of particular atomic norms should enable the scaling of atomic norm techniques to massive data sets.

Chapter 6

Convex Graph Invariants

6.1 Introduction

Graphs are useful in many applications throughout science and engineering as they offer a concise model for relationships among a large number of interacting entities. These relationships are often best understood using structural properties of graphs. *Graph invariants* play an important role in characterizing abstract structural features of a graph, as they do not depend on the labeling of the nodes of the graph. Indeed families of graphs that share common structural attributes are often specified via graph invariants. For example bipartite graphs can be defined by the property that they contain no cycles of odd length, while the family of regular graphs consists of graphs in which all nodes have the same degree. Such descriptions of classes of graphs in terms of invariants have found applications in areas as varied as combinatorics [48], network analysis in chemistry [21] and in biology [105], and in machine learning [93]. For instance the treewidth [123] of a graph is a basic invariant that governs the complexity of various algorithms for graph problems.

We begin by introducing three canonical problems involving structural properties of graphs, and the development of a unified solution framework to address these questions serves as motivation for our discussion throughout this chapter.

- **Graph deconvolution.** Suppose we are given a graph that is the combination of two known graphs overlaid on the same set of nodes. How do we recover the individual components from the composite graph? For example in Figure 6.1 we are given a composite graph that is formed by adding a cycle and the Clebsch graph. Given no extra knowledge of any labeling of the nodes, can we "deconvolve" the composite graph into the individual cycle/Clebsch graph components?

- **Graph generation.** Given certain structural constraints specified by invariants

how do we produce a graph that satisfies these constraints? A well-studied example is the question of constructing expander graphs. Another example may be that we wish to recover a graph given constraints, for instance, on certain subgraphs being forbidden, on the degree distribution, and on the spectral distribution.

- **Graph hypothesis testing.** Suppose we have two families of graphs, each characterized by some common structural properties specified by a set of invariants; given a new sample graph which of the two families offers a "better explanation" of the sample graph (see Figure 6.2)?

In Section 6.2 we describe these problems in more detail, and also give some concrete applications in network analysis and modeling in which such questions are of interest.

To efficiently solve problems such as these we wish to develop a collection of tractable computational tools. Convex relaxation techniques offer a candidate framework as they possess numerous favorable properties. Due to their powerful modeling capabilities, convex optimization methods can provide tractable formulations for solving difficult combinatorial problems exactly or approximately. Further convex programs may often be solved effectively using general-purpose off-the-shelf software. Finally one can also give conditions for the success of these convex relaxations based on standard optimality results from convex analysis.

Motivated by these considerations we introduce and study *convex graph invariants* in Section 6.3. These invariants are convex functions of the adjacency matrix of a graph. More formally letting A denote the adjacency matrix of a (weighted) graph, a convex graph invariant is a convex function f such that $f(A) = f(\Pi A \Pi^T)$ for all permutation matrices Π. Examples include functions of a graph such as the maximum degree, the MAXCUT value (and its semidefinite relaxation), the second smallest eigenvalue of the Laplacian (a concave invariant), and spectral invariants such as the sum of the k largest eigenvalues; see Section 6.3.3 for a more comprehensive list. As some of these invariants may possibly be hard to compute, we discuss in the sequel the question of approximating intractable convex invariants. We also study *invariant convex sets*, which are convex sets with the property that a symmetric matrix A is a member of such a set if and only if $\Pi A \Pi^T$ is also a member of the set for all permutations Π. Such convex sets are useful in order to impose various structural constraints on graphs. For example invariant convex sets can be used to express forbidden subgraph constraints (i.e., that a graph does not contain a particular subgraph such as a triangle), or require that a

graph be connected; see Section 6.3.4 for more examples. We compare the strengths and weaknesses of convex graph invariants versus more general non-convex graph invariants. Finally we also provide a robust optimization perspective of invariant convex sets. In particular we make connections between our work and the data-driven perspective on robust optimization studied in [14].

In order to systematically evaluate the expressive power of convex graph invariants we analyze *elementary* convex graph invariants, which serve as a basis for constructing arbitrary convex invariants. Given a symmetric matrix P, these elementary invariants (again, possibly hard to compute depending on the choice of P) are defined as follows:

$$\Theta_P(A) = \max_{\Pi} \operatorname{Tr}(P\Pi A\Pi^T), \tag{6.1}$$

where A represents the adjacency matrix of a graph, and the maximum is taken over all permutation matrices Π. It is clear that Θ_P is a convex graph invariant, because it is expressed as the maximum over a set of linear functions. Indeed several simple convex graph invariants can be expressed using functions of the form (6.1). For example $P = I$ gives us the total sum of the node weights, while $P = \mathbf{1}\mathbf{1}^T - I$ gives us twice the total (weighted) degree. Our main theoretical results in Section 6.3 can be summarized as follows: First we give a representation theorem stating that any convex graph invariant can be expressed as the supremum over elementary convex graph invariants (6.1) (see Theorem 6.3.1). Second we have a similar result stating that any invariant convex set can be expressed as the intersection of convex sets given by level sets of the elementary invariants (6.1) (see Proposition 6.3.1). These results follow as a consequence of the separation theorem from convex analysis. Finally we also show that for any two non-isomorphic graphs given by adjacency matrices A_1 and A_2, there exists a P such that $\Theta_P(A_1) \neq \Theta_P(A_2)$ (see Lemma 6.3.1). Hence convex graph invariants offer a *complete* set of invariants as they can distinguish between non-isomorphic graphs.

In Section 6.3.7 we discuss an important subclass of convex graph invariants, namely the set of convex *spectral invariants*. These are convex functions of symmetric matrices that depend only on the eigenvalues, and can equivalently be expressed as the set of convex functions of symmetric matrices that are invariant under conjugation by orthogonal matrices (note that convex graph invariants are only required to be invariant with respect to conjugation by permutation matrices) [42]. The properties of convex spectral invariants are well-understood, and they are useful in a number of practically relevant problems (e.g., characterizing the subdifferential of a unitarily invariant matrix

norm [142]). These invariants play a prominent role in our experimental demonstrations in Section 6.5.

As noted above convex graph invariants, and even elementary invariants, may in general be hard to compute. In Section 6.4 we investigate the question of approximately computing these invariants in a tractable manner. For many interesting special cases such as the MAXCUT value of a graph, or (the inverse of) the stability number, there exist well-known tractable semidefinite programming (SDP) relaxations that can be used as surrogates instead [72, 109]. More generally functions of the form of our elementary convex invariants (6.1) have appeared previously in the literature; see [32] for a survey. Specifically we note that evaluating the function $\Theta_P(A)$ for any fixed A, P is equivalent to solving the so-called Quadratic Assignment Problem (QAP), and thus we can employ various tractable linear programming, spectral, and SDP relaxations of QAP [32, 122, 147]. In particular we discuss recent work [43] on exploiting group symmetry in SDP relaxations of QAP, which is useful for approximately computing elementary convex graph invariants in many interesting cases.

Finally in Section 6.5 we return to the motivating problems described previously, and give solutions to these questions. These solutions are based on convex programming formulations, with convex graph invariants playing a fundamental role. We give theoretical conditions for the success of these convex formulations in solving the problems discussed above, and experimental demonstration for their effectiveness in practice. Indeed the framework provided by convex graph invariants allows for a *unified* investigation of our proposed solutions. As an example result we give a tractable convex program (in fact an SDP) in Section 6.5.1 to "deconvolve" the cycle and the Clebsch graph from a composite graph consisting of these components (see Figure 6.1); a salient feature of this convex program is that it only uses *spectral invariants* to perform the decomposition.

Summary of contributions We emphasize again the main contributions of this chapter. We begin by introducing three canonical problems involving structural properties of graphs. These problems arise in various applications (see Section 6.2), and serve as a motivation for our discussion in this chapter. In order to solve these problems we introduce convex graph invariants, and investigate their properties (see Section 6.3). Specifically we provide a representation theorem of convex graph invariants in terms of elementary invariants, and we make connections between these ideas and concepts from other areas such as robust optimization. Finally we describe tractable convex

programming solutions to the motivating problems based on convex graph invariants (see Section 6.5). Therefore, convex graph invariants provide a useful computational framework based on convex optimization for graph problems.

Related previous work We note that convex optimization methods have been used previously to solve various graph-related problems. We would particularly like to emphasize a body of work on convex programming formulations to optimize convex functions of the Laplacian eigenvalues of graphs [22, 23] subject to various constraints. Although our objective is similar in that we seek solutions based on convex optimization to graph problems, our work is different in several respects from these previous approaches. While the problems discussed in [22] explicitly involved the optimization of spectral functions, other graph problems such as those described in Section 6.2 may require non-spectral approaches (for example, hypothesis testing between two families of graphs that are isospectral, i.e., have the same spectrum, but are distinguished by other structural properties). As convex spectral invariants form a subset of convex graph invariants, the framework proposed in this chapter offers a larger suite of convex programming methods for graph problems. More broadly our work is the first to formally introduce and characterize convex graph invariants, and to investigate their properties as natural mathematical objects of independent interest.

Outline In Section 6.2 we give more details of the questions that motivate our study of convex graph invariants. Section 6.3 gives the definition of convex graph invariants and invariant convex sets, as well as several examples of these such functions and sets. We also discuss various properties of convex graph invariants in this section. In Section 6.4 we investigate the question of efficiently computing approximations to intractable convex graph invariants. We give detailed solutions using convex graph invariants to each of our motivating problems in Section 6.5, and we conclude with a brief discussion in Section 6.6.

■ 6.2 Applications

In this section we describe three problems involving structural properties of graphs, which serve as a motivation for our investigation of convex graph invariants. In Section 6.5 we give solutions to these problems using convex graph invariants.

■ 6.2.1 Graph deconvolution

Suppose we are given a graph that is formed by overlaying two graphs on the same set of nodes. More formally we have a graph whose adjacency matrix is formed by adding the adjacency matrices of two known graphs. However, we do not have any information about the relative labeling of the nodes in the two component graphs. Can we recover the individual components from the composite graph? As an example suppose we are given the combination of a cycle and a grid, or a cycle and the Clebsch graph, on the same set of nodes. Without any additional information about the labeling of the nodes, which may reveal the cycle/grid or cycle/Clebsch graph structure, the goal is to recover the individual components. Figure 6.1 gives a graphical illustration of this question. In general such decomposition problems may be ill-posed, and it is of interest to give conditions under which unique deconvolution is possible as well as to provide tractable computational methods to recover the individual components. In Section 6.5.1 we describe an approach based on convex optimization for graph deconvolution; for example this method decomposes the cycle and the Clebsch graph from a composite graph consisting of these components (see Figure 6.1) using only the spectral properties of the two graphs.

Well-known problems that have the flavor of graph deconvolution include the *planted clique* problem, which involves identifying hidden cliques embedded inside a larger graph, and the *clustering* problem in which the goal is to decompose a large graph into smaller densely connected clusters by removing just a few edges. Convex optimization approaches for solving such problems have been proposed recently [4, 5]. Graph deconvolution more generally may include other kinds of embedded structures beyond cliques.

Applications of graph deconvolution arise in network analysis in which one seeks to better understand a complex network by decomposing it into simpler components. Graphs play an important role in modeling, for example, biological networks [105] and social networks [59, 83], and lead to natural graph deconvolution problems in these areas. For instance graphs are useful for describing social exchange networks of interactions of multiple agents, and graph decompositions are useful for describing the structure of optimal bargaining solutions in such networks [89]. In a biological network setting, transcriptional regulatory networks of bacteria have been observed to consist of small subgraphs with specific structure (called motifs) that are connected together using a "backbone" [49]. Decomposing such regulatory networks into the component

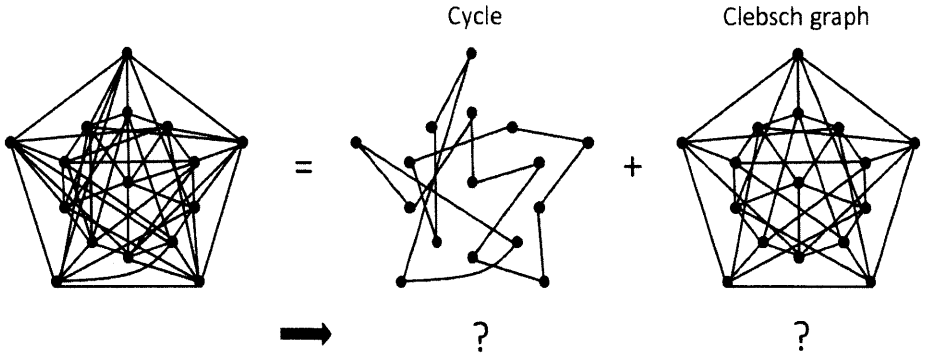

Figure 6.1. An instance of a deconvolution problem: Given a composite graph formed by adding the 16-cycle and the Clebsch graph, we wish to recover the individual components. The Clebsch graph is an example of a strongly regular graph on 16 nodes [70]; see Section 6.5.1 for more details about the properties of such graphs.

structures is useful for obtaining a better understanding of the high-level properties of the composite network.

■ 6.2.2 Generating graphs with desired structural properties

Suppose we wish to construct a graph with certain prescribed structural constraints. A very simple example may be the problem of constructing a graph in which each node has degree equal to two. A graph given by a single cycle satisfies this constraint. A less trivial problem is one in which the objective may be to build a connected graph with constraints on the spectrum of the adjacency matrix, the degree distribution, and the additional requirements that the graph be triangle-free and square-free. Of course such graph reconstruction problems may be infeasible in general, as there may be no graph consistent with the given constraints. Therefore it is of interest to derive suitable conditions under which this problem may be well-posed, and to develop a suitably flexible yet tractable computational framework to incorporate any structural information available about a graph.

A prominent instance of a graph construction problem that has received much attention is the question of generating expander graphs [81]. Expanders are, roughly speaking, sparse graphs that are well-connected, and they have found applications in numerous areas of computer science. Methods used to construct expanders range from random sampling approaches to deterministic constructions based on Ramanujan

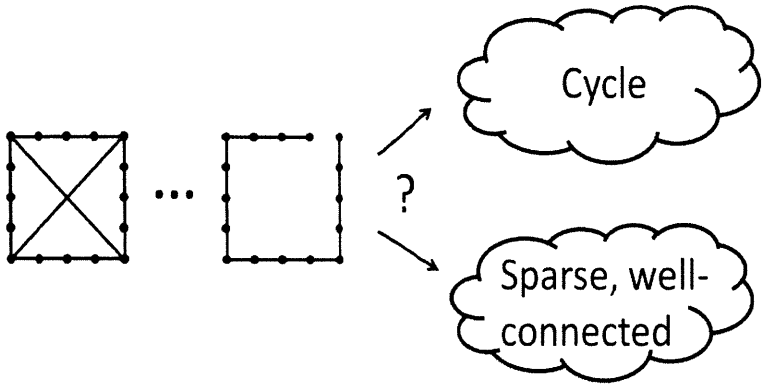

Figure 6.2. An instance of a hypothesis testing problem: We wish to decide which family of graphs offers a "better explanation" for a given candidate sample graph.

graphs [81]. In Section 6.5.2 we describe an approach based on convex optimization to generate sparse, weighted graphs with small degree and large spectral gap.

■ 6.2.3 Graph hypothesis testing

As our third problem we consider a more statistically motivated question. Suppose we have two families of graphs each characterized by some common structural properties specified by certain invariants. Given a new sample graph which of these two families offers a "better explanation" for the sample graph? For example as illustrated in Figure 6.2 we may have two families of graphs – one being the collection of cycles, and the other being the set of sparse, well-connected graphs. If a new sample graph is a path (i.e., a cycle with an edge removed), we would expect that the family of cycles should be a better explanation. On the other hand if the sample is a cycle plus some edges connecting diametrically opposite nodes, then the second family of sparse, well-connected graphs offers a more plausible fit. Notice that these classes of graphs may often be specified in terms of *different* sets of invariants, and it is of interest to develop a suitable framework in which we can incorporate diverse structural information provided about graph families.

We differentiate this problem from the well-studied question of *testing properties* of graphs [73]. Examples of property testing include testing whether a graph is 3-colorable, or whether it is close to being bipartite. An important goal in property testing is that one wishes to test for graph properties by only making a small number of "queries" of

a graph. We do not explicitly seek such an objective in our algorithms for hypothesis testing. We also note that hypothesis testing can be posed more generally than a yes/no question as in property testing, and as mentioned above the two families in hypothesis testing may be specified in terms of very different sets of invariants.

In order to address the hypothesis testing question in a statistical framework, we would need a statistical theory for graphs and appropriate error metrics with respect to which one could devise optimal decision rules. In Section 6.5.3 we discuss a computational approach to the hypothesis testing problem using convex graph invariants that gives good empirical performance, and we defer the issue of developing a formal statistical framework to future work.

■ 6.3 Convex Graph Invariants

In this section we define convex graph invariants, and discuss their properties. Throughout this chapter we denote as before the space of $n \times n$ symmetric matrices by $\mathbf{S}^n \simeq \mathbb{R}^{\binom{n+1}{2}}$. All our definitions of convexity are with respect to the space $\mathbf{S}^n$. We consider undirected graphs that do not have multiple edges and no self-loops; these are represented by adjacency matrices that lie in $\mathbf{S}^n$. Therefore a graph may possibly have node weights and edge weights. A graph is said to be *unweighted* if its node weights are zero, and if each edge has a weight of one (non-edges have a weight of zero); otherwise a graph is said to be *weighted.* Let $\mathbf{e}_i \in \mathbb{R}^n$ denote the vector with a one in the i'th entry and zero elsewhere, let I denote the $n \times n$ identity matrix, let $\mathbf{1} \in \mathbb{R}^n$ denote the all-ones vector, and let $J = \mathbf{1}\mathbf{1}^T \in \mathbf{S}^n$ denote the all-ones matrix. Further we let $\mathcal{A} = \{A : A \in \mathbf{S}^n,\ 0 \leq A_{i,j} \leq 1\ \forall i, j\}$; we will sometimes find it useful in our examples in Section 6.3.4 to restrict our attention to graphs with adjacency matrices in $\mathcal{A}$. Next let $\mathrm{Sym}(n)$ denote the symmetric group over n elements, i.e., the group of permutations of n elements. Elements of this group are represented by $n \times n$ permutation matrices. Let $\mathrm{O}(n)$ represent the orthogonal group of $n \times n$ orthogonal matrices. Finally given a vector $\mathbf{x} \in \mathbb{R}^n$ we recall that $\overline{\mathbf{x}}$ denotes the vector obtained by sorting the entries of $\mathbf{x}$ in descending order.

■ 6.3.1 Motivation: Graphs and adjacency matrices

Matrix representations of graphs in terms of adjacency matrices and Laplacians have been used widely both in applications as well as in the analysis of the structure of graphs

based on algebraic properties of these matrices [17]. For example the spectrum of the Laplacian of a graph reveals whether a graph is "diffusive" [81], or whether it is even connected. The degree sequence, which may be obtained from the adjacency matrix or the Laplacian, reveals whether a graph is regular, and it plays a role in a number of real-world investigations of graphs arising in social networks and the Internet.

Given a graph $\mathcal{G}$ defined on n nodes, a *labeling* of the nodes of $\mathcal{G}$ is a function ℓ that maps the nodes of $\mathcal{G}$ onto distinct integers in $\{1, \ldots, n\}$. An adjacency matrix $A \in \mathbf{S}^n$ is then said to *represent* or *specify* $\mathcal{G}$ if there exists a labeling ℓ of the nodes of $\mathcal{G}$ so that the weight of the edge between nodes i and j equals $A_{\ell(i)\ell(j)}$ for all pairs $\{i, j\}$ and the weight of node i equals $A_{\ell(i)\ell(i)}$ for all i. However an adjacency matrix representation A of the graph $\mathcal{G}$ is not unique. In particular $\Pi A \Pi^T$ also specifies $\mathcal{G}$ for all $\Pi \in \mathrm{Sym}(n)$. All these alternative adjacency matrices correspond to different labelings of the nodes of $\mathcal{G}$. *Thus the graph $\mathcal{G}$ is specified by the matrix A only up to a relabeling of the indices of A.* Our objective is to describe abstract structural properties of $\mathcal{G}$ that do *not* depend on a choice of labeling of the nodes. In order to characterize such *unlabeled* graphs in which the nodes have no distinct identity except through their connections to other nodes, it is important that any function of an adjacency matrix representation of a graph not depend on the particular choice of indices of A. Therefore we seek functions of adjacency matrices that are invariant under conjugation by permutation matrices, and denote such functions as *graph invariants.*

■ 6.3.2 Definition of convex invariants

A convex graph invariant is an invariant that is a convex function of the adjacency matrix of a graph. Specifically we have the following definition:

Definition 6.3.1. *A function $f : \mathbf{S}^n \to \mathbb{R}$ is a* convex graph invariant *if it is convex, and if for any $A \in \mathbf{S}^n$ it holds that $f(\Pi A \Pi^T) = f(A)$ for all permutation matrices $\Pi \in \mathrm{Sym}(n)$.*

Thus convex graph invariants are convex functions that are *constant over orbits* of the symmetric group acting on symmetric matrices by conjugation. As described above the motivation behind the invariance property is clear. The motivation behind the convexity property is that we wish to construct solutions based on convex programming formulations in order to solve problems such as those listed in Section 6.2. We present several examples of convex graph invariants in Section 6.3.3. We note that a *concave*

graph invariant is a real-valued function over $\mathbf{S}^n$ that is the negative of a convex graph invariant.

We also consider invariant convex sets, which are defined in an analogous manner to convex graph invariants:

Definition 6.3.2. *A set $C \subseteq \mathbf{S}^n$ is said to be an* invariant convex set *if it is convex and if for any $A \in C$ it is the case that $\Pi A \Pi^T \in C$ for all permutation matrices $\Pi \in \mathrm{Sym}(n)$.*

In Section 6.3.4 we present examples in which graphs can be constrained to have various properties by requiring that adjacency matrices belong to such convex invariant sets. We also make connections between robust optimization and invariant convex sets in Section 6.3.6.

In order to systematically study convex graph invariants, we analyze certain elementary invariants that serve as a basis for constructing arbitrary convex invariants. These elementary invariants are defined as follows:

Definition 6.3.3. *An* elementary convex graph invariant *is a function $\Theta_P : \mathbf{S}^n \to \mathbb{R}$ of the form*

$$\Theta_P(A) = \max_{\Pi \in \mathrm{Sym}(n)} \mathrm{Tr}(P \Pi A \Pi^T),$$

for any $P \in \mathbf{S}^n$.

It is clear that an elementary invariant is also a convex graph invariant, as it is expressed as the maximum over a set of convex functions (in fact linear functions). We describe various properties of convex graph invariants in Sections 6.3.5. One useful construction that we give is the expression of arbitrary convex graph invariants as suprema over elementary invariants. We also discuss convex spectral invariants in Section 6.3.7, which are convex functions of a symmetric matrix that depend purely on its spectrum. Finally an important point is that convex graph invariants may in general be hard to compute. In Section 6.4 we discuss this problem and propose further tractable convex relaxations for cases in which a convex graph invariant may be intractable to compute.

In the Appendix we describe convex functions defined on $\mathbb{R}^n$ that are invariant with respect to any permutation of the argument. Such functions have been analyzed previously, and we provide a list of their well-known properties. We contrast these properties with those of convex graph invariants throughout the rest of this section.

■ 6.3.3 Examples of convex graph invariants

We list several examples of convex graph invariants. As mentioned previously some of these invariants may possibly be difficult to compute, but we defer discussion of computational issues to Section 6.4. A useful property that we exploit in several of these examples is that a function defined as the supremum over a set of convex functions is itself convex [124].

Number of edges. The total number of edges (or sum of edge weights) is an elementary convex graph invariant with $P = \frac{1}{2}(\mathbf{1}\mathbf{1}^T - I)$.

Node weight. The maximum node weight of a graph, which corresponds to the maximum diagonal entry of the adjacency matrix of the graph, is an elementary convex graph invariant with $P = \mathbf{e}_1\mathbf{e}_1^T$. The maximum diagonal entry in *magnitude* of an adjacency matrix is a convex graph invariant, and can be expressed as follows with $P = \mathbf{e}_1\mathbf{e}_1^T$:

$$\text{max. node weight}(A) = \max\{\Theta_P(A), \Theta_{-P}(A)\}.$$

Similarly the sum of all the node weights, which is the sum of the diagonal entries of an adjacency matrix of a graph, can be expressed as an elementary convex graph invariant with P being the identity matrix.

Maximum degree. The maximum (weighted) degree of a node of a graph is also an elementary convex graph invariant with $P_{1,i} = P_{i,1} = 1$, $\forall i \neq 1$, and all the other entries of P set to zero.

Largest cut. The value of the largest weighted cut of a graph specified by an adjacency matrix $A \in \mathbf{S}^n$ can be written as follows:

$$\text{max. cut}(A) = \max_{\mathbf{y}\in\{-1,+1\}^n} \frac{1}{4}\sum_{i,j} A_{i,j}(1 - \mathbf{y}_i\mathbf{y}_j).$$

As this function is a maximum over a set of linear functions, it is a convex function of A. Further it is also clear that max. cut(A) = max. cut$(\Pi A \Pi^T)$ for all permutation matrices Π. Consequently the value of the largest cut of a graph is a convex graph invariant. We note here that computing this invariant is intractable in general. In practice one could instead employ the following well-known tractable SDP relaxation

[72], which is related to the MAXCUT value by an appropriate shift and rescaling:

$$\begin{array}{rll} f(A) = \min\limits_{X \in \mathbf{S}^n} & \mathrm{Tr}(XA) & \\ \text{s.t.} & X_{ii} = 1, \ \forall i & \\ & X \succeq 0. & \end{array} \tag{6.2}$$

As this relaxation is expressed as the minimum over a set of linear functions, it is a concave graph invariant. In Section 6.4.2 we discuss in greater detail tractable relaxations for invariants that are difficult to compute.

Isoperimetric number (Cheeger constant). The isoperimetric number, also known as the Cheeger constant [50], of a graph specified by adjacency matrix $A \in \mathbf{S}^n$ is defined as follows:

$$\text{isoperimetric number}(A) = \min_{U \subset \{1,\ldots,n\}, |U| \leq \frac{n}{2}, \mathbf{y} \in \mathbb{R}^n, \mathbf{y}_U = 1, \mathbf{y}_{U^c} = -1} \sum_{i,j} \frac{A_{i,j}(1 - \mathbf{y}_i \mathbf{y}_j)}{4|U|}.$$

Here $U^c = \{1, \ldots, n\} \backslash U$ denotes the complement of the set U, and $\mathbf{y}_U$ is the subset of the entries of the vector $\mathbf{y}$ indexed by U. As with the last example, it is again clear that this function is a concave graph invariant as it is expressed as the minimum over a set of linear functions. In particular it can be viewed as measuring the value of a "normalized" cut, and plays an important role in several aspects of graph theory [81].

Degree sequence invariants. Given a graph specified by adjacency matrix A (assume for simplicity that the node weights are zero), the weighted *degree sequence* is given by the vector $\mathbf{d}(A) = \overline{A\mathbf{1}}$, i.e., the vector obtained by sorting the entries of $A\mathbf{1}$ in descending order. It is easily seen that $\mathbf{d}(A)$ is a graph invariant. Consequently any function of $\mathbf{d}(A)$ is also a graph invariant. However our interest is in obtaining *convex* functions of the adjacency matrix A. An important class of functions of $\mathbf{d}(A)$ that are convex functions of A, and therefore are convex graph invariants, are of the form:

$$f(A) = \mathbf{v}^T \mathbf{d}(A),$$

for $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}_1 \geq \cdots \geq \mathbf{v}_n$. This function can also be expressed as the maximum over all permutations $\Pi \in \mathrm{Sym}(n)$ of the inner-product $\mathbf{v}^T \Pi A \mathbf{1}$. As described in the Appendix such linear *monotone* functionals can be used to express *all* convex functions over $\mathbb{R}^n$ that are invariant with respect to permutations of the argument. Consequently these monotone functions serve as building blocks for constructing all convex graph invariants that are functions of $\mathbf{d}(A)$.

Spectral invariants. Let the eigenvalues of the adjacency matrix A of a graph be denoted as $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$, and let $\lambda(A) = [\lambda_1(A), \ldots, \lambda_n(A)]$. These eigenvalues form the *spectrum* of the graph specified by A, and clearly remain unchanged under transformations of the form $A \to VAV^T$ for any orthogonal matrix $V \in \mathrm{O}(n)$ (and therefore for any permutation matrix). Hence any function of the spectrum of a graph is a graph invariant. Analogous to the previous example, an important class of spectral functions that are also *convex* are of the form:

$$f(A) = \mathbf{v}^T \lambda(A),$$

for $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}_1 \geq \cdots \geq \mathbf{v}_n$. We denote spectral invariants that are also convex functions as *convex spectral invariants.* As with convex invariants of the degree sequence, all convex spectral invariants can be constructed using monotone functions of the type described here (see the Appendix).

Second-smallest eigenvalue of Laplacian. This example is only meaningful for weighted graphs in which the node and edge weights are non-negative. For such a graph specified by adjacency matrix A, let $D_A = \mathrm{diag}(A\mathbf{1})$, where diag takes as input a vector and forms a diagonal matrix with the entries of the vector on the diagonal. The *Laplacian* of a graph is then defined as follows:

$$L_A = D_A - A.$$

If $A \in \mathbf{S}^n$ consists of nonnegative entries, then $L_A \succeq 0$. In this setting we denote the eigenvalues of L_A as $\lambda_1(L_A) \geq \cdots \geq \lambda_n(L_A)$. It is easily seen that $\lambda_n(L_A) = 0$ as the all-ones vector $\mathbf{1}$ lies in the kernel of L_A. The second-smallest eigenvalue $\lambda_{n-1}(L_A)$ of the Laplacian is a *concave* invariant function of A. It plays an important role as the graph specified by A is connected if and only if $\lambda_{n-1}(L_A) > 0$.

Inverse of Stability Number. A stable set of an unweighted graph $\mathcal{G}$ is a subset of the nodes of $\mathcal{G}$ such that no two nodes in the subset are adjacent. The stability number is the size of the largest stable set of $\mathcal{G}$, and is denoted by $\alpha(\mathcal{G})$. By a result of Motzkin and Straus [109], the inverse of the stability number can be written as follows:

$$\begin{aligned} \frac{1}{\alpha(\mathcal{G})} = \min_{\mathbf{x}} \quad & \mathbf{x}^T (I + A)\mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}_i \geq 0, \ \forall i, \quad \sum_i \mathbf{x}_i = 1. \end{aligned} \tag{6.3}$$

Here A is any adjacency matrix representing the graph $\mathcal{G}$. Although this formulation is for unweighted graphs with edge weights being either one or zero, we note that

the definition can in fact be *extended* to all weighted graphs, i.e., for graphs with adjacency matrix given by *any* $A \in \mathbf{S}^n$. Consequently, the inverse of this extended stability number of a graph is a concave graph invariant over $\mathbf{S}^n$ as it is expressed as the minimum over a set of linear functions. As this function is difficult to compute in general (because the stability number of a graph is intractable to compute), one could employ the following tractable relaxation:

$$\begin{aligned} f(A) = \min_{X \in \mathbf{S}^n} \quad & \mathrm{Tr}(X(I+A)) \\ \text{s.t.} \quad & X \geq 0, \quad X \succeq 0, \quad \mathbf{1}^T X \mathbf{1} = 1. \end{aligned} \tag{6.4}$$

This relaxation is also a concave graph invariant as it is expressed as the minimum over a set of affine functions.

■ 6.3.4 Examples of invariant convex sets

Next we provide examples of invariant convex sets. As described below constraints expressed using such sets are useful in order to require that graphs have certain properties. Note that a sublevel set $\{A : f(A) \leq \alpha\}$ for any convex graph invariant f is an invariant convex set. Therefore all the examples of convex graph invariants given above can be used to construct invariant convex set constraints.

Algebraic connectivity and diffusion. As mentioned in Section 6.3.3 a graph represented by adjacency matrix $A \in \mathcal{A}$ has the property that the second-smallest eigenvalue $\lambda_{n-1}(L_A)$ of the Laplacian of the graph is a concave graph invariant. The constraint set $\{A : A \in \mathcal{A},\ \lambda_{n-1}(L_A) \geq \epsilon\}$ for any $\epsilon > 0$ expresses the property that a graph must be *connected*. Further if we set ϵ to be relatively large, we can require that a graph has good diffusion properties.

Largest clique constraint. Let $K_k \in \mathbf{S}^n$ denote the adjacency matrix of an unweighted k-clique. Note that K_k is only nonzero within a $k \times k$ submatrix, and is zero-padded to lie in $\mathbf{S}^n$. Consider the following invariant convex set for $\epsilon > 0$:

$$\{A : A \in \mathcal{A},\ \Theta_{K_k}(A) \leq (k^2 - k) - \epsilon\}.$$

This constraint set expresses the property that a graph cannot have a clique of size k (or larger), with the edge weights of all edges in the clique being close to one. For example we can use this constraint set to require that a graph has no triangles (with large edge weights). It is important to note that triangles (and cliques more generally) are forbidden only with the qualification that all the edge weights in the triangle cannot

be close to one. For example a graph may contain a triangle with each edge having weight equal to $\frac{1}{2}$. In this case the function Θ_{K_3} evaluates to 3, which is much smaller than the maximum value of 6 that Θ_{K_3} can take for matrices in $\mathcal{A}$ that contain a triangle with edge weights equal to one.

Girth constraint. The girth of a graph is the length of the shortest cycle. Let $C_k \in \mathbf{S}^n$ denote the adjacency matrix of an unweighted k-cycle for $k \leq n$. As with the k-clique note that C_k is nonzero only within a $k \times k$ submatrix, and is zero-padded so that it lies in $\mathbf{S}^n$. In order to express the property that a graph has no small cycles, consider the following invariant convex set for $\epsilon > 0$:

$$\{A : A \in \mathcal{A},\ \Theta_{C_k}(A) \leq 2k - \epsilon\ \forall k \leq k_0\}.$$

Graphs belonging to this set cannot have cycles of length less than or equal to k_0, with the weights of edges in the cycle being close to one. Thus we can impose a lower bound on a weighted version of the girth of a graph.

Forbidden subgraph constraint. The previous two examples can be viewed as special cases of a more general constraint involving forbidden subgraphs. Specifically let A_k denote the adjacency matrix of an unweighted graph on k nodes that consists of E_k edges. As before A_k is zero-padded to ensure that it lies in $\mathbf{S}^n$. Consider the following invariant convex set for $\epsilon > 0$:

$$\{A : A \in \mathcal{A},\ \Theta_{A_k}(A) \leq 2E_k - \epsilon\}.$$

This constraint set requires that a graph not contain the subgraph given by the adjacency matrix A_k, with edge weights close to one.

Degree distribution. Using the notation described previously, let $\mathbf{d}(A) = \overline{A\mathbf{1}}$ denote the sorted degree sequence $(\mathbf{d}(A)_1 \geq \cdots \geq \mathbf{d}(A)_n)$ of a graph specified by adjacency matrix A. We wish to consider the set of all graphs that have degree sequence $\mathbf{d}(A)$. This set is in general not convex unless A represents a (weighted) regular graph, i.e., $\mathbf{d}(A) = \alpha\mathbf{1}$ for some constant α. Therefore we consider the *convex hull* of all graphs that have degree sequence given by $\mathbf{d}$:

$$\mathcal{D}(A) = \mathrm{conv}\{B : B \in \mathbf{S}^n,\ \overline{B\mathbf{1}} = \mathbf{d}(A)\}.$$

This set is in fact tractable to represent, and is given by the set of graphs whose degree sequence is *majorized* by $\mathbf{d}$:

$$\mathcal{D}(A) = \left\{B : B \in \mathbf{S}^n,\ \mathbf{1}^T B\mathbf{1} = \mathbf{1}^T\mathbf{d}(A),\ \sum_{i=1}^{k}(\overline{B\mathbf{1}})_i \leq \sum_{i=1}^{k}\mathbf{d}(A)_i\ \forall k = 1,\ldots,n-1\right\}.$$

By the majorization principle [11] another representation for this convex set is as the set of graphs whose degree sequence lies in the *permutahedron* generated by $\mathbf{d}$ [149]; the permutahedron generated by a vector is the convex hull of all permutations of the vector. The notion of majorization is sometimes also referred to as *Lorenz dominance* (see the Appendix for more details).

Spectral distribution. Let $\lambda(A)$ denote the spectrum of a graph represented by adjacency matrix A. As before we are interested in the set of all graphs that have spectrum $\lambda(A)$. This set is nonconvex in general, unless A is a multiple of the identity matrix in which case all the eigenvalues are the same. Therefore we consider the convex hull of all graphs (i.e., symmetric adjacency matrices) that have spectrum equal to $\lambda(A)$:

$$\mathcal{E}(A) = \text{conv}\{B : B \in \mathbf{S}^n,\ \lambda(B) = \lambda(A)\}.$$

This convex hull also has a tractable semidefinite representation analogous to the description above [11]:

$$\mathcal{E}(A) = \left\{ B : B \in \mathbf{S}^n,\ \text{Tr}(B) = \text{Tr}(A),\ \sum_{i=1}^{k} \lambda(B)_i \leq \sum_{i=1}^{k} \lambda(A)_i\ \forall k = 1, \ldots, n-1 \right\}.$$

Note that eigenvalues are specified in descending order, so that $\sum_{i=1}^{k} \lambda(B)_i$ represents the sum of the k-largest eigenvalues of B.

■ 6.3.5 Representation of convex graph invariants

All invariant convex sets and convex graph invariants can be represented using elementary convex graph invariants. In this section we describe both these representation results. Representation theorems in mathematics give expressions of complicated sets or functions in terms of simpler, basic objects. In functional analysis the Riesz representation theorem relates elements in a Hilbert space and its dual, by uniquely associating each element of the Hilbert space to a linear functional [128]. In probability theory de Finetti's theorem states that a collection of exchangeable random variables can be expressed as a mixture of independent, identically distributed random variables. In convex analysis every closed convex set can be expressed as the intersection of halfspaces [124]. In each of these cases representation theorems provide a powerful analysis tool as they give a *canonical* expression for complicated mathematical objects in terms of elementary sets/functions.

First we give a representation result for convex graph invariants. In order to get a flavor of this result consider the maximum absolute-value node weight invariant of Section 6.3.3, which is represented as the supremum over two elementary convex graph invariants. The following theorem states that in fact any convex graph invariant can be expressed as a supremum over elementary invariants:

Theorem 6.3.1. *Let f be any convex graph invariant. Then f can be expressed as follows:*

$$f(A) = \sup_{P \in \mathcal{P}} \Theta_P(A) - \alpha_P,$$

for $\alpha_P \in \mathbb{R}$ and for some subset $\mathcal{P} \subset \mathbf{S}^n$.

Proof. Since f is a convex function, it can be expressed as the supremum over linear functionals as follows:

$$f(A) = \sup_{P \in \mathcal{P} \subseteq \mathbf{S}^n} \operatorname{Tr}(PA) - \alpha_P,$$

for $\alpha_P \in \mathbb{R}$. This conclusion follows directly from the separation theorem in convex analysis [124]; in particular this description of the convex function f can be viewed as a specification in terms of supporting hyperplanes of the epigraph of f, which is a convex subset of $\mathbf{S}^n \times \mathbb{R}$. However as f is also a graph invariant, we have that $f(A) = f(\Pi A \Pi^T)$ for any permutation Π and for all $A \in \mathbf{S}^n$. Consequently for any permutation Π and for any $P \in \mathcal{P}$,

$$f(A) = f(\Pi A \Pi^T) \geq \operatorname{Tr}(P \Pi A \Pi^T) - \alpha_P.$$

Thus we have that

$$f(A) \geq \sup_{P \in \mathcal{P}} \Theta_P(A) - \alpha_P. \tag{6.5}$$

However it also clear that for each $P \in \mathcal{P}$

$$\Theta_P(A) - \alpha_P \geq \operatorname{Tr}(PA) - \alpha_P,$$

which allows us to conclude that

$$\sup_{P \in \mathcal{P}} \Theta_P(A) - \alpha_P \geq \sup_{P \in \mathcal{P}} \operatorname{Tr}(PA) - \alpha_P = f(A). \tag{6.6}$$

Combining equations (6.5) and (6.6) we have the desired result. □

Remark 6.3.2. *This result can be strengthened in the sense that one need only consider elements in $\mathcal{P}$ that lie in different equivalence classes up to conjugation by permutation*

matrices $\Pi \in \mathrm{Sym}(n)$. *In each equivalence class the representative functional is the one with the smallest value of* α_P. *This idea can be formalized as follows. Consider the group action* $\rho : (M, \Pi) \mapsto \Pi M \Pi^T$ *that conjugates elements in* $\mathbf{S}^n$ *by a permutation matrix in* $\mathrm{Sym}(n)$. *With this notation we may restrict our attention in Theorem 6.3.1 to* $\mathcal{P} \subset \mathbf{S}^n/\mathrm{Sym}(n)$, *where* $\mathbf{S}^n/\mathrm{Sym}(n)$ *represents the quotient space under the group action* ρ. *Such a mathematical object obtained by taking the quotient of a Euclidean space (or more generally a smooth manifold) under the action of a finite group is called an* orbifold. *With this strengthening one can show that there exists a* unique, minimal *representation set* $\mathcal{P} \subset \mathbf{S}^n/\mathrm{Sym}(n)$. *We however do not emphasize such refinements in subsequent results, and stick with the weaker statement that* $\mathcal{P} \subseteq \mathbf{S}^n$ *for notational and conceptual simplicity.*

As our next result we show that any invariant convex set can be represented as the intersection of sublevel sets of elementary convex graph invariants:

Proposition 6.3.1. *Let* $\mathcal{S} \subseteq \mathbf{S}^n$ *be an invariant convex set. Then there exists a representation of* $\mathcal{S}$ *as follows:*

$$\mathcal{S} = \bigcap_{P \in \mathcal{P}} \{A : A \in \mathbf{S}^n,\ \Theta_P(A) \leq \alpha_P\},$$

for some $\mathcal{P} \subseteq \mathbf{S}^n$ *and for* $\alpha_P \in \mathbb{R}$.

Proof. The proof of this statement proceeds in an analogous manner to that of Theorem 6.3.1, and is again essentially a consequence of the separation theorem in convex analysis. □

■ 6.3.6 A Robust Optimization view of invariant convex sets

Uncertainty arises in many real-world problems. An important goal in robust optimization (see [10] and the reference therein) is to translate formal notions of measures of uncertainty into convex constraint sets. Convexity is important in order to obtain optimization formulations that are tractable.

The representation of a graph via an adjacency matrix in $\mathbf{S}^n$ is inherently uncertain as we have no information about the specific labeling of the nodes of the graph. In this section we associate to each graph a convex polytope, which represents the best convex uncertainty set given a graph:

Definition 6.3.4. *Let $\mathcal{G}$ be a graph that is represented by an adjacency matrix $A \in \mathbf{S}^n$ (any choice of representation is suitable). The* convex hull *of the graph $\mathcal{G}$ is defined as the following convex polytope:*

$$\mathcal{C}(\mathcal{G}) = \text{conv}\{\Pi A \Pi^T : \Pi \in \text{Sym}(n)\}.$$

Recall that $\text{Sym}(n)$ is the symmetric group of $n \times n$ permutation matrices. One can check that the convex hull of a graph is an invariant convex set, and that its extreme points are the matrices $\Pi A \Pi^T$ for all $\Pi \in \text{Sym}(n)$. Note that this convex hull may in general be intractable to characterize; if these polytopes were tractable to characterize we would be able to solve the graph isomorphism problem in polynomial time.

The convex hull of a graph is the smallest convex set that contains all the adjacency matrices that represent the graph. Therefore $\mathcal{C}(\mathcal{G})$ is in some sense the "best convex characterization" of the graph $\mathcal{G}$. This notion is related to the concept of *risk measures* studied in [6], and the construction of convex uncertainty sets based on these risk measures studied in [14]. In particular we recall the following definition from [14]:

Definition 6.3.5. *Let $\mathcal{Z} = \{Z_1, \ldots, Z_k\}$ be any finite collection of elements with $Z_i \in \mathbf{S}^n$. Let $\mathbf{q} \in \mathbb{R}^k$ be a probability distribution, i.e., $\sum_i \mathbf{q}_i = 1$ and $\mathbf{q}_i \geq 0$, $\forall i$. Then the $\mathbf{q}$-permutohull is the polytope in $\mathbf{S}^n$ defined as follows:*

$$\mathcal{B}_{\mathbf{q}}(\mathcal{Z}) = \text{conv}\left\{\sum_i (\Pi \mathbf{q})_i Z_i : \Pi \in \text{Sym}(k)\right\}.$$

Convex uncertainty sets given by permutohulls emphasize a data-driven view of robust optimization as adopted in [14]. Specifically the only information available about an uncertain set in many settings is a finite collection of data vectors $\mathcal{Z}$, and the probability distribution $\mathbf{q}$ expresses preferences over such an unordered data set. Therefore given a data set and a probability distribution that quantifies uncertainty with respect to elements of this data set, the $\mathbf{q}$-permutohull is the smallest convex set expressing these uncertainty preferences. We note that an important property of a permutohull is that it is invariant with respect to relabeling of the data vectors in $\mathcal{Z}$.

The convex hull of a graph $\mathcal{C}(\mathcal{G})$ is a simple example of a permutohull $\mathcal{B}_{\mathbf{q}}(\mathcal{Z})$, with the distribution being $\mathbf{q} = (1, 0, \ldots, 0)$ and the set $\mathcal{Z} = \{\Pi A \Pi^T : \Pi \in \text{Sym}(n)\}$ where $A \in \mathbf{S}^n$ represents the graph $\mathcal{G}$. More complicated permutohulls of graphs may be of interest in several applications but we do not pursue these generalizations here, and instead focus on the case of the convex hull of a graph as defined above.

The convex hull of a graph is itself an invariant convex set by definition. Therefore we can appeal to Proposition 6.3.1 to give a representation of this set in terms of sublevel sets of elementary convex graph invariants. As our next result we show that the values of all elementary convex graph invariants of $\mathcal{G}$ can be used to produce such a representation:

Proposition 6.3.2. *Let $\mathcal{G}$ be a graph and let $A \in \mathbf{S}^n$ be an adjacency matrix representing $\mathcal{G}$. We then have that*

$$\mathcal{C}(\mathcal{G}) = \bigcap_{P \in \mathbf{S}^n} \{B : B \in \mathbf{S}^n,\ \Theta_P(B) \leq \Theta_P(A)\}.$$

Proof. One direction of inclusion in this result is easily seen. Indeed we have that for any $\Pi \in \mathrm{Sym}(n)$

$$\Pi A \Pi^T \in \bigcap_{P \in \mathbf{S}^n} \{B : B \in \mathbf{S}^n,\ \Theta_P(B) \leq \Theta_P(A)\}.$$

As the right-hand-side is a convex set it is clear that the convex hull $\mathcal{C}(\mathcal{G})$ belongs to the set on the right-hand-side:

$$\mathcal{C}(\mathcal{G}) \subseteq \bigcap_{P \in \mathbf{S}^n} \{B : B \in \mathbf{S}^n,\ \Theta_P(B) \leq \Theta_P(A)\}.$$

For the other direction suppose for the sake of a contradiction that we have a point $M \notin \mathcal{C}(\mathcal{G})$ but with $\Theta_P(M) \leq \Theta_P(A)$ for all $P \in \mathbf{S}^n$. As $M \notin \mathcal{C}(\mathcal{G})$ we appeal to the separation theorem from convex analysis [124] to produce a strict separating hyperplane between M and $\mathcal{C}(\mathcal{G})$, i.e., a $\tilde{P} \in \mathbf{S}^n$ such that

$$\mathrm{Tr}(\tilde{P}B) < \alpha, \forall B \in \mathcal{C}(\mathcal{G}), \quad \text{and} \quad \mathrm{Tr}(\tilde{P}M) > \alpha.$$

Further as $\mathcal{C}(\mathcal{G})$ is an invariant convex set, it must be the case that

$$\Theta_{\tilde{P}}(B) < \alpha, \forall B \in \mathcal{C}(\mathcal{G}).$$

On the other hand as $\mathrm{Tr}(\tilde{P}M) > \alpha$ we also have that $\Theta_{\tilde{P}}(M) > \alpha$. It is thus clear that

$$\Theta_{\tilde{P}}(A) < \alpha < \Theta_{\tilde{P}}(M),$$

which leads us to a contradiction and concludes the proof. □

Therefore elementary convex graph invariants are useful for representing all the "convex properties" of a graph. This result agrees with the intuition that the "maximum amount of information" that one can hope to obtain from convex graph invariants about a graph should be limited fundamentally by the convex hull of the graph.

As mentioned previously in many cases the convex hull of a graph may be intractable to characterize. One can obtain outer bounds to this convex hull by using a tractable subset of elementary convex graph invariants; therefore we may obtain tractable but weaker convex uncertainty sets than the convex hull of a graph. From Proposition 6.3.2 such approximations can be refined as we use additional elementary convex graph invariants. As an example the spectral convex constraint sets described in Section 6.3.4 provide a tractable relaxation that plays a prominent role in our experiments in Section 6.4.

■ 6.3.7 Comparison with spectral invariants

Convex functions that are invariant under certain group actions have been studied previously. The most prominent among these is the set of convex functions of symmetric matrices that are invariant under conjugation by orthogonal matrices [42]:

$$f(M) = f(VMV^T),\ \forall\ M \in \mathbf{S}^n,\ \forall\ V \in \mathrm{O}(n).$$

It is clear that such functions depend only on the spectrum of a symmetric matrix, and therefore we refer to them as *convex spectral invariants*:

$$f(M) = \tilde{f}(\lambda(M)),$$

where $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$. It is shown in [42] that f is convex if and only if $\tilde{f}$ is a convex function that is *symmetric* in its argument:

$$\tilde{f}(\mathbf{x}) = \tilde{f}(\Pi\mathbf{x}),\ \forall \mathbf{x} \in \mathbb{R}^n, \forall \Pi \in \mathrm{Sym}(n).$$

One can check that any convex spectral invariant can be represented as the supremum over monotone functionals of the spectrum of the form:

$$\tilde{f}(\mathbf{x}) = \mathbf{v}^T\overline{\mathbf{x}} - \alpha,$$

for $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}_1 \geq \cdots \geq \mathbf{v}_n$. See the Appendix for more details.

A convex spectral invariant is also a convex graph invariant as invariance with respect to conjugation by any orthogonal matrix is a stronger requirement than invariance

with respect to conjugation by any permutation matrix. As many convex spectral invariants are tractable to compute, they form an important subclass of convex graph invariants. In Section 6.4.1 we discuss a natural approximation to elementary convex graph invariants using convex spectral invariants by replacing the symmetric group $\mathrm{Sym}(n)$ in the maximization by the orthogonal group $\mathrm{O}(n)$. Finally one can define a spectrally invariant convex set $\mathcal{S}$ (analogous to invariant convex sets defined in Section 6.3.2) in which $M \in \mathcal{S}$ if and only if $VMV^T \in \mathcal{S}$ for all $V \in \mathrm{O}(n)$. Such sets are very useful in order to impose various spectral constraints on graphs, and often have tractable semidefinite representations.

■ 6.3.8 Convex versus non-convex invariants

There are many graph invariants that are not convex. In this section we give two examples that serve to illustrate the strengths and weaknesses of convex graph invariants. First consider the spectral invariant given by the fifth largest eigenvalue of a graph, i.e., $\lambda_5(A)$ for a graph specified by adjacency matrix A. This function is a graph invariant but it is not convex. However from Section 6.3.3 we have that the *sum* of the first five eigenvalues of a graph is a convex graph invariant. More generally any function of the form $v_1\lambda_1 + \cdots + v_5\lambda_5$ with $v_1 \geq \cdots \geq v_5$ is a convex graph invariant. Thus information about the fifth eigenvalue can be obtained in a "convex manner" only by including information about all the top five eigenvalues (or all the bottom $n-4$ eigenvalues). As a second example consider the (weighted) sum of the total number of triangles that occur as subgraphs in a graph. This function is again a non-convex graph invariant. However recall from the forbidden subgraph example in Section 6.3.4 that we can use elementary convex graph invariants to test whether a graph contains a triangle as a subgraph (with the edges of the triangle having large weights). Therefore, roughly speaking convex graph invariants can be used to decide whether a graph contains a triangle, while general non-convex graph invariants can provide more information about the total number of triangles in a graph. These examples demonstrate that convex graph invariants have certain limitations in terms of the type of information that they can convey about a graph.

The weaker form of information about a graph conveyed by convex graph invariants is nonetheless still useful in distinguishing between graphs. As the next result demonstrates convex graph invariants are strong enough to distinguish between non-isomorphic graphs. This lemma follows from a straightforward application of Proposi-

tion 6.3.2:

Lemma 6.3.1. *Let $\mathcal{G}_1, \mathcal{G}_2$ be two non-isomorphic graphs represented by adjacency matrices $A_1, A_2 \in \mathbf{S}^n$, i.e., there exists no permutation $\Pi \in \mathrm{Sym}(n)$ such that $A_1 = \Pi A_2 \Pi^T$. Then there exists a $P \in \mathbf{S}^n$ such that $\Theta_P(A_1) \neq \Theta_P(A_2)$.*

Proof. Assume for the sake of a contradiction that $\Theta_P(A_1) = \Theta_P(A_2)$ for all $P \in \mathbf{S}^n$. Then we have from Proposition 6.3.2 that $\mathcal{C}(\mathcal{G}_1) = \mathcal{C}(\mathcal{G}_2)$. As the extreme points of these polytopes must be the same, there must exist a permutation $\Pi \in \mathrm{Sym}(n)$ such that $A_1 = \Pi A_2 \Pi^T$. This leads to a contradiction. □

Hence for any two given non-isomorphic graphs there exists an elementary convex graph invariant that evaluates to different values for these two graphs. Consequently elementary convex graph invariants form a *complete* set of graph invariants as they can distinguish between any two non-isomorphic graphs.

■ 6.4 Computing Convex Graph Invariants

In this section we focus on efficiently computing and approximating convex graph invariants, and on tractable representations of invariant convex sets. We begin by studying the question of computing elementary convex graph invariants, before moving on to more general convex invariants.

■ 6.4.1 Elementary invariants and the Quadratic Assignment problem

As all convex graph invariants can be represented using only elementary invariants, we initially focus on computing the latter. Computing an elementary convex graph invariant $\Theta_P(A)$ for general A, P is equivalent to solving the so-called Quadratic Assignment Problem (QAP) [32]. Solving QAP is hard in general, because it includes as a special case the Hamiltonian cycle problem; if P is the adjacency matrix of the n-cycle, then for an unweighted graph specified by adjacency matrix A we have that $\Theta_P(A)$ is equal to $2n$ if and only if the graph contains a Hamiltonian cycle. However there are well-studied spectral and semidefinite relaxations for QAP, which we discuss next.

The *spectral relaxation* of $\Theta_P(A)$ is obtained by replacing the symmetric group $\mathrm{Sym}(n)$ in the definition by the orthogonal group $\mathrm{O}(n)$:

$$\Lambda_P(A) = \max_{V \in \mathrm{O}(n)} \mathrm{Tr}(PVAV^T). \tag{6.7}$$

Clearly $\Theta_P(A) \leq \Lambda_P(A)$ for all $A, P \in \mathbf{S}^n$. As one might expect $\Lambda_P(A)$ has a simple closed-form solution [67]:

$$\Lambda_P(A) = \lambda(P)^T \lambda(A), \tag{6.8}$$

where $\lambda(A), \lambda(P)$ are the eigenvalues of A, P sorted in descending order.

The spectral relaxation offers a simple bound, but is quite weak in many instances. Next we consider the well-studied *semidefinite relaxation* for the QAP, which offers a tighter relaxation [147]. The main idea behind the semidefinite relaxation is that we can linearize $\Theta_P(A)$ as follows:

$$\begin{aligned}
\Theta_P(A) &= \max_{\Pi \in \mathrm{Sym}(n)} \mathrm{Tr}(P\Pi A \Pi^T) \\
&= \max_{\mathbf{x} \in \mathbb{R}^{n^2}, \mathbf{x} = \mathrm{vec}(\Pi), \Pi \in \mathrm{Sym}(n)} \langle \mathbf{x}, (A \otimes P)\mathbf{x} \rangle \\
&= \max_{\mathbf{x} \in \mathbb{R}^{n^2}, \mathbf{x} = \mathrm{vec}(\Pi), \Pi \in \mathrm{Sym}(n)} \mathrm{Tr}((A \otimes P)\mathbf{x}\mathbf{x}^T).
\end{aligned}$$

Here $A \otimes P$ denotes the tensor product between A and P, and vec denotes the operation that stacks the columns of a matrix into a single vector. Consequently it is of interest to characterize the following convex hull:

$$\mathrm{conv}\{\mathbf{x}\mathbf{x}^T : \mathbf{x} \in \mathbb{R}^{n^2},\ \mathbf{x} = \mathrm{vec}(\Pi),\ \Pi \in \mathrm{Sym}(n)\}.$$

There is no known tractable characterization of this set, and by considering tractable approximations the semidefinite relaxation to $\Theta_P(A)$ is then obtained as follows:

$$\begin{aligned}
\Omega_P(A) = \max_{\mathbf{y} \in \mathbb{R}^{n^2},\ Y \in \mathbf{S}(n^2)} \quad & \mathrm{Tr}(P \otimes A) \\
\text{s.t.} \quad & \mathrm{Tr}((I \otimes (J - I))Y + ((J - I) \otimes I)Y) = 0 \\
& \mathrm{Tr}(Y) - 2\mathbf{y}^T \mathbf{1} = -n \\
& Y \geq 0,\ \begin{pmatrix} 1 & \mathbf{y}^T \\ \mathbf{y} & Y \end{pmatrix} \succeq 0.
\end{aligned} \tag{6.9}$$

We refer the reader to [147] for the detailed steps involved in the construction of this relaxation. This SDP relaxation gives an upper bound to $\Theta_P(A)$, i.e., $\Omega_P(A) \geq \Theta_P(A)$. One can show that if the extra rank constraint

$$\mathrm{rank} \begin{pmatrix} 1 & \mathbf{y}^T \\ \mathbf{y} & Y \end{pmatrix} = 1$$

is added to the SDP (6.9), then $\Omega_P(A) = \Theta_P(A)$. Therefore if the optimal value of the SDP (6.9) is achieved at some $\hat{\mathbf{y}}, \hat{Y}$ such that this rank-one constraint is satisfied, then the relaxation is tight, i.e., we would have that $\Omega_P(A) = \Theta_P(A)$.

While the semidefinite relaxation (6.9) can in principle be computed in polynomial-time, the size of the variable $Y \in \mathbf{S}(n^2)$ means that even moderate size problem instances are not well-suited to solution by interior-point methods. In many practical situations however, we often have that the matrix $P \in \mathbf{S}^n$ represents the adjacency matrix of some small graph on k nodes with $k \ll n$, i.e., P is nonzero only inside a $k \times k$ submatrix and is zero-padded elsewhere so that it lies in $\mathbf{S}^n$. For example as discussed in Section 6.3.4, P may represent the adjacency matrix of a triangle in a constraint expressing that a graph is triangle-free. In such cases computing or approximating $\Theta_P(A)$ may be done more efficiently as follows:

1. **Combinatorial enumeration.** For very small values of k it is possible to compute $\Theta_P(A)$ efficiently even by explicit combinatorial enumeration. The complexity of such a procedure scales as $\mathcal{O}(n^k)$. This approach may be suitable if, for example, P represents the adjacency matrix of a triangle.

2. **Symmetry reduction.** For larger values of k, combinatorial enumeration may no longer be appropriate. In these cases the special structure in P can be exploited to reduce the size of the SDP relaxation (6.9). Specifically, using the methods described in [43] it is possible to reduce the size of the matrix variables from $\mathcal{O}(n^2) \times \mathcal{O}(n^2)$ to size $\mathcal{O}(kn) \times \mathcal{O}(kn)$. More generally, it is also possible to exploit *group symmetry* in P to similarly reduce the size of the SDP (6.9) (see [43] for details).

■ 6.4.2 Other methods and computational issues

In many special cases in which computing convex graph invariants may be intractable, it is also possible to use other types of tractable semidefinite relaxations. As described in Section 6.3.3 the MAXCUT value and the inverse stability number of graphs are invariants that are respectively convex and concave. However both of these are intractable to compute, and as a result we must employ the SDP relaxations for these invariants as discussed in Section 6.3.3.

Another issue that arises in practice is the *representation* of invariant convex sets. As an example, let $f(A)$ denote the SDP relaxation of the MAXCUT value as defined

in (6.2). As $f(A)$ is a concave graph invariant, we may be interested in representing convex constraint sets as follows:

$$\{A : A \in \mathbf{S}^n,\ f(A) \geq \alpha\} = \{A : A \in \mathbf{S}^n,\ \mathrm{Tr}(XA) \geq \alpha\ \ \forall X \in \mathbf{S}^n \text{ s.t. } X_{ii} = 1,\ X \succeq 0\}.$$

In order to computationally represent such a set specified in terms of a universal quantifier, we appeal to convex duality. Using the standard dual formulation of (6.2), we have that:

$$\{A : A \in \mathbf{S}^n,\ f(A) \geq \alpha\} = \{A : A \in \mathbf{S}^n,\ \ \exists Y \text{ diagonal s.t. } A \succeq Y,\ \mathrm{Tr}(Y) \geq \alpha\}.$$

This reformulation provides a description in terms of existential quantifiers that is more suitable for practical representation. Such reformulations using convex duality are well-known, and can be employed more generally (e.g., for invariant convex sets specified by sublevel sets of the inverse stability number or its relaxations in Section 6.3.3)

■ 6.5 Using Convex Graph Invariants in Applications

In this section we give solutions to the stylized problems of Section 6.2 using convex graph invariants. In order to properly state our results we begin with a few definitions. All the convex programs in our numerical experiments are solved using a combination of the SDPT3 package [136] and the YALMIP parser [98]. Finally a key property of normal cones that we use in stating our results is that for any convex set $C \subseteq \mathbf{S}^n$, the normal cones at all the extreme points of C form a *partition*[1] of $\mathbf{S}^n$ [124].

■ 6.5.1 Application: Graph deconvolution

Given a combination of two graphs overlaid on the same set of nodes, the graph deconvolution problem is to recover the individual graphs (as introduced in Section 6.2.1).

Problem 1. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two graphs specified by particular adjacency matrices $A_1^*, A_2^* \in \mathbf{S}^n$. We are given the sum $A = A_1^* + A_2^*$, and the additional information that A_1^*, A_2^* correspond to particular realizations (labelings of nodes) of $\mathcal{G}_1, \mathcal{G}_2$. The goal is to recover A_1^* and A_2^* from A.*

See Figure 6.1 for an example illustrating this problem. The key unknown in this problem is the specific labeling of the nodes of $\mathcal{G}_1$ and $\mathcal{G}_2$ relative to each other in

[1]Note that there may be overlap on the boundaries of the normal cones at the extreme points, but these overlaps have smaller dimension than those of the normal cones.

the composite graph represented by A. As described in Section 6.3.6, the best convex constraints that express this uncertainty are the convex hulls of the graphs $\mathcal{G}_1, \mathcal{G}_2$. Therefore we consider the following natural solution based on convex optimization to solve the deconvolution problem:

Solution 1. *Recall that $\mathcal{C}(\mathcal{G}_1)$ and $\mathcal{C}(\mathcal{G}_2)$ are the convex hulls of the unlabeled graphs $\mathcal{G}_1, \mathcal{G}_2$ (which we are given), and that $\|\cdot\|_F$ denotes the Euclidean (Frobenius) norm. We propose the following convex program to recover A_1, A_2:*

$$\begin{aligned}(\hat{A}_1, \hat{A}_2) = \arg \min_{A_1, A_2 \in \mathbf{S}^n} \quad & \|A - A_1 - A_2\|_F \\ s.t. \quad & A_1 \in \mathcal{C}(\mathcal{G}_1),\ A_2 \in \mathcal{C}(\mathcal{G}_2).\end{aligned} \tag{6.10}$$

One could also use in the objective any other norm that is invariant under conjugation by permutation matrices. This program is convex, although it may not be tractable if the sets $\mathcal{C}(\mathcal{G}_1), \mathcal{C}(\mathcal{G}_2)$ cannot be efficiently represented. Therefore it may be desirable to use tractable convex relaxations C_1, C_2 of the sets $\mathcal{C}(\mathcal{G}_1), \mathcal{C}(\mathcal{G}_2)$, i.e., $\mathcal{C}(\mathcal{G}_1) \subseteq C_1 \subset \mathbf{S}^n$ and $\mathcal{C}(\mathcal{G}_2) \subseteq C_2 \subset \mathbf{S}^n$:

$$\begin{aligned}(\hat{A}_1, \hat{A}_2) = \arg \min_{A_1, A_2 \in \mathbf{S}^n} \quad & \|A - A_1 - A_2\|_F \\ s.t. \quad & A_1 \in C_1,\ A_2 \in C_2.\end{aligned} \tag{6.11}$$

Recall from Proposition 6.3.2 that we can represent $\mathcal{C}(\mathcal{G})$ using all the elementary convex graph invariants. Tractable relaxations to this convex hull may be obtained, for example, by just using spectral invariants, degree-sequence invariants, or any other subset of invariant convex set constraints that can be expressed efficiently. We give numerical examples later in this section. The following result gives conditions under which we can exactly recover A_1^*, A_2^* using the convex program (6.11):

Proposition 6.5.1. *Given the problem setup as described above, we have that $(\hat{A}_1, \hat{A}_2) = (A_1^*, A_2^*)$ is the* unique optimum *of* (6.11) *if and only if:*

$$T_{C_1}(A_1^*) \cap -T_{C_2}(A_2^*) = \{0\},$$

where $-T_{C_2}(A_2^*)$ denotes the negative of the tangent cone $T_{C_2}(A_2^*)$.

Proof. Note that in the setup described above (A_1^*, A_2^*) is an optimal solution of the convex program (6.11) as this point is feasible (since by construction $A_1^* \in \mathcal{C}(\mathcal{G}_1) \subseteq C_1$

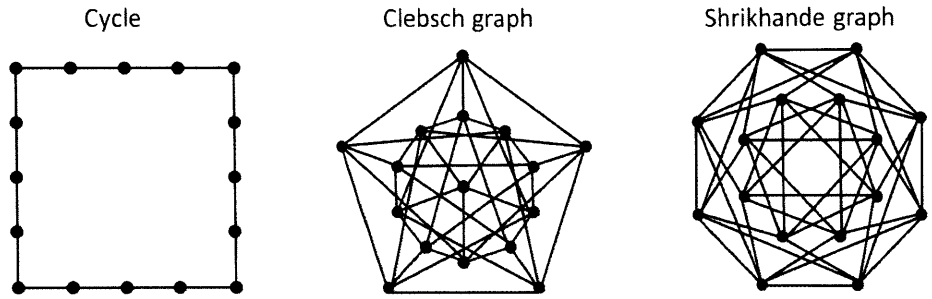

Figure 6.3. The three graphs used in the deconvolution experiments of Section 6.5.1. The Clebsch graph and the Shrikhande graph are examples of strongly regular graphs on 16 nodes [70]; see Section 6.5.1 for more details about the properties of such graphs.

and $A_2^* \in \mathcal{C}(\mathcal{G}_2) \subseteq C_2$), and the cost function achieves its minimum at this point. This result is concerned with (A_1^*, A_2^*) being the *unique* optimal solution.

For one direction suppose that $T_{C_1}(A_1^*) \cap -T_{C_2}(A_2^*) = \{0\}$. Then there exists no $Z_1 \in T_{C_1}(A_1^*), Z_2 \in T_{C_2}(A_2^*)$ such that $Z_1 + Z_2 = 0$ with $Z_1 \neq 0, Z_2 \neq 0$. Consequently every feasible direction from (A_1^*, A_2^*) into $C_1 \times C_2$ would increase the value of the objective. Thus (A_1^*, A_2^*) is the unique optimum of (6.11).

For the other direction suppose that (A_1^*, A_2^*) is the unique optimum of (6.11), and assume for the sake of a contradiction that $T_{C_1}(A_1^*) \cap -T_{C_2}(A_2^*)$ contains a nonzero element, which we'll denote by Z. There exists a scalar $\alpha > 0$ such that $A_1^* + \alpha Z \in C_1$ and $A_2^* - \alpha Z \in C_2$. Consequently $(A_1^* + \alpha Z, A_2^* - \alpha Z)$ is also a feasible solution that achieves the lowest possible cost of zero. This contradicts the assumption that (A_1^*, A_2^*) is the unique optimum. □

Thus we have that *transverse intersection* of the tangent cones $T_{C_1}(A_1^*)$ and $-T_{C_2}(A_2^*)$ is equivalent to *exact recovery* of (A_1^*, A_2^*) given the sum $A = A_1^* + A_2^*$. As $\mathcal{C}(\mathcal{G}_1) \subseteq C_1$ and $\mathcal{C}(\mathcal{G}_2) \subseteq C_2$, we have that $T_{\mathcal{C}(\mathcal{G}_1)}(A_1^*) \subseteq T_{C_1}(A_1^*)$ and $T_{\mathcal{C}(\mathcal{G}_2)} \subseteq T_{C_2}(A_2^*)$. These relations follow from the fact that the set of feasible directions from A_1^* and A_2^* into the respective convex sets is enlarged. Therefore the tangent cone transversality condition of Proposition 6.5.1 is generally more difficult to satisfy if we use relaxations C_1, C_2 to the convex hulls $\mathcal{C}(\mathcal{G}_1), \mathcal{C}(\mathcal{G}_2)$. Consequently we have a *tradeoff* between the complexity of solving the convex program, and the possibility of exactly recovering (A_1^*, A_2^*). However the following example suggests that it is possible to obtain tractable relaxations that still allow for perfect recovery.

Example. We consider the 16-cycle, the Shrikhande graph, and the Clebsch graph (see Figure 6.3), and investigate the deconvolution problem for all three pairings of these graphs. For illustration purposes suppose A_1^* is an adjacency matrix of the unweighted 16-node cycle denoted $\mathcal{G}_1$, and that A_2^* is an adjacency matrix of the 16-node Clebsch graph denoted $\mathcal{G}_2$ (see Figure 6.1). These adjacency matrices are random instances chosen from the set of all valid adjacency matrices that represent the graphs $\mathcal{G}_1, \mathcal{G}_2$. Given the sum $A = A_1^* + A_2^*$, we construct convex constraint sets C_1, C_2 as follows:

$$\begin{aligned} C_1 &= \mathcal{A} \cap \mathcal{E}(A_1^*) \\ C_2 &= \mathcal{A} \cap \mathcal{E}(A_2^*). \end{aligned}$$

Here $\mathcal{E}(A)$ represents the spectral constraints of Section 6.3.4. Therefore the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are characterized purely by their spectral properties. By running the convex program described above for 100 random choices of labelings of the vertices of the graphs $\mathcal{G}_1, \mathcal{G}_2$, we obtained *exact* recovery of the adjacency matrices (A_1^*, A_2^*) in all cases (see Table 6.1). *Thus we have exact decomposition based only on convex spectral constraints, in which the only invariant information used to characterize the component graphs $\mathcal{G}_1, \mathcal{G}_2$ are the spectra of $\mathcal{G}_1, \mathcal{G}_2$.* Similarly successful decomposition results using only spectral invariants are also seen in the cycle/Shrikhande graph deconvolution problem, and the Clebsch graph/Shrikhande graph deconvolution problem; Table 6.1 gives complete results.

The inspiration for using the Clebsch graph and the Shrikhande graph as examples for deconvolution is based on Proposition 6.5.1. Specifically, a graph for which the tangent cone with respect to the corresponding spectral constraint set $\mathcal{E}(A)$ (defined in Section 6.3.4) is small is well-suited to being deconvolved from other graphs using spectral invariants. This is because the tangent cone being smaller implies that the transversality condition of Proposition 6.5.1 is easier to satisfy. In order to obtain small tangent cones with respect to spectral constraint sets, we seek graphs that have many *repeated eigenvalues. Strongly regular graphs*, such as the Clebsch graph and the Shrikhande graph, are prominent examples of graphs with repeated eigenvalues as they have only three distinct eigenvalues. A strongly regular graph is an unweighted regular graph (i.e., each node has the same degree) in which every pair of adjacent vertices have the same number of common neighbors, and every pair of non-adjacent vertices have the same number of common neighbors [70]. We explore in more detail the properties of these and other graph classes in a separate report [35], where we characterize families

Underlying graphs	# successes in 100 random trials
The 16-cycle and the Clebsch graph	100
The 16-cycle and the Shrikhande graph	96
The Clebsch graph and the Shrikhande graph	94

Table 6.1. A summary of the results of graph deconvolution via convex optimization: We generated 100 random instances of each deconvolution problem by randomizing over the labelings of the components. The convex program uses only spectral invariants to characterize the convex hulls of the component graphs, as described in Section 6.5.1.

of graphs for which the transverse intersection condition of Proposition 6.5.1 provably holds for constraint sets C_1, C_2 constructed using tractable graph invariants.

■ 6.5.2 Application: Generating graphs with desired properties

We first consider the problem of constructing a graph with certain desired structural properties.

Problem 2. *Suppose we are given structural constraints on a graph in terms of a collection of (possibly nonconvex) graph invariants* $\{h_j(A) = \alpha_j\}$. *Can we recover a graph that is consistent with these constraints? For example we may be given constraints on the spectrum, the degree distribution, the girth, and the MAXCUT value. Can we construct some graph* $\mathcal{G}$ *that is consistent with this knowledge?*

This problem may be infeasible in that there may no graph consistent with the given information. We do not address this feasibility question here, and instead focus only on the computational problem of generating graphs that satisfy the given constraints assuming such graphs do exist. Next we propose a convex programming approach using invariant convex sets to construct a graph $\mathcal{G}$, specified by an adjacency matrix A, which satisfies the required constraints. Both the problem as well the solution can be suitably modified to include inequality constraints.

Solution 2. *We combine information from all the invariants to construct an invariant convex set* C. *Given a constraint of the form* $h_j(A) = \alpha_j$, *we consider the following convex set:*

$$C_j = \text{conv}\{A : A \in \mathbf{S}^n,\ h_j(A) = \alpha_j\}.$$

This set is convex by construction, and is an invariant convex set if h_j *is a graph invariant. If* h_j *is a convex graph invariant this set is equal to the sublevel set* $\{A :$

$A \in \mathbf{S}^n,\ h_j(A) \leq \alpha_j\}$. *Given a collection of constraints* $\{h_j(A) = \alpha_j\}$ *we then form an invariant convex constraint set as follows:*

$$C = \cap_j \ C_j.$$

Therefore any invariant information that is amenable to approximation as a convex constraint set can be incorporated in such a framework. For example constraints on the degree distribution or the spectrum can be naturally relaxed to tractable convex constraints, as described in Section 6.3.4. If the set C *as defined above is intractable to compute, one may further relax* C *to obtain efficient approximations. In many cases of interest a subset of the boundary of* C *corresponds to points at which* all *the constraints are active* $\{A : h_j(A) = \alpha_j\}$. *In order to recover one of these extreme points, we maximize a* random *linear functional defined by* $M \in \mathbf{S}^n$ *(with the entries in the upper-triangular part chosen to be independent and identically distributed to zero-mean, variance-one standard Gaussians) over the set* C*:*

$$\begin{aligned} \hat{A} = \arg\max_{A \in \mathbf{S}^n} \quad & \mathrm{Tr}(MA) \\ s.t. \quad & A \in C. \end{aligned} \tag{6.12}$$

This convex program is successful if $\hat{A}$ *is indeed an extreme point at which all the constraints* $\{h_j(A) = \alpha_j\}$ *are satisfied.*

Clearly this approach is well-suited for constructing constrained graphs only if the convex set C described in the solution scheme contains many extreme points at which all the constraints are satisfied. The next result gives conditions under which the convex program recovers an $\hat{A}$ that satisfies all the given constraints:

Proposition 6.5.2. *Consider the problem and solution setup as defined above. Define the set* N *as follows:*

$$N = \bigcup_{\{A \ :\ A \in C,\ h_j(A) = \alpha_j\ \forall j\}} N_C(A).$$

If $M \in N$ *then the optimum* $\hat{A}$ *of the convex program* (6.12) *satisfies all the specified constraints exactly. In particular if* M *is chosen uniformly at random as described above, then the probability of success is equal to the fraction of* $\mathbf{S}^n$ *covered by the union of the normal cones* N*.*

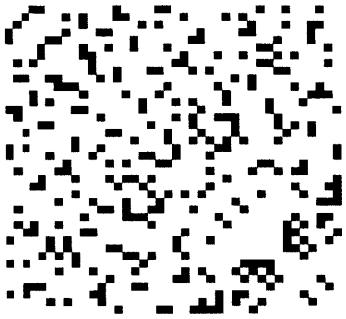

Figure 6.4. An adjacency matrix of a sparse, well-connected graph example obtained using the approach described in Section 6.5.2: The weights of this graph lie in the range $[0, 1]$, the black points represent edges with nonzero weight, and the white points denote absence of edges. The (weighted) degree of each node is 8, the average number of nonzero (weighted) edges per node is 8.4, the second-smallest eigenvalue of the Laplacian is 4, and the weighted diameter is 3.

Proof. The proof follows from standard results in convex analysis. In particular we appeal to the fact that a linear functional defined by M achieves its maximum at $\hat{A} \in C$ if and only if $M \in N_C(\hat{A})$. □

As a corollary of this result we observe that if the invariant information provided exactly characterizes the convex hull of a graph $\mathcal{G}$, then the set C above is the convex hull $\mathcal{C}(\mathcal{G})$ of the graph $\mathcal{G}$. In such cases the convex program given by (6.12) produces an adjacency matrix representing $\mathcal{G}$ with probability one. Next we provide the results of a simple experiment that demonstrates the effectiveness of our approach in generating sparse graphs with large spectral gap.

Example. In this example we aim to construct graphs on $n = 40$ nodes with adjacency matrices in $\mathcal{A}$ that have degree $d = 8$, node weights equal to zero, and the second-smallest eigenvalue of the Laplacian being larger than $\epsilon = 4$. The goal is to produce relatively *sparse* graphs that satisfy these constraints. The specified constraints can be used to construct a convex set as follows:

$$C = \{A : A \in \mathcal{A},\ \tfrac{1}{8}A\mathbf{1} = \mathbf{1},\ \lambda_{n-1}(L_A) \geq 4,\ A_{ii} = 0\ \forall i\}.$$

By maximizing 100 random linear functionals over this set we obtained graphs in all 100 cases with total degree equal to 8, and in 98 of the 100 cases with the minimum eigenvalue of the Laplacian equal to 4 (it is greater than 4 in the remaining two cases).

Interestingly the average number of edges with nonzero weight incident on each node is 8.8 over these 100 trials, thus providing very sparse graphs that are well-connected. Figure 6.4 gives an example of a graph generated randomly using this procedure; the average number of nonzero (weighted) edges per node of this graph is 8.4, and its (weighted) diameter is 3. Therefore this approach empirically yields sparse graphs that are well-connected (i.e., with a large spectral gap).

We would like to point out here a different approach to constructing well-connected graphs, which tries to add edges from a subset of candidate edges to maximize the second eigenvalue of the graph Laplacian [69]. An interesting question is to understand the structure of the extreme points of the set C in this example as the graph size and the degree (n, d) grow large, with ϵ held constant. For example it may be useful to compute the fraction of the normal cones at those extreme points corresponding to expander graphs. More generally it is of interest to give conditions on constraint sets under which the procedure described above is successful in providing graphs that satisfy all the constraints with high probability.

■ 6.5.3 Application: Graph hypothesis testing

Finally we give a solution to the hypothesis testing problem in which we have two families of graphs, and the goal is to decide which of these families offers a "better explanation" for a given candidate "sample" graph.

Problem 3. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote two families of graphs characterized in terms of invariants $\{h_j^1\}$ and $\{h_j^2\}$ respectively; for example, a family could be specified as some set of graphs that have similar spectral distributions, similar degree sequences, and similar girths. Given a graph $\mathcal{G}$, which of the two families $\mathcal{F}_1, \mathcal{F}_2$ of graphs is more similar to $\mathcal{G}$?*

We emphasize that the sets of invariants that characterize $\mathcal{F}_1, \mathcal{F}_2$ may in general be very different. Note that this question is not completely well-posed, as there may be different answers depending on one's notion of similarity. In order to address this point, we need to develop a statistical theory for graphs. In such a setting one could phrase this question formally as a statistical hypothesis testing problem with appropriate error metrics. Our focus in the present chapter is on proposing a convex optimization solution to the hypothesis testing based on convex graph invariants, and using a reasonable notion of similarity.

Solution 3. *Let $A \in \mathbf{S}^n$ be an adjacency matrix that represents the graph $\mathcal{G}$. We construct invariant convex sets C_1 and C_2 based on the sets of invariants $\{h_j^1\}, \{h_j^2\}$ in an analogous manner to the construction described in the solution to the graph construction problem of Section 6.5.2. As before one could employ further tractable relaxations of these sets if they are intractable to compute. Assuming that these convex constraint sets that summarize the families $\mathcal{F}_1$ and $\mathcal{F}_2$ are compact, we declare that $\mathcal{F}_1$ is closer to $\mathcal{G}$ than $\mathcal{F}_2$ if the following holds:*

$$\max_{M \in C_1} \operatorname{Tr}(AM) \quad \geq \quad \max_{M \in C_2} \operatorname{Tr}(AM). \tag{6.13}$$

Naturally we declare the opposite result if the inequality is switched. Computing the two sides in this test can be done via convex optimization, and this computation is tractable if C_1, C_2 are tractable to characterize.

Our choice of the function to be maximized over C_1, C_2 is motivated by a similar procedure in statistics and signal processing, which goes by the name of "matched filtering." Of course other (convex invariant) cost functions can also be optimized depending on one's notion of similarity. We point out two advantages of this approach to hypothesis testing. First the two families of graphs can be specified in terms of different sets of invariants, as seen in these examples. Second the optimal solutions of the convex programs in (6.13) in fact provide *approximations* to the graph $\mathcal{G}$ by elements in the families $\mathcal{F}_1, \mathcal{F}_2$. We give illustrations of these points in our examples, which we describe next.

Example. Let A_{cycle} denote the adjacency matrix of a 16-node unweighted cycle. As our first family we consider the set of cycles on 16 nodes. We approximate this family by the set of graphs that are triangle-free (in the sense described in Section 6.3.4), have degree equal to 2, and have the same spectrum as a 16-node unweighted cycle. Therefore the set C_1 is defined as follows:

$$C_1 = \{A : A \in \mathcal{A},\ A_{ii} = 0\ \forall i,\ \tfrac{1}{2}A\mathbf{1} = \mathbf{1},\ \Theta_{K_3}(A) \leq 4\} \cap \mathcal{E}(A_{\text{cycle}}).$$

As our second family, we consider sparse well-connected graphs on 16 nodes with maximum weighted degree less than or equal to 2.5, and with the second-smallest eigenvalue of the Laplacian bounded below by 1.1:

$$C_2 = \{A : A \in \mathcal{A},\ A_{ii} = 0\ \forall i,\ (A\mathbf{1})_i \leq 2.5\ \forall i,\ \lambda_{n-1}(L_A) \geq 1.1\}.$$

Applying the solution described above to a test graph given by a 16-node unweighted path graph (i.e., an unweighted cycle with an edge removed, see Figure 6.2), we find that the path graph is "closer" to the family $\mathcal{F}_1$ of cycles approximated by the set C_1 than it is to the family $\mathcal{F}_2$. This agrees with the intuition that a path graph is not well-connected, and is only one edge away from being a cycle. We also point out that the optimal solution to the convex program on the left-hand-side of the test (6.13) is in fact an unweighted 16-node cycle with the missing edge in the path graph added as an extra edge. Next we consider a different test graph – a 16-node cycle with two additional edges across diametrically opposite nodes, i.e., assuming we label the nodes of the 16-node cycle we add edges between nodes 1 and 9, and between nodes 5 and 13 (again see Figure 6.2). While this graph is only two edges away from being a cycle, the edges connecting far away nodes dramatically increase the connectivity of the graph. In this case we find using the convex programming hypothesis test (6.13) that the family $\mathcal{F}_2$ is in fact closer than $\mathcal{F}_1$ to the sample graph. Interestingly, the optimal solution to the convex program on the left-hand-side of the test (6.13) is again an unweighted 16-node cycle, this time with the two additional edges removed.

In order to thoroughly address the graph hypothesis testing problem, we need to develop a framework of statistical models over spaces of graphs. With a proper statistical framework in place we can evaluate the *probability of error* achieved by a hypothesis-testing algorithm with respect to a suitable error-metric, analogous to similar methods developed in other classical decision-theoretic problems in statistics. We defer these questions to a separate paper.

■ 6.6 Discussion

In this chapter we introduced and studied convex graph invariants, which are graph invariants that are convex functions of the adjacency matrix. Convex invariants form a rich subset of the set of all graph invariants, and they are useful in developing a unified computational framework based on convex optimization to solve a number of graph problems. In particular we described three canonical problems involving the structural properties of graphs, namely, graph construction given constraints, graph deconvolution of a composite graph into individual components, and graph hypothesis testing in which the objective is decide which of two given families of graphs offers a better explanation for a new sample graph. We presented convex optimization solutions

to all of these problems, with convex graph invariants playing a prominent role. These solutions provided attractive empirical performance, and the resulting convex programs are tractable and can be solved using general-purpose off-the-shelf software for moderate size instances.

Chapter 7

Conclusion

The central theme of this thesis is to provide solutions to address some of the challenges that arise in modeling the interactions among a large collection of variables. Here we describe the main contributions, and discuss some future research directions.

7.1 Summary of Contributions

Rank-Sparsity Uncertainty Principles and Matrix Decomposition

In Chapter 3 we studied the question of decomposing the sum of a sparse matrix and a low-rank matrix into the individual components. Such a decomposition problem arises in a number of applications in system identification, computational complexity, and statistical model selection. Indeed sparse-plus-low-rank matrix decomposition is central to Gaussian latent-variable graphical model selection addressed in Chapter 4. We proposed a tractable convex program to solve the decomposition problem, and gave conditions under which it exactly identifies the correct components. Fundamental to the analysis in Chapter 3 is a new *rank-sparsity uncertainty principle* relating the sparsity pattern of a matrix to its row and column spaces.

Latent Variable Graphical Model Selection via Convex Optimization

Latent variable model selection is a major challenge in statistics, and is also a problem of fundamental interest because the discovery of hidden causes affecting some observed phenomena is important in many scientific endeavors. Our main contribution in this area is a new convex optimization method with theoretical consistency guarantees for graphical model selection with latent variables. Specifically this convex program builds upon the framework in Chapter 3, and our analysis gives conditions under which the program consistently estimates model structure in the high-dimensional scaling regime.

The Convex Geometry of Linear Inverse Problems

The abstract mathematical formulations underlying many problems involving graphs and graphical models can in fact be viewed as instances of *inverse problems* in which we wish to learn/reconstruct structured graphs and simple statistical models given inexact or incomplete information. Chapter 5 develops tractable convex relaxations for a general class of inverse problems in which the objective is to recover certain "simple" models given a limited number of linear measurements. In situations when the underlying models have algebraic structure, the resulting convex programs can be solved or approximated by semidefinite programming. We provide sharp estimates of the number of generic measurements required for exact and robust recovery in a variety of settings. These estimates are based on computing certain Gaussian statistics related to the underlying model geometry.

Convex Graph Invariants

Finally we consider questions motivated by *statistical models over the space of graphs*, so that a graph itself is viewed as a sample drawn from a probability distribution defined over some set of graphs. Natural questions that arise in standard statistical settings can then be posed in a deterministic framework in this graph setting as well. For example we consider problems such as graph deconvolution, graph sampling, and graph hypothesis testing. In order to develop a unified computational framework to solve these problems, we introduce *convex graph invariants* in Chapter 6. We also discuss connections to other concepts such as majorization, robust optimization, and graph isomorphism.

■ 7.2 Future Directions

Special-Purpose Computational Methods

Many of the convex programs proposed in this thesis can be solved in polynomial-time using general-purpose software for moderate-size problem instances. However it is of interest to apply some of the convex programs (e.g., latent-variable graphical model selection in Chapter 4, or the computation of some subset of convex graph invariants in Chapter 6) to large-scale problems instances. Therefore special-purpose algorithms tailored to specific structured convex programs must be developed to scale to massive problem sizes.

Non-Gaussian Latent-Variable Modeling

The methods and analysis in Chapter 4 are relevant for Gaussian model selection. In many applications of interest, e.g., in computational biology, the random variables of interest are fundamentally non-Gaussian. Therefore it is important to develop a similar convex optimization formulation with consistency guarantees for latent-variable models with non-Gaussian variables, e.g., for categorical data.

Computational Approximations and Tradeoffs

Some of the convex programs proposed in Chapter 5 and in Chapter 6 cannot be solved in polynomial-time, and therefore we proposed in those chapters further convex relaxations which are tractable to solve. A basic question of interest in several settings is the tradeoff incurred due to these tractable relaxations. For example in Chapter 5 the tradeoff can be specified in terms of the increased number of linear measurements required for guaranteed recovery via convex optimization.

Non-Gaussian Linear Measurement Models

In Chapter 5 we analyze the recovery guarantees of convex relaxation methods in extracting structured models given linear measurements specified by random Gaussian functionals. While such an analysis is useful for general atomic sets, particular applications often necessitate the study of structured measurement matrices, e.g., partial Fourier measurements of sparse vectors or partial entrywise sampling of low-rank matrices. It is of interest to develop a unified framework based on a notion of *incoherence* that is general enough to encompass most interesting applications.

Conditions for Graph Deconvolution and Graph Generation

A further challenge that we are presently working to address is to provide theoretical guarantees on the performance of our convex programs described in Chapter 6. For example which families of graphs can be deconvolved via the tractable spectral relaxation? Which classes of structured graphs can be generated efficiently via convex optimization?

Appendix A

Proofs of Chapter 3

A.1 SDP Formulation

The problem (3.3) can be recast as a *semidefinite program* (SDP). Using the variational characterizations of the ℓ_1 norm and the nuclear norm from Chapter 2, (3.3) can be rewritten as

$$\begin{aligned} \min_{A,B,W_1,W_2,Z} \quad & \gamma \mathbf{1}_n^T Z \mathbf{1}_n + \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)) \\ \text{s.t.} \quad & \begin{pmatrix} W_1 & B \\ B' & W_2 \end{pmatrix} \succeq 0 \\ & -Z_{i,j} \leq A_{i,j} \leq Z_{i,j}, \quad \forall (i,j) \\ & A + B = C. \end{aligned} \tag{A.1}$$

Here, $\mathbf{1}_n \in \mathbb{R}^n$ refers to the vector that has 1 in every entry.

A.2 Proofs

Proof of Proposition 3.3.1

We begin by establishing that

$$\max_{N \in T(B^\star),\ \|N\| \leq 1} \|P_{\Omega(A^\star)}(N)\| < 1 \quad \Rightarrow \quad \Omega(A^\star) \cap T(B^\star) = \{0\}, \tag{A.2}$$

where $P_{\Omega(A^\star)}(N)$ denotes the projection onto the space $\Omega(A^\star)$. Assume for the sake of a contradiction that this assertion is not true. Thus, there exists $N \neq 0$ such that $N \in \Omega(A^\star) \cap T(B^\star)$. Scale N appropriately such that $\|N\| = 1$. Thus $N \in T(B^\star)$ with $\|N\| = 1$, but we also have that $\|P_{\Omega(A^\star)}(N)\| = \|N\| = 1$ as $N \in \Omega(A^\star)$. This leads to a contradiction.

Next, we show that

$$\max_{N \in T(B^\star),\ \|N\| \leq 1} \|P_{\Omega(A^\star)}(N)\| \leq \mu(A^\star)\xi(B^\star),$$

which would allow us to conclude the proof of this proposition. We have the following sequence of inequalities

$$\begin{aligned} \max_{N \in T(B^\star),\ \|N\| \leq 1} \|P_{\Omega(A^\star)}(N)\| &\leq \max_{N \in T(B^\star),\ \|N\| \leq 1} \mu(A^\star)\|P_{\Omega(A^\star)}(N)\|_\infty \\ &\leq \max_{N \in T(B^\star),\ \|N\| \leq 1} \mu(A^\star)\|N\|_\infty \\ &\leq \mu(A^\star)\xi(B^\star). \end{aligned}$$

Here the first inequality follows from the definition (3.2) of $\mu(A^\star)$ as $P_{\Omega(A^\star)}(N) \in \Omega(A^\star)$, the second inequality is due to the fact that $\|P_{\Omega(A^\star)}(N)\|_\infty \leq \|N\|_\infty$, and the final inequality follows from the definition (3.1) of $\xi(B^\star)$. □

Proof of Proposition 3.4.1

We first show that $(A^\star, B^\star)$ is *an* optimum of (3.3), before moving on to showing uniqueness. Based on subgradient optimality conditions applied at $(A^\star, B^\star)$, there must exist a dual Q such that

$$Q \in \gamma \partial \|A^\star\|_1 \quad \text{and} \quad Q \in \partial \|B^\star\|_*.$$

The second condition in this proposition guarantees the existence of a dual Q that satisfies *both* these conditions simultaneously (see (3.11) and (3.12)). Therefore, we have that $(A^\star, B^\star)$ is *an* optimum. Next we show that under the conditions specified in the lemma, $(A^\star, B^\star)$ is also a unique optimum. To avoid cluttered notation, in the rest of this proof we let $\Omega = \Omega(A^\star)$, $T = T(B^\star)$, $\Omega^c(A^\star) = \Omega^c$, and $T^\perp(B^\star) = T^\perp$.

Suppose that there is another feasible solution $(A^\star + N_A, B^\star + N_B)$ that is also a minimizer. We must have that $N_A + N_B = 0$ because $A^\star + B^\star = C = (A^\star + N_A) + (B^\star + N_B)$. Applying the subgradient property at $(A^\star, B^\star)$, we have that for *any* subgradient (Q_A, Q_B) of the function $\gamma\|A\|_1 + \|B\|_*$ at $(A^\star, B^\star)$

$$\gamma\|A^\star + N_A\|_1 + \|B^\star + N_B\|_* \geq \gamma\|A^\star\|_1 + \|B^\star\|_* + \langle Q_A, N_A \rangle + \langle Q_B, N_B \rangle. \tag{A.3}$$

Since (Q_A, Q_B) is a subgradient of the function $\gamma\|A\|_1 + \|B\|_*$ at $(A^\star, B^\star)$, we must have from (3.11) and (3.12) that

- $Q_A = \gamma \text{sign}(A^\star) + P_{\Omega^c}(Q_A)$, with $\|P_{\Omega^c}(Q_A)\|_\infty \leq \gamma$.
- $Q_B = UV' + P_{T^\perp}(Q_B)$, with $\|P_{T^\perp}(Q_B)\| \leq 1$.

Using these conditions we rewrite $\langle Q_A, N_A \rangle$ and $\langle Q_B, N_B \rangle$. Based on the existence of the dual Q as described in the lemma, we have that

$$\begin{aligned} \langle Q_A, N_A \rangle &= \langle \gamma \text{sign}(A^\star) + P_{\Omega^c}(Q_A), N_A \rangle \\ &= \langle Q - P_{\Omega^c}(Q) + P_{\Omega^c}(Q_A), N_A \rangle \\ &= \langle P_{\Omega^c}(Q_A) - P_{\Omega^c}(Q), N_A \rangle + \langle Q, N_A \rangle, \end{aligned} \tag{A.4}$$

where we have used the fact that $Q = \gamma \text{sign}(A^\star) + P_{\Omega^c}(Q)$. Similarly, we have that

$$\begin{aligned} \langle Q_B, N_B \rangle &= \langle UV' + P_{T^\perp}(Q_B), N_B \rangle \\ &= \langle Q - P_{T^\perp}(Q) + P_{T^\perp}(Q_B), N_B \rangle \\ &= \langle P_{T^\perp}(Q_B) - P_{T^\perp}(Q), N_B \rangle + \langle Q, N_B \rangle, \end{aligned} \tag{A.5}$$

where we have used the fact that $Q = UV' + P_{T^\perp}(Q)$. Putting (A.4) and (A.5) together, we have that

$$\begin{aligned} \langle Q_A, N_A \rangle + \langle Q_B, N_B \rangle &= \langle P_{\Omega^c}(Q_A) - P_{\Omega^c}(Q), N_A \rangle \\ &\quad + \langle P_{T^\perp}(Q_B) - P_{T^\perp}(Q), N_B \rangle \\ &\quad + \langle Q, N_A + N_B \rangle \\ &= \langle P_{\Omega^c}(Q_A) - P_{\Omega^c}(Q), N_A \rangle \\ &\quad + \langle P_{T^\perp}(Q_B) - P_{T^\perp}(Q), N_B \rangle \\ &= \langle P_{\Omega^c}(Q_A) - P_{\Omega^c}(Q), P_{\Omega^c}(N_A) \rangle \\ &\quad + \langle P_{T^\perp}(Q_B) - P_{T^\perp}(Q), P_{T^\perp}(N_B) \rangle. \end{aligned} \tag{A.6}$$

In the second equality, we used the fact that $N_A + N_B = 0$.

Since (Q_A, Q_B) is *any* subgradient of the function $\gamma\|A\|_1 + \|B\|_*$ at $(A^\star, B^\star)$, we have some freedom in selecting $P_{\Omega^c}(Q_A)$ and $P_{T^\perp}(Q_B)$ as long as they still satisfy the subgradient conditions $\|P_{\Omega^c}(Q_A)\|_\infty \leq \gamma$ and $\|P_{T^\perp}(Q_B)\| \leq 1$. We set $P_{\Omega^c}(Q_A) = \gamma \text{sign}(P_{\Omega^c}(N_A))$ so that $\|P_{\Omega^c}(Q_A)\|_\infty = \gamma$ and $\langle P_{\Omega^c}(Q_A), P_{\Omega^c}(N_A) \rangle = \gamma \|P_{\Omega^c}(N_A)\|_1$. Letting $P_{T^\perp}(N_B) = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ be the singular value decomposition of $P_{T^\perp}(N_B)$, we set $P_{T^\perp}(Q_B) = \tilde{U}\tilde{V}^T$ so that $\|P_{T^\perp}(Q_B)\| = 1$ and $\langle P_{T^\perp}(Q_B), P_{T^\perp}(N_B) \rangle = \|P_{T^\perp}(N_B)\|_*$.

With this choice of (Q_A, Q_B), we can simplify (A.6) as follows:

$$\begin{aligned}\langle Q_A, N_A\rangle + \langle Q_B, N_B\rangle \;\geq\; & (\gamma - \|P_{\Omega^c}(Q)\|_\infty)(\|P_{\Omega^c}(N_A)\|_1) \\ & +(1 - \|P_{T^\perp}(Q)\|)(\|P_{T^\perp}(N_B)\|_*).\end{aligned}$$

Since $\|P_{\Omega^c}(Q)\|_\infty < \gamma$ and $\|P_{T^\perp}(Q)\| < 1$, we have that $\langle Q_A, N_A\rangle + \langle Q_B, N_B\rangle$ is strictly positive unless $P_{\Omega^c}(N_A) = 0$ and $P_{T^\perp}(N_B) = 0$. Thus, $\gamma\|A^\star + N_A\|_1 + \|B^\star + N_B\|_* > \gamma\|A^\star\|_1 + \|B^\star\|_*$ if $P_{\Omega^c}(N_A) \neq 0$ and $P_{T^\perp}(N_B) \neq 0$. However, if $P_{\Omega^c}(N_A) = P_{T^\perp}(N_B) = 0$, then $P_\Omega(N_A) + P_T(N_B) = 0$ because we also have that $N_A + N_B = 0$. In other words, $P_\Omega(N_A) = -P_T(N_B)$. This can only be possible if $P_\Omega(N_A) = P_T(N_B) = 0$ (as $\Omega \cap T = \{0\}$), which in turn implies that $N_A = N_B = 0$. Therefore, $\gamma\|A^\star + N_A\|_1 + \|B^\star + N_B\|_* > \gamma\|A^\star\|_1 + \|B^\star\|_*$ unless $N_A = N_B = 0$. □

Proof of Theorem 3.4.1

As with the previous proof, we avoid cluttered notation by letting $\Omega = \Omega(A^\star)$, $T = T(B^\star)$, $\Omega^c(A^\star) = \Omega^c$, and $T^\perp(B^\star) = T^\perp$. One can check that

$$\xi(B^\star)\mu(A^\star) < \frac{1}{6} \Rightarrow \frac{\xi(B^\star)}{1 - 4\xi(B^\star)\mu(A^\star)} < \frac{1 - 3\xi(B^\star)\mu(A^\star)}{\mu(A^\star)}. \tag{A.7}$$

Thus, we show that if $\xi(B^\star)\mu(A^\star) < \frac{1}{6}$ then there exists a range of γ for which a dual Q with the requisite properties exists. Also note that plugging in $\xi(B^\star)\mu(A^\star) = \frac{1}{6}$ in the above range gives the strictly smaller range $[3\xi(B^\star), \frac{1}{2\mu(A^\star)}]$ for γ; for any choice of $p \in [0, 1]$ we have that $\gamma = \frac{(3\xi(B^\star))^p}{(2\mu(A^\star))^{1-p}}$ is always within the above range.

We aim to construct a dual Q by considering candidates in the direct sum $\Omega \oplus T$ of the tangent spaces. Since $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$, we can conclude from Proposition 3.3.1 that there exists a *unique* $\hat{Q} \in \Omega \oplus T$ such that $P_\Omega(\hat{Q}) = \gamma\text{sign}(A^\star)$ and $P_T(\hat{Q}) = UV'$ (recall that these are conditions that a dual must satisfy according to Proposition 3.4.1), as $\Omega \cap T = \{0\}$. The rest of this proof shows that if $\mu(A^\star)\xi(B^\star) < \frac{1}{6}$ then the projections of such a $\hat{Q}$ onto $T^\perp$ and onto Ω^c will be small, i.e., we show that $\|P_{\Omega^c}(\hat{Q})\|_\infty < \gamma$ and $\|P_{T^\perp}(\hat{Q})\| < 1$.

We note here that $\hat{Q}$ can be *uniquely* expressed as the sum of an element of T and an element of Ω, i.e., $\hat{Q} = Q_\Omega + Q_T$ with $Q_\Omega \in \Omega$ and $Q_T \in T$. The uniqueness of the splitting can be concluded because $\Omega \cap T = \{0\}$. Let $Q_\Omega = \gamma\text{sign}(A^\star) + \epsilon_\Omega$ and $Q_T = UV' + \epsilon_T$. We then have

$$P_\Omega(\hat{Q}) = \gamma\text{sign}(A^\star) + \epsilon_\Omega + P_\Omega(Q_T) = \gamma\text{sign}(A^\star) + \epsilon_\Omega + P_\Omega(UV' + \epsilon_T).$$

Since $P_\Omega(\hat{Q}) = \gamma \text{sign}(A^\star)$,

$$\epsilon_\Omega = -P_\Omega(UV' + \epsilon_T). \tag{A.8}$$

Similarly,

$$\epsilon_T = -P_T(\gamma \text{sign}(A^\star) + \epsilon_\Omega). \tag{A.9}$$

Next, we obtain the following bound on $\|P_{\Omega^c}(\hat{Q})\|_\infty$:

$$\begin{aligned} \|P_{\Omega^c}(\hat{Q})\|_\infty &= \|P_{\Omega^c}(UV' + \epsilon_T)\|_\infty \\ &\leq \|UV' + \epsilon_T\|_\infty \\ &\leq \xi(B^\star)\|UV' + \epsilon_T\| \\ &\leq \xi(B^\star)(1 + \|\epsilon_T\|), \end{aligned} \tag{A.10}$$

where we obtain the second inequality based on the definition of $\xi(B^\star)$ (since $UV' + \epsilon_T \in T$). Similarly, we can obtain the following bound on $\|P_{T^\perp}(\hat{Q})\|$

$$\begin{aligned} \|P_{T^\perp}(\hat{Q})\| &= \|P_{T^\perp}(\gamma \text{sign}(A^\star) + \epsilon_\Omega)\| \\ &\leq \|\gamma \text{sign}(A^\star) + \epsilon_\Omega\| \\ &\leq \mu(A^\star)\|\gamma \text{sign}(A^\star) + \epsilon_\Omega\|_\infty \\ &\leq \mu(A^\star)(\gamma + \|\epsilon_\Omega\|_\infty), \end{aligned} \tag{A.11}$$

where we obtain the second inequality based on the definition of $\mu(A^\star)$ (since $\gamma \text{sign}(A^\star) + \epsilon_\Omega \in \Omega$). Thus, we can bound $\|P_{\Omega^c}(\hat{Q})\|_\infty$ and $\|P_{T^\perp}(\hat{Q})\|$ by bounding $\|\epsilon_T\|$ and $\|\epsilon_\Omega\|_\infty$ respectively (using the relations (A.9) and (A.8)).

By definition of $\xi(B^\star)$ and using (A.8),

$$\begin{aligned} \|\epsilon_\Omega\|_\infty &= \|P_\Omega(UV' + \epsilon_T)\|_\infty \\ &\leq \|UV' + \epsilon_T\|_\infty \\ &\leq \xi(B^\star)\|UV' + \epsilon_T\| \\ &\leq \xi(B^\star)(1 + \|\epsilon_T\|), \end{aligned} \tag{A.12}$$

where the second inequality is obtained because $UV' + \epsilon_T \in T$. Similarly, by definition of $\mu(A^\star)$ and using (A.9)

$$\begin{aligned} \|\epsilon_T\| &= \|P_T(\gamma \text{sign}(A^\star) + \epsilon_\Omega)\| \\ &\leq 2\|\gamma \text{sign}(A^\star) + \epsilon_\Omega\| \\ &\leq 2\mu(A^\star)\|\gamma \text{sign}(A^\star) + \epsilon_\Omega\|_\infty \\ &\leq 2\mu(A^\star)(\gamma + \|\epsilon_\Omega\|_\infty), \end{aligned} \tag{A.13}$$

where the first inequality is obtained because $\|P_T(M)\| \leq 2\|M\|$, and the second inequality is obtained because $\gamma \text{sign}(A^\star) + \epsilon_\Omega \in \Omega$.

Putting (A.12) in (A.13), we have that

$$\begin{aligned} \|\epsilon_T\| &\leq 2\mu(A^\star)(\gamma + \xi(B^\star)(1 + \|\epsilon_T\|)) \\ \Rightarrow \|\epsilon_T\| &\leq \frac{2\gamma\mu(A^\star) + 2\xi(B^\star)\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}. \end{aligned} \tag{A.14}$$

Similarly, putting (A.13) in (A.12), we have that

$$\begin{aligned} \|\epsilon_\Omega\|_\infty &\leq \xi(B^\star)(1 + 2\mu(A^\star)(\gamma + \|\epsilon_\Omega\|_\infty)) \\ \Rightarrow \|\epsilon_\Omega\|_\infty &\leq \frac{\xi(B^\star) + 2\gamma\xi(B^\star)\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}. \end{aligned} \tag{A.15}$$

We now show that $\|P_{T^\perp}(\hat{Q})\| < 1$. Combining (A.15) and (A.11),

$$\begin{aligned} \|P_{T^\perp}(\hat{Q})\| &\leq \mu(A^\star)\left(\gamma + \frac{\xi(B^\star) + 2\gamma\xi(B^\star)\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) \\ &= \mu(A^\star)\left(\frac{\gamma + \xi(B^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) \\ &< \mu(A^\star)\left(\frac{\frac{1-3\xi(B^\star)\mu(A^\star)}{\mu(A^\star)} + \xi(B^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) \\ &= 1, \end{aligned}$$

since $\gamma < \frac{1-3\xi(B^\star)\mu(A^\star)}{\mu(A^\star)}$ by assumption.

Finally, we show that $\|P_{\Omega^c}(\hat{Q})\|_\infty < \gamma$. Combining (A.14) and (A.10),

$$\begin{aligned} \|P_{\Omega^c}(\hat{Q})\|_\infty &\leq \xi(B^\star)\left(1 + \frac{2\gamma\mu(A^\star) + 2\xi(B^\star)\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) \\ &= \xi(B^\star)\left(\frac{1 + 2\gamma\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) \\ &= \left[\xi(B^\star)\left(\frac{1 + 2\gamma\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right) - \gamma\right] + \gamma \\ &= \left[\frac{\xi(B^\star) + 2\gamma\xi(B^\star)\mu(A^\star) - \gamma + 2\gamma\xi(B^\star)\mu(A^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right] + \gamma \\ &= \left[\frac{\xi(B^\star) - \gamma(1 - 4\xi(B^\star)\mu(A^\star))}{1 - 2\xi(B^\star)\mu(A^\star)}\right] + \gamma \\ &< \left[\frac{\xi(B^\star) - \xi(B^\star)}{1 - 2\xi(B^\star)\mu(A^\star)}\right] + \gamma \\ &= \gamma. \end{aligned}$$

Here, we used the fact that $\frac{\xi(B^\star)}{1-4\xi(B^\star)\mu(A^\star)} < \gamma$ in the second inequality. □

Proof of Proposition 3.4.2

Based on the Perron-Frobenius theorem [82], one can conclude that $\|P\| \geq \|Q\|$ if $P_{i,j} \geq |Q_{i,j}|$, $\forall\, i, j$. Thus, we need only consider the matrix that has 1 in every location in the support set $\Omega(A)$ and 0 everywhere else. Based on the definition of the spectral norm, we can re-write $\mu(A)$ as follows:

$$\mu(A) = \max_{\|x\|_2=1, \|y\|_2=1} \sum_{(i,j)\in\Omega(A)} x_i y_j. \tag{A.16}$$

Upper bound For any matrix M, we have from the results in [130] that

$$\|M\|^2 \leq \max_{i,j} \quad \mathrm{r}_i \mathrm{c}_j, \tag{A.17}$$

where $\mathrm{r}_i = \sum_k |M_{i,k}|$ denotes the absolute row-sum of row i and $\mathrm{c}_j = \sum_k |M_{k,j}|$ denotes the absolute column-sum of column j. Let $M^{\Omega(A)}$ be a matrix defined as follows:

$$M^{\Omega(A)}_{i,j} = \begin{cases} 1, & (i,j) \in \Omega(A) \\ 0, & \text{otherwise.} \end{cases}$$

Based on the reformulation of $\mu(A)$ above (A.16), it is clear that

$$\mu(A) = \|M^{\Omega(A)}\|.$$

From the bound (A.17), we have that

$$\|M^{\Omega(A)}\| \leq \deg_{\max}(A).$$

Lower bound Now suppose that each row/column of A has *at least* $\deg_{\min}(A)$ nonzero entries. Using the reformulation (A.16) of $\mu(A)$ above, we have that

$$\mu(A) \geq \sum_{(i,j)\in\Omega(A)} \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} = \frac{|\mathrm{support}(A)|}{n} \geq \deg_{\min}(A).$$

Here we set $x = y = \frac{1}{\sqrt{n}}\mathbf{1}$, with $\mathbf{1}$ representing the all-ones vector, as candidates in the optimization problem (A.16). □

Proof of Proposition 3.4.3

Let $B = U\Sigma V^T$ be the SVD of B.

Upper bound We can upper-bound $\xi(B)$ as follows

$$\begin{aligned}
\xi(B) &= \max_{M \in T(B), \|M\| \leq 1} \|M\|_\infty \\
&= \max_{M \in T(B), \|M\| \leq 1} \|P_{T(B)}(M)\|_\infty \\
&\leq \max_{\|M\| \leq 1} \|P_{T(B)}(M)\|_\infty \\
&\leq \max_{M \text{ orthogonal}} \|P_{T(B)}(M)\|_\infty \\
&\leq \max_{M \text{ orthogonal}} \|P_U M\|_\infty + \max_{M \text{ orthogonal}} \|(I_{n\times n} - P_U) M P_V\|_\infty.
\end{aligned}$$

For the second inequality, we have used the fact that the maximum of a convex function over a convex set is achieved at one of the extreme points of the constraint set. The orthogonal matrices are the extreme points of the set of contractions (i.e., matrices with spectral norm ≤ 1). Note that for the non-square case we would need to consider partial isometries; the rest of the proof remains unchanged. We have used $P_{T(B)}(M) = P_U M + M P_V - P_U M P_V$ from (3.8) in the last inequality, where $P_U = UU^T$ and $P_V = VV^T$ denote the projections onto the spaces spanned by U and V respectively.

We have the following simple bound for $\|P_U M\|_\infty$ with M orthogonal:

$$\begin{aligned}
\max_{M \text{ orthogonal}} \|P_U M\|_\infty &= \max_{M \text{ orthogonal}} \max_{i,j} e_i^T P_U M e_j \\
&\leq \max_{M \text{ orthogonal}} \max_{i,j} \|P_U e_i\|_2 \, \|M e_j\|_2 \\
&= \max_i \|P_U e_i\|_2 \times \max_{M \text{ orthogonal}} \max_j \|M e_j\|_2 \\
&= \beta(U). \qquad \text{(A.18)}
\end{aligned}$$

Here we used the Cauchy-Schwartz inequality in the second line, and the definition of β from (3.13) in the last line.

Similarly, we have that

$$\begin{aligned}
\max_{M \text{ orthogonal}} \|(I_{n\times n} - P_U) M P_V\|_\infty &= \max_{M \text{ orthogonal}} \max_{i,j} e_i^T (I_{n\times n} - P_U) M P_V e_j \\
&\leq \max_{M \text{ orthogonal}} \max_{i,j} \|(I_{n\times n} - P_U) e_i\|_2 \, \|M P_V e_j\|_2 \\
&= \max_i \|(I_{n\times n} - P_U) e_i\|_2 \times \max_{M \text{ orthogonal}} \max_j \|M P_V e_j\|_2 \\
&\leq 1 \times \max_j \|P_V e_j\|_2 \\
&= \beta(V). \qquad \text{(A.19)}
\end{aligned}$$

Using the definition of $\text{inc}(B)$ from (3.14) along with (A.18) and (A.19), we have

that

$$\xi(B) \leq \beta(U) + \beta(V) \leq 2 \text{ inc}(B).$$

Lower bound Next we prove a lower bound on $\xi(B)$. Recall the definition of the tangent space $T(B)$ from (3.7). We restrict our attention to elements of the tangent space $T(B)$ of the form $P_U M = UU^T M$ for M orthogonal (an analogous argument follows for elements of the form $P_V M$ for M orthogonal). One can check that

$$\|P_U M\| = \max_{\|x\|_2=1, \|y\|_2=1} x^T P_U M y \leq \max_{\|x\|_2=1} \|P_U x\|_2 \max_{\|y\|_2=1} \|My\|_2 \leq 1.$$

Therefore,

$$\xi(B) \geq \max_{M \text{ orthogonal}} \|P_U M\|_\infty.$$

Thus, we only need to show that the inequality in line (2) of (A.18) is achieved by some orthogonal matrix M in order to conclude that $\xi(B) \geq \beta(U)$. Define the "most aligned" basis vector with the subspace U as follows:

$$i^* = \arg\max_i \|P_U e_i\|_2.$$

Let M be any orthogonal matrix with one of its columns equal to $\frac{1}{\beta(U)} P_U e_{i^*}$, i.e., a normalized version of the projection onto U of the most aligned basis vector. One can check that such a orthogonal matrix achieves equality in line (2) of (A.18). Consequently, we have that

$$\xi(B) \geq \max_{M \text{ orthogonal}} \|P_U M\|_\infty = \beta(U).$$

By a similar argument with respect to V, we have the lower bound as claimed in the proposition. □

Appendix B

Proofs of Chapter 4

B.1 Matrix Perturbation Bounds

Given a low-rank matrix we consider what happens to the invariant subspaces when the matrix is perturbed by a small amount. We assume without loss of generality that the matrix under consideration is square and symmetric, and our methods can be extended to the general non-symmetric non-square case. We refer the interested reader to [7,87] for more details, as the results presented here are only a brief summary of what is relevant for this Appendix. In particular the arguments presented here are along the lines of those presented in [7]. The appendices in [7] also provide a more refined analysis of second-order perturbation errors.

The resolvent of a matrix M is given by $(M - \zeta I)^{-1}$ [87], and it is well-defined for all $\zeta \in \mathbb{C}$ that do not coincide with an eigenvalue of M. If M has no eigenvalue with magnitude equal to η, then we have by the Cauchy residue formula that the projector onto the invariant subspace of a matrix M corresponding to all singular values smaller than η is given by

$$P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} (M - \zeta I)^{-1} d\zeta, \tag{B.1}$$

where $\mathcal{C}_\eta$ denotes the positively-oriented circle of radius η centered at the origin. Similarly, we have that the weighted projection onto the smallest singular values is given by

$$P^w_{M,\eta} = M P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} \zeta \, (M - \zeta I)^{-1} d\zeta, \tag{B.2}$$

Suppose that M is a low-rank matrix with smallest non-zero singular value σ, and let Δ be a perturbation of M such that $\|\Delta\|_2 \leq \kappa < \frac{\sigma}{2}$. We have the following identity for any $|\zeta| = \kappa$, which will be used repeatedly:

$$[(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} = -[M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1}. \tag{B.3}$$

We then have that

$$\begin{aligned} P_{M+\Delta,\kappa} - P_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{B.4}$$

Similarly, we have the following for $P^w_{M,\kappa}$:

$$\begin{aligned} P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \left\{ [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \left\{ [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \, [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} d\zeta \\ &\quad - \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \, [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{B.5}$$

Given these expressions, we have the following two results.

Proposition B.1.1. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest non-zero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 \leq \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \|\Delta\|_2.$$

Proof: This result follows directly from the expression (B.4), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned} \|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 &\leq \frac{1}{2\pi} \, 2\pi \, \kappa \, \frac{1}{\sigma - \kappa} \, \|\Delta\|_2 \, \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \|\Delta\|_2. \end{aligned}$$

Here, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \kappa}$ and $\|[(M+\Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. □

Next, we develop a similar bound for $P^w_{M,\kappa}$. Let $U(M)$ denote the invariant subspace of M corresponding to the non-zero singular values, and let $P_{U(M)}$ denote the projector onto this subspace.

Proposition B.1.2. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest non-zero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} - (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \leq \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})} \|\Delta\|_2^2.$$

Proof: One can check that

$$\frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta\, [M - \zeta I]^{-1}\Delta[M - \zeta I]^{-1} d\zeta = (I - P_{U(M)})\Delta(I - P_{U(M)}).$$

Next we use the expression (B.5), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned} \|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &- (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \\ &\leq \frac{1}{2\pi}\, 2\pi\, \kappa\, \kappa\, \frac{1}{\sigma-\kappa} \|\Delta\|_2 \frac{1}{\sigma-\kappa} \|\Delta\|_2 \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})} \|\Delta\|_2^2. \end{aligned}$$

As with the previous proof, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\kappa}$ and $\|[(M + \Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. □

We will use these expressions to derive bounds on the "twisting" between the tangent spaces at M and at $M + \Delta$ with respect to the rank variety.

■ B.2 Curvature of Rank Variety

For a symmetric rank-r matrix M, the projection onto the tangent space $T(M)$ (restricted to the variety of symmetric matrices with rank less than or equal to r) can be written in terms of the projection $P_{U(M)}$ onto the row space $U(M)$. For any matrix N

$$\mathcal{P}_{T(M)}(N) = P_{U(M)}N + NP_{U(M)} - P_{U(M)}NP_{U(M)}.$$

One can then check that the projection onto the normal space $T(M)^\perp$

$$\mathcal{P}_{T(M)^\perp}(N) = [I - \mathcal{P}_{T(M)}](N) = (I - P_{U(M)})\ N\ (I - P_{U(M)}).$$

Proof of Proposition 4.2.1: For any matrix N, we have that

$$\begin{aligned} [\mathcal{P}_{T(M+\Delta)} &- \mathcal{P}_{T(M)}](N) = \\ &[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]. \end{aligned}$$

Further, we note that for $\kappa < \frac{\sigma}{2}$

$$\begin{aligned} P_{U(M+\Delta)} - P_{U(M)} &= [I - P_{U(M)}] - [I - P_{U(M+\Delta)}] \\ &= P_{M,\kappa} - P_{M+\Delta,\kappa}, \end{aligned}$$

where $P_{M,\kappa}$ is defined in the previous section. Thus, we have the following sequence of inequalities for $\kappa = \frac{\sigma}{4}$:

$$\begin{aligned} \rho(T(M+\Delta), T(M)) &= \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] \\ &\qquad + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}]\|_2 \\ &\quad + \max_{\|N\|_2 \leq 1} \|[I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq 2\ \|P_{M+\Delta,\frac{\sigma}{4}} - P_{M,\frac{\sigma}{4}}\|_2 \\ &\leq \frac{2}{\sigma}\ \|\Delta\|_2, \end{aligned}$$

where we obtain the last inequality from Proposition B.1.1. □

Proof of Proposition 4.2.2: Since both M and $M + \Delta$ are rank-r matrices, we have that $\mathcal{P}^w_{M+\Delta,\kappa} = \mathcal{P}^w_{M,\kappa} = 0$. Consequently,

$$\begin{aligned} \|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 &= \|(I - P_{U(M)})\ \Delta\ (I - P_{U(M)})\|_2 \\ &\leq \frac{\|\Delta\|_2^2}{\sigma}, \end{aligned}$$

where we obtain the last inequality from Proposition B.1.2 with $\kappa = \frac{\sigma}{4}$. □

Proof of Lemma 4.3.1: Since $\rho(T_1, T_2) < 1$ one can check that the largest principal angle between T_1 and T_2 is strictly less than $\frac{\pi}{2}$. Consequently, the mapping $\mathcal{P}_{T_2} : T_1 \to T_2$ restricted to T_1 is bijective (as it is injective, and the spaces T_1, T_2 have the same dimension). Consider the maximum and minimum gain of the operator $\mathcal{P}_{T_2}$ restricted to T_1; for any $M \in T_1, \|M\|_2 = 1$:

$$\begin{aligned} \|\mathcal{P}_{T_2}(M)\|_2 &= \|M + [\mathcal{P}_{T_2} - \mathcal{P}_{T_1}](M)\|_2 \\ &\in [1 - \rho(T_1, T_2), 1 + \rho(T_1, T_2)]. \end{aligned}$$

Therefore, we can rewrite $\xi(T_2)$ as follows:

$$\begin{aligned}
\xi(T_2) &= \max_{N\in T_2, \|N\|_2\leq 1} \|N\|_\infty \\
&= \max_{N\in T_2, \|N\|_2\leq 1} \|\mathcal{P}_{T_2}(N)\|_\infty \\
&\leq \max_{N\in T_1, \|N\|_2\leq \frac{1}{1-\rho(T_1,T_2)}} \|\mathcal{P}_{T_2}(N)\|_\infty \\
&\leq \max_{N\in T_1, \|N\|_2\leq \frac{1}{1-\rho(T_1,T_2)}} [\|N\|_\infty + \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty] \\
&\leq \frac{1}{1 - \rho(T_1,T_2)} \left[\xi(T_1) + \max_{N\in T_1, \|N\|_2\leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty\right] \\
&\leq \frac{1}{1 - \rho(T_1,T_2)} \left[\xi(T_1) + \max_{\|N\|_2\leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2\right] \\
&\leq \frac{1}{1 - \rho(T_1,T_2)} [\xi(T_1) + \rho(T_1,T_2)] .
\end{aligned}$$

This concludes the proof of the lemma. □

■ B.3 Transversality and Identifiability

Proof of Lemma 4.3.3: We have that $\mathcal{A}^\dagger\mathcal{A}(S,L) = (S+L, S+L)$; therefore, $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L) = (S + \mathcal{P}_\Omega(L), \mathcal{P}_T(S) + L)$. We need to bound $\|S + \mathcal{P}_\Omega(L)\|_\infty$ and $\|\mathcal{P}_T(S) + L\|_2$. First, we have

$$\begin{aligned}
\|S + \mathcal{P}_\Omega(L)\|_\infty &\in [\|S\|_\infty - \|\mathcal{P}_\Omega(L)\|_\infty, \|S\|_\infty + \|\mathcal{P}_\Omega(L)\|_\infty] \\
&\subseteq [\|S\|_\infty - \|L\|_\infty, \|S\|_\infty + \|L\|_\infty] \\
&\subseteq [\gamma - \xi(T), \gamma + \xi(T)].
\end{aligned}$$

Similarly, one can check that

$$\begin{aligned}
\|\mathcal{P}_T(S) + L\|_2 &\in [-\|\mathcal{P}_T(S)\|_2 + \|L\|_2, \|\mathcal{P}_T(S)\|_2 + \|L\|_2] \\
&\subseteq [1 - 2\|S\|_2, 1 + 2\|S\|_2] \\
&\subseteq [1 - 2\gamma\mu(\Omega), 1 + 2\gamma\mu(\Omega)].
\end{aligned}$$

Thus, we can conclude that

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \in [1 - \chi(\Omega,T,\gamma), 1 + \chi(\Omega,T,\gamma)].$$

where $\chi(\Omega,T,\gamma)$ is defined in (4.7). □

Proof of Proposition 4.3.1: Before proving the two parts of this proposition we make a simple observation about $\xi(T')$ using the condition that $\rho(T,T') \leq \frac{\xi(T)}{2}$:

$$\begin{aligned}\xi(T') &\leq \frac{\xi(T)+\rho(T,T')}{1-\rho(T,T')} \\ &\leq \frac{\frac{3\xi(T)}{2}}{1-\frac{\xi(T)}{2}} \\ &\leq 3\xi(T).\end{aligned}$$

Here we used the property that $\xi(T) \leq 1$ in obtaining the final inequality. Consequently, noting that $\gamma \in [\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}]$ implies that

$$\chi(\Omega, T', \gamma) = \max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\} \leq \frac{\nu\alpha}{\beta(2-\nu)}. \tag{B.6}$$

Part 1: The proof of this step proceeds in a similar manner to that of Lemma 4.3.3. First we have for $S \in \Omega, L \in T'$ with $\|S\|_\infty = \gamma, \|L\|_2 = 1$:

$$\begin{aligned}\|\mathcal{P}_\Omega \mathcal{I}^*(S+L)\|_\infty &\geq \|\mathcal{P}_\Omega \mathcal{I}^* S\|_\infty - \|\mathcal{P}_\Omega \mathcal{I}^* L\|_\infty \\ &\geq \alpha\gamma - \|\mathcal{I}^* L\|_\infty \\ &\geq \alpha\gamma - \beta\xi(T').\end{aligned}$$

Next under the same conditions on S, L,

$$\begin{aligned}\|\mathcal{P}_{T'} \mathcal{I}^*(S+L)\|_2 &\geq \|\mathcal{P}_{T'} \mathcal{I}^* L\|_2 - \|\mathcal{P}_{T'} \mathcal{I}^* S\|_2 \\ &\geq \alpha - 2\|\mathcal{I}^* S\|_2 \\ &\geq \alpha - 2\beta\mu(\Omega)\gamma.\end{aligned}$$

Combining these last two bounds with (B.6), we conclude that

$$\begin{aligned}\min_{(S,L)\in\mathcal{Y},\ \|S\|_\infty=\gamma,\ \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S,L)) &\geq \alpha - \beta \max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\} \\ &\geq \alpha - \frac{\nu\alpha}{2-\nu} \\ &= \frac{2\alpha(1-\nu)}{2-\nu} \\ &\geq \frac{\alpha}{2},\end{aligned}$$

where the final inequality follows from the assumption that $\nu \in (0, \frac{1}{2}]$.

Part 2: Note that for $S \in \Omega, L \in T'$ with $\|S\|_\infty \leq \gamma, \|L\|_2 \leq 1$

$$\begin{aligned} \|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*(S+L)\|_\infty &\leq \|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*S\|_\infty + \|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*L\|_\infty \\ &\leq \delta\gamma + \beta\xi(T'). \end{aligned}$$

Similarly

$$\begin{aligned} \|\mathcal{P}_{T'^\perp}\mathcal{I}^*(S+L)\|_2 &\leq \|\mathcal{P}_{T'^\perp}\mathcal{I}^*S\|_2 + \|\mathcal{P}_{T'^\perp}\mathcal{I}^*L\|_2 \\ &\leq \delta + \beta\gamma\mu(\Omega). \end{aligned}$$

Combining these last two bounds with the bounds from the first part, we have that

$$\begin{aligned} \left\|\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}\left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}\right)^{-1}\right\|_{g_\gamma \to g_\gamma} &\leq \frac{\delta + \beta\max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\}}{\alpha - \beta\max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\}} \\ &\leq \frac{\delta + \frac{\nu\alpha}{2-\nu}}{\alpha - \frac{\nu\alpha}{2-\nu}} \\ &\leq \frac{(1-2\nu)\alpha + \frac{\nu\alpha}{2-\nu}}{\alpha - \frac{\nu\alpha}{2-\nu}} \\ &= 1 - \nu. \end{aligned}$$

This concludes the proof of the proposition. □

■ B.4 Proof of Main Result

Here we prove Theorem 4.4.1. Throughout this section we denote $m = \max\{1, \frac{1}{\gamma}\}$. Further $\Omega = \Omega(K_O^*)$ and $T = T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent spaces at the "true" sparse matrix $S^* = K_O^*$ and low-rank matrix $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. We assume that

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right] \tag{B.7}$$

We also let $E_n = \Sigma_O^n - \Sigma_O^*$ denote the difference between the true marginal covariance and the sample covariance. Finally we let $D = \max\{1, \frac{\nu\alpha}{3\beta(2-\nu)}\}$ throughout this section. For γ in the above range we note that

$$m \leq \frac{D}{\xi(T)}. \tag{B.8}$$

Standard facts that we use throughout this section are that $\xi(T) \leq 1$ and that $\|M\|_\infty \leq \|M\|_2$ for any matrix M.

We study the following convex program:

$$(\bar{S}_n, \bar{L}_n) = \arg\min_{S,L} \operatorname{Tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \quad \text{s.t.} \quad S - L \succ 0. \tag{B.9}$$

Comparing (B.9) with the convex program (4.9), the main difference is that we do not constraint the variable L to be positive semidefinite in (B.9) (recall that the nuclear norm of a positive semidefinite matrix is equal to its trace). However we show that the unique optimum $(\bar{S}_n, \bar{L}_n)$ of (B.9) under the hypotheses of Theorem 4.4.1 is such that $\bar{L}_n \succeq 0$ (with high probability). Therefore we conclude that $(\bar{S}_n, \bar{L}_n)$ is also the unique optimum of (4.9). The subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UDV^T$ is as follows:

$$N \in \partial\|M\|_* \quad \Leftrightarrow \quad \mathcal{P}_{T(M)}(N) = UV^T,\ \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1.$$

The proof of this theorem consists of a number of steps, each of which is analyzed in separate sections below. We explicitly keep track of the constants α, β, ν, ψ. The key ideas are as follows:

1. We show that if we solve the convex program (B.9) subject to the additional constraints that $S \in \Omega$ and $L \in T'$ for some T' "close to" T (measured by $\rho(T', T)$), then the error between the optimal solution $(\bar{S}_n, \bar{L}_n)$ and the underlying matrices (S^*, L^*) is small. This result is discussed in Appendix B.4.2.

2. We analyze the optimization problem (B.9) with the additional constraint that the variables S and L belong to the algebraic varieties of sparse and low-rank matrices respectively, and that the corresponding tangent spaces are close to the tangent spaces at (S^*, L^*). We show that under suitable conditions on the minimum nonzero singular value of the true low-rank matrix L^* and on the minimum magnitude nonzero entry of the true sparse matrix S^*, the optimum of this modified program is achieved at a *smooth* point of the underlying varieties. In particular the bound on the minimum nonzero singular value of L^* helps bound the curvature of the low-rank matrix variety locally around L^* (we use the results described in Appendix B.2). Further we also show that the tangent-spaces at the solution to this variety-constrained problem are close to the tangent spaces at the true underlying matrices (S^*, L^*). These results are described in Appendix B.4.3.

3. The next step is to show that the variety constraint can be linearized and changed to a tangent-space constraint (see Appendix B.4.4), thus giving us a *convex program.* Under suitable conditions this tangent-space constrained program also has an optimum that has the same support/rank as the true (S^*, L^*). Based on the previous step these tangent spaces in the constraints are close to the tangent spaces at the true (S^*, L^*). Therefore we use the first step to conclude that the resulting error in the estimate is small.

4. Finally we show that under the identifiability conditions of Section 4.3 these tangent-space constraints are inactive at the optimum (see Appendix B.4.7). Therefore we conclude with the statement that the optimum of the convex program (B.9) without any variety constraints is achieved at a pair of matrices that have the same support/rank as the true (S^*, L^*) (with high probability). Further the low-rank component of the solution is positive semidefinite, thus allowing us to conclude that the original convex program (4.9) also provides estimates that are consistent.

■ B.4.1 Bounded curvature of matrix inverse

Consider the Taylor series of the inverse of a matrix:

$$(M + \Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + R_{M^{-1}}(\Delta),$$

where

$$R_{M^{-1}}(\Delta) = M^{-1}\left[\sum_{k=2}^{\infty}(-\Delta M^{-1})^k\right].$$

This infinite sum converges for Δ sufficiently small. The following proposition provides a bound on the second-order term specialized to our setting:

Proposition B.4.1. *Suppose that γ is in the range given by* (B.7). *Let $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$ for $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and for any (Δ_S, Δ_L) with $\Delta_S \in \Omega$. Then we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}\mathcal{A}(\Delta_S, \Delta_L)) \leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)}.$$

Proof: We have that

$$\begin{aligned}
\|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\
&\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\
&\leq (1+\gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\
&\leq (1+\frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\
&\leq \frac{1}{2\psi},
\end{aligned}$$

where the second-to-last inequality follows from the range for γ (B.7), and the final inequality follows from the bound on $g_\gamma(\Delta_S, \Delta_L)$. Therefore,

$$\begin{aligned}
\|R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 &\leq \psi\sum_{k=2}^{\infty}(\|\Delta_S + \Delta_L\|_2\psi)^k \\
&\leq \psi^3\|\Delta_S + \Delta_L\|_2^2\frac{1}{1-\|\Delta_S + \Delta_L\|_2\psi} \\
&\leq 2\psi^3(1+\frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2 \\
&= 2\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2.
\end{aligned}$$

Here we apply the last two inequalities from above. Since the $\|\cdot\|_\infty$-norm is bounded above by the spectral norm $\|\cdot\|_2$, we have the desired result. □

■ B.4.2 Bounded errors

Next we analyze the following convex program subject to certain additional tangent-space constraints:

$$\begin{aligned}
(\hat{S}_\Omega, \hat{L}_{T'}) = \arg\min_{S,L} \operatorname{Tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.} \quad S - L \succ 0, \quad S \in \Omega, \quad L \in T',
\end{aligned} \tag{B.10}$$

for some subspace T'. We show that if T' is any tangent space to the low-rank matrix variety such that $\rho(T, T') \leq \frac{\xi(T)}{2}$, then we can bound the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})$. Let $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$ denote the orthogonal component of the true low-rank matrix, and recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the true marginal covariance and the sample covariance. The proof of the following result uses Brouwer's fixed-point theorem [113], and is inspired by the proof of a similar result in [119] for standard sparse graphical model recovery without latent variables.

Proposition B.4.2. *Let the error* (Δ_S, Δ_L) *in the solution of the convex program* (B.10) *(with* T' *such that* $\rho(T', T) \leq \frac{\xi(T)}{2}$*) be as defined above. Further let* $C_1 = \psi(1 + \frac{\alpha}{6\beta})$*, and define*

$$r = \max\left\{\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) + \lambda_n\right], \|\mathcal{C}_{T'}\|_2\right\}.$$

If we have that

$$r \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\},$$

for γ *in the range given by* (B.7)*, then*

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

Proof: Based on Proposition 4.3.1 we note that the convex program (B.10) is strictly convex (because the negative log-likelihood term has a strictly positive-definite Hessian due to the constraints involving transverse tangent spaces), and therefore the optimum is unique. Applying the optimality conditions of the convex program (B.10) at the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T'^\perp} \in T'^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{\Omega^\perp} \in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \quad \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{T'^\perp} \in \lambda_n\partial\|\hat{L}_{T'}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T'$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_\Omega, \quad \mathcal{P}_{T'}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_{T'},$$

where $Z_\Omega \in \Omega, Z_{T'} \in T'$ and $\|Z_\Omega\|_\infty = \lambda_n\gamma, \|Z_{T'}\|_2 \leq 2\lambda_n$ (we use here the fact that projecting onto a tangent space T' increases the spectral norm by at most a factor of two). Denoting $Z = [Z_\Omega, Z_{T'}]$, we conclude that

$$\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z, \tag{B.11}$$

with $g_\gamma(Z) \leq 2\lambda_n$. Since the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$ is unique, one can check using Lagrangian duality theory [124] that $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique solution of the equation (B.11). Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}$ in terms of the errors (Δ_S, Δ_L), we have using the Taylor series of the matrix inverse that

$$\begin{aligned}
\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} &= \Sigma_O^n - [\mathcal{A}(\Delta_S, \Delta_L) + (\Sigma_O^*)^{-1}]^{-1} \\
&= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L) \\
&= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L) + \mathcal{I}^*\mathcal{C}_T
\end{aligned} \tag{B.12}$$

Since T' is a tangent space such that $\rho(T', T) \leq \frac{\xi(T)}{2}$, we have from Proposition 4.3.1 that the operator $\mathcal{B} = \left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}\mathcal{I}^{*}\mathcal{A}\mathcal{P}_{\mathcal{Y}}\right)^{-1}$ from $\mathcal{Y}$ to $\mathcal{Y}$ is bijective and is well-defined. Now consider the following matrix-valued function from $(\delta_S, \delta_L) \in \mathcal{Y}$ to $\mathcal{Y}$:

$$F(\delta_S, \delta_L) = (\delta_S, \delta_L) - \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^{*}\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^{*}\mathcal{C}_{T'}] - Z\right\}.$$

A point $(\delta_S, \delta_L) \in \mathcal{Y}$ is a fixed-point of F if and only if $\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^{*}\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^{*}\mathcal{C}_{T'}] = Z$. Applying equations (B.11) and (B.12) above, we then see that the only fixed-point of F by construction is the "true" error $\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)$ restricted to $\mathcal{Y}$. The reason for this is that, as discussed above, $(\hat{S}_{\Omega}, \hat{L}_{T'})$ is the unique optimum of (B.10) and therefore is the *unique solution* of (B.11). Next we show that this unique fixed-point of F lies in the ball $\mathbb{B}_r = \{(\delta_S, \delta_L) \mid g_{\gamma}(\delta_S, \delta_L) \leq r, \quad (\delta_S, \delta_L) \in \mathcal{Y}\}$.

In order to prove this step, we resort to Brouwer's fixed point theorem [113]. In particular we show that the function F maps the ball $\mathbb{B}_r$ onto itself. Since F is a continuous function and $\mathbb{B}_r$ is a compact set, we can conclude the proof of this proposition. Simplifying the function F, we have that

$$F(\delta_S, \delta_L) = \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) - \mathcal{I}^{*}\mathcal{C}_{T'}] + Z\right\}.$$

Consequently, we have from Proposition 4.3.1 that

$$\begin{aligned} g_{\gamma}(F(\delta_S, \delta_L)) &\leq \frac{2}{\alpha}\, g_{\gamma}\left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^{*}\mathcal{C}_{T'}] - Z\right) \\ &\leq \frac{4}{\alpha}\left\{g_{\gamma}(\mathcal{A}^{\dagger}[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^{*}\mathcal{C}_{T'}]) + \lambda_n\right\} \\ &\leq \frac{r}{2} + \frac{4}{\alpha}\, g_{\gamma}(\mathcal{A}^{\dagger}R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))), \end{aligned}$$

where in the second inequality we use the fact that $g_{\gamma}(\mathcal{P}_{\mathcal{Y}}(\cdot, \cdot)) \leq 2g_{\gamma}(\cdot, \cdot)$ and that $g_{\gamma}(Z) \leq 2\lambda_n$, and in the final inequality we use the assumption on r.

We now focus on the term $g_{\gamma}(\mathcal{A}^{\dagger}R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L)))$:

$$\begin{aligned} \frac{4}{\alpha}\, g_{\gamma}(\mathcal{A}^{\dagger}R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))) &\leq \frac{8D\psi C_1^2(g_{\gamma}(\delta_S, \delta_L) + \|\mathcal{C}_{T'}\|_2)^2}{\xi(T)\alpha} \\ &\leq \frac{32D\psi C_1^2 r^2}{\xi(T)\alpha} \\ &\leq \frac{32D\psi C_1^2 r}{\xi(T)\alpha}\,\frac{\alpha\xi(T)}{64D\psi C_1^2} \\ &\leq \frac{r}{2}, \end{aligned}$$

where we have used the fact that $r \leq \frac{\alpha\xi(T)}{64D\psi C_1^2}$. Hence $g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq r$ by Brouwer's fixed-point theorem. Finally we observe that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) + \|\mathcal{C}_{T'}\|_2 \\ &\leq 2r. \end{aligned}$$

□

■ B.4.3 Solving a variety-constrained problem

In order to prove that the solution $(\bar{S}_n, \bar{L}_n)$ of (B.9) has the same sparsity pattern/rank as (S^*, L^*), we will study an optimization problem that explicitly enforces these constraints. Specifically, we consider the following *non-convex* constraint set:

$$\begin{aligned} \mathcal{M} \;=\; \{(S,L) \mid\; & S \in \Omega(S^*),\ \mathrm{rank}(L) \leq \mathrm{rank}(L^*), \\ & \|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2},\ g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n\} \end{aligned}$$

Recall that $S^* = K^*_O$ and $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. The first constraint ensures that the tangent space at S is the same as the tangent space at S^*; therefore the support of S is contained in the support of S^*. The second and third constraints ensure that L lives in the appropriate low-rank variety, but has a tangent space "close" to the tangent space T. The final constraint roughly bounds the sum of the errors $(S - S^*) + (L^* - L)$; note that this does not necessarily bound the individual errors. Notice that the only non-convex constraint is that $\mathrm{rank}(L) \leq \mathrm{rank}(L^*)$. We then have the following nonlinear program:

$$\begin{aligned} (\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}) = \arg\min_{S,L}\ & \mathrm{Tr}[(S - L)\ \Sigma^n_O] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.}\quad & S - L \succ 0,\quad (S,L) \in \mathcal{M}. \end{aligned} \tag{B.13}$$

Under suitable conditions this nonlinear program is shown to have a unique solution. Each of the constraints in $\mathcal{M}$ is useful for proving the consistency of the solution of the convex program (B.9). We show that under suitable conditions the constraints in $\mathcal{M}$ are actually inactive at the optimal $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$, thus allowing us to conclude that the solution of (B.9) is also equal to $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$; hence the solution of (B.9) shares the consistency properties of $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$. A number of interesting properties can be derived simply by studying the constraint set $\mathcal{M}$.

Proposition B.4.3. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *For* γ *in the range specified by* (B.7) *and letting* $C_2 = \frac{48}{\alpha} + \frac{1}{\psi^2}$, *we have that* $g_\gamma(\Delta_S, \Delta_L) \leq C_2\lambda_n$.

Proof: We have by the triangle inequality that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\mathcal{P}_\Omega(\Delta_S), \mathcal{P}_T(\Delta_L))) &\leq 11\lambda_n + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\mathcal{P}_{\Omega^\perp}(\Delta_S), \mathcal{P}_{T^\perp}(\Delta_L))) \\ &\leq 11\lambda_n + m\psi^2\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq 12\lambda_n, \end{aligned}$$

as $m \leq \frac{D}{\xi(T)}$. Therefore, we have that $g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq 24\lambda_n$, where $\mathcal{Y} = \Omega \times T$. Consequently, we can apply Proposition 4.3.1 to conclude that

$$g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq \frac{48\lambda_n}{\alpha}.$$

Finally, we use the triangle inequality again to conclude that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}(\Delta_S, \Delta_L)) \\ &\leq \frac{48\lambda_n}{\alpha} + m\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq C_2\lambda_n. \end{aligned}$$

□

This simple result immediately leads to a number of useful corollaries. For example we have that under a suitable bound on the minimum nonzero singular value of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, the constraint in $\mathcal{M}$ along the normal direction $T^\perp$ is locally inactive. Next we list several useful consequences of Proposition B.4.3.

Corollary B.4.1. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *Suppose* γ *is in the range specified by* (B.7), *and let* $C_3 = \left(\frac{6(2-\nu)}{\nu} + 1\right) C_2^2\psi^2 D$ *and* $C_4 = C_2 + \frac{3\alpha C_2^2(2-\nu)}{16(3-\nu)}$ *(where* C_2 *is as defined in Proposition B.4.3). Let the minimum nonzero singular value* σ *of* $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ *be such that* $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ *for* $C_5 = \max\{C_3, C_4\}$, *and suppose that the smallest magnitude nonzero entry of* S^* *is greater than* $\frac{C_6\lambda_n}{\mu(\Omega)}$ *for* $C_6 = \frac{C_2\nu\alpha}{\beta(2-\nu)}$. *Setting* $T' = T(L)$ *and* $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$, *we then have that:*

1. *L has rank equal to* $\text{rank}(L^*)$, *i.e., L is a smooth point of the variety of matrices with rank less than or equal to* $\text{rank}(L^*)$. *In particular L has the same inertia as* L^*.

2. $\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \leq \frac{\xi(T)\lambda_n}{19D\psi^2}$.

3. $\rho(T,T') \leq \frac{\xi(T)}{4}$.

4. $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$.

5. $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

6. $\text{sign}(S) = \text{sign}(S^*)$.

Proof: We note the following facts before proving each step. First $C_2 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2} \geq \frac{\xi(T)}{D\psi^2}$. Second $\xi(T) \leq 1$. Third we have from Proposition B.4.3 that $\|\Delta_L\|_2 \leq C_2\lambda_n$. Finally $\frac{6(2-\nu)}{\nu} \geq 18$ for $\nu \in (0, \frac{1}{2}]$. We prove each step separately.

For the first step, we note that

$$\sigma \geq \frac{C_3\lambda_n}{\xi(T)^2} \geq \frac{19C_2^2\psi^2 D\lambda_n}{\xi(T)^2} \geq \frac{19C_2\lambda_n}{\xi(T)} \geq 8C_2\lambda_n \geq 8\|\Delta_L\|_2.$$

Hence L is a smooth point with rank equal to $\text{rank}(L^*)$, and specifically has the same inertia as L^*.

For the second step, we use the fact that $\sigma \geq 8\|\Delta_L\|_2$ to apply Proposition 4.2.2:

$$\|\mathcal{P}_{T^\perp}(\Delta_L)\| \leq \frac{\|\Delta_L\|_2^2}{\sigma} \leq \frac{C_2^2\xi(T)^2\lambda_n^2}{C_3\lambda_n} \leq \frac{\xi(T)\lambda_n}{19D\psi^2}.$$

For the third step we apply Proposition 4.2.1 (by using the conclusion from above that $\sigma \geq 8\|\Delta_L\|_2$) so that

$$\rho(T,T') \leq \frac{2\|\Delta_L\|_2}{\sigma} \leq \frac{2C_2\xi(T)^2}{C_3} \leq \frac{2\xi(T)^2}{19C_2D\psi^2} \leq \frac{\xi(T)}{4}.$$

For the fourth step let σ' denote the minimum singular value of L. Consequently,

$$\sigma' \geq \frac{C_3\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq C_2\lambda_n\left[\frac{19C_2D\psi^2}{\xi(T)^2} - 1\right] \geq 8\|\Delta_L\|_2.$$

Using the same reasoning as in the proof of the second step, we have that

$$\|\mathcal{C}_{T'}\|_2 \leq \frac{\|\Delta_L\|_2^2}{\sigma'} \leq \frac{C_2^2\lambda_n^2}{(\frac{C_3}{\xi(T)^2} - C_2)\lambda_n} \leq \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2D\psi^2(\frac{6(2-\nu)}{\nu})} \leq \frac{\nu\xi(T)\lambda_n}{6(2-\nu)D\psi^2}.$$

Hence

$$g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq m\psi^2\|\mathcal{C}_{T'}\|_2 \leq \frac{\lambda_n\nu}{6(2-\nu)}.$$

For the fifth step the bound on σ' implies that

$$\sigma' \geq \frac{C_4\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq \frac{3C_2^2\alpha(2-\nu)}{16(3-\nu)}\lambda_n$$

Since $\sigma' \geq 8\|\Delta_L\|_2$, we have from Proposition 4.2.2 and some algebra that

$$\|\mathcal{C}_{T'}\|_2 \leq \frac{C_2^2\lambda_n^2}{\sigma'} \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.$$

For the final step since $\|\Delta_S\|_\infty \leq \gamma C_2\lambda_n$, the assumed lower bound on the minimum magnitude nonzero entry of S^* guarantees that $\text{sign}(S) = \text{sign}(S^*)$. □

Notice that this corollary applies to *any* $(S, L) \in \mathcal{M}$, and is hence applicable to *any solution* $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ of the $\mathcal{M}$-constrained program (B.13). For now we choose an arbitrary solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ and proceed. In the next steps we show that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is *the unique* solution to the convex program (B.9), thus showing that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is also the unique solution to (B.13).

■ B.4.4 From variety constraint to tangent-space constraint

Given the solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$, we show that the solution to the convex program (B.10) with the tangent space constraint $L \in T_\mathcal{M} \triangleq T(\hat{L}_\mathcal{M})$ is the same as $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ under suitable conditions:

$$\begin{aligned}(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = \arg\min_{S,L}\ & \text{Tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.}\quad & S - L \succ 0,\quad S \in \Omega,\quad L \in T_\mathcal{M}.\end{aligned} \tag{B.14}$$

Assuming the bound of Corollary B.4.1 on the minimum singular value of L^* the uniqueness of the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is assured. This is because we have from Proposition 4.3.1 and from Corollary B.4.1 that $\mathcal{I}^*$ is injective on $\Omega \oplus T_\mathcal{M}$. Therefore the Hessian of the convex objective function of (B.14) is strictly positive-definite at $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$.

We let $\mathcal{C}_\mathcal{M} = \mathcal{P}_{T_\mathcal{M}^\perp}(L^*)$. Recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the sample covariance matrix and the marginal covariance matrix of the observed variables.

Proposition B.4.4. *Let γ be in the range specified by* (B.7). *Suppose that the minimum nonzero singular value σ of $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is such that $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ (C_5 is defined in Corollary B.4.1). Suppose also that the minimum magnitude nonzero entry*

of S^* is greater than or equal to $\frac{C_6\lambda_n}{\mu(\Omega)}$ (C_6 is defined in Corollary B.4.1). Let $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Further suppose that

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.$$

Then we have that

$$(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = (\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}).$$

Proof: Note first that the condition on the minimum singular value of L^* in Corollary B.4.1 is satisfied. Therefore we proceed with the following two steps:

1. First we can change the non-convex constraint $\text{rank}(L) \leq \text{rank}(L^*)$ to the linear constraint $L \in T(\hat{L}_\mathcal{M})$. This is because the lower bound assumed for σ implies that L is a smooth point of the algebraic variety of matrices with rank less than or equal to $\text{rank}(L^*)$ (from Corollary B.4.1). Due to the convexity of all the other constraints and the objective, the optimum of this "linearized" convex program will still be $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$.

2. Next we can again apply Corollary B.4.1 (based on the bound on σ) to conclude that the constraint $\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2}$ is *locally inactive* at the point $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$.

Consequently, we have that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ can be written as the solution of a *convex program*:

$$\begin{aligned}(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}) = \arg\min_{S,L}\ & \text{Tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.}\ \ & S - L \succ 0, \quad S \in \Omega, \quad L \in T_\mathcal{M}, \\ & g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n.\end{aligned} \tag{B.15}$$

We now need to argue that the constraint $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n$ is also inactive in the convex program (B.15). We proceed by showing that the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ of the convex program (B.14) has the property that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})) < 11\lambda_n$, which concludes the proof of this proposition. We have from Corollary B.4.1 that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Since $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ by assumption, one

can verify that

$$\begin{aligned}\frac{8}{\alpha}\left[\lambda_n + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})\right] &\leq \frac{8\lambda_n}{\alpha}\left[1+\frac{\nu}{3(2-\nu)}\right] \\ &\leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} \\ &\leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

The last line follows from the assumption on λ_n. We also note that $\|C_{T_\mathcal{M}}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary B.4.1, which implies that $\|C_{T_\mathcal{M}}\|_2 \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}$. Letting $(\Delta_S, \Delta_L) = (S_\Omega - S^*, L^* - L_{T_\mathcal{M}})$, we can conclude from Proposition B.4.2 that $g_\gamma(\Delta_L, \Delta_S) \leq \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}$. Next we apply Proposition B.4.1 (as $g_\gamma(\Delta_L, \Delta_S) \leq \frac{1}{2C_1}$) to conclude that

$$\begin{aligned}g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) &\leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)} \\ &\leq \frac{2D\psi C_1^2}{\xi(T)}\frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}\frac{\alpha\xi(T)}{32D\psi C_1^2} \\ &\leq \frac{2(3-\nu)\lambda_n}{3(2-\nu)}. \qquad \text{(B.16)}\end{aligned}$$

From the optimality conditions of (B.14) one can also check that,

$$\begin{aligned}g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) &\leq 2\lambda_n + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger E_n) \\ &\leq 2[\lambda_n + g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})] \\ &\leq 4\left[\frac{2(3-\nu)\lambda_n}{3(2-\nu)}\right].\end{aligned}$$

Here we used (B.16) in the last inequality, and also that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ (as noted above from Corollary B.4.1) and that $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Therefore,

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq \frac{16\lambda_n}{3}, \qquad \text{(B.17)}$$

because $\nu \in (0, \frac{1}{2}]$. Based on Proposition 4.3.1 (the second part), we also have that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq (1-\nu)\frac{16\lambda_n}{3} \leq \frac{16\lambda_n}{3}. \qquad \text{(B.18)}$$

Summarizing steps (B.17) and (B.18),

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\Delta_S,\Delta_L)) &\leq g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S,\Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S,\Delta_L)) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \\ &\leq \frac{16\lambda_n}{3} + \frac{16\lambda_n}{3} + \frac{\lambda\nu}{6(2-\nu)} \\ &\leq \frac{32\lambda}{3} + \frac{\lambda_n}{18} \\ &< 11\lambda_n. \end{aligned}$$

This concludes the proof of the proposition. □

This proposition has the following important consequence.

Corollary B.4.2. *Under the assumptions of Proposition B.4.4 we have that* $\mathrm{rank}(\hat{L}_{T_\mathcal{M}}) = \mathrm{rank}(L^*)$ *and that* $T(\hat{L}_{T_\mathcal{M}}) = T_\mathcal{M}$. *Moreover,* $\hat{L}_{T_\mathcal{M}}$ *actually has the same inertia as* L^*. *We also have that* $\mathrm{sign}(\hat{S}_\Omega) = \mathrm{sign}(S^*)$.

■ B.4.5 Removing the tangent-space constraints

The following lemma provides a simple set of sufficient conditions under which the optimal solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ of (B.14) satisfies the optimality conditions of the convex program (B.9) (without the tangent space constraints).

Lemma B.4.1. *Let* $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ *be the solution to the tangent-space constrained convex program* (B.14). *Suppose that the assumptions of Proposition B.4.4 hold. If in addition we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}\mathcal{A}(\Delta_S,\Delta_L)) \leq \frac{\lambda_n\nu}{6(2-\nu)},$$

then $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ *is also the unique optimum of the convex program* (B.9).

Proof: Recall from Corollary B.4.2 that the tangent space at $\hat{L}_{T_\mathcal{M}}$ is equal to $T(L^*)$. Applying the optimality conditions of the convex program (B.14) at the optimum $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T_\mathcal{M}^\perp} \in T_\mathcal{M}^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1} + Q_{\Omega^\perp} \in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \quad \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1} + Q_{T_\mathcal{M}^\perp} \in \lambda_n\partial\|\hat{L}_{T_\mathcal{M}}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T_\mathcal{M}$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1}] = -\lambda_n\gamma\mathrm{sign}(S^*), \quad \mathcal{P}_{T_\mathcal{M}}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1}] = \lambda_n UV^T,$$

where $\hat{L}_{T_{\mathcal{M}}} = UDV^T$ is a reduced SVD of $\hat{L}_{T_{\mathcal{M}}}$. Denoting $Z = [-\lambda_n \gamma \text{sign}(S^*), \lambda_n UV^T]$, we conclude that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = Z, \tag{B.19}$$

with $g_\gamma(Z) = \lambda_n$. It is clear that the optimality condition of the convex program (B.9) (without the tangent-space constraints) on $\mathcal{Y}$ is satisfied. All we need to show is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}]) < \lambda_n. \tag{B.20}$$

Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}$ in terms of the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})$, we have that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} = E_n - R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) + \mathcal{I}^* \mathcal{A}(\Delta_S, \Delta_L).$$

Restating the condition (B.19) on $\mathcal{Y}$, we have that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L) = Z + \mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]. \tag{B.21}$$

(Recall that $\mathcal{C}_{T_{\mathcal{M}}} = \mathcal{P}_{T_{\mathcal{M}}^\perp}(L^*)$.) A sufficient condition to show (B.20) and complete the proof of this lemma is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) < \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]).$$

We prove this inequality next. Recall from Corollary B.4.1 that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$. Therefore, from equation (B.21) we can conclude that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) &\leq \lambda_n + 2(g_\gamma(\mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}])) \\ &\leq \lambda_n + 2 \left[\frac{3\lambda_n \nu}{6(2-\nu)} \right] \\ &\leq \frac{2\lambda_n}{2-\nu}. \end{aligned}$$

Here we used the bounds assumed on $g_\gamma(\mathcal{A}^\dagger E_n)$ and on $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L))$.

Applying the second part of Proposition 4.3.1, we have that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) &\leq \frac{2\lambda_n(1-\nu)}{2-\nu} \\ &\leq \lambda_n - \frac{\nu \lambda_n}{2-\nu} \\ &< \lambda_n - \frac{\nu \lambda_n}{2(2-\nu)} \\ &\leq \lambda_n - g_\gamma(\mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]) \\ &\leq \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*} \mathcal{A}(\Delta_S, \Delta_L) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]). \end{aligned}$$

This concludes the proof of the lemma. □

One can check that as $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is also the *unique* solution to the convex program (B.9) without the tangent-space constraints.

■ B.4.6 Probabilistic analysis

All the analysis described so far in this section has been completely deterministic in nature. Here we present the probabilistic component of our proof. Specifically, we study the rate at which the sample covariance matrix converges to the true covariance matrix. The following result from [41] plays a key role in our analysis:

Theorem B.4.1. *Given natural numbers n, p with $p \leq n$, let Γ be a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $s_1(\Gamma)$ and $s_p(\Gamma)$ of Γ are such that*

$$\max\left\{\Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\frac{p}{n}} + t\right], \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t\right]\right\} \leq \exp\left\{-\frac{nt^2}{2}\right\},$$

for any $t > 0$.

Using this result the next lemma provides a probabilistic bound between the sample covariance Σ_O^n formed using n samples and the true covariance Σ_O^* in spectral norm. This result is well-known, and we mainly discuss it here for completeness and also to show explicitly the dependence on $\psi = \|\Sigma_O^*\|_2$ (4.8).

Lemma B.4.2. *Let $\psi = \|\Sigma_O^*\|_2$. Given any $\delta > 0$ with $\delta \leq 8\psi$, let the number of samples n be such that $n \geq \frac{64p\psi^2}{\delta^2}$. Then we have that*

$$\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] \leq 2\exp\left\{-\frac{n\delta^2}{128\psi^2}\right\}.$$

Proof: Since the spectral norm is unitarily invariant, we can assume that Σ_O^* is diagonal without loss of generality. Let $\bar{\Sigma}^n = (\Sigma_O^*)^{-\frac{1}{2}}\Sigma_O^n(\Sigma_O^*)^{-\frac{1}{2}}$, and let $s_1(\bar{\Sigma}^n), s_p(\bar{\Sigma}^n)$ denote the largest/smallest singular values of $\bar{\Sigma}^n$. Note that $\bar{\Sigma}^n$ can be viewed as the sample covariance matrix formed from n independent samples drawn from a model with identity covariance, i.e., $\bar{\Sigma}^n = \Gamma\Gamma^T$ where Γ denotes a $p \times n$ matrix with i.i.d. Gaussian

entries that have zero-mean and variance $\frac{1}{n}$. We then have that

$$\begin{aligned}
\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] &\leq \Pr\left[\|\bar{\Sigma}^n - I\|_2 \geq \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\bar{\Sigma}^n) \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\bar{\Sigma}^n) \leq 1 - \tfrac{\delta}{\psi}\right] \\
&= \Pr\left[s_1(\Gamma)^2 \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\Gamma)^2 \leq 1 - \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \tfrac{\delta}{4\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \tfrac{\delta}{4\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + \tfrac{\delta}{8\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - \tfrac{\delta}{8\psi}\right] \\
&\leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.
\end{aligned}$$

Here we used the fact that $n \geq \frac{64p\psi^2}{\delta^2}$ in the fourth inequality, and we applied Theorem B.4.1 to obtain the final inequality by setting $t = \frac{\delta}{8\psi}$. □

The following corollary describes relates the number of samples required for an error bound to hold with probability $1 - 2\exp\{-p\}$.

Corollary B.4.3. *Let Σ_O^n be the sample covariance formed from n samples of the observed variables. Set $\delta_n = \sqrt{\frac{128p\psi^2}{n}}$. If $n \geq 2p$, then we have with probability greater than $1 - 2\exp\{-p\}$ that*

$$\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \leq \delta_n\right] \geq 1 - 2\exp\{-p\}.$$

Proof: We note that $n \geq 2p$ implies that $\delta_n \leq 8\psi$, and apply Lemma B.4.2. □

■ B.4.7 Putting it all together

In this section we tie together the results obtained thus far to conclude the proof of Theorem 4.4.1. We only need to show that the sufficient conditions of Lemma B.4.1 are satisfied. It follows directly from Corollary B.4.2 that the low-rank part $\hat{L}_{T_{\mathcal{M}}}$ is positive semidefinite, which implies that $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ is also the solution to the original regularized maximum-likelihood convex program (4.9) with the positive-semidefinite constraint. As usual set $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})$, and set $E_n = \Sigma_O^n - \Sigma_O^*$.

Assumptions: We specify here the constants that were suppressed in the statement of Theorem 4.4.1:

1. Let $C_7 = \frac{\alpha\nu}{32(3-\nu)D} \min\left\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\right\}$, and let the number of samples n be such that
$$n \geq \frac{p}{\xi(T)^4} \max\left\{\frac{128\psi^2}{C_7^2}, 2\right\}.$$

Note that $n \gtrsim \frac{p}{\xi(T)^4}$.

2. Set $\delta_n = \sqrt{\frac{128p\psi^2}{n}}$, and then set λ_n as follows:

$$\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}.$$

Note that $\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.

3. Let the minimum nonzero singular value σ of L^* be such that

$$\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2},$$

where C_5 is defined in Corollary B.4.1. Note that $\sigma \gtrsim \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}$.

4. Let the minimum magnitude nonzero entry θ of S^* be such that

$$\theta \geq \frac{C_6\lambda_n}{\mu(\Omega)},$$

where C_6 is defined in Corollary B.4.1. Note that $\theta \gtrsim \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}$.

Proof of Theorem 4.4.1: We condition on the event that $\|E_n\|_2 \leq \delta_n$, which holds with probability greater than $1 - 2\exp\{-p\}$ from Corollary B.4.3 as $n \geq 2p$ by assumption. We note that based on the bound on n, we also have that

$$\delta_n \leq \xi(T)^2\left[\frac{\alpha\nu}{32(3-\nu)D}\min\left\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\right\}\right].$$

In particular, these bounds imply that

$$\delta_n \leq \frac{\alpha\xi(T)\nu}{32(3-\nu)D}\min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\} \tag{B.22}$$

and that

$$\delta_n \leq \frac{\alpha^2\xi(T)^2\nu^2}{8192\psi C_1^2(3-\nu)^2D^2}. \tag{B.23}$$

Both these weaker bounds are used later.

Based on the assumptions above, the requirements of Lemma B.4.1 on the minimum nonzero singular value of L^* and the minimum magnitude nonzero entry of S^* are satisfied. We only need to verify the bounds on λ_n and $g_\gamma(\mathcal{A}^\dagger E_n)$ from Proposition B.4.4, and the bound on $g_\gamma(\mathcal{A}^\dagger R\mathcal{A}(\Delta_S, \Delta_L))$ from Lemma B.4.1.

First we verify the bound on λ_n. Based on the setting of λ_n above and bound on δ_n from (B.22), we have that

$$\begin{aligned}\lambda_n &= \frac{6D(2-\nu)\delta_n}{\xi(T)\nu} \\ &\leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

Next we combine the facts that $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$, and that $\|E_n\|_2 \leq \delta_n$ to conclude that

$$g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{D\delta_n}{\xi(T)} \leq \frac{\lambda_n \nu}{6(2-\nu)}.$$

Finally we provide a bound on the remainder by applying Propositions B.4.2 and B.4.1, which would satisfy the last remaining condition of Lemma B.4.1. In order to apply Proposition B.4.2, we note that

$$\begin{aligned}\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) + \lambda_n\right] &\leq \frac{8}{\alpha}\left[\frac{\nu}{3(2-\nu)} + 1\right]\lambda_n \\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} \\ &= \frac{32(3-\nu)D}{\alpha\xi(T)\nu}\delta_n \qquad \text{(B.24)} \\ &\leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

In the first inequality we used the fact that $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ (from above) and that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})$ is similarly bounded (from Corollary B.4.1 due to the bound on σ). In the second equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. In the final inequality we used the bound on δ_n from (B.22). This satisfies one of the requirements of Proposition B.4.2. The other condition on $\|\mathcal{C}_{T_\mathcal{M}}\|_2$ is also similarly satisfied due to the bound on σ from Corollary B.4.1. Specifically, we have that $\|\mathcal{C}_{T_\mathcal{M}}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary B.4.1, and use the same sequence of inequalities as above to satisfy the second requirement of Proposition B.4.2. Thus we conclude from Proposition B.4.2 and from (B.24) that

$$g_\gamma(\Delta_S, \Delta_L) \leq \frac{64(3-\nu)D}{\alpha\xi(T)\nu}\delta_n. \qquad \text{(B.25)}$$

This bound implies that $g_\gamma(\Delta_S, \Delta_L) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$, which proves the parametric consistency part of the theorem.

Since the bound (B.25) also satisfies the condition of Proposition B.4.1 (from the inequality following (B.24) above we see that $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$), we have that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger R(\Delta_S + \Delta_L)) &\leq \frac{2D\psi C_1^2}{\xi(T)} g_\gamma(\Delta_S, \Delta_L)^2 \\ &\leq \frac{2D\psi C_1^2}{\xi(T)} \left(\frac{64(3-\nu)D}{\alpha \xi(T)\nu} \right)^2 \delta_n^2 \\ &= \left[\frac{8192\psi C_1^2 (3-\nu)^2 D^2}{\alpha^2 \xi(T)^2 \nu^2} \delta_n \right] \frac{D\delta_n}{\xi(T)} \\ &\leq \frac{D\delta_n}{\xi(T)} \\ &= \frac{\lambda_n \nu}{6(2-\nu)}. \end{aligned}$$

In the final inequality we used the bound (B.23) on δ_n, and in the final equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. This concludes the algebraic consistency part of the theorem. □

Appendix C

Proofs of Chapter 5

C.1 Proof of Proposition 5.3.1

Proof. First note that the Gaussian width can be upper-bounded as follows:

$$w(\mathcal{C} \cap \mathbb{S}^{p-1}) \leq \mathbb{E}_{\mathbf{g}} \left[\sup_{\mathbf{z} \in \mathcal{C} \cap \mathcal{B}(0,1)} \mathbf{g}^T \mathbf{z} \right], \tag{C.1}$$

where $\mathcal{B}(0,1)$ denotes the unit Euclidean ball. The expression on the right hand side inside the expected value can be expressed as the optimal value of the following convex optimization problem for each $\mathbf{g} \in \mathbb{R}^p$:

$$\begin{array}{ll} \max_{\mathbf{z}} & \mathbf{g}^T \mathbf{z} \\ \text{s.t.} & \mathbf{z} \in \mathcal{C} \\ & \|\mathbf{z}\|^2 \leq 1 \end{array} \tag{C.2}$$

We now proceed to form the dual problem of (C.2) by first introducing the Lagrangian

$$\mathcal{L}(\mathbf{z}, \mathbf{u}, \gamma) = \mathbf{g}^T \mathbf{z} + \gamma(1 - \mathbf{z}^T \mathbf{z}) - \mathbf{u}^T \mathbf{z}$$

where $\mathbf{u} \in \mathcal{C}^*$ and $\gamma \geq 0$ is a scalar. To obtain the dual problem we maximize the Lagrangian with respect to $\mathbf{z}$, which amounts to setting

$$\mathbf{z} = \frac{1}{2\gamma}(\mathbf{g} - \mathbf{u}).$$

Plugging this into the Lagrangian above gives the dual problem

$$\begin{array}{ll} \min & \gamma + \frac{1}{4\gamma}\|\mathbf{g} - \mathbf{u}\|^2 \\ \text{s.t.} & \mathbf{u} \in \mathcal{C}^* \\ & \gamma \geq 0. \end{array}$$

Solving this optimization with respect to γ we find that $\gamma = \frac{1}{2}\|\mathbf{g} - \mathbf{u}\|$, which gives the dual problem to (C.2)

$$\begin{aligned} \min \quad & \|\mathbf{g} - \mathbf{u}\| \\ \text{s.t.} \quad & \mathbf{u} \in \mathcal{C}^* \end{aligned} \tag{C.3}$$

Under very mild assumptions about $\mathcal{C}$, the optimal value of (C.3) is equal to that of (C.2) (for example as long as $\mathcal{C}$ has a non-empty relative interior, strong duality holds). Hence we have derived

$$\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{z}\in\mathcal{C}\cap\mathcal{B}(0,1)} \mathbf{g}^T\mathbf{z}\right] = \mathbb{E}_{\mathbf{g}}\left[\operatorname{dist}(\mathbf{g},\mathcal{C}^*)\right]. \tag{C.4}$$

This equation combined with the bound (C.1) gives us the desired result. □

■ C.2 Proof of Theorem 5.3.3

Proof. We set $\beta = \frac{1}{\Theta}$. First note that if $\beta \geq \exp\{\frac{p}{9}\}$ then the width bound exceeds $\sqrt{p}$, which is the maximal possible value for the width of $\mathcal{C}$. Thus, we will assume throughout that $\beta \leq \exp\{\frac{p}{36}\}$.

Using Proposition 5.3.1 we need to upper bound the expected distance to the polar cone. Let $\mathbf{g} \sim \mathcal{N}(0, I)$ be a normally distributed random vector. Then the norm of $\mathbf{g}$ is independent from the angle of $\mathbf{g}$. That is, $\|\mathbf{g}\|$ is independent from $\mathbf{g}/\|\mathbf{g}\|$. Moreover, $\mathbf{g}/\|\mathbf{g}\|$ is distributed as a uniform sample on $\mathbb{S}^{p-1}$, and $\mathbb{E}_{\mathbf{g}}[\|\mathbf{g}\|] \leq \sqrt{p}$. Thus we have

$$\mathbb{E}_{\mathbf{g}}[\operatorname{dist}(\mathbf{g},\mathcal{C}^*)] \leq \mathbb{E}_{\mathbf{g}}[\|\mathbf{g}\| \cdot \operatorname{dist}(\mathbf{g}/\|\mathbf{g}\|, \mathcal{C}^* \cap \mathbb{S}^{p-1})] \leq \sqrt{p}\,\mathbb{E}_{\mathbf{u}}[\operatorname{dist}(\mathbf{u}, \mathcal{C}^* \cap \mathbb{S}^{p-1})] \tag{C.5}$$

where $\mathbf{u}$ is sampled uniformly on $\mathbb{S}^{p-1}$.

To bound the latter quantity, we will use isoperimetry. Suppose A is a subset of $\mathbb{S}^{p-1}$ and B is a spherical cap with the same volume as A. Let $N(A, r)$ denote the locus of all points in the sphere of Euclidean distance at most r from the set A. Let μ denote the Haar measure on $\mathbb{S}^{p-1}$ and $\mu(A; r)$ denote the measure of $N(A, r)$. Then spherical isoperimetry states that $\mu(A; r) \geq \mu(B; r)$ for all $r \geq 0$ (see, for example [95, 106]).

Let B now denote a spherical cap with $\mu(B) = \mu(\mathcal{C}^* \cap \mathbb{S}^{p-1})$. Then we have

$$\mathbb{E}_{\mathbf{u}}[\operatorname{dist}(\mathbf{u}, \mathcal{C}^* \cap \mathbb{S}^{p-1})] = \int_0^\infty \mathbb{P}[\operatorname{dist}(\mathbf{u}, \mathcal{C}^* \cap \mathbb{S}^{p-1}) > t]dt \tag{C.6}$$

$$= \int_0^\infty (1 - \mu(\mathcal{C}^* \cap \mathbb{S}^{p-1}; t))dt \tag{C.7}$$

$$\leq \int_0^\infty (1 - \mu(B; t))dt \tag{C.8}$$

where the first equality is the integral form of the expected value and the last inequality follows by isoperimetry. Hence we can bound the expected distance to the polar cone intersecting the sphere using only knowledge of the volume of spherical caps on $\mathbb{S}^{p-1}$.

To proceed let $v(\varphi)$ denote the volume of a spherical cap subtending a solid angle φ. An explicit formula for $v(\varphi)$ is

$$v(\varphi) = z_p^{-1} \int_0^{\varphi} \sin^{p-1}(\vartheta) d\vartheta \tag{C.9}$$

where $z_p = \int_0^{\pi} \sin^{p-1}(\vartheta) d\vartheta$ [88]. Let $\varphi(\beta)$ denote the minimal solid angle of a cap such that β copies of that cap cover $\mathbb{S}^{p-1}$. Since the geodesic distance on the sphere is always greater than or equal to Euclidean distance, if K is a spherical cap subtending ψ radians, $\mu(K;t) \geq v(\psi + t)$. Therefore

$$\int_0^{\infty} (1 - \mu(B;t)) dt \leq \int_0^{\infty} (1 - v(\varphi(\beta) + t)) dt \,. \tag{C.10}$$

We can proceed to simplify the right-hand-side integral:

$$\int_0^{\infty} (1 - v(\varphi(\beta) + t)) dt = \int_0^{\pi - \varphi(\beta)} (1 - v(\varphi(\beta) + t)) dt \tag{C.11}$$

$$= \pi - \varphi(\beta) - \int_0^{\pi - \varphi(\beta)} v(\varphi(\beta) + t) dt \tag{C.12}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^{\pi - \varphi(\beta)} \int_0^{\varphi(\beta) + t} \sin^{p-1} \vartheta d\alpha dt \tag{C.13}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^{\pi} \int_{\max(\vartheta - \varphi(\beta), 0)}^{\pi - \varphi(\beta)} \sin^{p-1} \vartheta dt d\alpha \tag{C.14}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^{\pi} \{\pi - \varphi(\beta) - \max(\vartheta - \varphi(\beta), 0)\} \sin^{p-1} \vartheta d\alpha \tag{C.15}$$

$$= z_p^{-1} \int_0^{\pi} \max(\vartheta - \varphi(\beta), 0) \sin^{p-1} \vartheta d\alpha \tag{C.16}$$

$$= z_p^{-1} \int_{\varphi(\beta)}^{\pi} (\vartheta - \varphi(\beta)) \sin^{p-1} \vartheta d\alpha \tag{C.17}$$

(C.14) follows by switching the order of integration and the rest of these equalities follow by straight-forward integration and some algebra.

Using the inequalities that $z_p \geq \frac{2}{\sqrt{p-1}}$ (see [95]) and $\sin(x) \leq \exp(-(x - \pi/2)^2/2)$

for $x \in [0, \pi]$, we can bound the last integral as

$$z_p^{-1} \int_{\varphi(\beta)}^{\pi} (\vartheta - \varphi(\beta)) \sin^{p-1} \vartheta d\alpha \leq \frac{\sqrt{p-1}}{2} \int_{\varphi(\beta)}^{\pi} (\vartheta - \varphi(\beta)) \exp\left(-\frac{p-1}{2}(\vartheta - \tfrac{\pi}{2})^2\right) d\vartheta \tag{C.18}$$

Performing the change of variables $a = \sqrt{p-1}(\vartheta - \frac{\pi}{2})$, we are left with the integral

$$\frac{1}{2} \int_{\sqrt{p-1}(\varphi(\beta)-\pi/2)}^{\sqrt{p-1}\pi/2} \left\{ \frac{a}{\sqrt{p-1}} + \left(\frac{\pi}{2} - \varphi(\beta)\right) \right\} \exp\left(-\frac{a^2}{2}\right) da \tag{C.19}$$

$$= -\frac{1}{2\sqrt{p-1}} \exp\left(-\frac{a^2}{2}\right) \Big|_{\sqrt{p-1}(\varphi(\beta)-\pi/2)}^{\sqrt{p-1}\pi/2} + \frac{\frac{\pi}{2} - \varphi(\beta)}{2} \int_{\sqrt{p-1}(\varphi(\beta)-\pi/2)}^{\sqrt{p-1}\pi/2} \exp\left(-\frac{a^2}{2}\right) da \tag{C.20}$$

$$\leq \frac{1}{2\sqrt{p-1}} \exp\left(-\frac{p-1}{2}(\pi/2 - \varphi(\beta))^2\right) + \sqrt{\frac{\pi}{2}}\left(\frac{\pi}{2} - \varphi(\beta)\right) \tag{C.21}$$

In this final bound, we bounded the first term by dropping the upper integrand, and for the second term we used the fact that

$$\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}\,. \tag{C.22}$$

We are now left with the task of computing a lower bound for $\varphi(\beta)$. We need to first reparameterize the problem. Let K be a spherical cap. Without loss of generality, we may assume that

$$K = \{x \in \mathbb{S}^{p-1} \,:\, x_1 \geq h\} \tag{C.23}$$

for some $h \in [0, 1]$. h is the *height* of the cap over the equator. Via elementary trigonometry, the solid angle that K subtends is given by $\pi/2 - \sin^{-1}(h)$. Hence, if $h(\beta)$ is the largest number such that β caps of height h cover $\mathbb{S}^{p-1}$, then $h(\beta) = \sin(\pi/2 - \phi(\beta))$.

The quantity $h(\beta)$ may be estimated using the following estimate from [25]. For $h \in [0, 1]$, let $\gamma(p, h)$ denote the volume of a spherical cap of $\mathbb{S}^{p-1}$ of height h.

Lemma C.2.1 ([25]). *For $1 \geq h \geq \frac{2}{\sqrt{p}}$,*

$$\frac{1}{10h\sqrt{p}}(1-h^2)^{\frac{p-1}{2}} \leq \gamma(p,h) \leq \frac{1}{2h\sqrt{p}}(1-h^2)^{\frac{p-1}{2}}\,. \tag{C.24}$$

Note that for $h \geq \frac{2}{\sqrt{p}}$,

$$\frac{1}{2h\sqrt{p}}(1-h^2)^{\frac{p-1}{2}} \leq \frac{1}{4}(1-h^2)^{\frac{p-1}{2}} \leq \frac{1}{4}\exp(-\tfrac{p-1}{2}h^2)\,. \tag{C.25}$$

So if

$$h = \sqrt{\frac{2\log(4\beta)}{p-1}} \tag{C.26}$$

then $h \leq 1$ because we have assumed $\beta \leq \frac{1}{4}\exp(\frac{4(p-1)}{\pi^3})$. Moreover, $h \geq \frac{2}{\sqrt{p}}$ and the volume of the cap with height h is less than or equal to $1/\beta$. That is

$$\varphi(\beta) \geq \pi/2 - \sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right). \tag{C.27}$$

Combining the estimate (C.21) with Proposition 5.3.1, and using our estimate for $\varphi(\beta)$, we get the bound

$$w(\mathcal{C}) \leq \frac{1}{2}\sqrt{\frac{p}{p-1}}\exp\left(-\tfrac{p-1}{2}\sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right)^2\right) + \sqrt{\frac{\pi p}{2}}\sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right) \tag{C.28}$$

This expression can be simplified by using the following bounds. First, $\sin^{-1}(x) \geq x$ lets us upper bound the first term by $\sqrt{\frac{p}{p-1}}\frac{1}{8\beta}$. For the second term, using the inequality $\sin^{-1}(x) \leq \frac{\pi}{2}x$ results in the upper bound

$$w(\mathcal{C}) \leq \sqrt{\frac{p}{p-1}}\left(\frac{1}{8\beta} + \frac{\pi^{3/2}}{2}\sqrt{\log(4\beta)}\right). \tag{C.29}$$

For $p \geq 9$ the upper bound can be expressed simply as $w(\mathcal{C}) \leq 3\sqrt{\log(4\beta)}$. We recall that $\beta = \frac{1}{\Theta}$, which completes the proof of the theorem. □

■ C.3 Direct Width Calculations

We first give the proof of Proposition 5.3.2.

Proof. Let $\mathbf{x}^\star$ be an s-sparse vector in $\mathbb{R}^p$ with ℓ_1 norm equal to 1, and let $\mathcal{A}$ denote the set of unit-Euclidean-norm one-sparse vectors. Let Δ denote the set of coordinates where $\mathbf{x}^\star$ is non-zero. Recall from Chapter 2 that the normal cone at $\mathbf{x}^\star$ with respect to the ℓ_1 ball is given by

$$N_{\mathcal{A}}(\mathbf{x}^\star) = \operatorname{cone}\{\mathbf{z} \in \mathbb{R}^p \ : \ \mathbf{z}_i = \operatorname{sgn}(\mathbf{x}_i^\star) \text{ for } i \in \Delta, \ |\mathbf{z}_i| \leq 1 \text{ for } i \in \Delta^c\} \tag{C.30}$$

$$= \{\mathbf{z} \in \mathbb{R}^p \ : \ \mathbf{z}_i = t\operatorname{sgn}(\mathbf{x}_i^\star) \text{ for } i \in \Delta, \ |\mathbf{z}_i| \leq t \text{ for } i \in \Delta^c \text{ for some } t > 0\}. \tag{C.31}$$

Here Δ^c represents the zero entries of $\mathbf{x}^\star$.

Given $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p)$, we would like to construct a $\mathbf{u} \in N_{\mathcal{A}}(\mathbf{x}^\star)$ that is close to $\mathbf{g}$. Pick $\mathbf{u}(\mathbf{g})$ as

$$\mathbf{u}_i(\mathbf{g}) = \begin{cases} \mathbf{g}_i & i \in \Delta^c \\ \|\mathbf{g}_{\Delta^c}\|_\infty \mathrm{sgn}(\mathbf{x}_i^\star) & i \in \Delta \end{cases} \tag{C.32}$$

That is, we set $\mathbf{u}(\mathbf{g})$ equal to $\mathbf{g}$ on Δ^c. On Δ, we set $\mathbf{u}(\mathbf{g})$ proportional to the sign of $\mathbf{x}^\star$, and scale this sign vector appropriately by the ℓ_∞ norm of $\mathbf{g}$ on Δ^c. For this choice, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{u}(\mathbf{g}) - \mathbf{g}\|^2] &= \mathbb{E}[\|\mathbf{u}_\Delta(\mathbf{g}) - \mathbf{g}_\Delta\|^2] && \text{(C.33)} \\ &= \mathbb{E}[\|\mathbf{u}_\Delta(\mathbf{g})\|^2] + \mathbb{E}[\|\mathbf{g}_\Delta\|^2] && \text{(C.34)} \\ &= s\mathbb{E}[\|\mathbf{g}_{\Delta^c}\|_\infty^2] + s && \text{(C.35)} \\ &\leq 2s\log(p-s) + 2s && \text{(C.36)} \end{aligned}$$

Here, the second equality holds because $\mathbf{g}_{\Delta^c}$ and $\mathbf{g}_\Delta$ are independent. The final inequality follows because the maximum squared magnitude of a sequence of $p-s$ normal random variables is bounded above by $2\log(p-s)+1$. By Corollary 5.3.1, this means that the ℓ_1 heuristic succeeds when n exceeds $2p(\log(p-s)+1)$.

For small values of s, we can tighten this result. The minimum squared distance to the normal cone at $\mathbf{x}^\star$ can be formulated as a one-dimensional convex optimization problem for arbitrary $\mathbf{z} \in \mathbb{R}^p$

$$\begin{aligned} \inf_{\mathbf{u} \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{z} - \mathbf{u}\|_2^2 &= \inf_{\substack{t \geq 0 \\ |\mathbf{u}_i| < t,\ i \in \Delta^c}} \sum_{i \in \Delta} (\mathbf{z}_i - t\mathrm{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j \in \Delta^c} (\mathbf{z}_j - \mathbf{u}_j)^2 && \text{(C.37)} \\ &= \inf_{t \geq 0} \sum_{i \in \Delta} (\mathbf{z}_i - t\mathrm{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j \in \Delta^c} \mathrm{shrink}(\mathbf{z}_j, t)^2 && \text{(C.38)} \end{aligned}$$

where

$$\mathrm{shrink}(z,t) = \begin{cases} z + t & z < -t \\ 0 & -t \leq z \leq t \\ z - t & z > t \end{cases} \tag{C.39}$$

is the ℓ_1-shrinkage function. Hence, for any fixed $t \geq 0$ independent of $\mathbf{g}$, we have

$$\mathbb{E}\left[\inf_{\mathbf{u}\in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g}-\mathbf{u}\|_2^2\right] \leq \mathbb{E}\left[\sum_{i\in\Delta}(\mathbf{g}_i - t\mathrm{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j\in\Delta^c}\mathrm{shrink}(\mathbf{g}_j,t)^2\right] \tag{C.40}$$

$$= s(1+t^2) + \mathbb{E}\left[\sum_{j\in\Delta^c}\mathrm{shrink}(\mathbf{g}_j,t)^2\right]. \tag{C.41}$$

Now we directly integrate the second term, treating each summand individually. For a zero-mean, unit-variance normal random variable g,

$$\mathbb{E}\left[\mathrm{shrink}(g,t)^2\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-t}(g+t)^2\exp(-g^2/2)dg + \frac{1}{\sqrt{2\pi}}\int_t^{\infty}(g-t)^2\exp(-g^2/2)dg \tag{C.42}$$

$$= \frac{2}{\sqrt{2\pi}}\int_t^{\infty}(g-t)^2\exp(-g^2/2)dg \tag{C.43}$$

$$= -\frac{2}{\sqrt{2\pi}}t\exp(-t^2/2) + \frac{2(1+t^2)}{\sqrt{2\pi}}\int_t^{\infty}\exp(-g^2/2)dg \tag{C.44}$$

$$\leq \frac{2}{\sqrt{2\pi}}\left(-t+\frac{1+t^2}{t}\right)\exp(-t^2/2) \tag{C.45}$$

$$= \frac{2}{\sqrt{2\pi}}\frac{1}{t}\exp(-t^2/2). \tag{C.46}$$

The first simplification follows because the shrink function and Gaussian distributions are symmetric about the origin. The second equality follows by integrating by parts. The inequality follows by a tight bound on the Gaussian Q-function

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty}\exp(-g^2/2)dg \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp(-x^2/2) \quad \text{for } x>0. \tag{C.47}$$

Using this bound, we get

$$\mathbb{E}\left[\inf_{\mathbf{u}\in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g}-\mathbf{u}\|_2^2\right] \leq s(1+t^2) + (p-s)\frac{2}{\sqrt{2\pi}}\frac{1}{t}\exp(-t^2/2) \tag{C.48}$$

Setting $t = \sqrt{2\log(p/s-1)-1}$ gives

$$\mathbb{E}\left[\inf_{\mathbf{z}\in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g}-\mathbf{z}\|_2^2\right] \leq 2s\left(\log\left(\frac{p-s}{s}\right)+1\right). \tag{C.49}$$

provided that $s \leq \frac{1}{1+e}p$. This bound on s arises because t must be greater than or equal to 0 and the second term in (C.48) is set to be less than $2s$.

□

Next we give the proof of Proposition 5.3.3.

Proof. Let $\mathbf{x}^\star$ be an $m_1 \times m_2$ matrix of rank r with singular value decomposition $U\Sigma V^*$, and let $\mathcal{A}$ denote the set of rank-one unit-Euclidean-norm matrices of size $m_1 \times m_2$. Without loss of generality, impose the conventions $m_1 \leq m_2$, Σ is $r \times r$, U is $m_1 \times r$, V is $m_2 \times r$, and assume the nuclear norm of $\mathbf{x}^\star$ is equal to 1.

Let $\mathbf{u}_k$ (respectively $\mathbf{v}_k$) denote the k'th column of U (respectively V). It is convenient to introduce the orthogonal decomposition $\mathbb{R}^{m_1 \times m_2} = \Delta \oplus \Delta^\perp$ where Δ is the linear space spanned by elements of the form $\mathbf{u}_k \mathbf{z}^T$ and $\mathbf{y}\mathbf{v}_k^T$, $1 \leq k \leq r$, where $\mathbf{z}$ and $\mathbf{y}$ are arbitrary, and $\Delta^\perp$ is the orthogonal complement of Δ. The space $\Delta^\perp$ is the subspace of matrices spanned by the family $(\mathbf{y}\mathbf{z}^T)$, where $\mathbf{y}$ (respectively $\mathbf{z}$) is any vector orthogonal to all the columns of U (respectively V). Recall from Chapter 2 that the normal cone of the nuclear norm ball at $\mathbf{x}^\star$ is given by the cone generated by the subdifferential at $\mathbf{x}^\star$:

$$N_{\mathcal{A}}(\mathbf{x}^\star) = \operatorname{cone}\left\{UV^T + W \in \mathbb{R}^{m_1 \times m_2} \ : \ W^T U = 0, \ \ WV = 0, \ \ \|W\|_{\mathcal{A}}^* \leq 1\right\} \tag{C.50}$$

$$= \left\{tUV^* + W \in \mathbb{R}^{m_1 \times m_2} \ : \ W^T U = 0, \ \ WV = 0, \ \ \|W\|_{\mathcal{A}}^* \leq t, \ \ t \geq 0\right\}. \tag{C.51}$$

Note that here $\|Z\|_{\mathcal{A}}^*$ is the operator norm, equal to the maximum singular value of Z [121].

Let G be a Gaussian random matrix with i.i.d. entries, each with mean zero and unit variance. Then the matrix

$$Z(G) = \|\mathcal{P}_{\Delta^\perp}(G)\| UV^* + \mathcal{P}_{\Delta^\perp}(G) \tag{C.52}$$

is in the normal cone at $\mathbf{x}^\star$. We can then compute

$$\mathbb{E}\left[\|G - Z(G)\|_F^2\right] = \mathbb{E}\left[\|\mathcal{P}_\Delta(G) - \mathcal{P}_\Delta(Z(G))\|_F^2\right] \tag{C.53}$$

$$= \mathbb{E}\left[\|\mathcal{P}_\Delta(G)\|_F^2\right] + \mathbb{E}\left[\|\mathcal{P}_\Delta(Z(G))\|_F^2\right] \tag{C.54}$$

$$= r(m_1 + m_2 - r) + r\,\mathbb{E}[\|\mathcal{P}_{\Delta^\perp}(G)\|^2]. \tag{C.55}$$

Here (C.54) follows because $\mathcal{P}_\Delta(G)$ and $\mathcal{P}_{\Delta^\perp}(G)$ are independent. The final line follows because $\dim(T) = r(m_1 + m_2 - r)$ and the Frobenius (i.e., Euclidean) norm of UV^* is $\|UV^*\|_F = \sqrt{r}$. Due to the isotropy of Gaussian random matrices, $\mathcal{P}_{\Delta^\perp}(G)$ is identically distributed as an $(m_1 - r) \times (m_2 - r)$ matrix with i.i.d. Gaussian entries each with mean zero and variance one. We thus know that

$$\mathbb{P}\left[\|\mathcal{P}_{\Delta^\perp}(G)\| \geq \sqrt{m_1 - r} + \sqrt{m_2 - r} + s\right] \leq \exp\left(-s^2/2\right) \tag{C.56}$$

(see, for example, [41]). To bound the latter expectation, we again use the integral form of the expected value. Letting $\mu_{T^\perp}$ denote the quantity $\sqrt{m_1 - r} + \sqrt{m_2 - r}$, we have

$$\mathbb{E}\left[\|\mathcal{P}_{\Delta^\perp}(G)\|^2\right] = \int_0^\infty \mathbb{P}\left[\|\mathcal{P}_{\Delta^\perp}(G)\|^2 > h\right] dh \tag{C.57}$$

$$\leq \mu_{T^\perp}^2 + \int_{\mu_{T^\perp}^2}^\infty \mathbb{P}\left[\|\mathcal{P}_{\Delta^\perp}(G)\|^2 > h\right] dh \tag{C.58}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \mathbb{P}\left[\|\mathcal{P}_{\Delta^\perp}(G)\|^2 > \mu_{T^\perp}^2 + t\right] dt \tag{C.59}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \mathbb{P}\left[\|\mathcal{P}_{\Delta^\perp}(G)\| > \mu_{T^\perp} + \sqrt{t}\right] dt \tag{C.60}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \exp(-t/2) dt \tag{C.61}$$

$$= \mu_{T^\perp}^2 + 2 \tag{C.62}$$

Using this bound in (C.55), we get that

$$\mathbb{E}\left[\inf_{Z \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|G - Z\|_F^2\right] \leq r(m_1 + m_2 - r) + r(\sqrt{m_1 - r} + \sqrt{m_2 - r})^2 + 2r \tag{C.63}$$

$$\leq r(m_1 + m_2 - r) + 2r(m_1 + m_2 - 2r) + 2r \tag{C.64}$$

$$\leq 3r(m_1 + m_2 - r) \tag{C.65}$$

where the second inequality follows from the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. We conclude that $3r(m_1 + m_2 - r)$ random measurements are sufficient to recover a rank r, $m_1 \times m_2$ matrix using the nuclear norm heuristic. □

Appendix D

Properties of Convex Symmetric Functions

A *convex symmetric function* is a convex function that is invariant with respect to a permutation of the argument:

Definition D.0.1. *A function* $g : \mathbb{R}^n \to \mathbb{R}$ *is a* convex symmetric function *if it is convex, and if for any* $\mathbf{x} \in \mathbb{R}^n$ *it holds that* $g(\Pi\mathbf{x}) = g(\mathbf{x})$ *for all permutation matrices* $\Pi \in \mathrm{Sym}(n)$.

The properties of such functions are well-known in the literature on convex analysis and optimization, and they arise in many applications. We briefly describe some of these properties and applications here.

An important class of convex symmetric functions is the set of linear functionals given by *monotone linear functionals*:

$$g(\mathbf{x}) = \mathbf{v}^T\overline{\mathbf{x}},$$

where $\mathbf{v}_1 \geq \cdots \geq \mathbf{v}_n$. Recall that $\overline{\mathbf{x}}$ is the vector obtained by sorting the entries of $\mathbf{x}$ in descending order. Monotone linear functionals can be used to express any convex symmetric function. Specifically, let $\mathcal{M} \subset \mathbb{R}^n$ represent the cone of monotone decreasing vectors in $\mathbb{R}^n$. Then for any convex symmetric function $g : \mathbb{R}^n \to \mathbb{R}$, we have that

$$g(\mathbf{x}) = \sup_{\mathbf{v}\in\mathcal{M}} \mathbf{v}^T\overline{\mathbf{x}} - \alpha_{\mathbf{v}}.$$

This statement is a simple consequence of the separation theorem from convex analysis [124]. Monotone linear functionals in turn can be expressed as the nonnegative *sum* of even more elementary functions called *distribution functions*, which are defined as

follows:

$$g_k(\mathbf{x}) = \sum_{i=1}^{k} (\overline{\mathbf{x}})_i.$$

These functions are closely related to the notion of conditional value-at-risk [125], which in turn is computed using quantiles of probability distributions.

Convex symmetric functions are intimately connected with the concept of *majorization* [104]. There are many equivalent characterizations of majorization [42,97], and we briefly mention some of these next. A vector $\mathbf{x} \in \mathbb{R}^n$ is said to majorize another vector $\mathbf{y} \in \mathbb{R}^n$ if

$$g_k(\mathbf{x}) \geq g_k(\mathbf{y}), \ \forall k = 1, \ldots, n-1 \quad \text{and} \quad g_n(\mathbf{x}) = g_n(\mathbf{y}).$$

The *permutahedron* of a vector $\mathbf{x} \in \mathbb{R}^n$ is the convex hull of all permutations of $\mathbf{x}$, and is given by the set of vectors in $\mathbb{R}^n$ that are majorized by $\mathbf{x}$. Thus, convex constraints given by distribution functions provide a simple characterization of the permutahedron generated by $\mathbf{x}$. Majorization is also closely related to the notion of *Lorenz dominance*; a (typically nonnegative) vector $\mathbf{x} \in \mathbb{R}^n$ is said to Lorenz-dominate $\mathbf{y} \in \mathbb{R}^p$ if $-\mathbf{x}$ is majorized by $-\mathbf{y}$. Lorenz dominance is used to measure the level of inequality in distributions, i.e., if a distribution $\mathbf{x}$ Lorenz-dominates a distribution $\mathbf{y}$ then $\mathbf{x}$ is "more equal" than $\mathbf{y}$ (see also the Gini coefficient, which is used to measure inequalities in countries).

A convex symmetric function is an example of a *Schur-convex function*, which is a function f such that $f(\mathbf{x}) \geq f(\mathbf{y})$ whenever $\mathbf{x}$ majorizes $\mathbf{y}$. Hence a Schur-convex function preserves order with respect to majorization. Consequently, such functions arise in many applications in which majorization plays a prominent role [104]. We note that the functions that are both convex and Schur-convex are exactly the convex symmetric functions.

A fairly similar set of results hold for convex functions of symmetric matrices that are invariant under conjugation of the argument by orthogonal matrices, i.e., convex functions $f : \mathbf{S}^n \to \mathbb{R}$ such that $f(VAV^T) = f(A)$ for all $A \in \mathbf{S}^n$ and for all $\Pi \in \mathrm{Sym}(n)$.

Bibliography

[1] AJA-FERNANDEZ, S., GARCIA, R., TAO, D., AND LI, X. (2009). *Tensors in Image Processing and Computer Vision (Advances in Pattern Recognition).* Springer.

[2] ALLMAN, E. S., MATIAS, C., AND RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statistics.* **37** 3099–3132.

[3] ALON, N. AND NAOR, A. (2006). Approximating the Cut-Norm via Grothendieck's Inequality. *SIAM Jour. on Comp.* **35** 787–803.

[4] AMES, B. AND VAVASIS, S. (2009). Nuclear norm minimization for the planted clique and biclique problems. *Preprint.* arXiv:0901.3348.

[5] AMES, B. AND VAVASIS, S. (2010). Convex optimization for the planted k-disjoint-clique problem. *Preprint.* arXiv:1008.2814.

[6] ARTZNER, P., DELBAEN, F., EBER, J., AND HEATH, D. (1999). Coherent measures of risk. *Math. Fin.* **9** 203–228.

[7] BACH, F. (2008). Consistency of trace norm minimization. *J. Mach. Lear. Res..* **9** 1019–1048.

[8] BARRON, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Tran. on Info. Theo.* **39** 930–945.

[9] BECKMANN, C. AND SMITH, S. (2005) Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage.* **25** 294-311.

[10] BenTal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization.* Princeton University Press.

[11] BenTal, A. and Nemirovski, A. (2001). *Lectures on Modern Convex Optimization.* SIAM.

[12] Bertsekas, D. (1996). *Constrained Optimization and Lagrange Multiplier Methods.* Athena Scientific.

[13] Bertsekas, D. P., Nedic, A. and Ozdaglar, A. E. (2003) *Convex Analysis and Optimization.* Athena Scientic, Belmont, MA.

[14] Bertsimas, D. and Brown, D. (2009). Constructing Uncertainty Sets for Robust Linear Optimization. *Oper. Res.* **57** 1483–1495.

[15] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statistics.* **36** 199–227.

[16] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statistics.* **36** 2577–2604.

[17] Biggs, N. (1994). *Algebraic Graph Theory.* Cambridge University Press.

[18] Bochnak, J., Coste, M., and Roy, M. (1988). Real Algebraic Geometry. *Springer.*

[19] Bollobás, B. (2001) *Random graphs.* Cambridge University Press.

[20] Bonchev, D. (1991). *Chemical Graph Theory: Introduction and Fundamentals.* Taylor and Francis.

[21] Bonsall, F. F. (1991). A General Atomic Decomposition Theorem and Banach's Closed Range Theorem. *The Quarterly Journal of Mathematics, 42(1):9-14.*

[22] Boyd, S. (2006). Convex Optimization of Graph Laplacian Eigenvalues. *Proc. Int'l. Cong. of Math.* **3** 1311–1319.

[23] Boyd, S., Diaconis, P., and Xiao, L. (2004). Fastest Mixing Markov Chain on a Graph. *SIAM Rev.* **46** 667–689.

[24] BOYD, S. P. AND VANDENBERGHE, L. (2004). *Convex optimization.* Cambridge University Press.

[25] BRIEDEN, A. AND GRITZMANN, P. AND KANNAN, R. AND KLEE, V. AND LOVASZ, L. AND SIMONOVITS, M. (1998). Approximation of Diameters: Randomization Doesn't Help. *Proc. of the 39th Annual Symp. on Foun. of Comp. Sci.* 244–251.

[26] CAI, J., CANDÈS, E., AND SHEN, Z. (2008). A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Jour. on Opt.* **20** 1956–1982.

[27] CAI, J., OSHER, S., AND SHEN, Z. (2008). Linearized Bregman Iterations for Compressed Sensing. *Technical report.*

[28] CANDÈS, E. AND PLAN, Y. (2009). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *to appear, IEEE Trans. Info. Theo.*

[29] CANDÈS, E. J., ROMBERG, J. AND TAO, T. (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Info. Theo.* **52** 489–509.

[30] CANDÈS, E. J. AND RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. of Comput. Math.* **9** 717–772.

[31] CANDÈS, E. J., LI, X., MA, Y. AND WRIGHT, J. (2009). Robust principal component analysis? *Preprint.*

[32] CELA, E. (1998). *The Quadratic Assignment Problem: Theory and Algorithms.* Springer.

[33] CHANDRASEKARAN, V., PARRILO, P., AND WILLSKY, A. (2010). Latent Variable Graphical Model Selection via Convex Optimization. *Preprint.*

[34] CHANDRASEKARAN, V., PARRILO, P., AND WILLSKY, A. (2010). Convex Graph Invariants. *Preprint.*

[35] CHANDRASEKARAN, V., PARRILO, P., AND WILLSKY, A. Graph Deconvolution via Semidefinite Programming. *In preparation.*

[36] CHANDRASEKARAN, V., RECHT, B., PARRILO, P., AND WILLSKY, A. (2010). The Convex Geometry of Linear Inverse Problems. *Preprint.*

[37] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P., AND WILLSKY, A. (2009). Rank-sparsity Incoherence for Matrix Decomposition. *SIAM Jour. on Optim.*, to appear.

[38] CODENOTTI, B. (2000). Matrix rigidity. *Lin. Alg. and its Appl.* **304** 181–192.

[39] COMBETTES, P. AND WAJS, V. (2005). Signal recovery by proximal forward-backward splitting. *Mult. Model. and Simu.* **4** 1168–1200.

[40] DAUBECHIES, I., DEFRIESE, M., AND DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. of Pure and Appl. Math.* **LVII** 1413–1457.

[41] DAVIDSON, K. R. AND SZAREK, S.J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces.* **I** 317–366.

[42] DAVIS, D. (1957). All convex invariants of Hermitian matrices. *Arch. der Math.* **8** 276–278.

[43] DE KLERK, E. AND SOTIROV, R. (2010). Exploiting group symmetry in semidefinite programming relaxations of the Quadratic Assignment Problem. *Math. Prog.*, **122** 225–246.

[44] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B.* **39** 1–38.

[45] DE SILVA, V. AND LIM, L. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Jour. on Mat. Analys.* **30** 1084-1127.

[46] DEVORE, R. AND TEMLYAKOV, V. (1996). Some remarks on greedy algorithms. *Adv. in Comp. Math.* **5** 173–187.

[47] DEZA, M. AND LAURENT, M. (1997). *Geometry of cuts and metrics.* Springer.

[48] DIESTEL, R. (2005). *Graph Theory.* Springer.

[49] DOBRIN, R., BEG, Q., BARABASI, A., AND OLTVAI, Z. (2004). Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinf.* **5**.

[50] DODZIUK, J. (1984). Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Amer. Math. Soc.* **284** 787-794.

[51] DONOHO, D. L. AND HUO, X. (2001) Uncertainty principles and ideal atomic decomposition. *IEEE Trans. on Info. Theo.* **47** 2845–2862.

[52] DONOHO, D. L. AND ELAD, M. (2003). Optimal Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ_1 Minimization. *Proc. of the Nat. Acad. of Sci.* **100** 2197–2202.

[53] DONOHO, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Math..* **59** 797–829.

[54] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Info. Theory.* **52** 1289–1306.

[55] DONOHO, D. AND TANNER, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA.* **102** 9446–9451.

[56] DONOHO, D. AND TANNER, J. (2010). Counting the Faces of Randomly-Projected Hypercubes and Orthants with Applications. *Disc. and Comp. Geom.* **43** 522–541.

[57] DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis 1: 290–330.*

[58] DYER, M., FRIEZE, A., AND KANNAN, R. (1991). A random polynomial-time algorithm for approximating the volume of convex bodies. *Jour. of the ACM.* **38**.

[59] EASLEY, D. AND KLEINBERG, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge University Press.

[60] ELIDAN, G., NACHMAN, I., AND FRIEDMAN, N. (2007). "Ideal Parent" structure learning for continuous variable Bayesian networks. *J. Mach. Lear. Res.*. **8** 1799–1833.

[61] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statistics.* **36** 2717–2756.

[62] FAN, J., FAN, Y., AND LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics.* **147** 186–197.

[63] FAZEL, M. AND GOODMAN, J. (1998). Approximations for Partially Coherent Optical Imaging Systems. *Tech. Report.* Department of Electrical Engineering, Stanford University.

[64] FAZEL, M. (2002). *Matrix Rank Minimization with Applications.* PhD thesis, Department of Electrical Engineering, Stanford University.

[65] FAZEL, M., HINDI, H. AND BOYD, S. (2003) Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. *Proc. of the Amer. Control Conf.*.

[66] FIGUEIREDO, M. AND NOWAK, R. (2003). An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. on Image Proc.*. **12** 906–916.

[67] FINKE, G., BURKARD, R., AND RENDL, F. (1987). Quadratic Assignment Problems. *Ann. of Disc. Math.* **31** 61–82.

[68] FUKUSHIMA, M. AND MINE, H. (1981). A generalized proximal point algorithm for certain non-convex minimization problems. *Int. Jour. of Sys. Sci.* **12** 989–1000.

[69] GHOSH, A. AND BOYD, S. (2006). Growing Well-Connected Graphs. *Proc. IEEE Conf. on Dec. and Cont.* 6605-6611.

[70] GODSIL, C. AND ROYLE, G. (2004). *Algebraic Graph Theory.* Springer-Verlag.

[71] GOEMANS, M. (2009). Smallest compact formulation for the permutahedron. *preprint.*

[72] GOEMANS, M. AND WILLIAMSON, D. (1995). Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *Jour. of the ACM.* **42** 1115–1145.

[73] GOLDREICH, O., GOLDWASSER, S., AND RON, D. (1996). Property testing and its connection to learning and approximation. *Proc. of Ann. Symp. on Foun. of Comp. Sci..*

[74] GOLUB, G. H. AND VAN LOAN, C. H. (1990). *Matrix computations.* The Johns Hopkins Univ. Press.

[75] GOODMAN, J. (2004). *Introduction to Fourier Optics*, Roberts and Company Publishers.

[76] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric aspects of functional analysis, Isr. Semin. 1986-87, Lect. Notes Math. 1317, 84-106.*

[77] GOUVEIA, J., PARRILO, P., AND THOMAS, R. (2010). Theta Bodies for Polynomial Ideals. *SIAM Jour. Optim.* **20** 2097–2118.

[78] HALE, T., YIN, W., AND ZHANG, Y. (2008). A fixed-point continuation method for ℓ_1-regularized minimization: Methodology and convergence. *SIAM Jour. on Opt.* **19** 1107–1130.

[79] HARRIS, J. (1995). *Algebraic Geometry: A First Course*, Springer-Verlag.

[80] HAUPT, J., BAJWA, W., RAZ, G., AND NOWAK, R. (2008). Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Tran. on Info. Theo.*, to appear.

[81] HOORY, S., LINIAL, N., AND WIGDERSON, A. (2006). Expander Graphs and their Applications. *Bull. Amer. Math. Soc.* **43** 439–561.

[82] HORN, R. A. AND JOHNSON, C. R. (1990). *Matrix analysis.* Cambridge University Press.

[83] JACKSON, M. (2008). *Social and Economic Networks.* Princeton University Press.

[84] JAGABATHULA, S. AND SHAH, D. (2010). Inferring Rankings Using Constrained Sensing. *Preprint*, arXiv:0910.0895.

[85] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statistics.* **29** 295–327.

[86] JONES, L. (1992). A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *Ann. of Stat.* **20** 608–613.

[87] KATO, T. (1995). *Perturbation theory for linear operators.* Springer.

[88] KLAIN, D. AND ROTA, G. (1997). *Introduction to geometric probability.* Cambridge University Press.

[89] KLEINBERG, J. AND TARDOS, E. (2008). Balanced Outcomes in Social Exchange Networks. *Proc. Symp. on Theo. of Comp.*

[90] KOLDA, T. (2001). Orthogonal Tensor Decompositions. *SIAM Jour. on Mat. Analys..* **23** 243–255.

[91] KOLDA, T. AND BADER, B. (2009). Tensor Decompositions and Applications. *SIAM Rev.* **51** 455–500.

[92] LAM, C. AND FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statistics.* **37** 4254–4278.

[93] LAURITZEN, S. L. (1996). *Graphical models.* Oxford University Press.

[94] LEDOIT, O. AND WOLF, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Analysis.* **88** 365–411.

[95] LEDOUX, M. (2000). *The Concentration of Measure Phenomenon.* American Mathematical Society.

[96] LEDOUX, M. AND TALAGRAND, M. (1991). *Probability in Banach Spaces.* Springer.

[97] LEWIS, A. (1995). The Convex Analysis of Unitarily Invariant Matrix Functions. *Jour. of Convex Analy.* **2** 173–183.

[98] LÖFBERG, J. (2004). YALMIP: A Toolbox for Modeling and Optimization in MATLAB. *Proceedings of the CACSD Conference, Taiwan.* Available from http://control.ee.ethz.ch/ joloef/yalmip.php.

[99] LOKAM, S. (1995). Spectral Methods for Matrix Rigidity with Applications to Size-Depth Tradeoffs and Communication Complexity. *36th IEEE Symp. on Found. of Comp. Sci. (FOCS).* 6–15.

[100] MA, S., GOLDFARB, D., AND CHEN, L. (2008). Fixed point and Bregman iterative methods for matrix rank minimization. *Preprint,* arXiv:0905.1643.

[101] MAHAJAN, M. AND SARMA, J. (2010). On the Compelxity of Matrix Rank and Rigidity. *Theo. of Comp. Sys.* **46** 9–26.

[102] MANGASARIAN, O. AND RECHT, B. (2009). Probability of Unique Integer Solution to a System of Linear Equations. *Preprint.*

[103] MARCENKO, V. A. AND PASTUR, L. A. (1967). Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb.* **1** 507–536.

[104] MARSHALL, A. AND OLKIN, I. (1979). *Inequalities: The Theory of Majorizations and Its Applications.* Academic Press.

[105] MASON, O. AND VERWOERD, M. (2007). Graph Theory and Networks in Biology. *IET Syst. Biol.* **1** 89-119.

[106] MATOUŠEK, J. (2002). *Lectures on Discrete Geometry.* Springer.

[107] MEINSHAUSEN, N. AND BUHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statistics.* **34** 1436–1462.

[108] MESBAHI, M. AND PAPAVASSILOPOULOS, G. P. (1997). On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Trans. on Auto. Cont.* **42** 239-243.

[109] MOTZKIN, T. AND STRAUS, E. (1965). Maxima for graphs and a new proof of a theorem of Turan. *Canad. J. Math.* **17** 533–540.

[110] NESTEROV, Y. (1997). Quality of semidefinite relaxation for nonconvex quadratic optimization. *Technical report.*

[111] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization.* Kluwer.

[112] NESTEROV, Y. (2007). Gradient methods for minimizing composite functions. *Preprint.*

[113] ORTEGA, J. M. AND RHEINBOLDT, W. G. (1970). *Iterative solution of nonlinear equations in several variables.* Academic Press.

[114] PARRILO, P. A. (2003). Semidefinite Programming Relaxations for Semialgebraic Problems. *Mathematical Programming Ser. B, Vol. 96, No.2, pp. 293-320.*

[115] PATI, Y. C. AND KAILATH, T. (1994). Phase-shifting masks for microlithography: Automated design and mask requirements. *Jour. of the Opt. Soc. of Amer. A.* bf 11.

[116] PISIER, G. (1981). Remarques sur un résultat non publié de B. Maurey. *Séminaire d'analyse fonctionnelle.* Ecole Polytechnique Centre de Mathematiques.

[117] POLAK, E. (1997).*Optimization: Algorithms and Consistent Approximations.* Springer.

[118] RAUHUT, H. (2009). Circulant and Toeplitz matrices in compressed sensing. *Proc. of SPARS'09.*

[119] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G., AND YU, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Preprint.*

[120] RECHT, B. (2009). *Personal Communication.*

[121] RECHT, B., FAZEL, M., AND PARRILO, P. A. (2010). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review.* **52** 471–501.

[122] RENDL, F. AND SOTIROV, R. (2007). Bounds for the Quadratic Assignment Problem using the bundle method. *Math. Prog.* **109** 505–524.

[123] ROBERTSON, N. AND SEYMOUR, P. (1984). Graph minors III: Planar tree-width. *Jour. of Comb. Theo., Series B.* **36** 49-64.

[124] ROCKAFELLAR, R. T. (1996). *Convex Analysis.* Princeton University Press.

[125] ROCKAFELLAR, R. AND URYASEV, S. (2000). Optimization of Conditional Value-at-Risk. *Jour. of Risk.* **2** 21–41.

[126] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E., AND ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Elec. J. Statistics.* **2** 494–515.

[127] RUDELSON, M. AND VERSHYNIN, R. (2006). Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. *CISS 2006 (40th Annual Conference on Information Sciences and Systems).*

[128] RUDIN, W. (1966). *Real and Complex Analysis.* McGraw-Hill.

[129] SANYAL, R., SOTTILE, F., AND STURMFELS, B. (2009) Orbitopes. *Preprint,* arXiv:0911.5436.

[130] SCHUR, I. (1911). Bemerkungen zur Theorie der beschrankten Bilinearformen mit unendlich vielen Veranderlischen. *Jour. fur Reine und Angew. Mathematik.* **140** 1-28.

[131] SONTAG, E. (1998). *Mathematical Control Theory.* Springer-Verlag, New York.

[132] SPEED, T. P. AND KIIVERI, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statistics.* **14** 138–150.

[133] SREBRO, N. AND SHRAIBMAN, A. (2005). Rank, Trace-Norm and Max-Norm. *18th Annual Conference on Learning Theory (COLT).*

[134] STOJNIC, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. *Preprint,* arXiv:0907.3666.

[135] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58** 267–288.

[136] K. C. TOH, M. J. TODD, AND R. H. TUTUNCU. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming.* Available from http://www.math.nus.edu.sg/ mattohkc/sdpt3.html.

[137] TOH, K. AND YUN, S. (2009). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Preprint.*

[138] VALIANT, L. G. (1977). Graph-theoretic arguments in low-level complexity. *6th Symp. on Math. Foun. of Comp. Sci.* 162-176.

[139] VANDENBERGHE, L. AND BOYD, S. (1996) Semidefinite Programming. *SIAM Review.* **38** 49–95.

[140] VARGA, R. S. (2000). *Matrix iterative analysis.* Springer-Verlag.

[141] WANG, C., SUN, D. AND TOH, K. C. (2009). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *Preprint.*

[142] WATSON, G. A. (1992).Characterization of the subdifferential of some matrix norms. *Lin. Alg. and Appl.* **170** 1039–1053.

[143] WITTEN, D. M., TIBSHIRANI, R. AND HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostat.* **10** 515–534.

[144] WRIGHT, S., NOWAK, R., AND FIGUEIREDO, M. (2009). Sparse Reconstruction by Separable Approximation. *IEEE Trans. on Sign. Proc.* **57** 2479–2493.

[145] WU, W. B. AND POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika.* **90** 831–844.

[146] YIN, W., OSHER, S., DARBON, J., AND GOLDFARB, D. (2007). Bregman Iterative Algorithms for Compessed Sensing and Related Problems. *Technical report.*

[147] ZHAO, Q., KARISCH, S., RENDL, F., AND WOLKOWICZ, H. (1998). Semidefinite Programming Relaxations for the Quadratic Assignment Problem. *Jour. of Comb. Opt..* **2** 71–109.

[148] ZHAO, P. AND YU, B. (2006). On model selection consistency of lasso. *J. Mach. Lear. Res..* **7** 2541–2567.

[149] ZIEGLER, G. (1995). *Lectures on Polytopes.* Springer.