# Convex relaxations for certain inverse problems on graphs

Afonso S. Bandeira

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Program in
Applied and Computational Mathematics
Adviser: Amit Singer

June 2015

# Abstract

Many maximum likelihood estimation problems are known to be intractable in the worst case. A common approach is to consider convex relaxations of the maximum likelihood estimator (MLE), and relaxations based on semidefinite programming (SDP) are among the most popular. This thesis focuses on a certain class of graph-based inverse problems, referred to as *synchronization-type problems*. These are problems where the goal is to estimate a set of parameters from pairwise information between them.

In this thesis, we investigate the performance of the SDP based approach for a range of problems of this type. While for many such problems, such as multi-reference alignment in signal processing, a precise explanation of their effectiveness remains a fascinating open problem, we rigorously establish a couple of remarkable phenomena.

For example, in some instances (such as community detection under the stochastic block model) the solution to the SDP matches the ground truth parameters (i.e. achieves exact recovery) for information theoretically optimal regimes. This is established by developing non-asymptotic bounds for the spectral norm of random matrices with independent entries.

On other instances (such as angular synchronization), the MLE itself tends to not coincide with the ground truth (although maintaining favorable statistical properties). Remarkably, these relaxations are often still tight (meaning that the solution of the SDP matches the MLE). For angular synchronization we establish this behavior by analyzing the solutions of certain randomized Grothendieck problems.

# Acknowledgements

First and foremost, I would like to thank my adviser Amit Singer. I was fortunate to have his wise guidance at every turn. I learned a lot during my time in Princeton, and a big part of it, I learned from Amit. I aspire to one day be as good of a mentor to my students as Amit was to me.

I was particularly fortunate to have a number of amazing mentors in Princeton. What I know now is a reflection of what they taught me. Besides Amit, I want to thank Moses Charikar, Emmanuel Abbe, and Ramon van Handel for their constant mentoring. A special thanks to Dustin Mixon who, as a senior graduate student, spent countless hours patiently mentoring a starting graduate student.

Without my incredible collaborators, none of this would have been possible. I learned something from every single one of them. A warm thanks to all!

To all my teachers in Coimbra and, in particular, to Projecto Delfos, a big thank you for showing me the beauty of Mathematics.

To Cara for not letting me take my work too seriously, but more importantly, for giving all of this that much more meaning.

And finally, to my amazing family, for their constant support.

To Isabel, José, and Matilde.

# Contents

# Chapter 1

# Introduction

Extracting information from data is of paramount importance in our society. This task is often posed in terms of estimation. Either as estimation of statistical parameters from observations or as recovery of a signal (or image) from measurements. Many of these statistical (or signal) recovery problems are formulated as optimization problems over a set of feasible parameters (or signals), where the objective to be optimized represents, in some way or another, how much the parameters explain the data. A prime example is the paradigm of maximum likelihood (ML) estimation.

This thesis focuses on *synchronization-type problems*. These are problems where the goal is to estimate a set of parameters from data concerning relations or interactions between pairs of them. A good example to have in mind is an important problem in computer vision, known as structure from motion: the goal is to build a three-dimensional model of an object from several two-dimensional photos of it taken from unknown positions. Although one cannot directly estimate the positions, one can compare pairs of pictures and gauge information on their relative positioning. The task of estimating the camera locations from this pairwise information is a synchronization-type problem. Another example, from signal processing, is multireference alignment, which is the problem of estimating a signal from measuring multiple arbitrarily shifted copies of it that are corrupted with noise. One of the most relevant instances of this type of problem arises in single particle reconstruction from cryo-Electron Microscopy: to resolve the global structure of a certain molecule by registering multiple images of the molecule at unknown orientations. Similarly, one can infer pairwise information by comparing pairs of images. The high levels of noise inherent to the imaging process make this problem particularly challenging. Several other relevant examples will be described in Section 1.2.

We will formulate each of these problems as an estimation problem on a graph $G = (V, E)$. More precisely, we will associate each data unit (say, a photo, a cryo-EM image, or a shifted signal) to a graph node $i \in V$. The problem can then be formulated as estimating, for each node $i \in V$, a group element $g_i \in \mathcal{G}$, where the group $\mathcal{G}$ is a group of transformations, such as translations, rotations, or permutations. The pairwise data, which we identify with edges of the graph $(i, j) \in E$, reveals information about the ratios $g_i(g_j)^{-1}$. In its simplest form, for each edge $(i, j) \in E$ of the graph, we have a noisy estimate of $g_i(g_j)^{-1}$ and the synchronization problem consists of estimating the individual group elements $g : V \to \mathcal{G}$ that are the most consistent with the edge estimates, often corresponding to the ML estimator.

Unfortunately, in most of the relevant instances, the parameter space is exponentially large and non-convex, often rendering an exact calculation of the estimator intractable for even moderately sized instances. This is a common ocurrence in many estimation problems, not special to synchronization-type problems. In such cases, it is common to settle for heuristics, such as expectation maximization, simulated annealing, genetic algorithms, or other global nonlinear optimization methods. Unfortunately, these methods often lack theoretical guarantees and it is not unusual for them to get stuck in local optima. To make matters worse, it is often hard to check whether the solution computed is the global optimum.

A popular alternative to these heuristics is the use of convex relaxations: to attempt optimizing (usually the log-likelihood) in a larger convex set that contains the parameter set of interest. The motivation for this approach is that one can, in many instances, solve the corresponding convex problems efficiently. The downside is that the solution obtained may not be in the original feasible set, forcing one to take an extra, potentially suboptimal, rounding step. An exceptional effort has been ongoing in theoretical computer science to design rounding schemes with worst-case approximation guarantees. Chapter 2 mostly treats this type of analysis.

The applied mathematics community has been focusing on a more optimistic approach; instead of trying to establish guarantees in a worst case paradigm, one tries to identify instances for which the solution to the convex relaxation matches the parameter of interest, achieving *exact recovery*. The motivation being that "the data is not the enemy" and that, despite the worst-case scenario hardness of the task at hand, it is not unusual for typical instances of the problem to be solvable efficiently. A particularly remarkable example (although not exactly an instance of convex relaxation as described above) is sparse recovery, where the realization that a sparse signal can be efficiently recovered by very few random linear measurements (with high probability) spawned the fruitful field of compressed sensing (see, for example, [65, 96]).

Convex relaxation based methods have an extra appeal: if the relaxation is *tight* (meaning that its solution is in the original set of feasible parameters) then one is sure that it coincides with the optimal solution for the original problem, thus creating a certificate of optimality. Fortunately, it seems that in many instances, convex relaxations do tend to be tight. In fact, several remarkable results exist showing that, under a particular probabilistic distribution of the input data, certain convex relaxations achieve, with high probability, exact recovery. The technique most commonly used in this type of analysis is to show that the "ground truth" parameter is the optimal solution to the convex problem, by leveraging convex duality.

Chapter 3 will study this phenomenon for a class of synchronization problems over the group of two elements, which includes the community detection problem with two communities.

There are however many cases in which, often because of noise on the data, one cannot hope to achieve exact recovery. In many such instances, one is still interested in computing the ML estimator, or more generally, in solving the original optimization problem. Unfortunately, the usual techniques to show tightness of the relaxation seem to break down. On the other hand, it seems that the tightness tendency of convex relaxations is still present. When dealing with semidefinite relaxations (described in Section 1.3), the relaxation being tight is usually equivalent to the matrix solution of the convex problem having a certain prescribed rank, so we refer to this fascinating phenomenon as "rank recovery" (in fact, in these cases, the original problem can be shown to be equivalent to the convex relaxation plus an extra non-convex rank constraint). This phenomenon is discussed in Chapter 5 and is, in Section 5.1, established for the setting of synchronization of in-plane rotations.

One of the crucial tools used in the analysis of semidefinite relaxations for many of the problems described above is *random matrix theory.* Since Eugene Wigner's remarkable finding, in the late 1950s [237], that the spectrum of a large class of random matrices with independent and identically distributed entries (i.i.d.) is, in high dimension, distributed essentially the same way,[1] the study of spectral properties of random matrices has spawned a panoply of fascinating research with important implications in many areas.[2] In much of the analysis carried out in this thesis one needs to estimate the largest eigenvalue, or sometimes the spectral norm, of certain, application dependent, random matrices. While the situation is very well understood when the entries are i.i.d., many important examples fall outside of this setting. On

[1] Under mild assumptions the distribution of the spectrum of such matrices converges to the so-called Wigner semicircle law [237]

[2] We refer the reader to the books [217, 22] for more on the subject of random matrices.

the other hand, there are tools capable of bounding this quantity for an amazingly broad class of examples, however, these inequalities are often asymptotically suboptimal (as it turns out to be the case in our applications). To address this issue, we devote Chapter 4 to study the spectral norm of a broad class of random matrices with independent, but not necessarily identically distributed, entries and develop sharp nonasymptotic estimates. These estimates play a crucial role in the analysis, described in Chapter 3, of semidefinite relaxations for synchronization over the group of two elements.

## 1.1 Organization of this thesis

This thesis is organized as follows: In the rest of this chapter we give a general formulation of synchronization problems, describe a number of important examples, and discuss several algorithmic approachs to the problem. Chapter 2 is mostly devoted to worst-case guarantees for both spectral and semidefinite programming based methods. In Chapter 3 we restrict ourselves to the group $\mathcal{G} \cong \mathbb{Z}_2$, and investigate exact recovery of the semidefinite relaxation. Chapter 4 establishes bounds on eigenvalues of certain random matrices that are heavily used in Chapter 3. Chapter 5 investigates instances on which the semidefinite relaxation has a tendency to be tight, albeit not achieving exact recovery. We discuss a larger class of problems, such as signal alignment, in Chapter 6 and conclude with several open problems in Chapter 7.

## 1.2 Synchronization problems: formulation and examples

We recall the formulation, described above, of a synchronization-type problem. Consider a graph $G = (V, E)$ and a group $\mathcal{G}$. One is given, for each edge of a graph

$G = (V, E)$, a noisy observation of $g_i g_j^{-1}$ and asked for find the node labels $g : V \to \mathcal{G}$, also referred to as the *group potential*, that most “agrees with measurements” (see Figure 1.1). Naturally, the measure of “agreement” is application specific. For the sake of generality, we will consider any type of edge measurement by thinking about an edge measurement as associating, to each value of $g_i g_j^{-1}$, a cost to be minimized (for example, minus the log-likelihood).

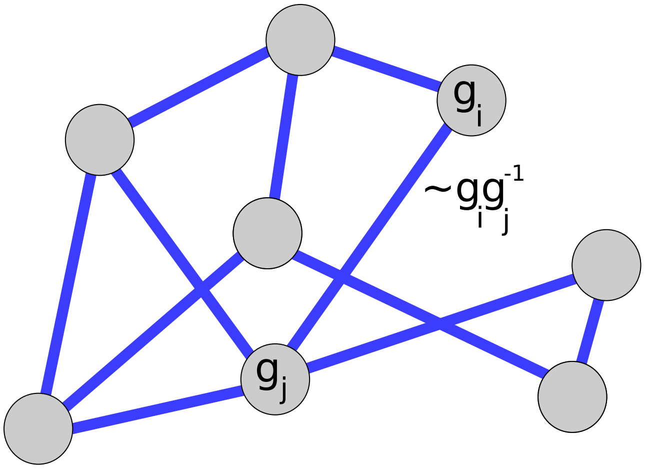

Figure 1.1: Given a graph $G = (V, E)$ and a group $\mathcal{G}$, the goal in synchronization-type problems is to estimate node labels $g : V \to \mathcal{G}$ from noisy edge measurements of offsets $g_i g_j^{-1}$.

**Problem 1.2.1.** *[General Synchronization-type problem] Given a graph $G = (V, E)$, a group $\mathcal{G}$, and, for each edge $(i, j) \in E$, a function $f_{ij} : \mathcal{G} \to \mathbb{R}$. The goal is to find the group potential $g : V \to \mathcal{G}$ that minimizes*

$$\min_{g:V\to\mathcal{G}} \sum_{(i,j)\in E} f_{ij}\left(g_i g_j^{-1}\right). \tag{1.1}$$

We remark that in the applications we consider compact groups $\mathcal{G}$. In fact, it is either finite or a (special) orthogonal group $O(d)$ or $SO(d)$. Moreover, when dealing with infinite groups, the functions $f_{ij}$ will be such that a minimizer of (1.1) exists. Note that there is always a global shift ambiguity for the solution of (1.1) as, for any $h \in \mathcal{G}$, the group potential $g : V \to \mathcal{G}$ and the one obtained by right-multiplying every element by $h$, have the same cost.

We now describe a number of relevant problems that can be formulated as synchornization problems, by an appropriate choice of $G = (V, E)$, $\mathcal{G}$, and functions $f_{ij}$. Many of these are direct applications of our techniques as motivated a lot of the developments in this thesis.

### 1.2.1 Angular Synchronization

We start with the example that motivated the use of the term *synchronization*. In 2011, Amit Singer [202] formulated the angular synchronization problem: to estimate $n$ unknown angles $\theta_1, \ldots, \theta_n$ from $m$ noisy measurements of their offsets $\theta_i - \theta_j \mod 2\pi$. This problem easily falls under the scope of synchronization-type problem by taking a graph with a node for each $\theta_i$, an edge associated with each measurement, and taking the group to be $\mathcal{G} \cong SO(2)$, the group of in-plane rotations. Some of its applications include time-synchronization of distributed networks [113], signal reconstruction from phaseless measurements (discussed in Section 2.2), and surface reconstruction problems in computer vision [10] and optics [191]. We will discuss this problem in greater detail in Section 5.1.

### 1.2.2 Orientation estimation in Cryo-EM

A particularly challenging application of this framework is the orientation estimation problem in Cryo-Electron Microscopy [204].

Cryo-EM is a technique used to determine the three-dimensional structure of biological macromolecules. The molecules are rapidly frozen in a thin layer of ice and imaged with an electron microscope, which gives 2-dimensional projections. One of

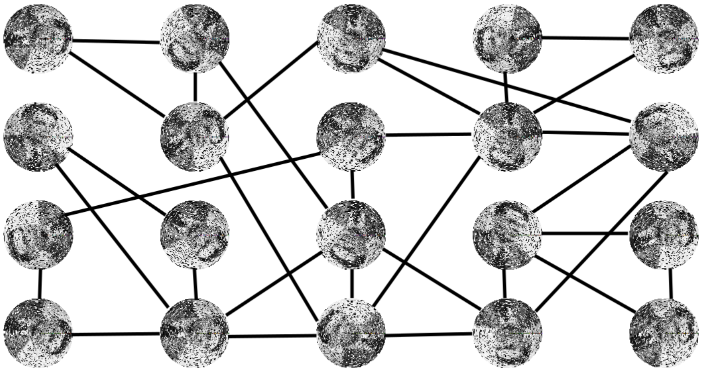

Figure 1.2: An example of an instance of a synchronization-type problem. Given noisy rotated copies of an image (corresponding to vertices of a graph), the goal is to recover the rotations. By comparing pairs of images (corresponding to edges of the graph), it is possible to estimate the relative rotations between them. The problem of recovering the rotation of each image from these relative rotation estimates is an instance of Angular synchronization.

the main difficulties with this imaging process is that these molecules are imaged at different unknown orientations in the sheet of ice and each molecule can only be imaged once (due to the destructive nature of the imaging process). More precisely, each measurement consists of a tomographic projection of a rotated (by an unknown rotation) copy of the molecule. The task is then to reconstruct the molecule density from many such measurements. As the problem of recovering the molecule density knowing the rotations fits in the framework of classical tomography—for which effective methods exist—we focus on determining the unknown rotations, the orientation estimation problem. In Section 6.3 we will briefly describe a mechanism to, from two such projections, obtain information between their orientation. The problem of finding the orientation of each projection from such pairwise information naturally fits in the framework of synchronization.

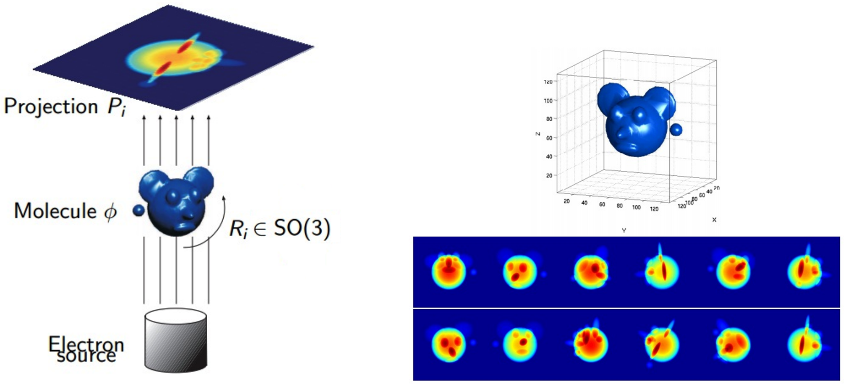

Figure 1.3: Illustration of the Cryo-EM imaging process: A molecule is imaged after being frozen at a random (unknown) rotation and a tomographic 2-dimensional projection is captured. Given a number of tomographic projections taken at unknown rotations, we are interested in determining such rotations with the objective of reconstructing the molecule density. Images courtesy of Amit Singer and Yoel Shkolnisky [204].

### 1.2.3 Synchronization over the orthogonal group

A natural extension to the angular synchronization framework is Synchronization over $O(d)$, the group of $d \times d$ orthogonal matrices [41, 233]. It plays a major role in an algorithm for the sensor network localization problem [89] and in structure from motion [158, 125]. The similar problem of synchronization over $SO(d)$, the group of rotations in $\mathbb{R}^d$ (or $d \times d$ orthogonal matrices with determinant 1), also has several applications, for example, the problem over $SO(3)$ can be used for global alignment of 3-d scans [224]. Chapter 2 will propose and analyze algorithmic frameworks for this problem.

### 1.2.4 Synchronization over $\mathbb{Z}_2$

When $d = 1$, $O(d) \cong \mathbb{Z}_2$ and Synchronization over $O(d)$ reduces to a binary version of the problem, treated in Chapter 3. This simplified version already includes many applications of interest. Similarly to before, given a graph $G = (V, E)$, the goal is recover unknown node labels $g : V \to \mathbb{Z}_2$ (corresponding to memberships to two clusters) from pairwise information. Each pairwise measurement either suggests the two involved nodes are in the same cluster or in different ones. The task of clustering the graph in order to agree, as much as possible, with these measurements is tightly connected to *correlation clustering* [44] and has applications to determining the orientation of a manifold [205].

In the case where all the measurements suggest that the involved nodes belong in different communities, or in other words

$$f_{ij}(-1) \leq f_{ij}(1),$$

for all edges $(i, j) \in E$, then Problem 1.2.1 essentially reduces to the `Max-Cut` problem. In fact, as we will see in Section 1.3, our algorithmic approach is motivated from that of Michel Goemans and David Williamson [114] for the `Max-Cut` problem.

### 1.2.5 Community detection

A remarkable example of this framework is the problem of community detection, or clustering, in a graph. Many real world graphs are known to have a community structure. A good example is the political blogosphere dataset [7]: the graph is composed by blogs and edges are drawn whenever one blog contains a link to another. It was observed [7] that blogs have a tendency to link more to blogs sharing political ideology, resulting in a political blogosphere graph where two communities have considerably more edges within them than across.

For many such graphs one is interested in recovering the communities from the connectivity of the graph. As we can read each connection (or link) as a suggestion that the two nodes should be in the same community, and the absence of such a connection (or link) as suggestion that they belong to different communities (see

Figure 1.4) it readily falls under the framework of Synchronization.

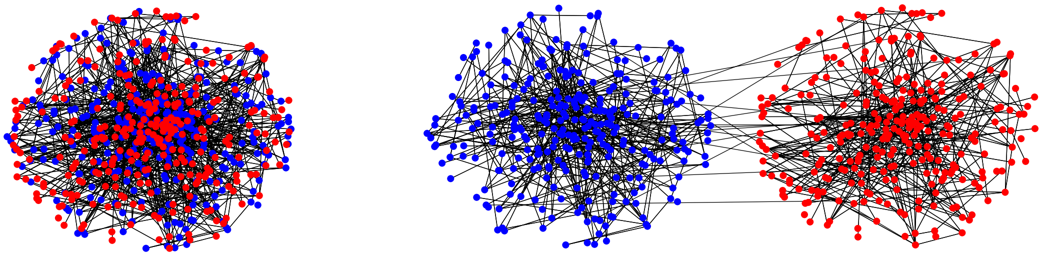

Figure 1.4: A graph generated form the stochastic block model (Definition 3.2.1) with 600 nodes and 2 communities, scrambled on the left and clustered on the right. The stochastic block model produces random graphs with community structure. In this example, pairs of nodes are connected independently with probability $p = 6/600$ if they are in the same community and with probability $q = 0.1/600$ if they are in different communities. The community detection problem consists in recovering the node structure on an unlabeled graph [3].

For the setting of two communities this problem is tightly connected with Synchronization over $\mathbb{Z}_2$ and will be the subject of Section 3.2. We will also briefly discuss the setting of multiple communities in Section 6.2.

### 1.2.6 Procrustes Problem

Given $n$ point clouds in $\mathbb{R}^d$ of $k$ points each, the orthogonal Procrustes problem [198] consists of finding $n$ orthogonal transformations that best simultaneously align the point clouds.

If the points are represented as the columns of matrices $A_1, \ldots, A_n$, where $A_i \in \mathbb{R}^{d\times k}$ then the orthogonal Procrustes problem consists of solving

$$\min_{O_1,\ldots,O_n \in O(d)} \sum_{i,j=1}^{n} ||O_i^T A_i - O_j^T A_j||_F^2. \tag{1.2}$$

Since $||O_i^T A_i - O_j^T A_j||_F^2 = \|A_i\|_F^2 + \|A_j\|_F^2 - 2\operatorname{Tr}\left((A_i A_j^T)^T O_i O_j^T\right)$, (1.2) has the same solution as the complementary version of the problem

$$\max_{O_1,\ldots,O_n\in O(d)} \sum_{i,j=1}^{n} \operatorname{Tr}\left((A_iA_j^T)^T O_i O_j^T\right), \tag{1.3}$$

which can easily be formulated in the framework of Problem 1.2.1 by taking the group to be $\mathcal{G} \cong O(d)$.

### 1.2.7 Signal Alignment

In signal processing, the multireference alignment problem [32] consists of recovering an unknown signal $u \in \mathbb{R}^L$ from $n$ observations of the form

$$y_i = R_{l_i} u + \sigma \xi_i,$$

where $R_{l_i}$ is a circulant permutation matrix that shifts $u$ by $l_i \in \mathbb{Z}_L$ coordinates, $\xi_i$ is a noise vector (which we will assume standard gaussian i.i.d. entries) and $l_i$ are unknown shifts.

If the shifts were known, the estimation of the signal $u$ would reduce to a simple denoising problem. For that reason, we will focus on estimating the shifts $\{l_i\}_{i=1}^n$. By comparing two observations $y_i$ and $y_j$ we can obtain information about the relative shift $l_i - l_j \mod L$. Using this intuition, in Section 6.1, we will formulate a quasi-MLE estimator for the multireference alignment problem as particular instance of Problem 1.2.1.

### 1.2.8 Other examples

Many other problems can be fruitfully considered through the synchronization framework. We briefly describe a short list below.

### Unique-Games Problem

The unique games conjecture, posed by Khot in 2002 [138], plays a central role in modern theoretical computer science.

**Conjecture 1.2.2.** *[Unique Games Conjecture] For $\delta, \varepsilon > 0$, it is impossible for a polynomial-time algorithm to distinguish between $\delta$-satisfiable and $(1-\varepsilon)$-satisfiable* Unique-Games *instances.*

A Unique-Games instance consists of a graph along with a permutation for each edge. The problem is to choose an assignment of labels to each vertex such that as many of the edge permutations as possible are satisfied (see Figure 1.5).

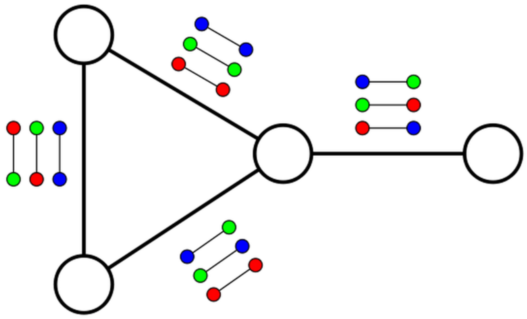

Figure 1.5: The Unique Games problem can be formulated as a coloring problem: Given a graph, a set of colors and, for each edge a permutation between the colors, the goal is to color the vertices to satisfy as many of the edge constraints as possible. The Unique Games Conjecture [138] states that, for any $\varepsilon, \delta > 0$, it is hard to distinguish between instances on which it is possible to satisfy a $(1-\varepsilon)$ fraction of the edge constraints from instances on which it is impossible to satisfy a $\delta$ fraction. Image courtesy of Wikipedia, uploaded by Thore Husfeldt.

One can view the permutations as giving pairwise information between the nodes labels. In fact, the semidefinite programming based approach that we will develop in Section 1.3.3 is in part motivated by the algorithmic approach used in the best known polynomial time approximation to the Unique Games Problem [75].

### Matching in Computer Graphics

This framework has been successfully used by Chen, Huang, and Guibas [130, 79] in the joint shape matching problem in Computer Graphics. The shape matching problem consists of estimating a map between feature points of two different shapes (perhaps corresponding to same object in different positions, and the goal is to identify which points of one of the shapes corresponds to which point in the other — see Figure 1.6). The idea of joint estimation is to use cycle consistency to help improve the maps. An estimation of this type for a pair of images can be thought of as a pairwise measurement between the labels of the feature points of the two images involved and the approach of join estimation can be formulated as a synchronization-type problem.

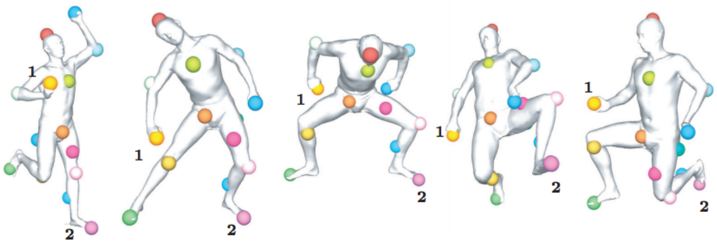

Figure 1.6: An important problem in Computer Graphics is that of Shape Matching: given different shapes of an object, the goal is to match feature points between shapes. Image courtesy of Qi-Xing Huang and Leonidas Guibas [130].

### Euclidean Clustering

The problem of clustering a point cloud in a Euclidean space can also be formulated as a synchronization problem. If the points are to be clustered in $L$ clusters then one can think of the task as assigning, to each data point $x_i$ a cluster label $l_i \in \mathbb{Z}_L$ based on pairwise distances between points. In fact, the min-sum k-clustering objective [239] can be formulated as an instance of Problem 1.2.1 by taking, for each pair of points $x_i$ and $x_j$, $f_{ij}(l_i - l_j)$ to be equal to the distance squared between $x_i$ and $x_j$ if $l_i \neq l_j$

mod $L$ and 0 otherwise. This will be discussed in more detailed in Section 6.2.

### Sensor network localization

The sensor network localization (or distance geometry) problem [89, 211], consists in estimating the position of a set of sensors from measurements of distances between pairs of them.

### Finding ground state of particles

Another important problem in this framework is in theoretical chemistry, the problem of finding the ground state of a set of atoms. This is achieved by minimizing a potential that depends on the locations of all particles. In many instances, the potential is given by summing, over all pairs, the Lennard-Jones potential [145] between two of the atoms (which depends only on which atoms they are, and their distance). By taking, for each pair of atoms $i$ and $j$, $f_{ij}$ to be a function of their distance representing the Lennard-Jones potential, we can formulate this problem as an instance of Problem 1.2.1.

Other instances include the ranking problem [133], camera motion estimation in computer vision [9], and many others.

## 1.3 Algorithmic approaches

In this section we will briefly describe several algorithmic approaches for these problems. Before describing both the spectral and semidefinite programming based approaches that we investigate in this thesis, we take make a few comments regarding model bias.

### 1.3.1 The model bias pitfall

In some of the problems described above, such as the multireference alignment of signals (or the orientation estimation problem in Cryo-EM), the alignment step is only a subprocedure of the estimation of the underlying signal (or the 3d density of the molecule). In fact, if the underlying signal was known, finding the shifts would be nearly trivial: for the case of the signals, one could simply use match-filtering to find the most likely shift $l_i$ for measurement $y_i$.

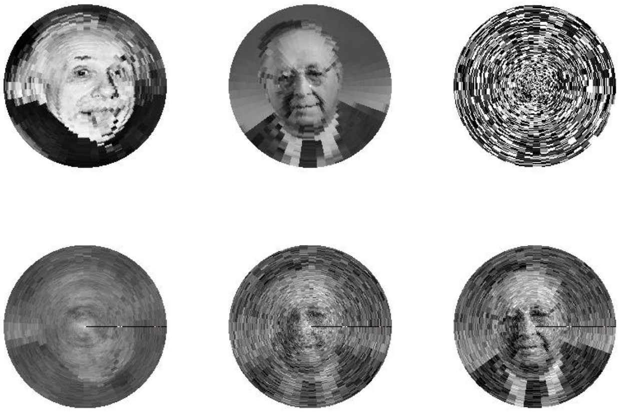

Figure 1.7: A simple experiment to illustrate the model bias phenomenon: Given a picture of the mathematician Hermann Weyl (second picture of the top row) we generate many images consisting of random rotations (we considered a discretization of the rotations of the plane) of the image with added gaussian noise. An example of one such measurements is the third image in the first row. We then proceeded to align these images to a reference consisting of a famous image of Albert Einstein (often used in the model bias discussions). After alignment, an estimator of the original image was constructed by averaging the aligned measurements. The result, first image on second row, clearly has more resemblance to the image of Einstein than to that of Weyl, illustration the model bias issue. One the other hand, the method proposed and analyzed in Chapter 6 produces the second image of the second row, which shows no signs of suffering from model bias. As a benchmark, we also include the reconstruction obtained by an oracle that is given the true rotations (third image in the second row).

When the true signal is not known, a common approach is to choose a reference signal $z$ that is not the true template but believed to share some properties with it. Unfortunately, this creates a high risk of model bias: the reconstructed signal $\hat{u}$ tends to capture characteristics of the reference $z$ that are not present on the actual original signal $u$ (see Figure 1.3.1 for an illustration of this phenomenon). This issue is well known among the biological imaging community [200, 127] (see, for example, [84] for a particularly recent discussion of it). As the experiment shown on Figure 1.3.1 suggests, the methods treated in this paper, based solely on pairwise information between observations, do not suffer from model bias as they do not use any information besides the data itself.

### 1.3.2 Spectral methods

Spectral graph partitioning, the use of the smallest eigenvector of the Laplacian matrix to partition a graph [120], is a standard technique for clustering with analysis both in worst case, through the celebrated Cheeger's inequality [15, 17], and for random instances [161]. The `Max-Cut` problem can also be tackled with a spectral method [219].

Analogously, for angular synchronization, one can construct a matrix, the Connection Laplacian [206, 202], whose eigenvectors associated with the smallest eigenvalues would provide the group potential if it were possible to satisfy all the edge constraints. Under a model of random noise that prevents such a solution, a good solution can still be obtained by rounding the smallest eigenvectors [202] (an analysis analogous to the analysis of spectral partitioning for random instances [161]). Similar spectral based algorithms were proposed in [89, 204] for $SO(3)$. In Section 2.1 we establish a Cheeger inequality for this operator, providing a worst-case analysis for the spectral approach to some synchronization-type problems.

We refer to Section 2.1 for both a more detailed description of spectral methods for $O(d)$ synchronization, and a description of Cheeger's inequality for spectral graph partitioning.

### 1.3.3 The semidefinite programming approach

Most of this thesis concerns the use of semidefinite relaxations to solve synchronization-type problems. The use of semidefinite relaxations in combinatorial optimization dates back to the late 1970s with the seminal work of László Lovász [150] in the so-called Lovász theta function, this approach was shortly after made algorithmic in [119]. In first half of the 1990s, interior point methods were adapted to solve semidefinite programs [13, 175], providing reasonably efficient methods to solve this type of problems. In 1995, Michel Goemans and David Williamson, devised the first approximation algorithm based on semidefinite programming [114]. Their algorithm gave the best known approximation ratio to the `Max-Cut` problem. Ever since, many approximation algorithms have been designed based on semidefinite programming. In fact, the algorithm we will analyze is greatly inspired by the semidefinite relaxation in [114]. Remarkably, an important conjecture of Khot [138] is known to imply that, for a large class of problems including `Max-Cut`, this approach produces optimal approximation ratios [185].

An approximation ratio [238] is a guarantee that, for any possible instance of the input, the algorithm outputs a solution whose performance is at least a certain fraction (the approximation ratio) of the optimal one, we will establish such a guarantee for the Procrustes problem (among others) in Section 2.3. However, the worst-case nature of this type of guarantee is often pessimistic. A popular alternative is to equip the input with a distribution and give guarantees for most inputs. More precisely, Chapter 3 is concerned with understanding when is it that the semidefinite relaxation approach gives exactly the correct answer (for most inputs). In Chapter 5 we investigate the tendency for a large class of semidefinite relaxations to be *tight*, i.e. the optimal solution of a semidefinite relaxation is the optimal solution of the original problem, and also establish such a phenomenon in instances for which exact recovery is unrealistic.

### A general semidefinite programming based approach for synchronization type problems

We now describe a general semidefinite programming based approach to solving Problem 1.2.1. This is inspired by the algorithms in [114, 75, 32] and we will revisit it in Chapter 6.

We will restrict our attention to compact groups $\mathcal{G}$. We start by considering a representation $\rho$ of $\mathcal{G}$ into $O(L)$, the group of $L \times L$ orthogonal matrices,

$$\rho : \mathcal{G} \to O(L). \tag{1.4}$$

The key point is that we want a representation $\rho$ for which, $f_{ij}\left(g_i g_j^{-1}\right)$ in (1.1) is a linear function of $\rho\left(g_i\right)\rho\left(g_j^{-1}\right)$. More precisely, identifying $\rho(g_i)$ with $\rho_i$, we need there to exist $C_{ij} \in \mathbb{R}^{L\times L}$ such that

$$f_{ij}\left(g_i g_j^{-1}\right) = \operatorname{Tr}\left(C_{ij}\rho_i\rho_j^T\right),$$

recall that $\rho_j^{-1} = \rho_j^T$, as $\rho_j \in O(L)$. Note that if $f_{ij}$ is an affine function instead, which will happen in some of the settings, it can be made linear by an additive shift which will not affect the optimizers of the problem.

We can then rewrite (1.1) as

$$\begin{array}{ll} \min & \sum_{(i,j)\in E} \operatorname{Tr}\left(C_{ij}\rho_i\rho_j^T\right) \\ \text{s. t.} & \rho_i \in \operatorname{Im}(\rho),\ \forall_i. \end{array} \tag{1.5}$$

Note that, if $\mathcal{G}$ is finite, such a representation $\rho$ always exists, perhaps by taking the representation of $\mathcal{G}$ to the group of permutations of its own elements (which always exists by Cayley's Theorem [190]).

We proceed by making use of the now standard lifting trick, to formulate (1.5) in terms of

$$X = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix} \begin{bmatrix} \rho_1^T & \rho_2^T & \cdots & \rho_n^T \end{bmatrix} \in \mathbb{R}^{NL \times NL}. \tag{1.6}$$

We use $X_{ij} \in \mathbb{R}^{L\times L}$ to denote the $(i,j)$-th $L \times L$ block of $X$.

It is clear that there exists $C \in \mathbb{R}^{NL\times NL}$ for which

$$\sum_{(i,j)\in E} \operatorname{Tr}\left(C_{ij}\rho_i\rho_j^T\right) = \operatorname{Tr}(CX),$$

meaning that we can rewrite (1.5) as

$$\begin{aligned} \min \quad & \operatorname{Tr}(CX) \\ \text{s. t.} \quad & X \text{ is of the form of (1.6)} \\ & \ \text{with } \rho_i \in \operatorname{Im}(\rho),\ \forall_i. \end{aligned} \tag{1.7}$$

Note that, if $X$ is of the form of (1.6) with $\rho_i \in \operatorname{Im}(\rho),\ \forall_i$, then we must have that $X \succeq 0$, $\operatorname{rank}(X) \leq L$, $X_{ii} = \rho_i\rho_i^T = I_{L\times L}$, and $X_{ij} = \rho_i\rho_j^T \in \operatorname{Im}(\rho)$. We will replace the constraints in (1.7) by these

$$\begin{aligned} \min \quad & \operatorname{Tr}(CX) \\ \text{s. t.} \quad & X_{ii} = I_{L\times L} \\ & X_{ij} \in \operatorname{Im}(\rho) \\ & X \succeq 0 \\ & \operatorname{rank}(X) \leq L. \end{aligned} \tag{1.8}$$

We then relax (1.8) by removing the non-convex rank constraint and by relaxing the $X_{ij} \in \mathrm{Im}(\rho)$ to the linear constraint $X_{ij} \in \mathrm{aff}\big(\mathrm{Im}(\rho)\big)$.

$$\begin{array}{ll} \min & \mathrm{Tr}(CX) \\ \text{s. t.} & X_{ii} = I_{L\times L} \\ & X_{ij} \in \mathrm{aff}\big(\mathrm{Im}(\rho)\big) \\ & X \succeq 0. \end{array} \tag{1.9}$$

(1.9) is a semidefinite program (SDP) and can be solved, up to arbitrary precision, in polynomial time using interior point methods [227].[3]

In the proceeding chapters we will analyze several instances of (1.9) in detail. For now we remark that when $\mathcal{G} \cong \mathbb{Z}_2$, the natural representation to consider is $\{\pm 1\}$ and so (1.9) coincides with the classical SDP for `Max-Cut` [114] which we will investigate, in detail, in Chapter 3. We note that, in this case, the constraint $X_{ij} \in \mathrm{aff}\big(\pm 1\big)$ simply corresponds to $-1 \leq X_{ij} \leq 1$ and is a redundant constraint.

Since we require the functions $f_{ij}$ to be linear in the representation, different representations will allow for different objective functions. On the other hand, the dimension $L$ will greatly affect the computational complexity of solving (1.9). This presents an important trade-off that will be discussed in Chapter 6 (see Remark 6.1.1). In the same Chapter, we will also describe how to exploit properties of certain representations to speed-up the solution of the SDP (see Remark 6.1.4).

## 1.4 Notation

We make use of several standard matrix and probability notation. For $M$ a matrix we denote its $k$-th smallest eigenvalue by $\lambda_k(M)$, largest eigenvalue by $\lambda_{\max}(M)$, its spectral norm by $\|M\|$ and its Frobenius norm by $\|M\|_F$. $\|M\|_{\mathrm{e},\infty} = \max_{i,j} |M_{ij}|$ is the entry-wise $\ell_\infty$ norm (largest entry in absolute value). Assuming further that

[3] For large scale problems it tends to be preferable to use methods based in alternating direction method of multipliers (ADMM) such as the one described in [236].

$M$ is positive semi-definite we define the $M$-inner product of vectors $x$ and $y$ as $\langle x, y\rangle_M = x^T M y$ (and say that two vectors are $M$-orthogonal if this inner product is zero). Also, we define the $M$-norm of $x$ as $\|x\|_M = \sqrt{\langle x, x\rangle_M}$. $\operatorname{diag}(M)$ is used to refer to a vector with the diagonal elements of $M$ as entries. For $x \in \mathbb{R}^n$ a vector, $\operatorname{diag}(x)$ denotes a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with $D_{ii} = x_i$. For a matrix $M$, $\operatorname{ddiag}(M) = \operatorname{diag}\left(\operatorname{diag}\left(M\right)\right)$. $\mathbf{1}_n$ denotes the all-ones vector in $\mathbb{R}^n$, we will drop the subscript whenever there is no risk of ambiguity. We define $D_M$ as a diagonal matrix whose diagonal entries are given by $D_{ii} = diag(M\,\mathbf{1}) = \sum_{j=1}^n M_{ij}$ and $L_M$ a Laplacian matrix defined as $L_M = D_M - M$.

For a scalar random variable $Y$, we refer to its $p$-norm as $\|Y\|_p = \left(\mathbb{E}|Y|^p\right)^{1/p}$ and infinity norm as $\|Y\|_\infty = \inf\{a : \ |Y| \le a \text{ a.s.}\}$. $a \lesssim b$ means that $a \le Cb$ for a universal constant $C$. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$. We write $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$.

Given a graph, $\deg(i)$ will be used to denote the degree of node $i$. In the case of the stochastic block model, $\deg_{in}(i)$ will be used for inner-cluster degree and $\deg_{out}(i)$ for outer-cluster degree.

For a complex scalar $a \in \mathbb{C}$, $\overline{a}$ denotes its complex conjugate and $|a| = \sqrt{a\overline{a}}$ its modulus. $\Re(M)$ and $\Im(M)$ extract, respectively, the real and imaginary parts of a matrix (or a vector, or a scalar). $\operatorname{aff}(S)$ denotes the affine hull of a set $S$ and $\operatorname{Im}(f)$ the image of a function $f$. $\mathbb{Z}_n$ denotes the cyclic group on $n$ elements and $[n] := \{1, \ldots, n\}$.

We say that an event $\mathcal{E}$ happens with high probability when

$$\mathbb{P}\left[\mathcal{E}\right] = 1 - n^{-\Omega(1)},$$

where $n$ is an underlying parameter that is thought of going to infinity, such as the dimension of the matrices or the number of nodes in the graphs being studied.

# Chapter 2

# Synchronization over groups of orthogonal matrices: Worst-case guarantees

## 2.1 A Cheeger inequality for the Connection Laplacian

This section is devoted to a spectral method for $O(d)$ Synchronization and a worst case guarantee for it, through a Cheeger inequality for the connection Laplacian [41]. Our approach takes inspiration from [202], who proposed a version of this method for angular synchronization, from techniques used to establish guarantees for a spectral method for `Max-Cut` [219], and from the celebrated Cheeger's inequality for partitioning [15, 17, 77]. This section is mostly based on [41].

### Cheeger's Inequality and the Graph Laplacian

Before considering the synchronization problem we will briefly present the classical graph Cheeger's inequality in the context of spectral partitioning. The material presented here is well known but it will help motivate the ideas that follow.

Let $G = (V, E)$ be an undirected weighted graph with $n$ vertices. We will now consider the problem of partitioning the vertices in two similarly sized sets in a way that minimizes the cut: the volume of edges across the subsets (of the partition). There are several ways to measure the performance of a particular partition of the graph. For now, we will consider the one known as the Cheeger constant. Given a partition $(S, S^c)$ of $V$ let

$$h_S := \frac{\mathrm{cut}(S)}{\min\{\mathrm{vol}(S), \mathrm{vol}(S^c)\}}, \tag{2.1}$$

where the value of the cut associated with $S$ is $cut(S) = \sum_{i\in S}\sum_{j\in S^c} w_{ij}$, its volume is $vol(S) = \sum_{i\in S} \deg(i)$, and $\deg(i) = \sum_{j\in V} w_{ij}$ is the weighted degree of vertex $i$. We want to partition the graph so that $h_S$ is minimized, and the minimum value is referred to as the Cheeger number of the graph, denoted $h_G = \min_{S\subset V} h_S$. Finding the optimal $S$ is known to be NP-hard, as it seems to require searching over an exponential number of possible partitions.

There is another way to measure the performance of a partition $(S, S^c)$ known as the normalized cut:

$$\mathrm{Ncut}(S) = \mathrm{cut}(S)\left(\frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(S^c)}\right).$$

As before, we want to find a subset with as small of an Ncut as possible. Note that the normalized cut and the Cheeger constant are closely related:

$$\frac{1}{2}\mathrm{Ncut}(S) \le h_S \le \mathrm{Ncut}(S).$$

Let us introduce a few important definitions. Let $A$ be the weighted adjacency matrix of $G$ and $D_G$ the degree matrix, a diagonal matrix with elements $\deg(i)$. If we consider a vector $f \in \mathbb{R}^n$ whose entries take only 2 possible values, one associated with vertices in $S$ and another in $S^c$, then the quadratic form $Q_f = \frac{1}{2}\sum_{ij} w_{ij}\left(f_i - f_j\right)^2$ is of fundamental importance as a measure of the cut between the sets. The symmetric positive semi-definite matrix that corresponds to this quadratic form, $L_A$, is known as the graph Laplacian of $G$. It is defined as $L_G = D_G - A$ and satisfies $v^T L_G v = Q_v$ for any $v \in \mathbb{R}^n$. It is also useful to consider the normalized graph Laplacian $\mathcal{L}_G = D_G^{-1/2} L_G D_G^{-1/2} = \mathrm{I} - D_G^{-1/2} W_G D_G^{-1/2}$, which is also a symmetric positive semi-definite matrix.

Let us represent a partition $(S, S^c)$ by a cut-function $f_S : V \to \mathbb{R}$ given by

$$f_S(i) = \begin{cases} \sqrt{\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}} & \text{if} \quad i \in S, \\ -\sqrt{\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}} & \text{if} \quad i \in S^c. \end{cases}$$

It is straightforward to show that $Q_{f_S} = f_S^T L_G f_S = \mathrm{Ncut}(S)$, $f_S^T D_G f_S = 1$, and $f_S^T D_G \mathbf{1} = 0$, where $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^n$. This is the motivation for a spectral method to approximate the minimum normalized cut problem. If we drop the constraint that $f$ needs to be a cut-function and simply enforce the properties established above then one would formulate the following relaxed problem

$$\min_{f:V\to\mathbb{R}, f^T D_G f = 1, f^T D_G \mathbf{1} = 0} f^T L_G f. \tag{2.2}$$

Since $\mathbf{1}^T L_G \mathbf{1} = 0$, we know by the Courant-Fisher formula that (2.2) corresponds to an eigenvector problem whose minimum is $\lambda_2(\mathcal{L}_G)$ and whose minimizer can be obtained by the corresponding eigenvector.

Since problem (2.2) is a relaxation of the minimum Ncut problem we automatically have $\frac{1}{2}\lambda_2(\mathcal{L}_G) \leq \frac{1}{2}\min_{S \subset V} \mathrm{Ncut} \leq h_G$. Remarkably one can show that the relaxation is not far from the partitioning problem. In fact, one can round the solution of (2.2) so that it corresponds to a partition $(S, S^c)$ of $G$, whose $h_S$ we can control. This is made precise by the following classical result in spectral graph theory (several different proofs for this inequality can be found in [81]):

**Theorem 2.1.1** (Cheeger Inequality [15, 17]). *Let $G = (V, E)$ be a graph and $\mathcal{L}_G$ its normalized graph Laplacian. Then*

$$\frac{1}{2}\lambda_2(\mathcal{L}_G) \leq h_G \leq \sqrt{2\lambda_2(\mathcal{L}_G)},$$

*where $h_G$ is the Cheeger constant of $G$. Furthermore, the bound is constructive: using the solution of the eigenvector problem one can produce partition $(S, S^c)$ that achieves the upper bound $\sqrt{2\lambda_2(\mathcal{L}_G)}$, such construction is referred to as the* spectral clustering *algorithm.*

On way to view spectral clustering is through the intuition that a random walk on a graph tends to be trapped in sections of the graph which have few connections to the rest of the vertices (this intuition is made more explicit in [156, 170]).

### Frustration, Vector-Valued Walks and the Connection Laplacian

If, in addition to a graph, we are given an orthogonal transformation $\rho_{ij} \in O(d)$ for each edge $(i, j) \in E$, we can consider a random walk that takes the transformations into account. One way of doing this is by defining a random walk that, instead of moving point masses, moves a vector from vertex to vertex and transforms it via the orthogonal transformation associated with the edge. One can similarly define a random walk that moves group elements on vertices. The Connection Laplacian was defined by Singer and Wu [206] to measure the convergence of such random walks. The construction requires that $\rho_{ji} = \rho_{ij}^{-1} = \rho_{ij}^T$. Define the symmetric matrix $A_{(C)} \in \mathbb{R}^{dn \times dn}$ so that the $(i, j)$-th $d \times d$ block is given by $(A_{(C)})_{ij} = w_{ij}\rho_{ij}$, where $w_{ij}$ is the weight of the edge $(i, j)$. Also, let $D_{(C)} \in \mathbb{R}^{dn \times dn}$, be the diagonal matrix such that $(D_{(C)})_{ii} = \deg(i) I_{d\times d}$. We assume $\deg(i) > 0$, for every $i$. The graph

Connection Laplacian $L_{(C)}$ is defined to be $L_{(C)} = D_{(C)} - A_{(C)}$, and the normalized graph Connection Laplacian is

$$\mathcal{L}_{(C)} = \mathrm{I} - D_{(C)}^{-1/2} A_{(C)} D_{(C)}^{-1/2}.$$

If $v : V \to \mathbb{S}^{d-1}$ assigns a unit vector in $\mathbb{R}^d$ to each vertex, we may think of $v$ as a vector in $dn$ dimensions. In this case the quadratic form

$$v^T L_{(C)} v = \sum_{(i,j)\in E} w_{ij} \left\| v_i - \rho_{ij} v_j \right\|^2 = \frac{1}{2} \sum_{i,j} w_{ij} \left\| v_i - \rho_{ij} v_j \right\|^2$$

is a measure of how well $v$ satisfies the edges. This will be zero if $v_i = \rho_{ij} v_j$ for all edges $(i,j)$. As $w_{ij} = 0$ when $(i,j) \notin E$, we can sum over all pairs of vertices without loss of generality. An assignment satisfying all edges will correspond to a stationary distribution in the vector-valued random walk.

Following our analogy with Cheeger's inequality for the normalized graph Laplacian, we normalize this measure by defining the *frustration* of $v$ as

$$\eta(v) = \frac{v^T L_1 v}{v^T D_1 v} = \frac{1}{2} \frac{\sum_{i,j} w_{ij} \left\| v_i - \rho_{ij} v_j \right\|^2}{\sum_i d_i \left\| v_i \right\|^2}. \tag{2.3}$$

We then define the $\mathbb{S}^{d-1}$ frustration constant of $G$ as

$$\eta_G = \min_{v : V \to \mathbb{S}^{d-1}} \eta(v). \tag{2.4}$$

The smallest eigenvalue of $\mathcal{L}_{(C)}$ provides a relaxation of $\eta_G$, as

$$\begin{aligned} \lambda_1(\mathcal{L}_{(C)}) &= \min_{z\in\mathbb{R}^{dn}} \frac{z^T \mathcal{L}_{(C)} z}{z^T z} = \min_{x\in\mathbb{R}^{dn}} \frac{(D_{(C)}^{\frac{1}{2}} x)^T \mathcal{L}_{(C)} (D_{(C)}^{\frac{1}{2}} x)}{(D_{(C)}^{\frac{1}{2}} x)^T (D_{(C)}^{\frac{1}{2}} x)} \\ &= \min_{x\in\mathbb{R}^{dn}} \frac{x^T L_{(C)} x}{x^T D_{(C)} x} = \min_{x : V \to \mathbb{R}^d} \eta(x). \end{aligned}$$

If there is a group potential $g : V \to O(d)$ that satisfies all the edges (which would again correspond to a stationary distribution for the $O(d)$-valued random walk), then we can obtain $d$ orthogonal vectors on which the quadratic form defined by $\mathcal{L}_{(C)}$ is zero. For each $1 \leq k \leq d$ we obtain one of these vectors by setting $v(i)$ to the $k$th column of $g(i)$ for all $i \in V$. In particular, this means that the columns of the matrices of the group potential that satisfies all of the edges lie in the nullspace of $\mathcal{L}_{(C)}$. Since $g(i) \in O(d)$ these vectors are orthogonal. If $G$ is connected, one can show that these are the only vectors in the nullspace of $\mathcal{L}_{(C)}$. This observation is the motivation for the use of a spectral algorithm for synchronization.

We define the frustration of a group potential $g : V \to O(d)$ to be

$$\nu(g) = \frac{1}{2d}\frac{1}{\mathrm{vol}(G)}\sum_{i,j} w_{ij}\|g_i - \rho_{ij}g_j\|_F^2. \tag{2.5}$$

We then define the $O(d)$ frustration constant of $G$ to be

$$\nu_G = \min_{g:V\to O(d)} \nu(g).$$

In Theorem 2.1.10, we prove that this frustration constant is small if and only if the sum of the first $d$ eigenvalues of $\mathcal{L}_{(C)}$ is small as well.

Since, $\|g_i - \rho_{ij}g_j\|_F^2 = \|g_i g_j^{-1} - \rho_{ij}\|_F^2$, the frustration constant is simply the minimum of Problem (1.1) when $\mathcal{G} \cong O(d)$ and $f_{ij}(g_i g_j^{-1}) = \|g_i g_j^{-1} - \rho_{ij}\|_F^2$.

### 2.1.1 Cheeger's type inequalities for the synchronization problem

We will now describe the the main results of this section. We present three spectral algorithms to solve three different formulations of synchronization problems and obtain for each a guarantee of performance in the form of a Cheeger's type inequality.

We will briefly summarize both the results and the ideas to obtain them, referring the reader to [41] for the rigorous proofs.

We start by considering the $\mathbb{S}^{d-1}$ synchronization problem [1]. This corresponds to finding, for each vertex $i$ of the graph, a vector $v_i \in \mathbb{S}^{d-1}$ in way that for each edge $(i, j)$ the vectors agree with the edges, meaning $v_i = \rho_{ij} v_j$. Since this might not always be possible we look for a function $v : V \to \mathbb{S}^{d-1}$ for which the frustration $\eta(v)$ is minimum (see (2.3)). Motivated by an algorithm to solve `Max-Cut` by Trevisan [219], we first consider a version of the problem for which we allows ourselves to synchronize only a subset of the vertices, corresponding to the partial synchronization in $\mathbb{S}^{d-1}$. We then move on to consider the full synchronization problem in $\mathbb{S}^{d-1}$.

Finally we will present our main result, an algorithm for $O(d)$ synchronization and a Cheeger-like inequality that equips it with a worst-case guarantee. Recall that the $O(d)$ synchronization corresponds to finding an assignment of an element $g_i \in O(d)$ to each vertex $i$ in a way that minimizes the discrepancy with the pairwise measurements $\rho_{ij} \sim g_i g_j^{-1}$ obtained for each edge. This corresponds to minimizing the $O(d)$ frustration, $\nu(g)$, (see (2.5)).

In the sequel, given $x \in \left(\mathbb{R}^d\right)^n$, we will denote by $x_i$ the $i$-th $d \times 1$ block of $x$ (that will correspond to the value of $x$ on the vertex $i$) and, for any $u > 0$ we define $x^u$ as

$$x_i^u = \begin{cases} \frac{x_i}{\|x_i\|} & \text{if} \quad \|x_i\|^2 \geq u, \\ 0 & \text{if} \quad \|x_i\|^2 < u. \end{cases} \tag{2.6}$$

[1] Note that the $\mathbb{S}^{d-1}$ synchronization problem differs slightly from the other synchronization problems considered in this thesis as the node labels do not belong to the same group as the edge measurements, but as we will see it will provide a fruitful first step towards a guarantee for $O(d)$ Synchronization. Moreover $\mathbb{S}^2$ synchronization corresponds to angular synchronization.

Furthermore, for $u = 0$ we denote $x^0$ by $\tilde{x}$, that is,

$$\tilde{x}_i = \begin{cases} \frac{x_i}{\|x_i\|} & \text{if} \quad x_i \neq 0, \\ 0 & \text{if} \quad x_i = 0. \end{cases} \tag{2.7}$$

### Partial synchronization in $\mathbb{S}^{d-1}$

The motivation for considering a spectral relaxation for the synchronization problem in $\mathbb{S}^{d-1}$ is the observation that $\lambda_1(\mathcal{L}_{(C)}) = \min_{x:V\to\mathbb{R}^d} \eta(x)$. In order to understand how tight the relaxation is we need to relate $\lambda_1(\mathcal{L}_{(C)})$ with $\eta_G = \min_{x:V\to\mathbb{S}^{d-1}} \eta(x)$.

Consider, however, the following example: a graph consisting of two disjoint components, one whose $\rho_{ij}$ measurements are perfectly compatible and another one on which they are not. Its graph Connection Laplacian would have a non-zero vector in its null space, corresponding to synchronizations on the compatible component and zero on the incompatible part (thus $\lambda_1(\mathcal{L}_{(C)}) = 0$). On the other hand, the constraint that $v$ has to take values on $\mathbb{S}^{d-1}$, will force it to try to synchronize the incompatible part thereby bounding $\eta_G$ away from zero. This example motivates a different formulation of the $\mathbb{S}^{d-1}$ synchronization problem where vertices are allowed not to be labeled (labeled with 0). We thus define the partial $\mathbb{S}^{d-1}$ frustration constant of $G$, as the minimum possible frustration value for such an assignment,

$$\eta_G^* = \min_{v:V\to\mathbb{S}^{d-1}\cup\{0\}} \eta(v). \tag{2.8}$$

We propose the following algorithm to solve the partial $\mathbb{S}^{d-1}$ synchronization problem.

**Algorithm 2.1.2.** *Given a graph $G = (V, E)$ and a function $\rho : E \to O(d)$, construct the normalized Connection Laplacian $\mathcal{L}_{(C)}$ and the degree matrix $D_{(C)}$. Compute $z$, the eigenvector corresponding to the smallest eigenvalue of $\mathcal{L}_{(C)}$. Let $x = D_{(C)}^{-\frac{1}{2}} z$. For each vertex index $i$, let $u_i = \|x_i\|$, and set $v^i : V \to \mathbb{S}^{d-1} \cup \{0\}$ as $v^i = x^{u_i}$, according*

*to (2.6). Output $v$ equal to the $v^i$ that minimizes $\eta\left(v^i\right)$.*

We refer the reader to [41] for a proof of the following Lemma.

**Lemma 2.1.3.** *Given $x \in \mathbb{R}^{dn}$ there exists $u > 0$ such that*

$$\eta(x^u) \leq \sqrt{10\eta(x)}.$$

*Moreover, if $d = 1$ the right-hand side can be replaced by $\sqrt{8\eta(x)}$.*

We note that it guarantees that the solution $v$ given by Algorithm 2.1.2 satisfies $\eta(v) \leq \sqrt{10\eta(x)}$. Since $x$ was computed so that $\eta(x) = \lambda_1(\mathcal{L}_{(C)})$, Algorithm 2.1.2 is guaranteed to output a solution $v$ such that

$$\eta(v) \leq \sqrt{10\lambda_1(\mathcal{L}_{(C)})}.$$

Note that $\lambda_1(\mathcal{L}_{(C)}) \leq \eta_G^*$, which is the optimum value for the partial $\mathbb{S}^{d-1}$ synchronization problem (see (2.8)). The idea to show that the rounding, from $x$, to the solution $v$ done by Algorithm 2.1.2 produces a solution with $\eta(v) \leq \sqrt{10\eta(x)}$ is to use the probabilistic method. One considers a random rounding scheme by rounding $x$ as in Algorithm 2.1.2 and (2.6) but thresholding at a random value $u$, drawn from a well-chosen distribution. One then shows that, in expectation, the frustration of the rounded vector is bounded by $\sqrt{10\eta(x)}$. This automatically ensures that there must exist a value $u$ that produces a solution with frustration bounded by $\sqrt{10\eta(x)}$. The rounding described in Algorithm 2.1.2 runs through all possible such roundings and is thus guaranteed to produce a solution satisfying the bound (we refer the reader to [41] for a proof of Lemma 2.1.3). An $O(1)$ version of this algorithm and analysis appeared in [219], when $\rho$ is the constant function equal to $-1$, in the context of the Max-Cut problem. In fact, if $d = 1$ the factor 10 can be substituted by 8 and the stronger inequality holds $\eta(v) \leq \sqrt{8\lambda_1(\mathcal{L}_{(C)})}$.

The above performance guarantee for Algorithm 2.1.2 automatically implies the following Cheeger-like inequality.

**Theorem 2.1.4.** *Let $G = (V, E)$ be a weighted graph. Given a function $\rho : E \to O(d)$, let $\eta_G^*$ be the partial $\mathbb{S}^{d-1}$ frustration constant of $G$ and $\lambda_1(\mathcal{L}_{(C)})$ the smallest eigenvalue of the normalized graph Connection Laplacian. Then*

$$\lambda_1(\mathcal{L}_{(C)}) \leq \eta_G^* \leq \sqrt{10\lambda_1(\mathcal{L}_{(C)})}. \tag{2.9}$$

*Furthermore, if $d = 1$, the stronger inequality holds, $\eta_G^* \leq \sqrt{8\lambda_1(\mathcal{L}_{(C)})}$.*

We note that Trevisan [219], in the context of MAX-CUT, iteratively performs this partial synchronization procedure in the subgraph composed of the vertices left unlabeled by the previous iteration, in order to label the entire graph. We, however, consider only one iteration.

### Full synchronization in $\mathbb{S}^{d-1}$

In this section we adapt Algorithm 2.1.2 to solve (full) synchronization in $\mathbb{S}^{d-1}$ and show performance guarantees, under reasonable conditions, by obtaining bounds for $\eta_G$, the frustration constant for synchronization in $\mathbb{S}^{d-1}$. The intuition given to justify the relaxation to partial $\mathbb{S}^{d-1}$ synchronization was based on the possibility of poor connectivity of the graph (small spectral gap). In this section we show that poor connectivity, as measured by a small spectral gap in the normalized graph Laplacian, is the only condition under which one can have large discrepancy between the frustration constants and the spectra of the graph Connection Laplacian. We will show that, as long as the spectral gap is bounded away from zero, one can in fact control the full frustration constants.

**Algorithm 2.1.5.** *Given a weighted graph $G = (V, E)$ and a function $\rho : E \to O(d)$, construct the normalized Connection Laplacian $\mathcal{L}_{(C)}$ and the degree matrix $D_{(C)}$.*

*Compute $z$, the eigenvector corresponding to the smallest eigenvalue of $\mathcal{L}_{(C)}$. Let $x = D_{(C)}^{-\frac{1}{2}} z$. Output the solution $\tilde{x} : V \to \mathbb{S}^{d-1} \cup \{0\}$ where each $\tilde{x}_i$ is defined as*

$$\tilde{x}_i = \frac{x_i}{\|x_i\|}.$$

*If $x_i = 0$, have $\tilde{x}_i$ be any vector in $\mathbb{S}^{d-1}$.*

Similarly to Algorithm 2.1.2, the following lemma guarantees that the solution $\tilde{x}$ given by Algorithm 2.1.5 satisfies $\eta(\tilde{x}) \leq 44 \frac{1}{\lambda_2(\mathcal{L}_G)} \eta(x)$.

**Lemma 2.1.6.** *For every $x \in \mathbb{R}^{dn}$, $\eta(\tilde{x}) \leq \frac{44}{\lambda_2(\mathcal{L}_0)} \eta(x)$.*

Again, since $x$ was computed so that $\eta(x) = \lambda_1(\mathcal{L}_{(C)})$, then Algorithm 2.1.2 is guaranteed to output a solution $v$ such that

$$\eta(v) \leq 44 \frac{\lambda_1(\mathcal{L}_{(C)})}{\lambda_2(\mathcal{L}_G)}.$$

Recall that, trivially, $\lambda_1(\mathcal{L}_{(C)}) \leq \eta_G$, which is the optimum value for the (full) $\mathbb{S}^{d-1}$ synchronization problem (see (2.4)). We refer the reader to [41] for the proof for Lemma 2.1.6. The idea here is to look at the vector of the local norms of $x$: $n_x \in \mathbb{R}^n$ where $n_x(i) = \|x_i\|$. It is not hard to show that $\frac{n_x^T L_G n_x}{n_x^T D_G n_x} \leq \eta(x)$, which means that, if $\eta(x)$ is small then $n_x$ cannot vary much between two vertices that share an edge. Since $\lambda_2(\mathcal{L}_G)$ is large one can show that such a vector needs to be close to constant, which means that the norms of $x$ across the vertices are similar. If the norms were all the same then the rounding $v_i = \frac{x_i}{\|v_i\|}$ would not affect the value of $\eta(\cdot)$, we take this slightly further by showing that if the norms are similar then we can control how much the rounding affects the penalty function.

The above performance guarantee for Algorithm 2.1.5 automatically implies another Cheeger-like inequality.

**Theorem 2.1.7.** *Let $G = (V, E)$ be a graph. Given a function $\rho : E \to O(d)$, let $\eta_G$ be the $\mathbb{S}^{d-1}$ frustration constants of $G$, $\lambda_1(\mathcal{L}_{(C)})$ the smallest eigenvalue of the normalized graph Connection Laplacian and $\lambda_2(\mathcal{L}_G)$ the second smallest eigenvalue of the normalized graph Laplacian. Then,*

$$\lambda_1(\mathcal{L}_{(C)}) \leq \eta_G \leq 44 \frac{\lambda_1(\mathcal{L}_{(C)})}{\lambda_2(\mathcal{L}_G)}.$$

### The $O(d)$ synchronization problem

We present now the main contribution of this section, a spectral algorithm for $O(d)$ synchronization together with a Cheeger-type inequality that provides a worst-case performance guarantee for the algorithm.

Before presenting the algorithm let us note the differences between this problem and the $\mathbb{S}^{d-1}$ synchronization problem, presented above. For the $\mathbb{S}^{d-1}$ case, the main difficulty that we faced in trying to obtain candidate solutions from eigenvectors was the local unit norm constraint. This is due to the fact that the synchronization problem requires its solution to be a function from $V$ to $\mathbb{S}^{d-1}$, corresponding to a vector in $\mathbb{R}^{dn}$ whose vertex subvectors have unit norm, while the eigenvector, in general, does not satisfy such a constraint. Nevertheless, the results in the previous section show that, by simply rounding the eigenvector, one does not lose more than a linear term, given that the graph Laplacian has a spectral gap bounded away from zero.

However, the $O(d)$ synchronization setting is more involved. The reason being that, besides the local normalization constraint, there is also a local orthogonality constraint (at each vertex, the $d$ vectors have to be orthogonal so that they can be the columns of an orthogonal matrix). For $\mathbb{S}^{d-1}$ we locally normalized the vectors, by choosing for each vertex the unit vector closest to $x_i$. For $O(d)$ synchronization we will pick, for each vertex, the orthogonal matrix closest (in the Frobenius norm) to the matrix $\left[x_i^1 \cdots x_i^d\right]$, where $x_i^j$ corresponds to the $d-$dimensional vector assigned to vertex $i$ by the $j$'th eigenvector. This rounding can be achieved by the Polar decomposition. Given a $d \times d$ matrix $X$, the matrix $\mathcal{P}(X)$, solution of $\min_{O \in O(d)} \|O - X\|_F$, is one of the components of the Polar decomposition of $X$ (see [128, 147] and references therein). We note that $\mathcal{P}(X)$ can be computed efficiently through the SVD decomposition of $X$. In fact, given the SVD of $X$, $X = U\Sigma V^T$, the closest orthogonal matrix to $X$ is given by $\mathcal{P}(X) = UV^T$ (see [128]). This approach is made precise in the following spectral algorithm for $O(d)$-synchronization.

**Algorithm 2.1.8.** *Given a weighted graph $G = (V, E)$ and a function $\rho : E \to O(d)$, construct the normalized Connection Laplacian $\mathcal{L}_{(C)}$ and the degree matrix $D_{(C)}$. Compute $z^1, \dots, z^d$, the first d eigenvectors corresponding to the d smallest eigenvalues of $\mathcal{L}_{(C)}$. Let $x^j = D_{(C)}^{-\frac{1}{2}} z^j$, for each $j = 1, \dots, d$. Output the solution $g : V \to O(d)$ where each $g_i$ is defined as*

$$g_i = \mathcal{P}(X_i),$$

*where $X_i = \left[x_i^1 \cdots x_i^d\right]$ and $\mathcal{P}(X_i)$ is the closest orthogonal matrix of $X_i$, which can be computed via the SVD of $X_i$, if $X_i = U_i \Sigma_i V_i^T$, then $\mathcal{P}(X_i) = U_i V_i^T$. If $X_i$ is singular*[2] *simply set $\mathcal{P}(X_i)$ to be $\mathrm{I}_d$.*

Similarly to how the performance of the $\mathbb{S}^{d-1}$ synchronization algorithms was obtained, the following lemma bounds the effect of the rounding step in Algorithm 2.1.8.

**Lemma 2.1.9.** *Given $x^1, \dots, x^d \in \mathbb{R}^{dn}$ such that $\langle x^k, x^l \rangle_{D_1} = 0$ for all $k \neq l$, consider the potential $g : V \to O(d)$ given as $g_i = \mathcal{P}\left(X_i\right)$ where $X_i = \left[x_i^1 \cdots x_i^d\right]$ and $\mathcal{P}(X)$ is the closest (in the Frobenius norm) orthogonal matrix of $X$. If $X_i$ is singular $\mathcal{P}(X_i)$ is simply set to be $\mathrm{I}_d$. Then,*

$$\nu(g) \leq \left(2d^{-1} + 2^{10} d^3\right) \frac{1}{\lambda_2(\mathcal{L}_0)} \sum_{i=1}^{d} \eta\left(x^i\right).$$

[2] In this case the uniqueness of $\mathcal{P}(X_i)$ is not guaranteed and thus the map is not well-defined.

Before rounding, the frustration of the solution $\left[x^1 \cdots x^d\right]$ is $\frac{1}{d}\sum_{i=1}^d \eta\left(x^i\right)$. Lemma 2.1.9 guarantees that the solution $g$ obtained by the rounding in Algorithm 2.1.8 satisfies $\nu(g) \leq 1026d^3 \frac{1}{\lambda_2(\mathcal{L}_G)} \sum_{i=1}^d \eta\left(x^i\right)$. Because of how the vectors $x^1, \ldots, x^d$ were built, $\sum_{i=1}^d \eta\left(x^i\right) = \sum_{i=1}^d \lambda_i(\mathcal{L}_{(C)})$, and this means that the solution $g$ computed by Algorithm 2.1.8 satisfies

$$\nu(g) \leq 1026d^3 \frac{1}{\lambda_2(\mathcal{L}_G)} \sum_{i=1}^d \lambda_i(\mathcal{L}_{(C)}).$$

This performance guarantee automatically implies our main result, a Cheeger inequality for the Connection Laplacian.

**Theorem 2.1.10.** *Let $\lambda_i(\mathcal{L}_{(C)})$ and $\lambda_i(\mathcal{L}_G)$ denote the i-th smallest eigenvalues of the normalized Connection Laplacian $\mathcal{L}_{(C)}$ and the normalized graph Laplacian $\mathcal{L}_G$ respectively. Let $\nu_G$ denote the frustration constant for $O(d)$ Synchronization. Then,*

$$\frac{1}{d}\sum_{i=1}^d \lambda_i(\mathcal{L}_{(C)}) \leq \nu_G \leq 1026d^3 \frac{1}{\lambda_2(\mathcal{L}_G)} \sum_{i=1}^d \lambda_i(\mathcal{L}_{(C)}).$$

Note that, once again, the lower bound is trivially obtained by noting that the eigenvector problem is a relaxation of the original synchronization problem.

Although we refer the reader to [41] for a proof of Lemma 2.1.9, we give a brief intuitive explanation of how the result is obtained.

As discussed above, the performance guarantee for Algorithm 2.1.5 relies on a proper understanding of the effect of the rounding step. In particular we showed that if $\lambda_2(\mathcal{L}_G)$ is small, then locally normalizing the candidate solution (which corresponds to the rounding step) has an effect over the penalty function that we can control. The case of $O(d)$ Synchronization is dealt with similarly. Instead of local normalization, the rounding step for Algorithm 2.1.8 is based on the polar decomposition. We start by understanding when the polar decomposition is stable (in the sense of changing the penalty function on a given edge) and see that this is the case when the candidate solution $X_i \in \mathbb{R}^{d\times d}$ is not close to being singular. The idea then is to show that only a small portion (which will depend on $\sum_{i=1}^{d} \lambda_i(\mathcal{L}_{(C)})$ and $\lambda_2(\mathcal{L}_G)$) of the graph can have candidate solutions $X_i$ close to singular and use that to show that, overall, we can bound the harmful contribution potentially caused by the rounding procedure on the penalty function.

### 2.1.2 Tightness of results

Let us consider the ring graph on $n$ vertices $G_n = (V_n, E_n)$ with $V_n = [n]$ and $E = \{(i, (i+1) \bmod n), \, i \in [n]\}$ with the edge weights all equal to 1 and $\rho : V \to O(d)$ as $\rho_{(n,1)} = -\mathrm{I}$ and $\rho = \mathrm{I}$ for all other edges. Define $x \in \mathbb{R}^{dn}$ by $x_k = \left[2\frac{k}{n} - 1, 0, \dots, 0\right]^T$. It is easy to check that $\eta(x) = \mathcal{O}(n^{-2})$ and that, for any $u > 0$, if $x^u \not\equiv 0$, there will have to be at least one edge that is not compatible with $x^u$, implying $\eta(x^u) \geq \frac{1}{2n}$. This shows that the 1/2 exponent in Lemma 2.1.3 is needed. In fact, by adding a few more edges to the graph $G_n$ one can also show the tightness of Theorem 2.1.4: Consider the "rainbow" graph $H_n$ that is constructed by adding to $G_n$, for each non-negative integer $k$ smaller than $n/2$, an edge between vertex $k$ and vertex $n-k$ with $\rho_{(k,n-k)} = -\mathrm{I}$. The vector $x$ still satisfies $\eta(x) = \mathcal{O}(n^{-2})$, however, for any non-zero vector $v : V \to \mathbb{S}^{d-1} \cup \{0\}$, it is not hard to show that $\eta(v)$ has to be of order at least $n^{-1}$, meaning that $\eta_G^*$ is $\Omega(\sqrt{\lambda_1(\mathcal{L}_{(C)})})$. This also means that, even if considering $\eta_G^*$, one could not get a linear bound (as provided by Lemma 2.1.6) without the control on $\lambda_2(\mathcal{L}_G)$.

Theorem 2.1.10 provides a non-trivial bound only if $\lambda_2(\mathcal{L}_G)$ is sufficiently large. It is clear that if one wants to bound full frustration constants, a dependency on $\lambda_2(\mathcal{L}_G)$ is needed. It is, nevertheless, non-obvious that this dependency is still needed if we consider partial versions of $O(d)$ frustration constants, $\vartheta_G^*$ or $\nu_G^*$. This can, however, be illustrated by a simple example in $O(2)$; consider a disconnected graph $G$ with two sufficiently large complete components, $G^1 = (V^1, E^1)$ and $G^2 = (V^2, E^2)$. For each edge let $\rho_{i,j} = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. It is clear that the vectors $x^1$ and $x^2$ defined such that $x_i^1 = [0, 1_{V^1}(i)]^T$ and $x_i^2 = [0, 1_{V^2}(i)]^T$ are orthogonal to each other and lie in the null space of the graph Connection Laplacian of $G$. This implies that $\lambda_2(\mathcal{L}_{(C)}) = 0$. On the other hand, it is straightforward to check that $\nu_G^*$ is not zero because it is impossible to perfectly synchronize the graph (or any of the components, for that matter).

## 2.2 Phase Retrieval as an application

### 2.2.1 Introduction

In this section we will describe how Algorithm 2.1.5 and Lemma 2.1.6 can be used to provide, and analyze, an algorithm to solve the *phase retrieval* problem. This section is based on the material in [11, 39, 33]. We start by a brief description and history of the problem.

Given a collection of vectors $\Phi := \{\varphi_\ell\}_{\ell=1}^N \subseteq \mathbb{C}^M$ and a signal $x \in \mathbb{C}^M$, consider measurements of the form

$$z_\ell := |\langle x, \varphi_\ell \rangle|^2 + \nu_\ell, \tag{2.10}$$

where $\nu_\ell$ is noise; we call these noisy *intensity measurements*, the purpose of phase retrieval is to estimate $x$ from these measurements.

Several areas of imaging science, such as X-ray crystallography [123, 162, 163], diffraction imaging [63], astronomy [90] and optics [232], use measurements of this form with the intent of reconstructing the original signal. Note that in the measurement process (2.10), we inherently lose some information about $x$. Indeed, for every $\omega \in \mathbb{C}$ with $|\omega| = 1$, we see that $x$ and $\omega x$ produce the same intensity measurements. Thus, the best one can hope to do with the intensity measurements of $x \in \mathbb{C}^M$ is reconstruct the class $[x] \in \mathbb{C}^M/\!\sim$, where $\sim$ is the equivalence relation of being identical up to a global phase factor.

In practice, phase retrieval falls short of determining the original signal up to global phase. First of all, the intensity measurement process that is used, say $\mathcal{A} : \mathbb{C}^M/{\sim} \to \mathbb{R}^N_{\geq 0}$ with $\mathcal{A}(x) = |\Phi^* x|^2$ (entrywise) and viewing $\Phi = [\varphi_1 \cdots \varphi_N]$, often lacks injectivity, making it impossible to reconstruct uniquely. Moreover, the phase retrieval algorithms that are used in practice take alternating projections onto the column space of $\Phi^*$ (to bring phase to the measurements) and onto the nonconvex set of vectors $y$ whose entry magnitudes match the intensity measurements $|\Phi^* x|^2$ (to maintain fidelity in the magnitudes) [105, 111, 117]. Unfortunately, the convergence of these algorithms (and various modifications thereof) is particularly sensitive to the choice of initial phases [157].

These deficiencies have prompted two important lines of research in phase retrieval:

(i) For which measurement designs $\Phi$ is $[x] \mapsto |\Phi^* x|^2$ injective?

(ii) For which injective designs can $[x]$ be reconstructed stably and efficiently?

A first step toward solving (i) is determining how large $N$ must be in order for $\mathcal{A}$ to be injective. It remains an open problem to find the smallest such $N$, it was conjectured in [31] that $N \geq 4M - 4$ is necessary. Indeed, embedding results in differential geometry give that $N \geq (4 + o(1))M$ is necessary [24, 126]. As for sufficiency, recently Conca et al. [85] show that for almost every choice of $\Phi$, $\mathcal{A}$ is injective whenever $N \geq 4M - 4$, improving over a result by Balan, Casazza and Edidin [28] establishing a similar result for $N \geq 4M - 2$. Before then, notable constructions with few measurements were given in [50, 165]. Just a few months ago Vinzant [229] disproved the $4M - 4$ conjecture with an injective construction for $M = 4$ and $N = 11 < 12 = 4 \times 4 - 4$. Though the community has investigated various conditions for injectivity, very little is known about how to stably and efficiently reconstruct in the injective case. In fact, some instances of the phase retrieval problem are known to be NP-complete [196], and so any general reconstruction process is necessarily inefficient, assuming P $\neq$ NP.

This leads one to attempt provably stable and efficient reconstruction from measurements of the form (2.10) with particular ensembles $\Phi$. Until recently, this was only known to be possible in cases where $N = \Omega(M^2)$ [27]. By contrast, the state of the art comes from Candès, Strohmer and Voroninski [67], who use semidefinite programming to stably reconstruct from $N = \mathcal{O}(M \log M)$ Gaussian-random measurements, recently improved to $N = \mathcal{O}(M)$ in [64]. There is other work along this vein [69, 141, 231] which also uses semidefinite programming and provides related guarantees. The method described in this section will bypass semidefinite programming by leveraging the spectral approach to angular synchronization (Algorithm 2.1.5).

While (i) and (ii) above describe what has been a more theoretical approach to phase retrieval, practitioners have meanwhile considered various alternatives to the intensity measurement process (2.10). A common theme among these alternatives is the use of interference to extract more information about the desired signal. For example, *holography* interferes the signal of interest $x \in \mathbb{C}^M$ with a known reference signal $y \in \mathbb{C}^M$, taking measurements of the form $|F(x + \omega y)|^2$, where $\omega \in \mathbb{C}$ has unit modulus and $F$ denotes the Fourier transform [98]; three such measurements (i.e., $3M$ scalar measurements) suffice for injectivity [234]. Alternatively, *spectral phase interferometry for direct electric-field reconstruction (SPIDER)* interferes the signal of interest $x \in \mathbb{C}^M$ with time- and frequency-shifted versions of itself $Sx \in \mathbb{C}^M$, taking measurements of the form $|F(x + Sx)|^2$ [131]; while popular in practice for ultrashort pulse measurement, SPIDER fails to accurately resolve the relative phase of well-separated frequency components [136]. Another interesting approach is *ptychography*, in which overlapping spatial components $P_i x, P_j x \in \mathbb{C}^M$ are interfered with each other, and measurements have the form $|F(P_i x + P_j x)|^2$ [189]. Recently, *vectorial phase retrieval* was proposed, in which two unknown signals $x, y \in \mathbb{C}^M$ are interfered with each other, and the measurements are $|Fx|^2$, $|Fy|^2$, $|F(x+y)|^2$ and $|F(x+\mathrm{i}y)|^2$ [187]; furthermore, [186] gives that almost every pair of signals is uniquely determined by these $4M$ scalar measurements, in which case both signals can be reconstructed using the polarization identity. While practitioners seem to have identified interference as an instrumental technique for quality phase retrieval, the reconstruction algorithms which are typically used, much like the classical algorithms in [105, 111, 117], are iterative and lack convergence guarantees (e.g., [98, 155], though [186, 187] are noteworthy exceptions).

Returning to measurements of the form (2.10), this paper combines ideas from both state-of-the-art theory and state-of-the-art practice by proposing an exchange of sorts: If you already have $\mathcal{O}(M \log M)$ Gaussian-random measurements vectors (as prescribed in [67]), then we offer a faster reconstruction method with a stable performance guarantee, but at the price of $\mathcal{O}(M \log M)$ additional (non-adaptive) measurements. These new measurement vectors are interferometry-inspired combinations of the originals, and the computational speedups gained in reconstruction come from our use of different spectral methods. While the ideas here can be applied for phase retrieval of 2-D images, we focus on the 1-D case for simplicity.

We also note that these techniques can be adapted to the setting of masked Fourier measurements, thereby mimicking the illumination methodology of [69] (for the sake of brevity we do not describe those adaptation here but refer the reader to [33]). They have also been adapted to the setting on which the signal to be recovered is known to be sparse [39]. This strongly suggests that these techniques can be leveraged to tackle a wide variety of practical instances of the phase retrieval problem.

To help motivate our measurement design and phase retrieval procedure, we start by considering the simpler, noiseless case. In this case, the success of our method follows from a neat trick involving the polarization identity along with some well-known results in the theory of expander graphs. Afterwards, we will briefly describe how to modify the method to obtain provable stability in the noisy case; here, Algorithm 2.1.5 and Lemma 2.1.6 play a crucial role.

### 2.2.2 The noiseless case

In this section, we provide a new measurement design and reconstruction algorithm for phase retrieval. Here, we specifically address the noiseless case, in which $\nu_\ell$ in (2.10) is zero for every $\ell = 1, \ldots, N$; this case will give some intuition for a more stable version of our techniques, which we discuss later. In the noiseless case, we will use on the order of the fewest measurements possible, namely $N = \mathcal{O}(M)$, where $M$ is the dimension of the signal.

Before stating our measurement design and phase retrieval procedure, we motivate both with some discussion. Take a finite set $V$, and suppose we take intensity measurements of $x \in \mathbb{C}^M$ with a set $\Phi_V := \{\varphi_i\}_{i\in V}$ that spans $\mathbb{C}^M$. Again, we wish to recover $x$ up to a global phase factor. Having $|\langle x, \varphi_i\rangle|$ for every $i \in V$, we claim it suffices to determine the relative phase between $\langle x, \varphi_i\rangle$ and $\langle x, \varphi_j\rangle$ for all pairs $i \neq j$. Indeed, if we had this information, we could arbitrarily assign some nonzero coefficient $c_i = |\langle x, \varphi_i\rangle|$ to have positive phase. If $\langle x, \varphi_j\rangle$ is also nonzero, then it has well-defined relative phase

$$\rho_{ij} := \big(\tfrac{\langle x,\varphi_i\rangle}{|\langle x,\varphi_i\rangle|}\big)^{-1}\tfrac{\langle x,\varphi_j\rangle}{|\langle x,\varphi_j\rangle|}, \tag{2.11}$$

which determines the coefficient by multiplication: $c_j = \rho_{ij}|\langle x, \varphi_j\rangle|$. Otherwise when $\langle x, \varphi_j\rangle = 0$, we naturally take $c_j = 0$, and for notational convenience, we arbitrarily take $\rho_{ij} = 1$. From here, the original signal's equivalence class $[x] \in \mathbb{C}^M/\!\sim$ can be identified by applying the canonical dual frame $\{\tilde{\varphi}_j\}_{j\in V}$, namely the Moore-Penrose pseudoinverse, of $\Phi_V$:

$$\sum_{j\in V} c_j\tilde{\varphi}_j = \sum_{j\in V}\rho_{ij}|\langle x,\varphi_j\rangle|\tilde{\varphi}_j = \big(\tfrac{\langle x,\varphi_i\rangle}{|\langle x,\varphi_i\rangle|}\big)^{-1}\sum_{j\in V}\langle x,\varphi_j\rangle\tilde{\varphi}_j = \big(\tfrac{\langle x,\varphi_i\rangle}{|\langle x,\varphi_i\rangle|}\big)^{-1}x \in [x]. \quad (2.12)$$

Having established the utility of the relative phase between coefficients, we now seek some method of extracting this information. To this end, we turn to a special version of the polarization identity:

**Lemma 2.2.1** (Mercedes-Benz Polarization Identity)**.** *Take* $\zeta := \mathrm{e}^{2\pi\mathrm{i}/3}$. *Then for any* $a, b \in \mathbb{C}$,

$$\bar{a}b = \frac{1}{3}\sum_{k=0}^{2}\zeta^k|a+\zeta^{-k}b|^2. \quad (2.13)$$

*Proof.* We start by expanding the right-hand side of (2.13):

$$\text{RHS} := \frac{1}{3}\sum_{k=0}^{2}\zeta^k|a+\zeta^{-k}b|^2 = \frac{1}{3}\sum_{k=0}^{2}\zeta^k\big(|a|^2+2\Re(\zeta^{-k}\bar{a}b)+|b|^2\big) = \frac{2}{3}\sum_{k=0}^{2}\zeta^k\Re(\zeta^{-k}\bar{a}b).$$

Multiplying, we find

$$\Re(\zeta^{-k}\bar{a}b) = \Re(\zeta^{-k})\Re(\bar{a}b) - \Im(\zeta^{-k})\Im(\bar{a}b) = \Re(\zeta^{k})\Re(\bar{a}b) + \Im(\zeta^{k})\Im(\bar{a}b).$$

We substitute this into our expression for RHS:

$$\Re(\text{RHS}) = \frac{2}{3}\bigg[\Re(\bar{a}b)\sum_{k=0}^{2}\big(\Re(\zeta^k)\big)^2 + \Im(\bar{a}b)\sum_{k=0}^{2}\Re(\zeta^k)\Im(\zeta^k)\bigg],$$
$$\Im(\text{RHS}) = \frac{2}{3}\bigg[\Re(\bar{a}b)\sum_{k=0}^{2}\Re(\zeta^k)\Im(\zeta^k) + \Im(\bar{a}b)\sum_{k=0}^{2}\big(\Im(\zeta^k)\big)^2\bigg].$$

Finally, we apply the following easy-to-verify identities:

$$\sum_{k=0}^{2}\big(\Re(\zeta^k)\big)^2 = \sum_{k=0}^{2}\big(\Im(\zeta^k)\big)^2 = \frac{3}{2}, \qquad \sum_{k=0}^{2}\Re(\zeta^k)\Im(\zeta^k) = 0,$$

which yield RHS $= \bar{a}b$. □

The above polarization identity can also be proved by viewing $\{\zeta^k\}_{k=0}^2$ as a Mercedes-Benz frame in $\mathbb{R}^2$ and $\frac{2}{3}\sum_{k=0}^2\zeta^k\Re(\zeta^{-k}u)$ as the corresponding reconstruction formula for $u\in\mathbb{C}=\mathbb{R}^2$. We can now use this polarization identity to determine relative phase (2.11):

$$\overline{\langle x,\varphi_i\rangle}\langle x,\varphi_j\rangle=\frac{1}{3}\sum_{k=0}^{2}\zeta^k\big|\langle x,\varphi_i\rangle+\zeta^{-k}\langle x,\varphi_j\rangle\big|^2=\frac{1}{3}\sum_{k=0}^{2}\zeta^k\big|\langle x,\varphi_i+\zeta^k\varphi_j\rangle\big|^2. \tag{2.14}$$

Thus, if in addition to $\Phi_V$ we measure with $\{\varphi_i+\zeta^k\varphi_j\}_{k=0}^2$, we can use (2.14) to determine $\overline{\langle x,\varphi_i\rangle}\langle x,\varphi_j\rangle$ and then normalize to get the relative phase:

$$\rho_{ij}:=\big(\tfrac{\langle x,\varphi_i\rangle}{|\langle x,\varphi_i\rangle|}\big)^{-1}\tfrac{\langle x,\varphi_j\rangle}{|\langle x,\varphi_j\rangle|}=\tfrac{\overline{\langle x,\varphi_i\rangle}\langle x,\varphi_j\rangle}{|\overline{\langle x,\varphi_i\rangle}\langle x,\varphi_j\rangle|}, \tag{2.15}$$

provided both $\langle x,\varphi_i\rangle$ and $\langle x,\varphi_j\rangle$ are nonzero. To summarize our discussion of reconstructing a single signal, if we measure with $\Phi_V$ and $\{\varphi_i+\zeta^k\varphi_j\}_{k=0}^2$ for every pair $i,j\in V$, then we can recover $[x]$. However, such a method uses $|V|+3\binom{|V|}{2}$ measurements, and since $\Phi_V$ must span $\mathbb{C}^M$, we necessarily have $|V|\geq M$ and thus a total of $\Omega(M^2)$ measurements. Note that a nearly identical formulation of these $\Omega(M^2)$ measurements appears in Theorem 5.2 of [27]. The proof of this result shows how one can adaptively appeal to only $\mathcal{O}(M)$ of the measurements to perform phase retrieval, suggesting that most of these measurements are actually unnecessary. However, since the $\mathcal{O}(M)$ measurements that end up being used are highly dependent on the signal being measured, one cannot blindly restrict to a particular subcollection of $\mathcal{O}(M)$ measurement vectors a priori without forfeiting injectivity.

In pursuit of $\mathcal{O}(M)$ measurements, take some simple graph $G=(V,E)$, arbitrarily assign a direction to each edge, and only take measurements with $\Phi_V$ and $\Phi_E:=\bigcup_{(i,j)\in E}\{\varphi_i+\zeta^k\varphi_j\}_{k=0}^2$. To recover $[x]$, we again arbitrarily assign some nonzero vertex measurement to have positive phase, and then we propagate relative phase information along the edges by multiplication to determine the phase of the other vertex measurements relative to the original vertex measurement:

$$\rho_{ik} = \rho_{ij}\rho_{jk}. \tag{2.16}$$

However, if $x$ is orthogonal to a given vertex vector, then that measurement is zero, and so relative phase information cannot propagate through the corresponding vertex; indeed, such orthogonality has the effect of removing the vertex from the graph, and for some graphs, this will prevent recovery. For example, if $G$ is a star, then $x$ could be orthogonal to the vector corresponding to the internal vertex, whose removal would render the remaining graph edgeless. That said, we should select $\Phi_V$ and $G$ so as to minimize the impact of orthogonality with vertex vectors.

First, we can take $\Phi_V$ to be *full spark*, that is, $\Phi_V$ has the property that every subcollection of $M$ vectors spans. Full spark frames appear in a wide variety of applications. Explicit deterministic constructions of them are given in [12, 184]. For example, we can select the first $M$ rows of the $|V| \times |V|$ discrete Fourier transform matrix, and take $\Phi_V$ to be the columns of the resulting $M \times |V|$ matrix; in this case, the fact that $\Phi_V$ is full spark follows from the Vandermonde determinant formula. In our application, $\Phi_V$ being full spark will be useful for two reasons. First, this implies that $x \neq 0$ is orthogonal to at most $M-1$ members of $\Phi_V$, thereby limiting the extent of $x$'s damage to our graph. Additionally, $\Phi_V$ being full spark frees us from requiring the graph to be connected after the removal of vertices; indeed, any remaining component of size $M$ or more will correspond to a subcollection of $\Phi_V$ that spans, meaning it has a dual frame to reconstruct with. It remains to find a graph of $\mathcal{O}(M)$ vertices and edges that maintains a size-$M$ component after the removal of any $M-1$ vertices.

To this end, we consider a well-studied family of sparse graphs known as *expander graphs*. We choose these graphs for their notably strong connectivity properties. There is a combinatorial definition of expander graphs, but we will focus on the spectral definition. Given a $d$-regular graph $G$ of $n$ vertices, consider its adjacency matrix $A$, and recall that its normalized Laplacian is given by $\mathcal{L}_G := I - \frac{1}{d}A$. We are particularly interested in the eigenvalues of the Laplacian: $0 = \lambda_1 \leq \cdots \leq \lambda_n$. The second eigenvalue $\lambda_2$ of the Laplacian is called the *spectral gap* of the graph, and, as it was clear from Cheeger's inequality in Section 2.1, this value is particularly useful in evaluating the graph's connectivity. We say $G$ has *expansion* $\delta$ if $\{\lambda_2, \ldots, \lambda_n\} \subseteq [1-\delta, 1+\delta]$; note that since $1 - \delta \leq \lambda_2$, small expansion implies large spectral gap. Furthermore, a family of $d$-regular graphs $\{G_i\}_{i=1}^{\infty}$ is a *spectral expander family* if there exists $c < 1$ such that every $G_i$ has expansion $\delta(G_i) \leq c$. Since $d$ is constant over an expander family, expanders with many vertices have particularly few edges. There are many results which describe the connectivity of expanders, but the following is particularly relevant to the application that follows:

**Lemma 2.2.2** (Spectral gap grants connectivity [124])**.** *Consider a $d$-regular graph $G$ of $n$ vertices with spectral gap $\lambda_2$. For all $\varepsilon \leq \frac{\lambda_2}{6}$, removing any $\varepsilon dn$ edges from $G$ results in a connected component of size $\geq (1 - \frac{2\varepsilon}{\lambda_2})n$.*

Note that removing $\varepsilon n$ vertices from a $d$-regular graph necessarily removes $\leq \varepsilon dn$ edges, and so this lemma directly applies. For our application, we want to guarantee that the removal of any $M-1$ vertices maintains a size-$M$ component. To do this, we will ensure both (i) $M - 1 \leq \varepsilon n$ and (ii) $M - 1 < (1 - \frac{2\varepsilon}{\lambda_2})n$, and then invoke the above lemma. Note that since $n \geq M \geq 2$,

$$\varepsilon \leq \frac{\lambda_2}{6} \leq \frac{\mathrm{Tr}[L]}{6(n-1)} = \frac{n}{6(n-1)} \leq \frac{1}{3} < \frac{2}{3} \leq 1 - \frac{2\varepsilon}{\lambda_2},$$

where the last inequality is a rearrangement of $\varepsilon \leq \frac{\lambda_2}{6}$. Thus $\varepsilon n < (1 - \frac{2\varepsilon}{\lambda_2})n$, meaning

(i) implies (ii), and so it suffices to have $M \leq \varepsilon n + 1$. Overall, we use the following criteria to pick our expander graph: Given the signal dimension $M$, use a $d$-regular graph $G = (V, E)$ of $n$ vertices with spectral gap $\lambda_2$ such that $M \leq (\frac{\lambda_2}{6})n + 1$. Then by the previous discussion, the total number of measurements is $N = |V| + 3|E| = (\frac{3}{2}d + 1)n$. If we think of the degree $d$ as being fixed, then the number of vertices $n$ in the graph is proportional to the total number of measurements $N$ (this is the key distinction from the previous complete-graph case).

Recall that we seek $N = \mathcal{O}(M)$ measurements. To minimize the redundancy $\frac{N}{M}$ for a fixed degree $d$, we would like a maximal spectral gap $\lambda_2$, and it suffices to seek minimal spectral expansion $\delta$. Spectral graph families known as *Ramanujan graphs* are asymptotically optimal in this sense; taking $\mathcal{G}_n^d$ to be the set of connected $d$-regular graphs with $\geq n$ vertices, Alon and Boppana (see [15]) showed that for any fixed $d$,

$$\lim_{n \to \infty} \inf_{G \in \mathcal{G}_n^d} \delta(G) \geq \frac{2\sqrt{d-1}}{d},$$

while Ramanujan graphs are defined to have spectral expansion $\leq \frac{2\sqrt{d-1}}{d}$. To date, Ramanujan graphs have only been constructed for certain values of $d$. One important construction was given by Lubotzky, Phillips, and Sarnak [152], which produces a Ramanujan family whenever $d - 1 \equiv 1 \bmod 4$ is prime. Among these graphs, we get the smallest redundancy $\frac{N}{M}$ when $M = \lfloor (1 - \frac{2\sqrt{d-1}}{d})\frac{n}{6} + 1 \rfloor$ and $d = 6$:

$$\frac{N}{M} \leq \frac{(\frac{3}{2}d + 1)n}{(1 - \frac{2\sqrt{d-1}}{d})\frac{n}{6}} = 45\big(3 + \sqrt{5}\big) \approx 235.62.$$

Thus, in such cases, our techniques allow for phase retrieval with only $N \leq 236M$ measurements. However, the number of vertices in each Ramanujan graph from [152] is of the form $q(q^2 - 1)$ or $\frac{q(q^2-1)}{2}$, where $q \equiv 1 \bmod 4$ is prime, and so any bound on redundancy $\frac{N}{M}$ using these graphs will only be valid for particular values of $M$.

In order to get $N = \mathcal{O}(M)$ in general, we use the fact that random graphs are nearly Ramanujan with high probability. In particular, for every $\varepsilon > 0$ and even $d$, a random $d$-regular graph has spectral expansion $\delta \leq \frac{2\sqrt{d-1}+\varepsilon}{d}$ with high probability as $n \to \infty$ [107]. Thus, picking $\varepsilon$ and $d$ to satisfy $\frac{2\sqrt{d-1}+\varepsilon}{d} < 1$, we may take $M = \lfloor (1 - \frac{2\sqrt{d-1}+\varepsilon}{d})\frac{n}{6} + 1 \rfloor$ to get

$$\frac{N}{M} \leq \frac{(\frac{3}{2}d+1)n}{(1-\frac{2\sqrt{d-1}+\varepsilon}{d})\frac{n}{6}},$$

and this choice will satisfy $M \leq (\frac{\lambda_2}{6})n + 1$ with high probability. To see how small this redundancy is, note that taking $\varepsilon = 0.1$ and $d = 8$ gives $N \leq 240M$. While the desired expansion properties of a random graph are only present with high probability, estimating the spectral gap is inexpensive, and so it is computationally feasible to verify whether a randomly drawn graph is good enough. Moreover, $n$ can be any sufficiently large integer, and so the above bound is valid for all sufficiently large $M$, i.e., our procedure can perform phase retrieval with $N = \mathcal{O}(M)$ measurements in general.

Combining this with the above discussion, we have the following measurement design and phase retrieval procedure:

**Measurement Design A (noiseless case)**

- Fix $d > 2$ even and $\varepsilon \in (0, d - 2\sqrt{d-1})$.
- Given $M$, pick some $d$-regular graph $G = (V, E)$ with spectral gap $\lambda_2 \geq \lambda' := 1 - \frac{2\sqrt{d-1}+\varepsilon}{d}$ and $|V| = \lceil \frac{6}{\lambda'}(M-1) \rceil$, and arbitrarily direct the edges.
- Design the measurements $\Phi := \Phi_V \cup \Phi_E$ by taking $\Phi_V := \{\varphi_i\}_{i \in V} \subseteq \mathbb{C}^M$ to be full spark and $\Phi_E := \bigcup_{(i,j)\in E}\{\varphi_i + \zeta^k \varphi_j\}_{k=0}^{2}$.

**Phase Retrieval Procedure A (noiseless case)**

- Given $\{|\langle x, \varphi\rangle|^2\}_{\varphi\in\Phi}$, delete the vertices $i \in V$ with $|\langle x, \varphi_i\rangle|^2 = 0$.

- In the remaining induced subgraph, find a connected component of $\geq M$ vertices $V'$.

- Pick a vertex in $V'$ to have positive phase and propagate/multiply relative phases (2.16), which are calculated by normalizing (2.14), see (2.15).

- Having $\{\langle x,\varphi_i\rangle\}_{i\in V'}$ up to a global phase factor, find the least-squares estimate of $[x]$ by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i\in V'}$, see (2.12).

Note that this phase retrieval procedure is particularly fast. Indeed, if we use $E\subseteq V^2$ to store $G$, then we can delete vertices $i\in V$ with $|\langle x,\varphi_i\rangle|^2=0$ by deleting the edges for which (2.14) is zero, which takes $\mathcal{O}(|E|)$ time. Next, if the members of $E$ are ordered lexicographically, the remaining subgraph can be easily partitioned into connected components in $\mathcal{O}(|E|)$ time by collecting edges with common vertices, and then propagating relative phase in the largest component is performed in $\mathcal{O}(|E|)$ time using a depth- or breadth-first search. Overall, we only use $\mathcal{O}(M)$ time before the final least-squares step of the phase retrieval procedure, which happens to be the bottleneck, depending on the subcollection $\Phi_{V'}$. In general, we can find the least-squares estimate in $\mathcal{O}(M^3)$ time using Gaussian elimination, but if $\Phi_{V'}$ has special structure (e.g., it is a submatrix of the discrete Fourier transform matrix), then one might exploit that structure to gain speedups (e.g., use the fast Fourier transform in conjunction with an iterative method). Regardless, our procedure reduces the nonlinear phase retrieval problem to the much simpler problem of solving an overdetermined *linear* system.

While this measurement design and phase retrieval procedure is particularly efficient, it certainly lacks stability. Perhaps most notably, we have not imposed anything on $\Phi_V$ that guarantees stability with inverting $\Phi_{V'}$; indeed, we have merely enforced linear independence between vectors, while stability will require well-conditioning. Another noteworthy source of instability is our method of phase propagation (which is a form of angular synchronization). The process described above, in the presence of noise, will naturally accumulates error; we will remediate this issue by making use of the spectral method described in Section 2.1 (Algorithm 2.1.5). These adaptations while turning the process stable, will do it at the price of a log factor in the number of measurements: $N = \mathcal{O}(M \log M)$.[3]

### 2.2.3 The noisy case

We now consider a noise-robust version of the measurement design and phase retrieval procedure of the previous section. In the end, the measurement design will be nearly identical: vertex measurements will be independent complex Gaussian vectors (thereby being full spark with probability 1), and the edge measurements will be the same sort of linear combinations of vertex measurements. Our use of randomness in this version will enable the vertex measurements to simultaneously satisfy two important conditions with high probability: *projective uniformity with noise* and *numerical erasure robustness*. Before defining these conditions, we motivate them by considering a noisy version of our phase retrieval procedure.

Recall that our noiseless procedure starts by removing the vertices $i \in V$ for which $|\langle x, \varphi_i \rangle|^2 = 0$. Indeed, since we plan to propagate relative phase information along edges, these 0-vertices are of no use, as relative phase with these vertices is not well defined. Since we calculate relative phase by normalizing (2.14), we see that relative phase is sensitive to perturbations when (2.14) is small, meaning either $\langle x, \varphi_i \rangle$ or $\langle x, \varphi_j \rangle$ is small. As such, while 0-vertices provide no relative phase information in the noiseless case, small vertices provide *unreliable* information in the noisy case, and so we wish to remove them accordingly (alternatively, one might use weights according to one's confidence in the information, but we decided to use hard thresholds to simplify the analysis). However, we also want to ensure that there are only a few

[3] We do not think this log factor is necessary, but we leave this pursuit for future work.

small vertices. In the noiseless case, we limit the number of 0-vertices by using a full spark frame; in the noisy case, we make use of a new concept we call *projective uniformity*:

**Definition 2.2.3.** *The $\alpha$-projective uniformity of $\Phi = \{\varphi_i\}_{i=1}^{n} \subseteq \mathbb{C}^M$ is given by*

$$\mathrm{PU}(\Phi;\alpha) = \min_{\substack{x\in\mathbb{C}^M\\ \|x\|=1}} \max_{\substack{\mathcal{I}\subseteq\{1,\ldots,n\}\\ |\mathcal{I}|\geq\alpha n}} \min_{i\in\mathcal{I}} |\langle x,\varphi_i\rangle|^2.$$

In words, projective uniformity gives the following guarantee: For every unit-norm signal $x$, there exists a collection of vertices $\mathcal{I} \subseteq V$ of size at least $\alpha|V|$ such that $|\langle x,\varphi_i\rangle|^2 \geq \mathrm{PU}(\Phi_V;\alpha)$ for every $i \in \mathcal{I}$. As such, projective uniformity effectively limits the total number of small vertices possible, at least before the measurements are corrupted by noise. However, the phase retrieval algorithm will only have access to noisy versions of the measurements, and so we must account for this subtlety in our procedure. In an effort to isolate the reliable pieces of relative phase information, Algorithm 2.2.4 removes the vertices corresponding to small noisy edge combinations (2.14).

**Algorithm 2.2.4.** *[Pruning for reliability]*

```
Input: Graph $G = (V,E)$, function $f\colon E \to \mathbb{R}$ such that $f(i,j) = |\overline{\langle x,\varphi_i\rangle}\langle x,\varphi_j\rangle + \varepsilon_{ij}|$, pruning parameter $\alpha$
Output: Subgraph $H$ with a larger smallest edge weight
Initialize $H \leftarrow G$

For $i = 1$ to $\lfloor(1-\alpha)|V|\rfloor$ do:
Find the minimizer $(i,j) \in E$ of $f$ $H \leftarrow H \setminus \{i,j\}$
```

We now explain why only reliable pieces of relative phase information will remain after running the above algorithm, provided $\Phi_V$ has sufficient projective uniformity. The main idea is captured in the following:

**Lemma 2.2.5.** *Define $\|\theta\|_{\mathbb{T}} := \min_{k\in\mathbb{Z}} |\theta - 2\pi k|$ for all angles $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Then for any $z, \varepsilon \in \mathbb{C}$,*

$$\|\arg(z+\varepsilon) - \arg(z)\|_{\mathbb{T}} \leq \pi\frac{|\varepsilon|}{|z|}.$$

For the sake of brevity we refer the reader to [11] for a proof.

By taking $z = \overline{\langle x, \varphi_i\rangle}\langle x, \varphi_j\rangle + \varepsilon_{ij}$ and $\varepsilon = -\varepsilon_{ij}$, we can use this lemma to bound the relative phase error we incur when normalizing $z$. In fact, consider the minimum of $f$ when Algorithm 2.2.4 is complete. Since the algorithm deletes vertices from $G$ according to the input signal $x$, this minimum will vary with $x$; let PUN denote the smallest possible minimum value. Then the relative phase error incurred with $(i, j) \in E$ is no more than $\pi|\varepsilon_{ij}|/\text{PUN}$, regardless of the signal measured. Indeed, our use of *projective uniformity with noise* (i.e., PUN) is intended to bound the instability that comes with normalizing small values of (2.14). PUN can be bounded below by using the projective uniformity of $\Phi_V$, and furthermore, a complex Gaussian $\Phi_V$ has projective uniformity with overwhelming probability. We refer the reader to [11] for proofs of these statements.

After applying Algorithm 2.2.4, our graph will have slightly fewer vertices, but the remaining edges will correspond to reliable pieces of relative phase information. Recall that we plan to use this information on the edges to determine phases for the vertices, and we want to do this in a stable way. As we will do this through Algorithm 2.1.5 whose guarantees (Lemma 2.1.6) depend on the connectivity of the graph, we will require a high level of connectivity in the graph.

As such, we seek to remove a small proportion of vertices so that the remaining graph is very connected, i.e., has large spectral gap. To do this, we will iteratively remove sets of vertices that are poorly connected to the rest of the graph. These sets will be identified using spectral clustering, described in Section 2.1, for which Theorem 2.1.1 gives a performance guarantee.

We note that, when implement spectral clustering, the bottleneck is computing an eigenvector. For our application, we will iteratively apply spectral clustering to identify small collections of vertices which are poorly connected to the rest of the graph and then remove them to enhance connectivity (Algorithm 2.2.6).

**Algorithm 2.2.6.** *[Pruning for connectivity]*

```
Input: Graph $G = (V, E)$, pruning parameter $\tau$
Output: Subgraph $H$ with spectral gap $\lambda_2(H) \geq \tau$
Initialize $H \leftarrow G$
While $\lambda_2(H) < \tau$ do:
Perform spectral clustering to identify a small set of vertices $S$
$H \leftarrow H \setminus S$
```

We refer the reader to [11] for a guarantee that, for a particular choice of threshold $\tau$, Algorithm 2.2.6 recovers a level of connectivity that may have been lost when pruning for reliability in Algorithm 2.2.4, and it does so by removing only a small proportion of the vertices.

At this point, we have pruned our graph so that the measured relative phases are reliable and the vertex phases can be stably reconstructed. We are now in a position to use Algorithm 2.1.5 to reconstruct these vertex phases from the measured relative phases, as it is an instance of the angular synchronization problem.

Lemma 2.1.6 guarantees that the estimates for the phases of the inner products $\{\langle x, \varphi_i \rangle\}_{i \in V'}$, produced by Algorithm 2.1.5 have a small frustration. However, in this case we need a guarantee, not in terms of frustration, but in terms of how close these estimates are to the true phases. This will be achieved by the following theorem. We refer the reader to [11] for a proof.

**Theorem 2.2.7.** *Consider a graph $G = (V, E)$ with spectral gap $\tau > 0$, and define $\|\theta\|_{\mathbb{T}} := \min_{k \in \mathbb{Z}} |\theta - 2\pi k|$ for all angles $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Consider the matrix $A_{(C)}$ defined* *as*

$$A_1[i,j] = \frac{\overline{\langle x, \varphi_i \rangle}\langle x, \varphi_j \rangle + \varepsilon_{ij}}{|\overline{\langle x, \varphi_i \rangle}\langle x, \varphi_j \rangle + \varepsilon_{ij}|}, \tag{2.17}$$

*when* $\{i,j\} \in E$, *and* $A_{(C)}[i,j] = 0$ *otherwise.*

*Then, Algorithm 2.1.5 outputs* $u \in \mathbb{C}^{|V|}$ *with unit-modulus entries such that, for some phase* $\theta \in \mathbb{R}/2\pi\mathbb{Z}$,

$$\sum_{i \in V} \big\| \arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta \big\|_{\mathbb{T}}^2 \leq \frac{C\|\varepsilon\|^2}{\tau^2 P^2},$$

*where* $P := \min_{\{i,j\} \in E} |\overline{\langle x, \varphi_i \rangle}\langle x, \varphi_j \rangle + \varepsilon_{ij}|$ *and* $C$ *is a universal constant.*

To reiterate, Algorithm 2.1.5 will produce estimates for the phases of the inner products $\{\langle x, \varphi_i \rangle\}_{i \in V'}$. Also, we can take square roots of the vertex measurements $\{|\langle x, \varphi_i \rangle|^2 + \nu_i\}_{i \in V'}$ to estimate $\{|\langle x, \varphi_i \rangle|\}_{i \in V'}$. Then we can combine these to estimate $\{\langle x, \varphi_i \rangle\}_{i \in V'}$.

However, note that the largest of these inner products will be most susceptible to noise in the corresponding phase estimate. As such, we remove a small fraction of these largest vertices so that the final collection of vertices $V''$ has size $\kappa|V|$, where $V$ was the original vertex set, and $\kappa$ is sufficiently close to 1.

Now that we have estimated the phases of $\{\langle x, \varphi_i \rangle\}_{i \in V''}$, we wish to reconstruct $x$ by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i \in V''}$. However, since $V''$ is likely a strict subset of $V$, it can be difficult in general to predict how stable the pseudoinverse will be. Fortunately, a recent theory of *numerically erasure-robust frames (NERFs)* makes this prediction possible: If the members of $\Phi_V$ are independent Gaussian vectors, then with high probability, every submatrix of columns $\Phi_{V''}$ with $\kappa = |V''|/|V|$ sufficiently large has a stable pseudoinverse [104]. This concludes the phase retrieval procedure, briefly outlined below together with the measurement design.

**Measurement Design B (noisy case)**

- Fix $d > 2$ even and $\varepsilon \in (0, d - 2\sqrt{d-1})$.
- Given $M$, pick some $d$-regular graph $G = (V, E)$ with spectral gap $\lambda_2 \geq \lambda' := 1 - \frac{2\sqrt{d-1}+\varepsilon}{d}$ and $|V| = cM \log M$ for $c$ sufficiently large, and arbitrarily direct the edges.
- Design the measurements $\Phi := \Phi_V \cup \Phi_E$ by taking $\Phi_V := \{\varphi_i\}_{i \in V} \subseteq \mathbb{C}^M$ to have independent entries with distribution $\mathbb{C}\mathcal{N}(0, \frac{1}{M})$ and $\Phi_E := \bigcup_{(i,j)\in E}\{\varphi_i + \zeta^k \varphi_j\}_{k=0}^{2}$.

**Phase Retrieval Procedure B (noisy case)**

- Given $\{|\langle x, \varphi_\ell \rangle|^2 + \nu_\ell\}_{\ell=1}^{N}$, prune the graph $G$, keeping only reliable vertices (Algorithm 2.2.4).
- Prune the remaining induced subgraph for connectivity, producing the vertex set $V'$ (Algorithm 2.2.6).
- Estimate the phases of the vertex measurements using angular synchronization.
- Remove the vertices with the largest measurements, keeping only $|V''| = \kappa |V|$.
- Having estimates for $\{\langle x, \varphi_i \rangle\}_{i \in V''}$ up to a global phase factor, find the least-squares estimate of $[x]$ by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i \in V''}$, see (2.12).

Having established our measurement design and phase retrieval procedure for the noisy case, we now present the following guarantee of stable performance:

**Theorem 2.2.8.** *Pick $N \sim CM \log M$ with $C$ sufficiently large, and take $\{\varphi_\ell\}_{\ell=1}^{N} = \Phi_V \cup \Phi_E$ defined in Measurement Design B. Then there exist constants $C', K > 0$*

*such that the following guarantee holds for all $x \in \mathbb{C}^M$ with overwhelming probability: Consider measurements of the form*

$$z_\ell := |\langle x, \varphi_\ell \rangle|^2 + \nu_\ell.$$

*If the noise-to-signal ratio satisfies* $\mathrm{NSR} := \frac{\|\nu\|}{\|x\|^2} \leq \frac{C'}{\sqrt{M}}$*, then Phase Retrieval Procedure B produces an estimate $\tilde{x}$ from $\{z_\ell\}_{\ell=1}^N$ with squared relative error*

$$\frac{\|\tilde{x} - \mathrm{e}^{\mathrm{i}\theta} x\|^2}{\|x\|^2} \leq K \sqrt{\frac{M}{\log M}}\ \mathrm{NSR}$$

*for some phase $\theta \in [0, 2\pi)$.*

The interested reader is directed to [11] for a proof of this guarantee. Before concluding this section, we evaluate the result. Note that the norms of the $\varphi_\ell$'s tend to be $\mathcal{O}(1)$, and so the noiseless measurements $|\langle x, \varphi_\ell \rangle|^2$ tend to be of size $\mathcal{O}(\|x\|^2/M)$. Also, in the worst-case scenario, the noise annihilates our measurements $\nu_\ell = -|\langle x, \varphi_\ell \rangle|^2$, rendering the signal $x$ unrecoverable; in this case, $\|\nu\| = \mathcal{O}(\|x\|^2 \sqrt{(\log M)/M})$ since $N = CM \log M$. In other words, if we allowed the noise-to-signal ratio to scale slightly larger than $C'/\sqrt{M}$ (i.e., by a log factor), then it would be impossible to perform phase retrieval in the worst case. As such, the above guarantee is optimal in some sense. Furthermore, since $\sqrt{M/\log M}\ \mathrm{NSR} = \mathcal{O}(1/\sqrt{\log M})$ by assumption, the result indicates that our phase retrieval process exhibits more stability as $M$ grows large.

In comparison to the state of the art, namely, the work of Candès, Strohmer and Voroninski [67], the most visible difference between our stability results is how we choose to scale the measurement vectors. Indeed, the measurement vectors of the present paper tend to have norm $\mathcal{O}(1)$, whereas the measurement vectors of Candès et al. are all scaled to have norm $\sqrt{M}$; considering the statement of Theorem 2.2.8 is riddled with square roots of $M$, either choice of scaling is arguably natural. For the sake of a more substantial comparison, their result (Theorem 1.2) gives that if the $\varphi_\ell$'s are uniformly sampled from the sphere of radius $\sqrt{M}$, and if $\|\nu\| \leq \epsilon$, then with high probability, there exists $C_0 > 0$ such that

$$\frac{\|\tilde{x} - \mathrm{e}^{\mathrm{i}\theta}x\|}{\|x\|} \leq C_0 \min\left\{1, \frac{\epsilon}{\|x\|^2}\right\}$$

for some phase $\theta \in [0, 2\pi)$; here, $\tilde{x}$ is the estimate which comes from PhaseLift. (This result actually suffered from the subtlety that it does not hold for all signals $x \in \mathbb{C}^M$ simultaneously, but this was later rectified in a sequel [64].) Note that the 1 in the minimum takes effect when $\|x\|^2 < \epsilon$, meaning it is possible that $\nu_\ell = -|\langle x, \varphi_\ell\rangle|^2$ (corresponding to our worst-case scenario, above); as such, this part of the guarantee is not as interesting. The other part of the guarantee is particularly interesting: Ignoring the $\sqrt{M/\log M}$ factor in Theorem 2.2.8 and identifying $\epsilon/\|x\|^2$ with NSR, we see that the main difference between the two guarantees is that Candès et al. bound relative error in terms of NSR rather than bounding *squared* relative error. In this sense, their result is stronger. On the other hand, they take $\epsilon$ as an input to their reconstruction algorithm (PhaseLift), whereas our method is agnostic to the size of $\nu$. In particular, our guarantee is continuous in the sense that relative error vanishes with $\nu$.

## 2.3 The little Grothendieck problem over the orthogonal group

This section (mostly based on [36]) will provide an approximation algorithm for the Procrustes problem by rewriting it as a little Grothendieck problem over $O(d)$, the group of orthogonal matrices. For $d = 1$, the little Grothendieck problem [18] in combinatorial optimization is written as

$$\max_{x_i\in\{\pm 1\}} \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}x_ix_j, \tag{2.18}$$

where $C$ is a $n\times n$ positive semidefinite matrix real matrix. It can be viewed as an instance of Problem 1.2.1 by taking $\mathcal{G}=\mathbb{Z}_2$ and $f_{ij}\left(g_ig_j^{-1}\right)=-C_{ij}x_ix_j^{-1}$. Note that, if $C$ is a Laplacian matrix of a graph then (2.18) is equivalent to the `Max-Cut` problem. The seminal paper of Goemans and Williamson [114] provides a semidefinite relaxation for (2.18):

$$\max_{\substack{X_i\in\mathbb{R}^n\\ \|X_i\|^2=1}} \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}X_i^TX_j. \tag{2.19}$$

It is readily seen that (2.19) is equivalent to a semidefinite program and can be solved, to arbitrary precision, in polynomial time [227]. In the same paper [114] it is shown that a simple rounding technique is guaranteed to produce a solution whose objective value is, in expectation, at least a multiplicative factor $\frac{2}{\pi}\min_{0\leq\theta\leq\pi}\frac{\theta}{1-\cos\theta}\approx 0.878$ of the optimum.

A few years later, Nesterov [174] showed an approximation ratio of $\frac{2}{\pi}$ for the general case of an arbitrary positive semidefinite $C\succeq 0$ using the same relaxation as [114]. This implies, in particular, that the value of (2.18) can never be smaller than $\frac{2}{\pi}$ times the value of (2.19). Interestingly, such an inequality was already known from the influential work of Alexander Grothendieck on norms of tensor products of Banach spaces [118] (see [181] for a survey on this).

Several more applications have since been found for the Grothendieck problem (and variants), and its semidefinite relaxation. Alon and Naor [18] showed applications to estimating the cut-norm of a matrix; Ben-Tal and Nemirovski [46] showed applications to control theory; Briet, Buhrman, and Toner [58] explored connections with quantum non-locality; and many more (see [16]).

In this section, we will focus on a natural generalization of problem (2.18), the little Grothendieck problem over the orthogonal group [36], where the variables are now elements of the orthogonal group $O(d)$, instead of $\{\pm 1\}$. More precisely, given $C \in \mathbb{R}^{dn\times dn}$ a positive semidefinite matrix, we consider the problem

$$\max_{O_1,\dots,O_n \in O(d)} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Tr}\left(C_{ij}^T O_i O_j^T\right), \tag{2.20}$$

where $C_{ij}$ denotes the $(i,j)$-th $d \times d$ block of $C$. Similarly for the $d = 1$ case, it can be viewed as an instance of Problem 1.2.1 by taking $\mathcal{G} = O(d)$ and $f_{ij}\left(g_i g_j^{-1}\right) = -\operatorname{Tr}\left(C_{ij} O_i O_j^{-1}\right)$.

As we will see in Section 2.3.1, several problems can be written in the form (2.20), such as the Procrustes problem [198, 173, 210] described in Section 1.2.6 and Global Registration [76]. Moreover, the approximation ratio we obtain for (2.20) translates into the same approximation ratio for these applications, improving over the best previously known approximation ratio of $\frac{1}{2\sqrt{2}}$, given by [171] for these problems.

We also note that (2.20) coincides, up to an additive shift, with Synchronization over $O(d)$, unfortunately the nature of approximation ratios render them less meaningful after an additive shift.

Problem (2.20) belongs to a wider class of problems considered by Nemirovski [173] called QO-OC (Quadratic Optimization under Orthogonality Constraints), which itself is a subclass of QC-QP (Quadratically Constrained Quadratic Programs). Please refer to Section 2.3.1 for a more detailed comparison with the results of Nemirovski. More recently, Naor et al. [171] propose an efficient rounding for the non commutative Grothendieck inequality that provides an approximation algorithm for a vast set of problems involving orthogonality constraints, including problems of the form of (2.20). However, although our result only holds for a subclass of the problems considered in [171] it has a better approximation ratio and appears to consist of a smaller relaxation.

Similarly to (2.19) we formulate a semidefinite relaxation we name the *Orthogonal-Cut* SDP:

$$\max_{\substack{X_iX_i^T=I_{d\times d}\\ X_i\in\mathbb{R}^{d\times dn}}} \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Tr}\left(C_{ij}^T X_i X_j^T\right). \tag{2.21}$$

Problem (2.21) is equivalent to the semidefinite program

$$\max_{\substack{G\in\mathbb{R}^{dn\times dn}\\ G_{ii}=I_{d\times d},\ G\succeq 0}} \operatorname{Tr}(CG), \tag{2.22}$$

and so can be solved efficiently [227].

One of the main contributions of this paper is showing that Algorithm 2.3.3 gives a constant factor approximation to (2.20), with an optimal approximation ratio for our relaxation. It consists of a simple generalization of the rounding in [114] applied to (2.21).

**Theorem 2.3.1.** *Let $C \succeq 0$ and real. Let $V_1, \ldots, V_n \in O(d)$ be the (random) output of the orthogonal version of Algorithm 2.3.3. Then*

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Tr}\left(C_{ij}^T V_i V_j^T\right)\right] \geq \alpha_{\mathbb{R}}(d)^2 \max_{O_1,\ldots,O_n\in O(d)} \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Tr}\left(C_{ij}^T O_i O_j^T\right),$$

*where $\alpha_{\mathbb{R}}(d)$ is the constant defined below.*

**Definition 2.3.2.** *Let $G_{\mathbb{R}} \in \mathbb{R}^{d\times d}$ be a gaussian random matrix with i.i.d real valued entries $\mathcal{N}\left(0, d^{-1}\right)$ We define*

$$\alpha_{\mathbb{R}}(d) := \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d} \sigma_j(G_{\mathbb{R}})\right],$$

*where $\sigma_j(G)$ is the $j$th singular value of $G$.*

Although we do not have a complete understanding of the behavior of $\alpha_{\mathbb{R}}(d)$ as functions of $d$, we can, for each $d$ separately, compute a closed form expression. For

$d = 1$ we recover the sharp $\alpha_{\mathbb{R}}(1)^2 = \frac{2}{\pi}$ results of Nesterov [174]. One can also show that $\lim_{d\to\infty} \alpha_{\mathbb{R}}(d)^2 = \left(\frac{8}{3\pi}\right)^2$ which is larger than $\frac{2}{\pi}$. We find the fact that the approximation ratio seems to get better, as the dimension increases, quite intriguing. One might naively think that the problem for a specific $d$ can be formulated as a degenerate problem for a larger $d$, however this does not seem to be true. Unfortunately, we were unable to provide a proof for the monotonicity of $\alpha_{\mathbb{R}}(d)$ and leave it as an open problem (Conjecture 2.3.5). Nevertheless, lower bounds that have the right asymptotics can be shown. In particular, we refer the reader to [36] for a proof that these approximation ratios are uniformly bounded below by the approximation ratio given in [171]. In fact, the approximation ratios $\alpha_{\mathbb{R}}(d)^2$ are optimal, in the integrality gap since: we refer the reader to [36] for a constructions of $C$ for which the quotient of the value of (2.21) over (2.20) confirms the optimality of $\alpha_{\mathbb{R}}(d)^2$. In a nutshell this is done by adapting the classical construction for the $d = 1$ case (see, e.g., [18]).[4]

The (randomized) approximation algorithm we propose to solve (2.20) is as follows.

**Algorithm 2.3.3.** *Compute $X_1, \ldots, X_n \in \mathbb{R}^{d\times nd}$ a solution to (2.21). Let $R$ be a $nd \times d$ gaussian random matrix whose entries are real i.i.d. $\mathcal{N}(0, \frac{1}{d})$. The approximate solution for (2.20) is now computed as*

$$V_i = \mathcal{P}(X_i R).$$

Recall that $\mathcal{P}(X) = \operatorname{argmin}_{Z\in O(d)} \|Z - X\|_F$, which can be easily computed via the singular value decomposition of $X = U\Sigma V^T$ as $\mathcal{P}(X) = UV^T$ (see [102, 135, 128]).

[4] As the reader can see in [36], when adapting the construction there is an extra difficult that can solved by using the Lowner-Heinz Theorem on operator convexity [71]

### 2.3.1 Applications

Problem (2.20) can describe several problems of interest. As examples, we describe below how it encodes a complementary version of the orthogonal Procrustes problem (see Section 1.2.6) and the problem of Global Registration over Euclidean Transforms.

#### Orthogonal Procrustes

Recall the setting of Orthogonal Procrustes described in Section 1.2.6. Given $n$ point clouds in $\mathbb{R}^d$ of $k$ points each, the orthogonal Procrustes problem [198] consists of finding $n$ orthogonal transformations that best simultaneously align the point clouds. If the points are represented as the columns of matrices $A_1, \dots, A_n$, where $A_i \in \mathbb{R}^{d\times k}$ then the orthogonal Procrustes problem consists of solving

$$\min_{O_1,\dots,O_n \in O(d)} \sum_{i,j=1}^{n} ||O_i^T A_i - O_j^T A_j||_F^2. \tag{2.23}$$

Since $||O_i^T A_i - O_j^T A_j||_F^2 = \|A_i\|_F^2 + \|A_j\|_F^2 - 2\operatorname{Tr}\left((A_i A_j^T)^T O_i O_j^T\right)$, (2.23) has the same solution as the complementary version of the problem

$$\max_{O_1,\dots,O_n \in O(d)} \sum_{i,j=1}^{n} \operatorname{Tr}\left((A_i A_j^T)^T O_i O_j^T\right). \tag{2.24}$$

Since $C \in \mathbb{R}^{dn\times dn}$ given by $C_{ij} = A_i A_j^T$ is positive semidefinite, problem (2.24) is encoded by (2.20) and Algorithm 2.3.3 provides a solution with an approximation ratio guaranteed (Theorem 2.3.1) to be at least $\alpha_{\mathbb{R}}(d)^2$.

The algorithm proposed in Naor et al. [171] gives an approximation ratio of $\frac{1}{2\sqrt{2}}$, smaller than $\alpha_{\mathbb{R}}(d)^2$, for (2.24). Nemirovski [173] proposed a different semidefinite relaxation (with a matrix variable of size $d^2n \times d^2n$ instead of $dn \times dn$ as in (2.21)) for the orthogonal Procrustes problem. In fact, his algorithm approximates the slightly different problem

$$\max_{O_1,\ldots,O_n\in O(d)} \sum_{i\neq j} \operatorname{Tr}\left((A_iA_j^T)^T O_i O_j^T\right), \tag{2.25}$$

which is an additive constant (independent of $O_1,\ldots,O_n$) smaller than (2.24). The best known approximation ratio for this semidefinite relaxation, due to So [210], is $\mathcal{O}\left(\frac{1}{\log(n+k+d)}\right)$. Although an approximation to (2.25) would technically be stronger than an approximation to (2.24), the two quantities are essentially the same provided that the point clouds are indeed perturbations of orthogonal transformations of the same original point cloud, as is the case in most applications (see [171] for a more thorough discussion on the differences between formulations (2.24) and (2.25)).

### Global Registration over Euclidean Transforms

The problem of global registration over Euclidean rigid motions is an extension of orthogonal Procrustes. In global registration, one is required to estimate the positions $x_1,\ldots,x_k$ of $k$ points in $\mathbb{R}^d$ and the unknown rigid transforms of $n$ local coordinate systems given (perhaps noisy) measurements of the local coordinates of each point in some (though not necessarily all) of the local coordinate systems. The problem differs from Procrustes in two aspects: First, for each local coordinate system, we need to estimate not only an orthogonal transformation but also a translation in $\mathbb{R}^d$. Second, each point may appear in only a subset of the coordinate systems. Despite those differences, it is shown in [76] that global registration can also be reduced to the form (2.20) with a matrix $C$ that is positive semidefinite.

More precisely, denoting by $P_i$ the subset of points that belong to the $i$-th local coordinate system ($i = 1\ldots n$), and given the local coordinates

$$x_l^{(i)} = O_i^T\left(x_k - t_i\right) + \xi_{il}$$

of point $x_l \in P_i$ (where $O_i$ denotes the orthogonal transformation, $t_i$ a translation and $\xi_{il}$ a noise term). The idea is then to minimize the function

$$\phi = \sum_{i=1}^{n} \sum_{l\in P_i} \left\| x_l - (O_i x_l^{(i)} + t_i) \right\|^2 ,$$

over $x_l, t_i \in \mathbb{R}^d, O_i \in O(d)$. It is not difficult to see that the optimal $x_l^\star$ and $t_i^\star$ can be written in terms of $O_1, \ldots, O_n$. Substituting them back into $\phi$, the authors in [76] reduce the previous optimization to solving

$$\max_{O_i\in O(d)} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Tr}\left( \left[BL^\dagger B^T\right]_{ij} O_i O_j^T \right) , \tag{2.26}$$

where $L$ is a certain Laplacian matrix and $L^\dagger$ is its pseudo inverse (see [76]). This means that $BL^\dagger B^T \succeq 0$, and (2.26) is of the form of (2.20).

### 2.3.2 Analysis of the approximation algorithm

We now prove Theorem 2.3.1. As (2.21) is a relaxation of problem (2.20) its maximum is necessarily at least as large as the one of (2.20). This means that Theorem 2.3.1 is a direct consequence of the following Theorem.

**Theorem 2.3.4.** *Let $C \succeq 0$ and real. Let $X_1, \ldots, X_n$ be a feasible solution to (2.21). Let $V_1, \ldots, V_n \in O(d)$ be the output of the (random) rounding procedure described in Algorithm 2.3.3. Then*

$$\mathbb{E}\left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Tr}\left( C_{ij}^T V_i V_j^T \right) \right] \geq \alpha_{\mathbb{R}}(d)^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Tr}\left( C_{ij}^T X_i X_j^T \right) ,$$

*where $\alpha_{\mathbb{R}}(d)$ is the constant in Definition 2.3.2.*

Before proving Theorem 2.3.4 we present a sketch of the proof for the case $d = 1$ (and real). The argument is known as the Rietz method (See [18]):

Let $X_1, \ldots, X_n \in \mathbb{R}^{1\times n}$ be a feasible solution to (2.21), meaning that $X_i X_i^T = 1$.

Let $R \in \mathbb{R}^{n\times 1}$ be a random matrix with i.i.d. standard gaussian entries. Our objective is to compare $\mathbb{E}\left[\sum_{i,j}^{n} C_{ij}\operatorname{sign}(X_iR)\operatorname{sign}(X_jR)\right]$ with $\sum_{i,j}^{n} C_{ij}X_iX_j^T$. The main observation is that although $\mathbb{E}\left[\operatorname{sign}(X_iR)\operatorname{sign}(X_jR)\right]$ is not a linear function of $X_iX_j^T$, the expectation $\mathbb{E}\left[\operatorname{sign}(X_iR)X_jR\right]$ is. In fact $\mathbb{E}\left[\operatorname{sign}(X_iR)X_jR\right] = \alpha_{\mathbb{R}}(1)X_iX_j^T = \sqrt{\frac{2}{\pi}}X_iX_j^T$ - which follows readily by thinking of $X_i, X_j$ as their projections on a 2 dimensional plane. We use this fact (together with the positiveness of $C$) to show our result. The idea is to build the matrix $S \succeq 0$,

$$S_{ij} = \left(X_iR - \sqrt{\frac{\pi}{2}}\operatorname{sign}(X_iR)\right)\left(X_jR - \sqrt{\frac{\pi}{2}}\operatorname{sign}(X_jR)\right).$$

Since both $C$ and $S$ are PSD, $\operatorname{Tr}(CS) \geq 0$, which means that

$$0 \leq \mathbb{E}\left[\sum_{ij} C_{ij}(X_iR - \sqrt{\frac{\pi}{2}}\operatorname{sign}(X_iR))(X_jR - \sqrt{\frac{\pi}{2}}\operatorname{sign}(X_jR))\right].$$

Combining this with the observation above and the fact that $\mathbb{E}\left[X_iRX_jR\right] = X_iX_j^T$, we have

$$\mathbb{E}\sum_{i,j}^{n} C_{ij}\operatorname{sign}(X_iR)\operatorname{sign}(X_jR) \geq \frac{2}{\pi}\sum_{i,j}^{n} C_{ij}X_iX_j^T.$$

*Proof.* [of Theorem 2.3.4] Let $R \in \mathbb{R}^{nd\times d}$ be a gaussian random matrix with i.i.d entries $\mathcal{N}\left(0,\frac{1}{d}\right)$. We want to lower bound

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Tr}\left(C_{ij}^TV_iV_j^T\right)\right] = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Tr}\left(C_{ij}^T\mathcal{P}(U_iR)\mathcal{P}(U_jR)^T\right)\right].$$

Similarly to the $d=1$ case, one of the main ingredients of the proof is the fact that, for any $M, N \in \mathbb{R}^{d\times dn}$ such that $MM^T = NN^T = I_{d\times d}$,

$$\mathbb{E}\left[\mathcal{P}(MR)(NR)^T\right] = \mathbb{E}\left[(MR)\mathcal{P}(NR)^T\right] = \alpha_{\mathbb{R}}(d)MN^T.$$

We refer the reader to [36] for a proof of this fact.

Just as above, we define the positive semidefinite matrix $S \in \mathbb{R}^{dn \times dn}$ whose $(i,j)$-th block is given by

$$S_{ij} = \left(U_iR - \alpha_{\mathbb{R}}(d)^{-1}\mathcal{P}(U_iR)\right)\left(U_jR - \alpha_{\mathbb{R}}(d)^{-1}\mathcal{P}(U_jR)\right)^T.$$

We have

$$\begin{aligned}
\mathbb{E}S_{ij} &= \mathbb{E}\left[U_iR(U_jR)^T - \alpha_{\mathbb{R}}(d)^{-1}\mathcal{P}(U_iR)(U_jR)^T\right.\\
&\qquad \left. -\alpha_{\mathbb{R}}(d)^{-1}U_iR\mathcal{P}(U_jR)^T + \alpha_{\mathbb{R}}(d)^{-2}\mathcal{P}(U_iR)\mathcal{P}(U_jR)^T\right]\\
&= U_i\mathbb{E}\left[RR^T\right]U_j^T - \alpha_{\mathbb{R}}(d)^{-1}\mathbb{E}\left[\mathcal{P}(U_iR)(U_jR)^T\right]\\
&\qquad -\alpha_{\mathbb{R}}(d)^{-1}\mathbb{E}\left[U_iR\mathcal{P}(U_jR)^T\right] + \alpha_{\mathbb{R}}(d)^{-2}\mathbb{E}\left[V_iV_j^T\right]\\
&= U_iU_j^T - U_iU_j^T - U_iU_j^T + \alpha_{\mathbb{R}}(d)^{-2}\mathbb{E}\left[V_iV_j^T\right]\\
&= \alpha_{\mathbb{R}}(d)^{-2}\mathbb{E}\left[V_iV_j^T\right] - U_iU_j^T.
\end{aligned}$$

By construction $S \succeq 0$. Since $C \succeq 0$, $\mathrm{Tr}(CS) \geq 0$, which means that

$$0 \leq \mathbb{E}\left[\mathrm{Tr}\left(CS\right)\right] = \mathrm{Tr}\left(C\mathbb{E}[S]\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Tr}\left(C_{ij}^T\left(\alpha_{\mathbb{R}}(d)^{-2}\mathbb{E}\left[V_iV_j^T\right] - U_iU_j^T\right)\right).$$

Thus,

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Tr}\left(C_{ij}^TV_iV_j^T\right)\right] \geq \alpha_{\mathbb{R}}(d)^2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Tr}\left(C_{ij}^TU_iU_j^T\right).$$

□

### 2.3.3 The approximation ratio $\alpha_{\mathbb{R}}(d)^2$

The approximation ratio we obtain (Theorem 2.3.1) for Algorithm 2.3.3 is given by $\alpha_{\mathbb{R}}(d)^2$, which is defined as the average singular value of a $d \times d$ Gaussian matrix $G$ with i.i.d $\mathcal{N}(0, \frac{1}{d})$ entries. These singular values correspond to the square root of the eigenvalues of a Wishart matrix $W = GG^T$, which are well-studied objects (see, e.g.,

[201] or [87]).

For $d = 1$, this corresponds to the expected value of the absolute value of standard gaussian (real or complex) random variable. Hence,

$$\alpha_{\mathbb{R}}(1) = \sqrt{\frac{2}{\pi}},$$

meaning that, for $d = 1$, we recover the approximation ratio of $\frac{2}{\pi}$, of Nesterov [174].

For any $d \geq 1$, the marginal distribution of an eigenvalue of the Wishart matrix $W = GG^T$ is known [148, 87, 146]. Denoting by $p_d$ the marginal distribution, we have

$$\alpha_{\mathbb{R}}(d) = \frac{1}{d^{1/2}} \int_0^{\infty} x^{1/2} p_d(x) dx. \tag{2.27}$$

One can easily evaluate $\lim_{d\to\infty} \alpha_{\mathbb{R}}(d)$ by noting that the distribution of the eigenvalues of the Wishart matrix we are interested in, as $d \to \infty$, converges in probability to the Marchenko Pastur distribution [201] with density

$$\mathrm{mp}(x) = \frac{1}{2\pi x}\sqrt{x(4-x)}\mathbf{1}_{[0,4]}.$$

This immediately gives,

$$\lim_{d\to\infty} \alpha_{\mathbb{R}}(d) = \int_0^4 \sqrt{x}\frac{1}{2\pi x}\sqrt{x(4-x)}dx = \frac{8}{3\pi}.$$

Although one could potentially obtain lower bounds for $\alpha_{\mathbb{R}}^2(d)$ from results on the rate of convergence to $\mathrm{mp}(x)$ [116], the sharpest known lower bounds are obtained by writing $p_d(x)$ in terms of Laguerre polynomials and estimating these [36]. In fact, it can be shown [36] that

$$\alpha_{\mathbb{R}}(d) \geq \frac{8}{3\pi} - \frac{9.07}{d}.$$

This strongly suggests the following conjecture (supported by numerical computations).

**Conjecture 2.3.5.** *Let $\alpha_{\mathbb{R}}(d)$ be the average singular value of a $d \times d$ matrix with random i.i.d. $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries (see Definition 2.3.2). Then, for all $d \geq 1$,*

$$\alpha_{\mathbb{R}}(d+1) \geq \alpha_{\mathbb{R}}(d).$$

### Extensions

In some applications, such as the Common Lines problem [204] (related to the problem described in Section 1.2.2), one is interested in a more general version of (2.20) where the variables take values in the Stiefel manifold $O(d,r)$, the set of matrices $O \in \mathbb{R}^{d\times r}$ such that $OO^T = I_{d\times d}$. This motivates considering a generalized version of (2.20) formulated as, for $r \geq d$,

$$\max_{O_1,\dots,O_n \in O(d,r)} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Tr}\left(C_{ij}^T O_i O_j^T\right), \tag{2.28}$$

for $C \succeq 0$. The special case $d = 1$ was formulated and studied in [58, 60, 59] in the context of quantum non-locality and quantum XOR games. Note that in the special case $r = nd$, (2.28) reduces to (2.21) and is equivalent to a semidefinite program. In [36] a simple adaption of Algorithm 2.3.3, is proposed and shown to have sharp approximation ratios.

Another important extension is to consider the little Grothendieck problem over the unitary group $U(d)$, the group of complex valued matrices $U \in \mathbb{C}^{d\times d}$ such that $UU^* = I$. The results above hold almost verbatim for this setting and are considered in [36]. Moreover, for the case of $U(1) \cong SO(2)$, which is related to angular synchronize, the approximation ratio obtained with these techniques matched the sharp approximation ratio in [209].

## 2.4 The Row-by-Row method for the little Grothendieck SDP

We use this section to briefly discuss methods to solve the SDP (2.19) as it is a part of many of the algorithms presented in this thesis.

While there are polynomial time methods, based on interior points, to solve (2.19) up to arbitrary accuracy [227], these methods tend to be slow in practice and not suitable for large scale instances. Many alternatives exist, for example the Alternating Direction Method of Multipliers (ADMM) approach [236].

For $d = 1$, there is a dedicated row-by-row method [235] that optimizes 2.19 each row at a time. In what follows, we will show an adaptation of this method for any $d \geq 1$.

Recall the formulation of (2.19):

$$\begin{array}{c} \max \operatorname{Tr}(CX) \\ \text{s.t. } X \succeq 0 \text{ and } X_{ii} = I_{d\times d} \end{array},$$

for $C \in \mathbb{R}^{nd\times nd}$ symmetric.

In a nutshell, the idea of the row-by-row method [235], is to initialize $X$ as any feasible point and then update $X$, each row at once, by finding the optimal values for that row that keep the feasibility of $X$. In the special case of problem (2.19) one can initialize with $X = I_{dn\times dn}$. We proceed by describing the row subproblem (without loss of generality we assume that we are changing the first row). Let $X$ be the current iterate (which is a feasible for (2.19)) and we write

$$X_0 = \begin{bmatrix} 1 & y^T \\ y & B \end{bmatrix},$$

with $y \in \mathbb{R}^{nd-1}$. Furthermore, we write:

$$B = \begin{bmatrix} I_{(d-1)\times(d-1)} & b_1^T \\ b_1 & B' \end{bmatrix},$$

where $b_1 \in \mathbb{R}^{nd-1 \times d-1}$.

$$\begin{aligned} &\max c^T y \\ \text{s.t. } &\begin{bmatrix} 1 & y^T \\ y & B \end{bmatrix} \succeq 0 \text{ and } y[1:d-1] = 0, \end{aligned} \tag{2.29}$$

where $c \in \mathbb{R}^{nd-1}$ is twice the last $nd-1$ terms of the first column of $C$.

A solution to the similar problem:

$$\begin{aligned} &\max c^T y \\ \text{s.t. } &\begin{bmatrix} 1 & y^T \\ y & B \end{bmatrix} \succeq 0, \end{aligned} \tag{2.30}$$

can be computed by a simple matrix-vector product (the idea is to use the Schur complement, see [235]) and is given by:

$$y = \frac{1}{\sqrt{c^T B c}} Bc, \tag{2.31}$$

if $c^T Bc > 0$ and $y = 0$, otherwise.

The idea for the adaptions is simple: The solution of (2.29) does not depend on the first $d-1$ entries of $c$ so we will design them in such a way that the solution of (2.30) satisfies $y[1:d-1] = 0$ which will ensure that the solution of the two problems match and so we can solve (2.29) with a matrix-vector product.

Let us rewrite $y = [y_0^T y'^T]^T$, where $y' \in \mathbb{R}^{(n-1)d}$ and (2.31) as:

$$\begin{bmatrix} y_0 \\ y' \end{bmatrix} = \frac{1}{\sqrt{c^T Bc}} \begin{bmatrix} I_{(d-1)\times(d-1)} & b_1^T \\ b_1 & B' \end{bmatrix} \begin{bmatrix} c_0 \\ c' \end{bmatrix} = \begin{bmatrix} c_0 + b_1^T c' \\ b_1 c_0 + B'c \end{bmatrix}. \tag{2.32}$$

This means we should set $c_0 = -b_1^T c'$ to get $y_0 = 0$. In this case,

$$y' = \frac{1}{\sqrt{c^T Bc}} \left(B' - b_1 b_1^T\right) c'.$$

We also have

$$c^T Bc = [\big( - b_1^T c'\big)^T c'^T] \begin{bmatrix} 0 \\ \left(B' - b_1 b_1^T\right) c' \end{bmatrix} = c'^T \left(B' - b_1 b_1^T\right) c'.$$

This shows the following Theorem.

**Theorem 2.4.1.** *The solution to the subproblem (2.29) is given by* $y = [0_{d-1}^T y'^T]^T$ *where* $y'$ *can be obtained by a matrix-vector product:*

$$y' = \frac{1}{\sqrt{c'^T \left(B' - b_1 b_1^T\right) c'}} \left(B' - b_1 b_1^T\right) c',$$

*if* $c'^T \left(B' - b_1 b_1^T\right) c' > 0$*, and* $y' = 0$ *otherwise.*

# Chapter 3

# Exact recovery for random instances: Synchronization over $\mathbb{Z}_2$

## 3.1 Semidefinite relaxations for Synchronization-type problems over $\mathbb{Z}_2$

In this chapter we will treat synchronization-type problems (Problem 1.2.1) over $\mathbb{Z}_2$, the group of two elements [1, 2, 3, 29, 88]. In particular we will investigate the tendency for, in this setting, the semidefinite relaxation (1.9) to achieve exact recovery, i.e. the solution of the semidefinite programming (1.9) corresponds desired group potential. Most of this chapter is based on [1, 2, 3, 29]. Let us recall the setting.

**Problem 3.1.1.** *[General Synchronization-type problem in $\mathbb{Z}_2$] Given a graph $G = (V, E)$, and, for each edge $(i, j) \in E$, a function $f_{ij} : \mathbb{Z}_2 \to \mathbb{R}$. The goal is to find the group potential $g : V \to \mathbb{Z}_2$ that minimizes*

$$\min_{g:V\to\mathbb{Z}_2} \sum_{(i,j)\in E} f_{ij}\left(g_i g_j^{-1}\right). \tag{3.1}$$

We will identify $\mathbb{Z}_2$ with ±1 (the operation being multiplication). Given $f_{ij}$ we define $f_{ij}^{(+)} = f_{ij}(1)$ and $f_{ij}^{(-)} = f_{ij}(-1)$. Let $x_i \in \pm 1$, we note that

$$\begin{aligned} f_{ij}\left(x_i x_j^{-1}\right) = f_{ij}\left(x_i x_j\right) &= f_{ij}^{(+)} \frac{1 + x_i x_j}{2} + f_{ij}^{(-)} \frac{1 - x_i x_j}{2} \\ &= \left(\frac{f_{ij}^{(+)} - f_{ij}^{(-)}}{2}\right) x_i x_j + \frac{f_{ij}^{(+)} + f_{ij}^{(-)}}{2}. \end{aligned}$$

This means that, by setting $Y \in \mathbb{R}^{n\times n}$ to satisfy $Y_{ij} = -\left(\frac{f_{ij}^{(+)} - f_{ij}^{(-)}}{2}\right)$ whenever $(i,j) \in E$ and $Y_{ij} = 0$ otherwise, (3.1) has the same optimizers as

$$\begin{array}{ll} \max & \sum_{ij} Y_{ij} x_i x_j \\ \text{s. t.} & x_i \in \{\pm 1\}. \end{array} \tag{3.2}$$

We note the similarities between this formulation and (1.5). In this case, the SDP relaxation in (1.9) can be rewritten (as an optimization problem over matrices $X \in \mathbb{R}^{n\times n}$ as:

$$\begin{array}{ll} \max & \operatorname{Tr}(YX) \\ \text{s. t.} & X_{ii} = 1 \\ & -1 \leq X_{ij} \leq 1 \\ & X \succeq 0. \end{array} \tag{3.3}$$

Note however that the constraint $-1 \leq X_{ij} \leq 1$ is redundant, and so (3.3) it is equivalent to:

$$\begin{array}{ll} \max & \operatorname{Tr}(YX) \\ \text{s. t.} & X_{ii} = 1 \\ & X \succeq 0. \end{array} \tag{3.4}$$

This SDP corresponds exactly to (2.19) and the one proposed in [114]. Note that if the unique optimal solution of (3.4) is a rank 1 matrix $X = xx^T$ this means that $x$ is the optimizer of (3.1) and we can compute it; since we can easily compute the solution of (3.4), $X = xx^T$ and recover $x$ by taking the leading eigenvector of

$X$. In what follows we will assume that $Y$ corresponds to measurements with some distribution and that there is an underlying group potential $z \in \{\pm 1\}^n$ that we aim to recover. We will understand when is it the case that $X = zz^T$ is the unique optimal solution of (3.4). Such understanding will allows to propose information theoretically optimal algorithms for a recovery in the stochastic block model in two communities (Section 3.2) and a certain class of inverse problems on Erdős-Rényi graphs (Section 3.3). Before going into those problems, we start with a simpler illustrative instance.

### 3.1.1 A simpler problem: $\mathbb{Z}_2$ Synchronization with Gaussian noise.

Given a noise level $\sigma$ and a vector $z \in \{\pm 1\}^n$, that we wish to recover, suppose we given noisy measurements

$$Y_{ij} = z_i z_j + \sigma W_{ij},$$

for each pair $(i, j)$, where $W_{ij}$ are i.i.d. standard Gaussian random variables (with $W_{ij} = W_{ji}$). A version of this problem, over the complex numbers, will be treated in Section 5.1. Our objective is to devise an algorithm that recovers the correct $z$ with high probability. By definition, the maximum a posteriori (MAP) estimator maximizes the probability of recovering the correct variable $z$. Given that we have no a priori information on $z$ we assume a uniform prior, in that case the MAP estimator coincides with the Maximum Likelihood Estimator (MLE) for $z$. The latter is the solution of (3.2). We wish to understand when it is that $X = zz^T$ is the unique minimizer of the SDP relaxation (3.4).

A fruitful way of approaching this relies on duality. The dual of (3.4) is given by:

$$\begin{aligned} \min \quad & \operatorname{Tr}(D) \\ \text{s.t.} \quad & D \text{ is diagonal} \\ & D - Y \succeq 0. \end{aligned} \tag{3.5}$$

Weak duality guarantees that if $X$ and $D$ are feasible solutions of respectively (3.4) and (3.5) then $\operatorname{Tr}(YX) \leq \operatorname{Tr}(D)$. Indeed, since $X$ and $D-Y$ are both positive semidefinite, we must have

$$0 \leq \operatorname{Tr}\left[(D-Y)X\right] = \operatorname{Tr}(D) - \operatorname{Tr}(YX). \tag{3.6}$$

This means that if we are able to find a so-called dual certificate, a matrix $D$ feasible for (3.5) for which $\operatorname{Tr}(D) = \operatorname{Tr}(Yxx^T)$, then it guarantees that $X = xx^T$ is an optimal solution of (3.4). To guarantee uniqueness it suffices to further ensure that $\lambda_2(D-Y) > 0$. In fact, if there existed another optimal solution $X$, by (3.6), one would have $\operatorname{Tr}\left[(D-Y)X\right] = 0$ which can be shown to imply, together with the feasibility of $X$, that $X = xx^T$ (see, for example, [1]). This establishes the following Lemma.

**Lemma 3.1.2.** *[Dual Certificate] Let $Y$ be a symmetric $n \times n$ matrix and $x \in \{\pm 1\}^n$. If there exists a diagonal matrix $D$, such that $\operatorname{Tr}(D) = x^T Y x$, $D - Y \succeq 0$, and $\lambda_2(D-Y) > 0$ then $X = xx^T$ is the unique optimal solution of (3.4).*

Taking a candidate dual certificate $D$ whose diagonal elements are given by

$$D_{ii} = \sum_{j=1}^{n} Y_{ij} x_i x_j.$$

Note that $D = D_{[\operatorname{diag}(x) Y \operatorname{diag}(x)]}$. It is easy to see that $\operatorname{Tr}(D) = x^T Y x$ and $(D-Y)x = 0$ which gives the following Lemma.

**Lemma 3.1.3.** *Let $Y$ be a symmetric $n \times n$ matrix and $x \in \{\pm 1\}^n$. Let $D$ be the diagonal matrix defined as $D = D_{[\operatorname{diag}(x) Y \operatorname{diag}(x)]}$. As long as*

$$\lambda_2(D - Y) > 0,$$

*$X = xx^T$ is the unique optimal solution of* (3.4)*.*

Note that these guarantees, (Lemmas 3.1.2 and 3.1.3) do not depend on the matrix $Y$ or the distribution from which it is drawn.

Let us return to the setting on which $Y = zz^T + \sigma W$, where $W$ is a standard Wigner matrix: a symmetric matrix with i.i.d. standard Gaussian entries. We want to determine for which values of $\sigma$ one excepts $X = zz^T$ to be, with high probability, the solution of (3.4), as we are interested not only to compute the MLE but also for it to coincide with the planted vector $z$ we want to recover. Since $\operatorname{diag}(z) W \operatorname{diag}(z) \sim W$ we can, without loss of generality, take $z = \mathbf{1}$. In that case, we are interested in understanding when

$$\lambda_2 \left( D_{\left[\mathbf{1}\,\mathbf{1}^T + \sigma W\right]} - \left(\mathbf{1}\,\mathbf{1}^T + \sigma W\right) \right) > 0. \tag{3.7}$$

Since

$$D_{\left[\mathbf{1}\,\mathbf{1}^T + \sigma W\right]} - \left(\mathbf{1}\,\mathbf{1}^T + \sigma W\right) = \left(n I_{n\times n} - \mathbf{1}\,\mathbf{1}^T\right) - \sigma\left(-D_W + W\right) = L_{\mathbf{1}\,\mathbf{1}^T} - \sigma L_{[-W]},$$

and $\mathbf{1}$ is always in the nullspace of any Laplacian matrix, it is not difficult to see that (3.7) is equivalent to

$$\lambda_{\max}\left(L_{[-W]}\right) < \frac{n}{\sigma}. \tag{3.8}$$

The triangular inequality tells us that $\lambda_{\max}\left(L_{[-W]}\right) \leq \lambda_{\max}\left(-D_W\right) + \|W\|$. It is well known that, for any $\varepsilon > 0$, $\|W\| \leq (2 + \varepsilon)\sqrt{n}$ with high probability (see, for example, Theorem II.11 in [91]). On the other hand,

$$\lambda_{\max}\left(-D_W\right) = \max_{i\in[n]}\left[-\left(D_W\right)_{ii}\right],$$

which is the maximum of $n$ Gaussian random variables each with variance $n$. A simple union bound yields that, for any $\varepsilon > 0$, $\lambda_{\max}\left(D_{[-W]}\right) < \sqrt{(2+\varepsilon)n\log n}$ with high probability. This readily implies an exact recovery guarantee for the $\mathbb{Z}_2$ Synchronization with Gaussian noise.

**Proposition 3.1.4.** *Let $z \in \{\pm 1\}^n$ and $Y = zz^T + \sigma W$ where $W$ is a symmetric matrix with i.i.d. standard Gaussian entries. If there exists $\varepsilon > 0$ such that $\sigma < \sqrt{\frac{n}{(2+\varepsilon)\log n}}$ then, with high probability, $X = zz^T$ is the unique solution to the Semidefinite Program* (3.4)*.*

Let us investigate the optimality of this upper bound on $\sigma$. If the diagonal elements of $D_{[-W]}$ were independent[1], their distribution would be known to indeed concentrate around $\sqrt{2\log n}$, suggesting that

$$\|W\| \ll \lambda_{\max}\left(D_{[-W]}\right), \tag{3.9}$$

which would imply

$$\lambda_{\max}\left(L_{[-W]}\right) = \left[1 + o(1)\right]\lambda_{\max}\left(D_{[-W]}\right). \tag{3.10}$$

Both of these statements can be rigorously shown to be true. While a simple adaptation of the proof of Theorem 4.2.1 can establish (3.9) and (3.10) we omit their proofs for the sake of brevity, but emphasize that in this particular setting (where $W$ is a standard Wigner matrix), one does not need the whole strength of Theorem 4.2.1 as

[1] The diagonal entries of $D_W$ are not independent because each pair of sums shares a term $W_{ij}$ as a summand.

simple elementary proofs exist.

This would suggest that, in rough terms, the success of the relaxation (3.4) depends mostly on whether $\lambda_{\max}\left(D_{[-W]}\right) < \frac{n}{\sigma}$, which is equivalent to

$$\max_{i\in[n]}\left[-\sigma\sum_{j=1}^{n}W_{ij}\right] < n, \tag{3.11}$$

which can be interpreted as a bound on the amount of noise per row of $Y$. We argue next that this type of upper bound is indeed necessary for any method to succeed at recovering $z$ from $Y$.

Once again, let us consider $z = \mathbf{1}$ without loss of generality. Let us consider an oracle version of problem on which one is given the correct label of every single node except of node $i$. It is easy to see that the maximum likelihood estimator for $z_i$ on this oracle problem is given by

$$\text{sign}\left[\sum_{j\in[n]\setminus i}Y_{ij}\right] = \text{sign}\left[n-1+\sigma\sum_{j\in[n]\setminus i}W_{ij}\right],$$

which would give the correct answer if and only if

$$-\sigma\sum_{j\in[n]\setminus i}W_{ij} < n-1. \tag{3.12}$$

This means that if

$$\max_{i\in[n]}\left[-\sigma\sum_{j\in[n]\setminus i}W_{ij}\right] > n-1, \tag{3.13}$$

one does not expect the MLE to succeed (with high probability) at recovering $z$ from $Y = zz^T + \sigma W$. This means that (with a uniform prior on $z$) no method is able to recover $z$ with high probability. Note the similarity between (3.11) and (3.13). This strongly suggest the optimality of the semidefinite programming based approach (3.4).

These optimality arguments can be made rigorous. In fact, in Sections 3.2 and 3.3, we will establish precise optimality results of these type, for the applications we are interested in. The main ingredient (3.9) in the rough argument above was the realization that the spectral norm of $W$ is, with high probability, asymptotically smaller than the largest diagonal entry of $D_{[-W]}$. This provides strong motivation for Theorems 4.2.1 and 4.2.2, which establish precisely this fact for a large class of matrices with independent off-diagonal entries. Empowered with this result, we will be able to establish optimality for the semidefinite programming approach to solve the problems of $\mathbb{Z}_2$ Synchronization in Erdős-Rényi graphs and recovery in the stochastic block model, where the underlying random matrices have much less well understood distributions. Modulo the use of Theorem 4.2.1, the arguments used will be very reminiscent of the the ones above.

It is pertinent to compare this approach with the one of using noncommutative Khintchine inequality, or the related matrix concentration inequalities [220, 221], to estimate the spectral norms in question. Unfortunately, those general purpose methods are, in our case, not fine enough to give us satisfactory results. One illustration of their known suboptimality is the fact that the upper bound they give for $\|W\|$ is of order $\sqrt{n \log n}$, which does not allow to establish (3.9), a crucial step in the argument. In fact, the looseness of these bounds is reflected in the suboptimal guarantees obtained in [1, 2, 3]. Theorems 4.2.1 and 4.2.2 are able to establish a phenomenon of the type of (3.9) by relying on sharp estimates for the spectral norm of matrices with independent entries described in Chapter 4. Although Theorem 4.2.1 will only be established in Chapter 4 we include it below as we will use it in this Chapter.

### Theorem 4.2.1 — see Chapter 4 for a proof

Let $L$ be an $n \times n$ symmetric random Laplacian matrix (i.e. satisfying $L1 = 0$) with centered independent off-diagonal entries such that $\sum_{j \in [n] \setminus i} \mathbb{E} L_{ij}^2$ is equal for every $i$.

Define $\sigma$ and $\sigma_\infty$ as

$$\sigma^2 = \sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2 \quad \text{and} \quad \sigma_\infty^2 = \max_{i\neq j} \|L_{ij}\|_\infty^2 .$$

If there exists $c > 0$ such that

$$\sigma \geq c\,(\log n)^{\frac{1}{2}}\,\sigma_\infty, \tag{3.14}$$

then there exists $c_1$, $C_1$, $\beta_1$, all positive and depending only on $c$, such that

$$\lambda_{\max}(L) \leq \left(1 + \frac{C_1}{(\log n)^{\frac{1}{2}}}\right) \max_i L_{ii}$$

with probability at least $1 - c_1 n^{-\beta_1}$.

Empowered with Theorem 4.2.1, we will give sharp guarantees for certain algorithms to solve the problems of $\mathbb{Z}_2$ Synchronization on an Erdős-Rényi graph and community detection in the Stochastic Block Model.

## 3.2 The Stochastic Block Model with two communities

The problem of community detection, or clustering, in a graph is a central one in countless applications. Unfortunately, even the simplified version of finding a partition of the graph into two balanced vertex sets that minimizes the number of edges across the partition, referred to as minimum bisection, is known to be NP-hard. Nevertheless, certain heuristics are known to work well for typical realizations of random graph models that exhibit a community structure [161, 51, 103]. A particularly popular example of such a random graph model is the Stochastic Block Model with two communities.

**Definition 3.2.1.** *[Stochastic Block Model with two communities] Given $n$ even, and $0 \leq p, q \leq 1$, we say that a random graph $G$ is drawn from $\mathcal{G}(n, p, q)$, the Stochastic Block Model with two communities, if $G$ has $n$ nodes, divided in two clusters of $\frac{n}{2}$ nodes each, and for each pair of vertices $i, j$, $(i, j)$ is an edge of $G$ with probability $p$ if $i$ and $j$ are in the same cluster and $q$ otherwise, independently from any other edge.*

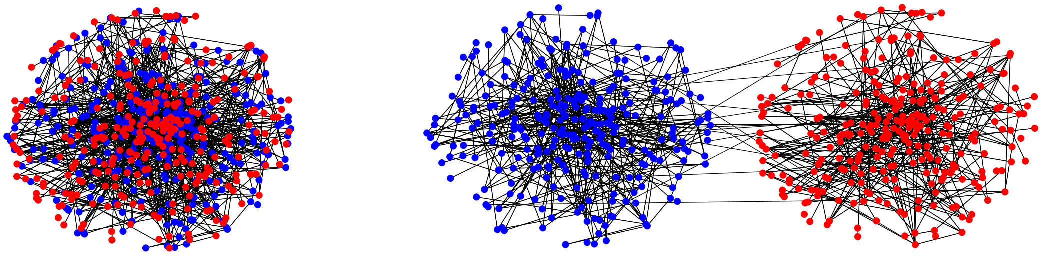

Figure 3.1: A graph generated form the stochastic block model with 600 nodes and 2 communities, scrambled on the left and clustered on the right. Nodes in this graph connect with probability $p = 6/600$ within communities and $q = 0.1/600$ across communities [3].

We will focus on the setting $p > q$. The problem of recovering, from a realization $G \sim \mathcal{G}(n, p, q)$, the original partition of the underlying vertices gained popularity when Decelle et al. [92] conjectured, for the constant average degree regime, a fascinating phase transition. More precisely, if $p = \frac{a}{n}$ and $q = \frac{b}{n}$ with $a > b$ constants, it was conjectured that as long as

$$(a - b)^2 > 2(a + b),$$

it is possible to make an estimate of the original partition that correlates with the true partition, and that below this threshold it is impossible to do so. This conjecture was later proven in a remarkable series of works by Mossel et al. [169, 168] and Massoulie [160]. Instead of settling for an estimate that correlates with the true partition, we will focus on exactly recovering the partition.

Of course, one can only hope to recover the communities up to a global flip of the labels, in other words, only the partition can be recovered. Hence we use the terminology *exact recovery* or simply *recovery* when the partition is recovered correctly with high probability (w.h.p.). When $p = q$, it is clearly impossible to recover the communities, whereas for $p > q$ or $p < q$, one may hope to succeed in certain regimes. While this is a toy model, it captures some of the central challenges for community detection. We will focus on the setting of $p > q$.

There has been a significant body of literature on the recovery property for the stochastic block model with two communities $\mathcal{G}(n, p, q)$, ranging from computer science and statistics literature to machine learning literature. We provide next a partial[2] list of works that obtain bounds on the connectivity parameters to ensure recovery with various algorithms:

| | | |
|---|---|---|
| [62] Bui et al. '84 | min-cut method | $p = \Omega(1/n), q = o(n^{-1-4/((p+q)n)})$ |
| [100] Dyer, Frieze '89 | min-cut via degrees | $p - q = \Omega(1)$ |
| [51] Boppana '87 | spectral method | $(p-q)/\sqrt{p+q} = \Omega(\sqrt{\log(n)/n})$ |
| [208] Snijders, Nowicki '97 | EM algorithm | $p - q = \Omega(1)$ |
| [132] Jerrum, Sorkin '98 | Metropolis algorithm | $p - q = \Omega(n^{-1/6+\varepsilon})$ |
| [86] Condon, Karp '99 | augmentation algorithm | $p - q = \Omega(n^{-1/2+\varepsilon})$ |
| [72] Carson, Impagliazzo '01 | hill-climbing algorithm | $p - q = \Omega(n^{-1/2}\log^4(n))$ |
| [161] Mcsherry '01 | spectral method | $(p-q)/\sqrt{p} \geq \Omega(\sqrt{\log(n)/n})$ |
| [48] Bickel, Chen '09 | N-G modularity | $(p-q)/\sqrt{p+q} = \Omega(\log(n)/\sqrt{n})$ |
| [5] Rohe, Chatterjee, Yu '11 | spectral method | $p - q = \Omega(1)$ |

While these algorithmic developments are impressive, we next argue how they do not reveal the sharp behavioral transition that takes place in this model. In particular, we will obtain an improved bound that is shown to be tight.

### 3.2.1 Information theoretic lower bound

We start by providing information-theoretic lower bounds (this part is based in material in [3] and we note that similar lower bounds were obtained independently by Mossel et al. [167]). In Section 3.2.2 we will analyze the effectiveness of the semidefi-

[2] The approach of McSherry was recently simplified and extended in [230].

nite relaxation, analogous to (3.4).

From a random graph perspective, note that recovery requires the graph to be at least connected (with high probability), hence we require at least logarithmic average inner degree (see Theorem 4.2.6), meaning that $p \geq \Omega\left(\frac{\log n}{n}\right)$. This motivates us to consider the regime:

$$p = \frac{\alpha \log n}{n} \quad \text{and} \quad q = \frac{\beta \log n}{n}.$$

Our objective is too understand for which values of $\alpha$ and $\beta$ is it possible to exactly recover the original partition (up to a global flip). We note that the estimator that maximizes the probability of reconstructing the communities correctly is the Maximum A Posteriori (MAP) estimator. Since we have no a priori information on the community assignment, we consider a uniform prior which renders MAP equivalent to the MLE (see [3]). Hence if MLE fails in reconstructing the communities with high probability when $n$ diverges, there is no algorithm (efficient or not) which can succeed with high probability. However, ML amounts to finding a balanced cut (a bisection) of the graph which minimizes the number of edges across the cut (in the case $a > b$), i.e., the min-bisection problem, which is well-known to be NP-hard. Hence ML can be used[3] to establish the fundamental limit but does not provide an efficient algorithm, which we consider in a second stage.

**Theorem 3.2.2.** *Let $\alpha > \beta \geq 0$. Let $G \sim \mathcal{G}(n,p,q)$, the Stochastic Block Model with two communities, with*

$$p = \frac{\alpha \log n}{n} \quad \textit{and} \quad q = \frac{\beta \log n}{n},$$

*with $\alpha > \beta$. If*

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}, \tag{3.15}$$

[3] ML was also used for the SBM in [80], requiring however poly-logarithmic degrees for the nodes.

*or equivalently*

$$\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} < 1,$$

*then the MLE fails in recovering the communities with probability bounded away from zero.*

*Moreover, there is a node with less connections to each cluster than to the other cluster.*

Before proving Theorem 3.2.2 we note that (3.15) is sharp [3, 167]. In fact, in Section 3.2.2 we will show, Corollary 3.2.12, that above this treshold not only the MLE achieves exact recovery but also the SDP relaxation analogue of (3.4).

Note that the best bounds from the table above are obtained from [51] and [161], which allow for recovery in the regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, obtaining the conditions $(\alpha-\beta)^2 > 64(\alpha+\beta)$ in [161] and $(\alpha-\beta)^2 > 72(\alpha+\beta)$ in [51]. Hence, although these works reach the scaling for $n$ where the threshold takes place, they do not obtain the right threshold behaviour in terms the parameters $\alpha$ and $\beta$.

Let $A$ and $B$ denote the two communities, each with $\frac{n}{2}$ nodes. Also, let

$$\gamma(n) = \log^3 n, \; \delta(n) = \frac{\log n}{\log\log n},$$

and let $H$ be a fixed subset of $A$ of size $\frac{n}{\gamma(n)}$. We define the following events:

$$\begin{cases} F & = & \text{maximum likelihood fails} \\ F_A & = & \exists_{i\in A}: \text{ i is connected to more nodes in B than in A} \\ \Delta & = & \text{no node in H is connected to at least } \delta(n) \text{ other nodes in H} \\ F_H^{(j)} & = & \text{node } j\in H \text{ satisfies } E(j, A\setminus H) + \delta(n) \leq E(j,B) \\ F_H & = & \cup_{j\in H} F_H^{(j)}, \end{cases} \tag{3.16}$$

where $E(\cdot,\cdot)$ is the number of edges between two sets. Note that we identify nodes of our graph with integers with a slight abuse of notation when there is no risk of confusion.

We also define

$$\rho(n) = \mathbb{P}\left(F_H^{(i)}\right) \tag{3.17}$$

**Lemma 3.2.3.** *If $\mathbb{P}(F_A) \geq \frac{2}{3}$ then $\mathbb{P}(F) \geq \frac{1}{3}$.*

*Proof.* By symmetry, the probability of a failure in $B$ is also at least $\frac{2}{3}$ so, by union bound, with probability at least $\frac{1}{3}$ both failures will happen simultaneously which implies that the MLE fails. □

**Lemma 3.2.4.** *If $\mathbb{P}(F_H) \geq \frac{9}{10}$ then $\mathbb{P}(F) \geq \frac{1}{3}$.*

*Proof.* It is easy to see that $\Delta \cap F_H \Rightarrow F_A$. Also, a straighfoward calculation (that we defer to [3]) gives

$$\mathbb{P}(\Delta) \geq \frac{9}{10}. \tag{3.18}$$

Hence,

$$\mathbb{P}(F_A) \geq \mathbb{P}(F_H) + \mathbb{P}(\Delta) - 1 \geq \frac{8}{10} > \frac{2}{3},$$

which together with Lemma 3.2.3 concludes the proof. □

**Lemma 3.2.5.** *Recall the definitions in (3.16) and (3.17). If*

$$\rho(n) > n^{-1}\gamma(n)\log(10)$$

*then, for sufficiently large $n$, $\mathbb{P}(F) \geq \frac{1}{3}$.*

*Proof.* We will use Lemma 3.2.4 and show that if $\rho(n) > n^{-1}\gamma(n)\log(10)$ then $\mathbb{P}(F_H) \geq \frac{9}{10}$, for sufficiently large $n$.

$F_H^{(i)}$ are independent and identically distributed random variables so

$$\begin{aligned}\mathbb{P}\left(F_H\right) &= \mathbb{P}\left(\cup_{i\in H} F_H^{(i)}\right) = 1 - \mathbb{P}\left(\cap_{i\in H}\left(F_H^{(i)}\right)^c\right) \\ &= 1 - \left(1 - \mathbb{P}\left(F_H^{(i)}\right)\right)^{|H|} = 1 - (1-\rho(n))^{\frac{n}{\gamma(n)}}\end{aligned}$$

This means that $\mathbb{P}\left(F_H\right) \geq \frac{9}{10}$ is equivalent to $(1-\rho(n))^{\frac{n}{\gamma(n)}} \leq \frac{1}{10}$. If $\rho(n)$ is not $o(1)$ than the inequality is obviously true, if $\rho(n) = o(1)$ then,

$$\lim_{n\to\infty} (1-\rho(n))^{\frac{n}{\gamma(n)}} = \lim_{n\to\infty} (1-\rho(n))^{\frac{1}{\rho(n)}\rho(n)\frac{n}{\gamma(n)}} = \lim_{n\to\infty} \exp\left(-\rho(n)\frac{n}{\gamma(n)}\right) \leq \frac{1}{10},$$

where the last inequality used the hypothesis $\rho(n) > n^{-1}\gamma(n)\log(10)$. □

**Definition 3.2.6.** [Definition 3 in [3]] *Let $N$ be a natural number, $p, q \in [0,1]$, and $\delta$, we define*

$$T(N,p,q,\delta) = \mathbb{P}\left(\sum_{i=1}^{N}(Z_i - W_i) \geq \delta\right), \tag{3.19}$$

*where $W_1, \ldots, W_N$ are i.i.d. Bernoulli($p$) and $Z_1, \ldots, Z_N$ are i.i.d. Bernoulli($q$), independent of $W_1, \ldots, W_N$.*

We borrow, from [3], an estimate for $T(N,p,q,\delta)$.

**Lemma 3.2.7.** *Let $\alpha > \beta > 0$ constants not depending on $n$ and let $T(N,p,q,\delta)$ be as in Definition 3.2.6. Then*

$$-\log T\left(\frac{n}{2}, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \frac{\log(n)}{\log\log(n)}\right) \leq \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o\left(\log(n)\right). \tag{3.20}$$

*Proof of Theorem 3.2.7.* From the definitions in (3.16) and (3.17) we have

$$\rho(n) = \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}-\frac{n}{\gamma(n)}} W_i \geq \frac{\log(n)}{\log\log(n)}\right) \tag{3.21}$$

where $W_1, \ldots, W_N$ are i.i.d. Bernoulli$\left(\frac{\alpha\log(n)}{n}\right)$ and $Z_1, \ldots, Z_N$ are i.i.d. Bernoulli$\left(\frac{\beta\log(n)}{n}\right)$, all independent. Since

$$\mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}-\frac{n}{\gamma(n)}} W_i \geq \frac{\log(n)}{\log\log(n)}\right) \geq \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i \geq \frac{\log(n)}{\log\log(n)}\right), \tag{3.22}$$

we get

$$-\log\rho(n) \leq -\log T\left(n/2, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \frac{\log(n)}{\log\log(n)}\right), \tag{3.23}$$

and Lemma 3.2.7 implies

$$-\log\rho(n) \leq \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o(\log(n)). \tag{3.24}$$

Hence $\rho(n) > n^{-1}\gamma(n)\log(10)$, and the conclusion follows from Lemma 3.2.5. □

### 3.2.2 Exact recovery for the Stochastic Block Model with two communities via semidefinite relaxation

We shift our attention to the algorithmic question. Recall Definition 3.2.1, and let $G \sim G(n,p,q)$ and $g \in \{\pm 1\}^n$ be a vector that is 1 in one of the clusters and $-1$ in the other. We are interested in efficient algorithms that succeed at recovering $g$, with high probability, all the way up to the information theoretical treshold (3.15).

The maximum likelihood estimator for $g$ is given by

$$\begin{array}{ll} \max & x^T B x \\ \text{s.t.} & x \in \mathbb{R}^n \\ & x_i^2 = 1, \\ & \sum_{i=1}^n x_i = 0, \end{array} \tag{3.25}$$

where $B$ is the signed adjacency of $G$, meaning that $B_{ij} = 1$ if $(i,j)$ is an edge of $G$ and $B_{ij} = -1$ otherwise. Note that $B = 2A - \left(\mathbf{1}\mathbf{1}^T - I\right)$, where $A$ is the adjacency matrix. We will drop the balanced constraint $\sum_{i=1}^n x_i = 0$, arriving at (3.2) for $Y = B$. The intuitive justification is that there are enough $-1$ entries in $B$ to discourage unbalanced constraints. We will consider the semidefinite relaxation (3.4).

$$\begin{array}{ll} \max & \operatorname{Tr}\left[\left(2A - \left(\mathbf{1}\mathbf{1}^T - I\right)\right) X\right] \\ \text{s.t.} & X_{ii} = 1 \\ & X \succeq 0. \end{array} \tag{3.26}$$

We want to understand when is it that $X = gg^T$ is the unique solution of (3.26). Lemma 3.1.3 shows that $gg^T$ is indeed the unique solution of (3.26) as long as the second smallest eigenvalue of

$$D_{\left[\operatorname{diag}(g)\left(2A - \left(\mathbf{1}\mathbf{1}^T - I\right)\right)\operatorname{diag}(g)\right]} - \left[2A - \left(\mathbf{1}\mathbf{1}^T - I\right)\right], \tag{3.27}$$

is strictly positive.

Let us introduce a new matrix.

**Definition 3.2.8.** $[\Gamma_{\text{SBM}}]$ *Given a graph $G$ drawn from the stochastic block model with two clusters,*

$$\Gamma_{\text{SBM}} = \mathcal{D}_+ - \mathcal{D}_- - A,$$

*where $\mathcal{D}_+$ is a diagonal matrix of inner degrees, $\mathcal{D}_-$ is a diagonal matrix of outer degrees and $A$ is the adjacency matrix of the graph.*

It is easy to see that $D_{[\mathrm{diag}(g)A\mathrm{diag}(g)]} = \mathcal{D}_+ - \mathcal{D}_-$. In fact,

$$D_{\left[\mathrm{diag}(g)(2A-\left(\mathbf{1}\,\mathbf{1}^T-I\right))\mathrm{diag}(g)\right]} - \left[2A - \left(\mathbf{1}\,\mathbf{1}^T - I\right)\right] = 2\Gamma_{\mathrm{SBM}} + \mathbf{1}\,\mathbf{1}^T,$$

which means that $gg^T$ is the unique solution of (3.26) as long as $\lambda_2\left(\Gamma_{\mathrm{SBM}} + \mathbf{1}\,\mathbf{1}^T\right) > 0$.

Note that

$$\begin{aligned}\mathbb{E}\left[2\Gamma_{\mathrm{SBM}} + \mathbf{1}\,\mathbf{1}^T\right] &= 2\left(\left(\frac{n}{2}p - \frac{n}{2}q\right)I_{n\times n} - \left(\frac{p+q}{2}\mathbf{1}\,\mathbf{1}^T + \frac{p-q}{2}gg^T\right)\right) + \mathbf{1}\,\mathbf{1}^T \\ &= n\left(p-q\right)\left(I_{n\times n} - \frac{gg^T}{n}\right) + n\left(1-(p+q)\right)\frac{\mathbf{1}\,\mathbf{1}^T}{n}.\end{aligned}$$

If we suppose that $p < \frac{1}{2}$, we have $1-(p+q) > p-q$ the second smallest eigenvalue of $\mathbb{E}\left[2\Gamma_{\mathrm{SBM}} + \mathbf{1}\,\mathbf{1}^T\right]$ is $n\left(p-q\right)$. This establishes the following Lemma.

**Lemma 3.2.9.** *Let $n \geq 4$ be even and let $G$ be drawn from $G(n,p,q)$ with edge probabilities $p < \frac{1}{2}$ and $q < p$. As long as*

$$\lambda_{\max}\left(-\Gamma_{\mathrm{SBM}} + \mathbb{E}\left[\Gamma_{\mathrm{SBM}}\right]\right) < \frac{n}{2}(p-q),$$

*the Semidefinite program (3.26) for the stochastic block model problem achieves exact recovery, meaning that $gg^T$ is its unique solution.*

Estimating this largest eigenvalue using Theorem 4.2.1, we obtain the following theorem.

**Theorem 3.2.10.** *Let $n \geq 4$ be even and let $G$ be drawn from $G(n,p,q)$. As long as $\frac{\log n}{3n} < p < \frac{1}{2}$ and $q < p$, then there exists $\Delta > 0$ such that, with high probability, the* *following holds: If,*

$$\min_i \left(\deg_{in}(i) - \deg_{out}(i)\right) \geq \frac{\Delta}{\sqrt{\log n}} \mathbb{E}\left[\deg_{in}(i) - \deg_{out}(i)\right] \qquad (3.28)$$

*then the semidefinite program* (3.26) *achieves exact recovery.*

*Proof.* The idea is again to apply Theorem 4.2.1. One obstacle is that $\Gamma_{\text{SBM}}$ is not a Laplacian matrix. Let $g$ denote the vector that is 1 in a cluster and $-1$ in the other, and let $\text{diag}(g)$ denote a diagonal matrix with the entries of $g$ on the diagonal. We define

$$\Gamma'_{\text{SBM}} = \text{diag}(g)\Gamma_{\text{SBM}}\text{diag}(g).$$

Note that $\Gamma'_{\text{SBM}}$ is a Laplacian and both the eigenvalues and diagonal elements of $\mathbb{E}\left[\Gamma'_{\text{SBM}}\right] - \Gamma'_{\text{SBM}}$ are the as $\mathbb{E}\left[\Gamma_{\text{SBM}}\right] - \Gamma_{\text{SBM}}$.

We apply Theorem 4.2.1 to $L = -\Gamma'_{\text{SBM}} + \mathbb{E}\left[\Gamma'_{\text{SBM}}\right]$. Note that $L$ has independent off-diagonal entries and

$$\begin{aligned} \sum_{j \in [n] \setminus i} \mathbb{E}\left[L_{ij}^2\right] &= \left(\frac{n}{2} - 1\right)\left(p - p^2\right) + \frac{n}{2}\left(q - q^2\right) \geq \frac{n}{8} p \geq \frac{\log n}{24} \\ &\geq \frac{\log n}{24}(1-q) = \frac{\log n}{24} \max_{i \neq j} \left\| L_{ij}^2 \right\|_\infty . \end{aligned}$$

Hence, there exists a constant $\Delta'$ such that, with high probability,

$$\lambda_{max}\left(-\Gamma'_{\text{SBM}} + \mathbb{E}\left[\Gamma'_{\text{SBM}}\right]\right) \leq \left(1 + \frac{\Delta'}{\sqrt{\log n}}\right) \max_{i \in [n]} \left[-(\Gamma'_{\text{SBM}})_{ii} + \mathbb{E}\left[(\Gamma'_{\text{SBM}})_{ii}\right]\right],$$

which is equivalent to

$$\lambda_{max}\left(-\Gamma_{\text{SBM}} + \mathbb{E}\left[\Gamma_{\text{SBM}}\right]\right) \leq \left(1 + \frac{\Delta'}{\sqrt{\log n}}\right) \max_{i \in [n]} \left[-(\Gamma_{\text{SBM}})_{ii} + \mathbb{E}\left[(\Gamma_{\text{SBM}})_{ii}\right]\right]. \qquad (3.29)$$

We just need to show that, there exists $\Delta > 0$ such that, if (3.28) holds, then

$$\left(1 + \frac{\Delta'}{\sqrt{\log n}}\right) \max_{i\in[n]} \left[-(\Gamma_{\text{SBM}})_{ii} + \mathbb{E}\left[(\Gamma_{\text{SBM}})_{ii}\right]\right] < \frac{n}{2}(p-q) - p. \tag{3.30}$$

Note that $(\Gamma_{\text{SBM}})_{ii} = \deg_{in}(i) - \deg_{out}(i)$ and

$$\mathbb{E}\left[\deg_{in}(i) - \deg_{out}(i)\right] = \frac{n}{2}(p-q) - p.$$

Condition (3.28) can thus be rewriten as

$$\max_{i\in[n]} \left[-(\Gamma_{\text{SBM}})_{ii} + \mathbb{E}\left[(\Gamma_{\text{SBM}})_{ii}\right]\right] \leq \left[1 - \frac{\Delta}{\sqrt{\log n}}\right] \left(\frac{n}{2}(p-q) - p\right).$$

The Theorem is then proven by noting that, for any $\Delta'$, there exists $\Delta$ such that

$$\left[1 - \frac{\Delta}{\sqrt{\log n}}\right] \left(\frac{n}{2}(p-q) - p\right) \leq \left[1 + \frac{\Delta'}{\sqrt{\log n}}\right]^{-1} \left(\frac{n}{2}(p-q) - p\right).$$

□

As a corollary of this theorem we can establish a sharp threshold for exact recovery for the stochastic block model of two clusters solving a problem posed in [3]. We note that this problem was simultaneously solved by the parallel research efforts of Hajek et al. [121].

We first show a lemma concerning $\min_i \left(\deg_{in}(i) - \deg_{out}(i)\right)$, analogous to Lemma 4.2.4.

**Lemma 3.2.11.** *Let $G$ be a random graph with $n$ nodes drawn accordingly to the stochastic block model on two communities with edge probabilities $p$ and $q$. Let $p = \frac{\alpha \log n}{n}$ and $q = \frac{\beta \log n}{n}$, where $\alpha > \beta$ are constants. Then for any constant $\Delta > 0$,*

*1. If*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{3.31}$$

*then, with high probability,*

$$\min_i \left(\deg_{in}(i) - \deg_{out}(i)\right) \geq \frac{\Delta}{\sqrt{\log n}} \mathbb{E}\left[\deg_{in}(i) - \deg_{out}(i)\right].$$

*2. On the other hand, if*

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}, \tag{3.32}$$

*then, with high probability,*

$$\min_i \left(\deg_{in}(i) - \deg_{out}(i)\right) < 0,$$

*and exact recovery is impossible.*

Part (2) is Theorem 3.2.2, included in order to emphasize the dichotomy. Before proving this lemma we note how, together with Theorem 3.2.10, this immediately implies the following Corollary.

**Corollary 3.2.12.** *Let $G$ be a random graph with $n$ nodes drawn accordingly to the stochastic block model on two communities with edge probabilities $p$ and $q$. Let $p = \frac{\alpha \log n}{n}$ and $q = \frac{\beta \log n}{n}$, where $\alpha > \beta$ are constants. Then, as long as*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{3.33}$$

*the semidefinite program* (3.26) *coincides with the true partition with high probability. On the other hand, if*

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}, \tag{3.34}$$

*then no procedure can suceed with high probability at recovering the true partition.*

In order to establish Lemma 3.2.11 we will borrow an estimate from [3].

**Lemma 3.2.13.** *Recall Definition 3.2.6. Let $\alpha$, $\beta$, and $\Delta'$ be constants. Then,*

$$T\left(\frac{n}{2}, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, -\Delta' \sqrt{\log n}\right) \leq \exp\left[-\left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} - \delta(n)\right)\log n\right],$$

*with* $\lim_{n\to\infty} \delta(n) = 0$.

*Proof.* The proof of this Lemma is obtained by straightforward adaptations to the proof of Lemma 8 in [3].

□

We are now ready to prove Lemma 3.2.11.

*Proof.* [of Lemma 3.2.11]

Let $\alpha > \beta$ be constants satisfying condition (3.32). Given $\Delta > 0$, we want to show that, with high probability

$$\min_i \left(\deg_{in}(i) - \deg_{out}(i)\right) \geq \frac{\Delta}{\sqrt{\log n}} \frac{n}{2}(p-q). \tag{3.35}$$

Let us fix $i$ throughout the rest of the proof. It is clear that we can write

$$\deg_{in}(i) - \deg_{out}(i) = \left(\sum_{i=1}^{\frac{n}{2}-1} W_i\right) - \left(\sum_{i=1}^{n/2} Z_i\right) = \sum_{i=1}^{n/2} (W_i - Z_i) + Z_{\frac{n}{2}},$$

where $W_1, \ldots, W_m$ are i.i.d. Bernoulli($p$) and $Z_1, \ldots, Z_m$ are i.i.d. Bernoulli($q$), independent of $W_1, \ldots, W_m$. Hence, since

$$\frac{\Delta}{\sqrt{\log n}}\left(\frac{n}{2}(p-q)\right) = \Delta\sqrt{\log n}\left(\frac{\alpha-\beta}{2}\right),$$

the probability of $\deg_{in}(i) - \deg_{out}(i) < \frac{\Delta}{\sqrt{\log n}}\left(\frac{n}{2}(p-q)\right)$ is equal to

$$\mathbb{P}\left[\sum_{i=1}^{n/2} (Z_i - W_i) - Z_{\frac{n}{2}} > -\Delta\sqrt{\log n}\left(\frac{\alpha-\beta}{2}\right)\right]$$

which is upper bounded by,

$$\mathbb{P}\left[\sum_{i=1}^{n/2}(Z_i - W_i) > -\Delta\sqrt{\log n}\left(\frac{\alpha-\beta}{2}\right)\right].$$

Take $\Delta' = \Delta\left(\frac{\alpha-\beta}{2}\right) + 1$ and recall Definition 3.2.6, then

$$\begin{aligned}\mathbb{P}&\left[\deg_{in}(i) - \deg_{out}(i) < \frac{\Delta}{\sqrt{\log n}}\frac{n}{2}(p-q)\right]\\ &\leq T\left(\frac{n}{2}, \frac{\alpha\log n}{n}, \frac{\beta\log n}{n}, -\Delta'\sqrt{\log n}\right)\\ &\leq \exp\left[-\left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} - \delta(n)\right)\log n\right],\end{aligned}$$

where $\lim_{n\to infty}\delta(n) = 0$, and the last inequality used Lemma 3.2.13.

Via a simple union bound, it is easy to see that,

$$\begin{aligned}\mathbb{P}&\left[\min_i\left(\deg_{in}(i) - \deg_{out}(i)\right) < \frac{\Delta}{\sqrt{\log n}}\frac{n}{2}(p-q)\right]\\ &\leq \exp\left[-\left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} - 1 - \delta(n)\right)\log n\right],\end{aligned}$$

which means that, as long as $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$, (3.35) holds with high probability. Straightforward algebraic manipulations show that (3.31) implies this condition, concluding the proof of the Corollary.

□

## 3.3 $\mathbb{Z}_2$ Synchronization with outlier noise

In this section we consider the $\mathbb{Z}_2$ Synchronization with outlier noise and with an underlying noncomplete graph $G = (V, E)$ [1, 2] (Recall the problem in Section 3.1.1). More precisely, given an underlying graph $G$ with $n$ nodes, the task consists in recovering a binary vector $z \in \{\pm 1\}^n$ from noisy measurements $Y_{ij}$ of $z_i z_j$, for each edge

$(i,j) \in E$,

$$Y_{ij} = \begin{cases} z_i z_j & \text{with probability} \quad 1-\varepsilon \\ -z_i z_j & \text{with probability} \quad \varepsilon, \end{cases}$$

where $\varepsilon < \frac{1}{2}$ represents the noise level. We take $Y_{ij} = 0$ for $(i,j) \notin E$. We are interested in understanding the efficacy of the semidefinite program (3.4) at recovering $z$. We will focus in the low signal-to-noise ratio (SNR) regime, $\varepsilon \to \frac{1}{2}$ (as $n \to \infty$). Before analyzing the semidefinite relaxation based algorithm we present information theoretical lower bounds to serve as a benchmark. We omit the proof[4] and refer the interested reader to [1, 2].

**Theorem 3.3.1.** *[[1, 2]] Let $G = (V, E)$ be the underlying graph and consider $\varepsilon \to \frac{1}{2}$.*

*Let $0 < \tau < 2/3$ and let $d$ be the average degree of $G$. If $d \leq n^{\tau}$ then, recovery with high probability is possible only if*

$$\frac{d}{\log n} \geq 2\frac{1 - 3\tau/2}{(1-2\varepsilon)^2} + o\left(\frac{1}{(1-2\varepsilon)^2}\right). \tag{3.36}$$

We remark that an information theoretical lower bound was independently obtained by Chen and Goldsmith [78]. However, while considering a more general problem, the results they obtain are only optimal up to polylog factors.

### 3.3.1 Underlying Erdős-Rényi graph

We will now focus our attention to the setting on which the underlying graph $G$ is an Erdős-Rényi graph $G \sim G(n,p)$. Recall that, for an integer $n$ and an edge probability parameter $0 \leq p \leq 1$, the Erdős-Rényi graph model [101] $\mathcal{G}(n,p)$ is a random graph on $n$ nodes where each one of the $\binom{n}{2}$ edges appears independently with probability $p$.

We are interested in understanding for which values of $p$ and $\varepsilon$ is it possible to

[4] it shares many ideas with the proof of Theorem 3.2.2.

exactly recover $z$ using the semidefinite relaxation approach. It is easy to see that, just like in the example in Section 3.1.1, the maximum likelihood estimator is given by (3.2). Similarly, we consider its semidefinite relaxation (3.4) and investigate when $X = zz^T$ is the unique solution of (3.4).

It is easy to see that $Y$ is given by

$$Y = \operatorname{diag}(z)\left(A_G - 2A_H\right)\operatorname{diag}(z),$$

where $A_G$ is the adjacency matrix of the underlying graph and $A_H$ is the adjacency of the graph consisting of the corrupted edges. In this case we want conditions on $\varepsilon$ and $p$ under which $zz^T$ is the unique solution to:

$$\begin{array}{ll} \max & \operatorname{Tr}\left[\operatorname{diag}(z)\left(A_G - 2A_H\right)\operatorname{diag}(z)X\right] \\ \text{s.t.} & X_{ii} = 1 \\ & X \succeq 0. \end{array} \tag{3.37}$$

Lemma 3.1.3 states that $zz^T$ is indeed the unique solution as long as the second smallest eigenvalue of

$$D_{A_G - 2A_H} - \operatorname{diag}(z)\left(A_G - 2A_H\right)\operatorname{diag}(z) = D_G - 2D_H - \operatorname{diag}(z)\left(A_G - 2A_H\right)\operatorname{diag}(z) \tag{3.38}$$

is strictly positive. As $\operatorname{diag}(z)\left(D_G - 2D_H\right)\operatorname{diag}(z) = D_G - 2D_H$ and conjugating by $\operatorname{diag}(z)$ does not alter the eigenvalues, the second smallest eigenvalue of (3.38) being strictly positive is equivalent to

$$\lambda_2\left(D_G - A_G - 2\left(D_H - A_H\right)\right) > 0. \tag{3.39}$$

Since $D_G - A_G - 2\left(D_H - A_H\right) = L_G - 2L_H$, where $L_G$ and $L_H$ are the Laplacians of, respectively, $G$ and $H$, we define $L_{\text{Synch}}$ and write the condition in terms of $L_{\text{Synch}}$.

**Definition 3.3.2.** $[L_{\text{Synch}}]$ *In the setting described above,*

$$L_{\text{Synch}} = L_G - 2L_H,$$

*where $G$ is the graph of all measurements and $H$ is the graph of wrong measurements.*

Then, (3.39) is equivalent to $\lambda_2\left(L_{\text{Synch}}\right) > 0$. The following Lemma readily follows by noting that $\mathbb{E}\left[L_{\text{Synch}}\right] = np(1-2\,\varepsilon)I_{n\times n} - p(1-2\,\varepsilon)\,\mathbf{1}\,\mathbf{1}^T$.

**Lemma 3.3.3.** *Consider the $\mathbb{Z}_2$ Synchronization problem defined above and $L_{\text{Synch}}$ defined in Definition 3.3.2. As long as*

$$\lambda_{\max}\left(-L_{\text{Synch}} + \mathbb{E}\left[L_{\text{Synch}}\right]\right) < np(1-2\,\varepsilon),$$

*the Semidefinite program* (3.37) *achieves exact recovery.*

We will estimate this largest eigenvalue using Theorem 4.2.1.

Let us define, for a node $i$, $\deg_+(i)$ as the number of non-corrupted edges incident to $i$ and $\deg_-(i)$ as the number of corrupted edges incident to $i$ We start by obtaining the following theorem.

**Theorem 3.3.4.** *As long as $n > 2$, $p > \frac{\log n}{2n}$ and $p(1-2\,\varepsilon)^2 \leq \frac{1}{2}$, there exists $\Delta > 0$ such that, with high probability, the following holds: If*

$$\min_{i\in[n]}\left[\deg_+(i) - \deg_-(i)\right] \geq \frac{\Delta}{\sqrt{\log n}}\mathbb{E}\left[\deg_+(i) - \deg_-(i)\right], \tag{3.40}$$

*then the semidefinite program* (3.37) *achieves exact recovery.*

*Proof.* [of Theorem 3.3.4]

The idea is to apply Theorem 4.2.1 to $L = -L_{\text{Synch}} + \mathbb{E}\left[L_{\text{Synch}}\right]$. Note that $L$ has independent off-diagonal entries and

$$\begin{aligned}\sum_{j\in[n]\setminus i}\mathbb{E}\left[L_{ij}^2\right] &= (n-1)\left(p-p^2(1-2\,\varepsilon)^2\right)\geq\frac{1}{4}np\geq\frac{1}{8}\log n\\ &\geq \frac{1+p(1-2\,\varepsilon)}{8(1+\sqrt{2})}\log n=\frac{\log n}{8(1+\sqrt{2})}\max_{i\neq j}\left\|L_{ij}^2\right\|_\infty .\end{aligned}$$

Hence, there exists a constant $\Delta'$ such that, with high probability,

$$\lambda_{max}\left(-L_{\text{Synch}}+\mathbb{E}\left[L_{\text{Synch}}\right]\right)\leq\left(1+\frac{\Delta'}{\sqrt{\log n}}\right)\max_{i\in[n]}\left[-(L_{\text{Synch}})_{ii}+\mathbb{E}\left[(L_{\text{Synch}})_{ii}\right]\right].$$

We just need to show that, there exists $\Delta>0$ such that, if (3.40) holds, then

$$\left(1+\frac{\Delta'}{\sqrt{\log n}}\right)\max_{i\in[n]}\left[-(L_{\text{Synch}})_{ii}+\mathbb{E}\left[(L_{\text{Synch}})_{ii}\right]\right]<np(1-2\,\varepsilon). \tag{3.41}$$

Recall that $(L_{\text{Synch}})_{ii}=\deg_+(i)-\deg_-(i)$ and $\mathbb{E}(L_{\text{Synch}})_{ii}=(n-1)p(1-2\,\varepsilon)$. We can rewrite (3.41) as

$$\min_{i\in[n]}(L_{\text{Synch}})_{ii}>(n-1)p(1-2\,\varepsilon)-np(1-2\,\varepsilon)\left(1+\frac{\Delta'}{\sqrt{\log n}}\right)^{-1}.$$

Straightforward algebraic manipulations show that there exists a constant $\Delta$ such that

$$(n-1)p(1-2\,\varepsilon)-np(1-2\,\varepsilon)\left(1+\frac{\Delta'}{\sqrt{\log n}}\right)^{-1}\leq\frac{\Delta}{\sqrt{\log n}}\mathbb{E}\left[\deg_+(i)-\deg_-(i)\right],$$

proving the Theorem.

□

We note that, if $p\leq\frac{\log n}{2n}$, then Theorem 4.2.6 implies that, with high probability, the underlying graph is disconnected implying impossibility of exact recovery. We also note that if we do not have

$$\min_{i\in[n]} \left[\deg_+(i) - \deg_-(i)\right] \geq 0, \tag{3.42}$$

then the maximum likelihood does not match the ground truth, rendering exact recovery unrealistic[5]. The optimality of this analysis hinges upon the fact that the right-hand side of (3.40) is asymptotically smaller than the expectation of $\deg_+(i) - \deg_-(i)$, suggesting that (3.40) and (3.42) have similar probabilities and the same phase transition.

The next Theorem establishes the optimality of the semidefinite programming based approach in a particular regime, solving a problem raised in [1, 2]. While it is clear that one can use Theorem 3.3.4 to establish similar results for many other regimes (for some, through estimates similar to the ones in Lemma 3.2.13), the main purpose of this section is not to perform a detailed analysis of this problem but rather to illustrate the efficacy of these semidefinite relaxations and of its analysis through means of Theorem 4.2.1. The independent parallel research efforts of Hajek et al. [122] address other regimes for this particular problem, we refer the interested reader there.

**Corollary 3.3.5.** *As long as $\varepsilon < \frac{1}{2}$ and $p(1-2\,\varepsilon)^2 \leq \frac{1}{2}$, there exists a constant $K$ for which the following holds: If there exists $\delta > 0$ such that*

$$(n-1)p \geq (1+\delta)\frac{2}{(1-2\,\varepsilon)^2}\left[1 + \frac{K}{\sqrt{\log n}} + \frac{5}{3}(1-2\,\varepsilon)\right]\log n, \tag{3.43}$$

*then the Semidefinite program* (3.37) *achieves exact recovery with high probability.*

Before proving this corollary we note that Theorem 3.3.1 ensure that the threshold in Corollary 3.3.5 is optimal for, at least, an interesting range of values of $\varepsilon$. Empowered with Theorem 3.3.4, the proof of this corollary becomes rather elementary.

[5] Recall that, if we assume a uniform prior, the MLE is the method that maximizes the probability of exact recovery

*Proof.* [of Corollary 3.3.5]

This corollary will be established with a simple use of Bernstein's inequality.

Our goal is to show that, given $\Delta$, there exists a $K$ and $\delta$ such that, under the hypothesis of the Corollary,

$$\min_{i\in[n]}\left[\deg_+(i)-\deg_-(i)\right]\geq\frac{\Delta}{\sqrt{\log n}}\mathbb{E}\left[\deg_+(i)-\deg_-(i)\right],$$

holds with high probability. This implies, via Theorem 3.3.4, that the semidefinite program (3.37) achieves exact recovery with high probability.

We will consider $n$ to be large enough. We start by noting that it suffices to show that there exists $\delta>0$ such that, for each $i\in[n]$ separately,

$$\mathbb{P}\left[\deg_+(i)-\deg_-(i)<\frac{\Delta}{\sqrt{\log n}}\mathbb{E}\left[\deg_+(i)-\deg_-(i)\right]\right]\leq n^{-(1+\delta)}. \tag{3.44}$$

Indeed, (3.44) together with a union bound over the $n$ nodes of the graph would establish the Corollary.

Throughout the rest of the proof we will fix $i\in[n]$ and use $\deg_+$ and $\deg_-$ to denote, respectively, $\deg_+(i)$ and $\deg_-(i)$. It is easy to see that

$$\deg_+-\deg_-=(n-1)p(1-2\,\varepsilon)-\sum_{j=1}^{n-1}x_j,$$

where $x_j$ are i.i.d. centered random variables with distribution

$$x_j=\begin{cases}-1+p(1-2\,\varepsilon) & \text{with probability} \quad p(1-\varepsilon)\\ 1+p(1-2\,\varepsilon) & \text{with probability} \quad p\,\varepsilon\\ p(1-2\,\varepsilon) & \text{with probability} \quad 1-p.\end{cases}$$

For any $t > 0$ Bernstein's inequality gives

$$\mathbb{P}\left[\sum_{j=1}^{n-1} x_j > t\right] \leq \exp\left(-\frac{t^2/2}{(n-1)\mathbb{E}x_j^2 + \frac{t}{3}\|x_j\|_\infty}\right).$$

Taking $t = \left[1 - \frac{\Delta}{\sqrt{\log n}}\right](n-1)p(1-2\,\varepsilon)$ gives

$$\begin{aligned}
&\mathbb{P}\left[\deg_+ - \deg_- < \frac{\Delta}{\sqrt{\log n}}\mathbb{E}\left[\deg_+ - \deg_-\right]\right] \\
&\leq \exp\left(-\frac{\left(\left[1 - \frac{\Delta}{\sqrt{\log n}}\right](n-1)p(1-2\,\varepsilon)\right)^2/2}{(n-1)\mathbb{E}x_j^2 + \frac{\left(\left[1-\frac{\Delta}{\sqrt{\log n}}\right](n-1)p(1-2\,\varepsilon)\right)}{3}\|x_j\|_\infty}\right) \\
&= \exp\left(-\frac{\left[1 - \frac{\Delta}{\sqrt{\log n}}\right]^2(n-1)p(1-2\,\varepsilon)^2/2}{\frac{1}{p}\mathbb{E}x_j^2 + \frac{\left(\left[1-\frac{\Delta}{\sqrt{\log n}}\right](1-2\,\varepsilon)\right)}{3}\|x_j\|_\infty}\right)
\end{aligned}$$

Condition (3.43) (for a $K$ to be determined later) guarantees that

$$(n-1)p(1-2\,\varepsilon)^2/2 \geq (1+\delta)\left[1 + \frac{K}{\sqrt{\log n}} + \frac{5}{3}(1-2\,\varepsilon)\right]\log n,$$

meaning that we just need to show that there exists $K > 0$ for which

$$\frac{\left[1 - \frac{\Delta}{\sqrt{\log n}}\right]^2\left(1 + \frac{K}{\sqrt{\log n}} + \frac{5}{3}(1-2\,\varepsilon)\right)}{\frac{1}{p}\mathbb{E}x_j^2 + \frac{\left(\left[1-\frac{\Delta}{\sqrt{\log n}}\right](1-2\,\varepsilon)\right)}{3}\|x_j\|_\infty} \geq 1.$$

Note that $\frac{1}{p}\mathbb{E}x_j^2 = 1 + p(1-2\,\varepsilon) \leq 1 + (1-2\,\varepsilon)$ and $\|x_j\|_\infty = 1 + p(1-2\,\varepsilon) \leq 2$, implying that

$$\frac{1}{p}\mathbb{E}x_j^2 + \frac{\left(\left[1 - \frac{\Delta}{\sqrt{\log n}}\right](1-2\,\varepsilon)\right)}{3}\|x_j\|_\infty \leq 1 + \frac{5}{3}(1-2\,\varepsilon).$$

Also, $\left[1 - \frac{\Delta}{\sqrt{\log n}}\right]^2 \geq 1 - \frac{2\Delta}{\sqrt{\log n}}$. The corollary is then proved by noting that there exists $K > 0$ such that

$$\frac{K}{\sqrt{\log n}} \geq 2K\frac{\Delta}{\log n} + \frac{2\Delta}{\sqrt{\log n}}\left(1 + \frac{5}{3}(1 - 2\,\varepsilon)\right).$$

□

### 3.3.2 Underlying deterministic graph

We now briefly discuss the case on which the underlying graph $G$ is any deterministic $d$-regular graph. It is natural to give recovery guarantees in terms of the Cheeger constant (2.1): A graph with a small minimum cut consists of two rather disconnected components so that there is a good chance that most edges in the small cut are corrupted, which would render recovery unrealistic. In fact, a sufficient condition for recovery in terms of the Cheeger constant is obtained in [1].

**Theorem 3.3.6.** *[[1]] Cosnder a $d$-regular underlying graph $G = (V, E)$ with Cheeger constant $h_G$. If $\varepsilon \to \frac{1}{2}$, as long as*

$$\frac{d}{\log n} > \frac{2}{h_G\,(1 - 2\varepsilon)^2} + o\left(\frac{1}{h_G\,(1 - 2\varepsilon)^2}\right). \tag{3.45}$$

*then exact recovery with high probability is possible.*

Unfortunately, deriving a necessary condition that bounds the Cheeger constant away from zero is impossible. Indeed, suppose the base-graph consists of two equally sized components, which are connected by $\log n$ edges. Moreover, assume the two graphs that are obtained by disconnecting the two components have Cheeger constant $h_G$ and minimum degree $c \log n$, where $c$ is some positive constant for which the sufficient condition (3.45) of Theorem 3.3.6 holds. Then, Theorem 3.3.6 implies that each component can be recovered correctly (up to an inevitable offset between the two graphs). Moreover, with high probability less than half of the $\log n$ edges that connect the two components are corrupted by noise. Hence, the MLE indeed recovers the correct vertex-variables up to a global shift. But the Cheeger constant of the graph satisfies $h_g \leq 2/\left(cn\right)$ and thus converges to zero as $n$ approaches infinity. This leaves the interesting open question of investigating a characteristic of the graph that captures how easy it is to solve (on it) the type of inverse problems considered here.

In [1] a condition is also derived for exact recovery through the semidefinite program (3.37). It uses a a different measure of connectivity for $G$, in terms of the eigenvalues of its normalized Laplacian. Recall that, if $G$ is a d-regular graph, $\mathcal{L}_G = I_{d\times d} - \frac{1}{d}A_G$ is its normalized Laplacian. The Theorem reads as follows.

**Theorem 3.3.7.** *Let $G$ be a d-regular graph, and $\mathcal{L}_G$ its normalized Laplacian. Let $\varepsilon \to \frac{1}{2}$. As long as*

$$\frac{d}{\log n} \geq 8\frac{1}{\lambda_2\left(\mathcal{L}_G\right)^2}(1+\delta)\left[\frac{1}{(1-2\,\varepsilon)^2} + o\left(\frac{1}{(1-2\,\varepsilon)^2}\right)\right].$$

*the SDP achieves exact recovery with probability at least $1 - n^{\delta}$.*

*If, furthermore, $\lambda_2\left(\mathcal{L}_G\right) = 1 - o(1)$ and $\lambda_{\max}\left(\mathcal{L}_G\right) = 1 + o(1)$ the condition can be replaced by:*

$$\frac{d}{\log n} \geq 4(1+\delta)\frac{1}{(1-2\,\varepsilon)^2} + o\left(\frac{1}{(1-2\,\varepsilon)^2}\right). \qquad (3.46)$$

**Remark 3.3.8.** *The case where $\lambda_2\left(\mathcal{L}_G\right) = 1 - o(1)$ and $\lambda_{\max}\left(\mathcal{L}_G\right) = 1 + o(1)$ is of particular interest as this is satisfied for random d-regular graphs as, for every $\delta > 0$, $\max\left\{1 - \lambda_2\left(\mathcal{L}_G\right), \left|1 - \lambda_{\max}\left(\mathcal{L}_G\right)\right|\right\} \leq 2\frac{\sqrt{d-1}+\delta}{d}$ with high probability [183, 107]. Also, if $G$ is a d-regular Ramanujan expander, then $\max\left\{1 - \lambda_2\left(\mathcal{L}_G\right), \left|1 - \lambda_{\max}\left(\mathcal{L}_G\right)\right|\right\} \leq 2\frac{\sqrt{d-1}}{d}$.*

Theorem 3.3.7 and Theorem 3.3.6 can be compared using Cheeger's inequality (Theorem 2.1.1). In fact, when $\varepsilon \to \frac{1}{2}$, the sufficient condition (3.45) in Theorem 3.3.6

is implied by

$$\frac{d}{\log n} > \frac{4}{\lambda_2\left(\mathcal{L}_G\right)(1-2\varepsilon)^2} + o\left(\frac{1}{\lambda_2\left(\mathcal{L}_G\right)(1-2\varepsilon)^2}\right). \tag{3.47}$$

We believe the discrepancy in the leading constant between (3.47) and (3.46) is due to the analysis in [1], in particular the use of Matrix Bernstein's inequality [220]. We suspect that the gap can be improved by adapting the analysis described above for the Erdős-Rényi setting (via Theorem 4.2.1) and leave such an improvement for future work.

# Chapter 4

# A detour through random matrix theory

## 4.1 Spectral norm of matrices with independent entries

In this chapter we take a detour through *random matrix theory* that will culminate with Theorem 4.2.1 (stated and proved in Section 4.2), which played a crucial role in the analysis of semidefinite programming–based algorithms carried out in Chapter 3.

We start in this section by investigating the spectral norm of a large class of matrices. The results of this section, which is based on [42], are of independent interest. A simple example is when $W$ a standard Wigner matrix, that is, a symmetric $n \times n$ matrix whose entries are i.i.d. standard Gaussians then, as we discussed in Section 3.1.1, $\|W\|/\sqrt{n} \to 2$. Furthermore, the requirement that the entries need to be Gaussian random variables can be relaxed to very mild conditions [109, 26, 22, 217] (this is expected from the well-known fact that the empirical spectral density converges to the Wigner semicircle law supported in $[-2, 2]$). The corresponding result for rectangular matrices with i.i.d. entries is even older [110].

More recently, there has been considerable interest in structured random matrices where the entries are no longer identically distributed. As the combinatorial methods that are used for this purpose typically exploit the specific structure of the entries, precise asymptotic results on the spectral norm of structured matrices must generally be obtained on a case-by-case basis (see, for example, [212, 213]).

In order to gain a deeper understanding of the spectral norm of structured matrices, it is natural to ask whether one can find a unifying principle that captures at least the correct scale of the norm in a general setting, that is, in the absence of specific structural assumptions. This question is most naturally phrased in a nonasymptotic setting: given a random matrix $X$ can we obtain upper and lower bounds on $\|X\|$, in terms of natural parameters that capture the structure of $X$, that differ only by universal constants? Nonasymptotic bounds on the norm of a random matrix have long been developed in a different area of probability that arises from problems in geometric functional analysis, and have had a significant impact on various areas of pure and applied mathematics [91, 193, 228]. Unfortunately, as we will shortly see, the best known general results along these lines fail to capture the correct scale of the spectral norm of structured matrices except in extreme cases.

In this section, we investigate the norm of random matrices with independent entries. Consider for concreteness the case of Gaussian matrices (this section's main results will extend to more general distributions of the entries). Let $X$ be the $n \times n$ symmetric random matrix with entries $X_{ij} = g_{ij}b_{ij}$, where $\{g_{ij} : i \geq j\}$ are independent standard Gaussian random variables and $\{b_{ij} : i \geq j\}$ are given scalars.

Perhaps the most useful known nonasymptotic bound on the spectral norm $\|X\|$ can be obtained as a consequence of the noncommutative Khintchine inequality of Lust-Piquard and Pisier [180], or alternatively (in a much more elementary fashion) from the "matrix concentration" method that has been widely developed in recent years [176, 220, 221]. This yields the following inequality in our setting:

$$\mathbb{E}\|X\| \lesssim \sigma\sqrt{\log n} \qquad \text{with} \qquad \sigma := \max_i \sqrt{\sum_j b_{ij}^2}.$$

Unfortunately, as it was discussed in Section 3.1.1, this inequality already fails to be sharp in the simplest case of Wigner matrices: here $\sigma = \sqrt{n}$, so that the resulting bound $\mathbb{E}\|X\| \lesssim \sqrt{n \log n}$ falls short of the correct scaling[1] $\mathbb{E}\|X\| \sim \sqrt{n}$. On the other hand, the logarithmic factor in this bound is necessary: if $X$ is the diagonal matrix with independent standard Gaussian entries, then $\sigma = 1$ and $\mathbb{E}\|X\| \sim \sqrt{\log n}$. We therefore conclude that while the noncommutative Khintchine bound is sharp in extreme cases, it fails to capture the structure of the matrix $X$ in a satisfactory manner.

A different bound on $\|X\|$ can be obtained by a method due to Gordon (see [91]) that exploits Slepian's comparison lemma for Gaussian processes, or alternatively from a simple $\varepsilon$-net argument [228, 217]. This yields the following inequality:

$$\mathbb{E}\|X\| \lesssim \sigma_*\sqrt{n} \qquad \text{with} \qquad \sigma_* := \max_{ij} |b_{ij}|.$$

While the parameter $\sigma_*$ that appears in this bound is often much smaller than $\sigma$, the dimensional scaling of this bound is much worse than in the noncommutative Khintchine bound. In particular, while this bound captures the correct $\sqrt{n}$ rate for Wigner matrices, it is vastly suboptimal in almost every other situation (for example, in the diagonal matrix example considered above).

Further nonasymptotic bounds on $\|X\|$ have been obtained in the present setting by Latała [142] and by Riemer and Schütt [188]. In most examples, these bounds provide even worse rates than the noncommutative Khintchine bound. Seginer [199]

[1] Section 3.1.1 describes an example where this suboptimality is critical, as the extra logarithmic term prevents one from establishing the qualitative phenomenon expressed in (3.9).

obtained a slight improvement on the noncommutative Khintchine bound that is specific to the special case where the random matrix has uniformly bounded entries (see Section 4.1.5 below). None of these results provides a sharp understanding of the scale of the spectral norm for general structured matrices.

In what follows we will develop a new family of nonasymptotic bounds on the spectral norm of structured random matrices that prove to be optimal in a surprisingly general setting. Our main bounds are of the form

$$\mathbb{E}\|X\| \lesssim \sigma + \sigma_* \sqrt{\log n}, \tag{4.1}$$

which provides a sort of interpolation between the two bounds discussed above. For example, the following is one of the main results of this section.

**Theorem 4.1.1.** *Let $X$ be the $n \times n$ symmetric matrix with $X_{ij} = g_{ij} b_{ij}$, where $\{g_{ij} : i \geq j\}$ are i.i.d. $\sim N(0,1)$ and $\{b_{ij} : i \geq j\}$ are given scalars. Then*

$$\mathbb{E}\|X\| \leq (1+\varepsilon)\left\{2\sigma + \frac{6}{\sqrt{\log(1+\varepsilon)}}\sigma_* \sqrt{\log n}\right\}$$

*for any $0 < \varepsilon \leq 1/2$, where $\sigma, \sigma_*$ are as defined above.*

Let us emphasize two important features of this result.

- It is almost trivial to obtain a matching lower bound of the form

  $$\mathbb{E}\|X\| \gtrsim \sigma + \sigma_* \sqrt{\log n}$$

  that holds as long as the coefficients $b_{ij}$ are not too inhomogeneous (Section 4.1.4). This means that Theorem 4.1.1 captures the optimal scaling of the expected norm $\mathbb{E}\|X\|$ under surprisingly minimal structural assumptions.

- In the case of Wigner matrices, Theorem 4.1.1 yields a bound of the form

  $$\mathbb{E}\|X\| \leq (1+\varepsilon)2\sqrt{n} + o(\sqrt{n})$$

  for arbitrarily small $\varepsilon > 0$. Thus Theorem 4.1.1 not only captures the correct scaling of the spectral norm, but even recovers the precise asymptotic behavior $\|X\|/\sqrt{n} \to 2$ as $n \to \infty$.

In view of these observations, it seems that Theorem 4.1.1 is essentially the optimal result of its kind: there is little hope to accurately capture inhomogeneous models where Theorem 4.1.1 is not sharp in terms of simple parameters such as $\sigma, \sigma_*$ (see Remark 4.1.17). On the other hand, we can now understand the previous bounds as extreme cases of Theorem 4.1.1. The noncommutative Khintchine bound matches Theorem 4.1.1 when $\sigma/\sigma_* \lesssim 1$: this case is minimal as $\sigma/\sigma_* \geq 1$. Gordon's bound matches Theorem 4.1.1 when $\sigma/\sigma_* \gtrsim \sqrt{n}$: this case is maximal as $\sigma/\sigma_* \leq \sqrt{n}$. In intermediate regimes, Theorem 4.1.1 yields a strictly better scaling.

While we will focus most of the exposition on Theorem 4.1.1 and its proof, our methods are not restricted to this particular setting. In fact, we will state a number of extensions of Theorem 4.1.1 while refering the reader to [42] for the proofs of these resuls.

One of the nice features of Theorem 4.1.1 is that its proof explains very clearly why the result is true. Once the idea has been understood, the technical details prove to be of minimal difficulty, which suggests that the "right" approach has been found. Let us briefly illustrate the idea behind the proof in the special case where the coefficients $b_{ij}$ take only the values $\{0, 1\}$ (this setting guided our intuition, though the ultimate proof is no more difficult in the general setting). We can then interpret the matrix of coefficients $(b_{ij})$ as the adjacency matrix of a graph $G$ on $n$ points, and we have $\sigma_* = 1$ and $\sigma = \sqrt{k}$ where $k$ is the maximal degree of $G$.

Following a classical idea in random matrix theory, we use the fact that the spectral norm $\|X\|$ is comparable to the quantity $\mathrm{Tr}[X^p]^{1/p}$ for $p \sim \log n$. If one writes out the expression for $\mathbb{E}\,\mathrm{Tr}[X^p]$ in terms of the coefficients, it is readily seen that controlling this quantity requires us to count the number of cycles in $G$ for which every edge is visited an even number of times. One might expect that the graph $G$ of degree $k$ that possesses the most such cycles is the complete graph on $k$ points. If this were the case, then one could control $\mathbb{E}\,\mathrm{Tr}[X^p]$ by $\mathbb{E}\,\mathrm{Tr}[Y^p]$ where $Y$ is a Wigner matrix of dimension $k$. This intuition is almost, but not entirely, correct: while a $k$-clique typically possesses more distinct topologies of cycles, each cycle of a given topology can typically be embedded in more ways in a regular graph on $n$ points than in a $k$-clique. Careful bookkeeping shows that the latter can be accounted for by choosing a slightly larger Wigner matrix of dimension $k+p$. We therefore obtain a comparison theorem between the spectral norm of $X$ and the spectral norm of a $(k+p)$-dimensional Wigner matrix, which is of the desired order $\sqrt{k+p} \sim \sqrt{k} + \sqrt{\log n}$ for $p \sim \log n$. We can now conclude by using standard ideas from probability in Banach spaces to obtain sharp nonasymptotic bounds on the norm of the resulting Wigner matrix, avoiding entirely any combinatorial complications. (A purely combinatorial approach would be nontrivial as very high moments of Wigner matrices can appear in this argument.)

We conclude the introduction by noting that both the noncommutative Khintchine inequality and Gordon's bound can be formulated in a more general context beyond the case of independent entries. Whether the conclusion of Theorem 4.1.1 extends to this situation is a natural question of considerable interest. This fascinating direction for future research is further discussed in Section 7.2.7.

### 4.1.1 Proof of Theorem 4.1.1

The main idea behind the proof of Theorem 4.1.1 is the following comparison theorem.

**Proposition 4.1.2.** *Let $Y_r$ be the $r \times r$ symmetric matrix such that $\{(Y_r)_{ij} : i \geq j\}$*

*are independent $N(0,1)$ random variables, and suppose that $\sigma_* \le 1$. Then*

$$\mathbb{E}\,\mathrm{Tr}[X^{2p}] \le \frac{n}{\lceil\sigma^2\rceil + p}\,\mathbb{E}\,\mathrm{Tr}\left[Y^{2p}_{\lceil\sigma^2\rceil+p}\right] \qquad \textit{for every } p \in \mathbb{N}.$$

Let us begin by completing the proof of Theorem 4.1.1 given this result. We need the following lemma, which is a variation on standard ideas (cf. [91]).

**Lemma 4.1.3.** *Let $Y_r$ be the $r \times r$ symmetric matrix such that $\{(Y_r)_{ij} : i \ge j\}$ are independent $N(0,1)$ random variables. Then for every $p \ge 2$*

$$\mathbb{E}[\|Y_r\|^{2p}]^{1/2p} \le 2\sqrt{r} + 2\sqrt{2p}.$$

*Proof.* We begin by noting that

$$\|Y_r\| = \lambda_+ \vee \lambda_-, \qquad \lambda_+ := \sup_{v\in\mathbb{S}^{r-1}} \langle v, Y_r v\rangle, \qquad \lambda_- = -\inf_{v\in\mathbb{S}^{r-1}} \langle v, Y_r v\rangle,$$

where $\mathbb{S}^{r-1}$ is the unit sphere in $\mathbb{R}^r$. We are therefore interested in the supremum of the Gaussian process $\{\langle v, Y_r v\rangle\}_{v\in\mathbb{S}^{r-1}}$, whose natural distance can be estimated as

$$\mathbb{E}|\langle v, Y_r v\rangle - \langle w, Y_r w\rangle|^2 \le 2\sum_{i,j}\{v_i v_j - w_i w_j\}^2 \le 4\|v-w\|^2$$

(using $1 - x^2 \le 2(1-x)$ for $x \le 1$). The right-hand side of this expression is the natural distance of the Gaussian process $\{2\langle v, g\rangle\}_{v\in S}$, where $g$ is the standard Gaussian vector in $\mathbb{R}^r$. Therefore, Slepian's lemma [53, Theorem 13.3] implies

$$\mathbb{E}\lambda_+ = \mathbb{E}\sup_{v\in\mathbb{S}^{r-1}} \langle v, Y_r v\rangle \le 2\mathbb{E}\sup_{v\in\mathbb{S}^{r-1}} \langle v, g\rangle = 2\mathbb{E}\|g\| \le 2\sqrt{r}.$$

Moreover, note that $\lambda_+$ and $\lambda_-$ have the same distribution (as evidently $Y_r$ and $-Y_r$

have the same distribution). Therefore, using the triangle inequality for $\|\cdot\|_{2p}$,

$$\begin{aligned}\mathbb{E}[\|Y_r\|^{2p}]^{1/2p} &= \|\lambda_+ \vee \lambda_-\|_{2p} \\ &\le \mathbb{E}\lambda_+ + \|\lambda_+ \vee \lambda_- - \mathbb{E}\lambda_+\|_{2p} \\ &= \mathbb{E}\lambda_+ + \|(\lambda_+ - \mathbb{E}\lambda_+) \vee (\lambda_- - \mathbb{E}\lambda_-)\|_{2p}.\end{aligned}$$

It follows from Gaussian concentration [53, Theorem 5.8 and Theorem 2.1] that

$$\mathbb{E}[(\lambda_+ - \mathbb{E}\lambda_+)^{2p} \vee (\lambda_- - \mathbb{E}\lambda_-)^{2p}] \le p!4^{p+1} \le (2\sqrt{2p})^{2p}$$

for $p \ge 2$. Putting together the above estimates completes the proof. □

*Proof of Theorem 4.1.1.* We can clearly assume without loss of generality that the matrix $X$ is normalized such that $\sigma_* = 1$. For $p \ge 2$, we can estimate

$$\begin{aligned}\mathbb{E}\|X\| &\le \mathbb{E}[\mathrm{Tr}[X^{2p}]]^{1/2p} \\ &\le n^{1/2p}\,\mathbb{E}[\|Y_{\lceil\sigma^2\rceil+p}\|^{2p}]^{1/2p} \\ &\le n^{1/2p}\{2\sqrt{\lceil\sigma^2\rceil + p} + 2\sqrt{2p}\}\end{aligned}$$

by Proposition 4.1.2 and Lemma 4.1.3, where we used $\mathrm{Tr}[Y_r^{2p}] \le r\|Y_r\|^{2p}$. This yields

$$\begin{aligned}\mathbb{E}\|X\| &\le e^{1/2\alpha}\{2\sqrt{\lceil\sigma^2\rceil + \lceil\alpha\log n\rceil} + 2\sqrt{2\lceil\alpha\log n\rceil}\} \\ &\le e^{1/2\alpha}\{2\sigma + 2\sqrt{\alpha\log n + 2} + 2\sqrt{2\alpha\log n + 2}\}\end{aligned}$$

for the choice $p = \lceil\alpha\log n\rceil$. If $n \ge 2$ and $\alpha \ge 1$, then $2 \le 3\log 2 \le 3\alpha\log n$, so

$$\mathbb{E}\|X\| \le e^{1/2\alpha}\{2\sigma + 6\sqrt{2\alpha\log n}\}.$$

Defining $e^{1/2\alpha} = 1 + \varepsilon$ and noting that $\varepsilon \le 1/2$ implies $\alpha \ge 1$ yields the result provided that $n \geq 2$ and $p \geq 2$. The remaining cases are easily dealt with separately. The result holds trivially in the case $n = 1$. On the other hand, the case $p = 1$ can only occur when $\alpha \log n \leq 1$ and thus $n \leq 2$. In this case can estimate directly $\mathbb{E}\|X\| \leq \sqrt{n(\lceil \sigma^2 \rceil + p)} \leq \sigma\sqrt{2} + 2$ using Proposition 4.1.2. □

**Remark 4.1.4.** *Note that we use the moment method only to prove the comparison theorem of Proposition 4.1.2; as will be seen below, this requires only trivial combinatorics. All the usual combinatorial difficulties of random matrix theory are circumvented by Lemma 4.1.3, which exploits the theory of Gaussian processes.*

**Remark 4.1.5.** *The constant* 6 *in the second term in Theorem 4.1.1 arises from crude rounding in our proof. While this constant can be somewhat improved for large $n$, our proof cannot yield a sharp constant here: it can be verified in the example of the diagonal matrix $b_{ij} = \mathbf{1}_{i=j}$ that the constant $\sqrt{2}$ in the precise asymptotic $\mathbb{E}\|X\| \sim \sqrt{2 \log n}$ cannot be recovered from our general proof. We therefore do not insist on optimizing this constant, but rather state the convenient bound in Theorem 4.1.1 which holds for any $n$. In contrast to the constant in the second term, the constant in the first term is sharp.*

We now turn to the proof of Proposition 4.1.2. Let us begin by recalling some standard observations. The quantity $\mathbb{E}\,\mathrm{Tr}[X^{2p}]$ can be expanded as

$$\mathbb{E}\,\mathrm{Tr}[X^{2p}] = \sum_{u_1,\ldots,u_{2p} \in [n]} b_{u_1u_2} b_{u_2u_3} \cdots b_{u_{2p}u_1}\, \mathbb{E}[g_{u_1u_2} g_{u_2u_3} \cdots g_{u_{2p}u_1}].$$

Let $G_n = ([n], E_n)$ be the complete graph on $n$ points, that is, $E_n = \{\{u, u'\} : u, u' \in [n]\}$ (note that we have included self-loops). We will identify any $\mathbf{u} = (u_1, \ldots, u_{2p}) \in [n]^{2p}$ with a cycle $u_1 \to u_2 \to \cdots \to u_{2p} \to u_1$ in $G_n$ of length $2p$. If we denote by $n_i(\mathbf{u})$ the number of distinct edges that are visited precisely $i$ times by the cycle $\mathbf{u}$, then we can write (here $g \sim N(0,1)$)

$$\mathbb{E}\,\mathrm{Tr}[X^{2p}] = \sum_{\mathbf{u}\in[n]^{2p}} b_{u_1u_2}b_{u_2u_3}\cdots b_{u_{2p}u_1}\prod_{i\geq 1}\mathbb{E}[g^i]^{n_i(\mathbf{u})}.$$

A cycle $\mathbf{u}$ is called *even* if it visits each distinct edge an even number of times, that is, if $n_i(\mathbf{u}) = 0$ whenever $i$ is odd. As $\mathbb{E}[g^i] = 0$ when $i$ is odd, it follows immediately that the sum in the above expression can be restricted to even cycles.

The *shape* $\mathbf{s}(\mathbf{u})$ of a cycle $\mathbf{u}$ is obtained by relabeling the vertices in order of appearance. For example, the cycle $7 \to 3 \to 5 \to 4 \to 3 \to 5 \to 4 \to 3 \to 7$ has shape $1 \to 2 \to 3 \to 4 \to 2 \to 3 \to 4 \to 2 \to 1$. We denote by

$$\mathcal{S}_{2p} := \{\mathbf{s}(\mathbf{u}) : \mathbf{u} \text{ is an even cycle of length } 2p\}$$

the collection of shapes of even cycles, and we define the collection of even cycles with given shape $\mathbf{s}$ starting (and ending) at a given point $u$ as

$$\Gamma_{\mathbf{s},u} := \{\mathbf{u}\in[n]^{2p} : \mathbf{s}(\mathbf{u}) = \mathbf{s},\ u_1 = u\}$$

for any $u \in [n]$ and $\mathbf{s} \in \mathcal{S}_{2p}$. Clearly the edge counts $n_i(\mathbf{u})$ depend only on the shape $\mathbf{s}(\mathbf{u})$ of $\mathbf{u}$, and we can therefore unambiguously write $n_i(\mathbf{s})$ for the number of distinct edges visited $i$ times by any cycle with shape $\mathbf{s}$. We then obtain

$$\mathbb{E}\,\mathrm{Tr}[X^{2p}] = \sum_{u\in[n]}\sum_{\mathbf{s}\in\mathcal{S}_{2p}}\prod_{i\geq 1}\mathbb{E}[g^i]^{n_i(\mathbf{s})}\sum_{\mathbf{u}\in\Gamma_{\mathbf{s},u}} b_{u_1u_2}b_{u_2u_3}\cdots b_{u_{2p}u_1}.$$

Finally, given any shape $\mathbf{s} = (s_1,\ldots,s_{2p})$, we denote by $m(\mathbf{s}) = \max_i s_i$ the number of distinct vertices visited by any cycle with shape $\mathbf{s}$.

Now that we have set up a convenient bookkeeping device, the proof of Proposition 4.1.2 is surprisingly straightforward. It relies on two basic observations.

**Lemma 4.1.6.** *Suppose that $\sigma_* \le 1$. Then we have for any $u \in [n]$ and $\mathbf{s} \in \mathcal{S}_{2p}$*

$$\sum_{\mathbf{u} \in \Gamma_{\mathbf{s},u}} b_{u_1 u_2} b_{u_2 u_3} \cdots b_{u_{2p} u_1} \le \sigma^{2(m(\mathbf{s})-1)}.$$

*In particular, it follows that*

$$\mathbb{E}\,\mathrm{Tr}[X^{2p}] \le n \sum_{\mathbf{s} \in \mathcal{S}_{2p}} \sigma^{2(m(\mathbf{s})-1)} \prod_{i \ge 1} \mathbb{E}[g^i]^{n_i(\mathbf{s})}.$$

*Proof.* Fix an initial point $u$ and shape $\mathbf{s} = (s_1, \ldots, s_{2p})$. Let

$$i(k) = \inf\{j : s_j = k\}$$

for $1 \le k \le m(\mathbf{s})$. That is, $i(k)$ is the first time in any cycle of shape $\mathbf{s}$ at which its $k$th distinct vertex is visited (of course, $i(1) = 1$ by definition).

Now consider any cycle $\mathbf{u} \in \Gamma_{\mathbf{s},u}$. As the cycle is even, the edge $\{u_{i(k)-1}, u_{i(k)}\}$ must be visited at least twice for every $2 \le k \le m(\mathbf{s})$. On the other hand, as the vertex $u_{i(k)}$ is visited for the first time at time $i(k)$, the edge $\{u_{i(k)-1}, u_{i(k)}\}$ must be distinct from the edges $\{u_{i(\ell)-1}, u_{i(\ell)}\}$ for all $\ell < k$. We can therefore estimate

$$\begin{aligned}\sum_{\mathbf{u} \in \Gamma_{\mathbf{s},u}} b_{u_1 u_2} b_{u_2 u_3} \cdots b_{u_{2p} u_1} &\le \sum_{\mathbf{u} \in \Gamma_{\mathbf{s},u}} b^2_{u u_{i(2)}} b^2_{u_{i(3)-1} u_{i(3)}} \cdots b^2_{u_{i(m(\mathbf{s}))-1} u_{i(m(\mathbf{s}))}} \\ &= \sum_{v_2 \ne \cdots \ne v_{m(\mathbf{s})}} b^2_{u v_2} b^2_{v_{s_{i(3)-1}} v_3} \cdots b^2_{v_{s_{i(m(\mathbf{s}))-1}} v_{m(\mathbf{s})}},\end{aligned}$$

where we used that $\max_{ij} |b_{ij}| = \sigma_* \le 1$. As $s_{i(k)-1} < k$ by construction, it is readily seen that the quantity on the right-hand side is bounded by $\sigma^{2(m(\mathbf{s})-1)}$. □

**Lemma 4.1.7.** *Let $Y_r$ be defined as in Proposition 4.1.2. Then for any $r > p$*

$$\mathbb{E}\,\mathrm{Tr}[Y_r^{2p}] = r \sum_{\mathbf{s} \in \mathcal{S}_{2p}} (r-1)(r-2) \cdots (r - m(\mathbf{s}) + 1) \prod_{i \ge 1} \mathbb{E}[g^i]^{n_i(\mathbf{s})}.$$

*Proof.* In complete analogy with the identity for $\mathbb{E}\,\mathrm{Tr}[X^{2p}]$, we can write

$$\mathbb{E}\,\mathrm{Tr}[Y_r^{2p}] = \sum_{\mathbf{s}\in\mathcal{S}_{2p}} |\{\mathbf{u} \in [r]^{2p} : \mathbf{s}(\mathbf{u}) = \mathbf{s}\}| \prod_{i\geq 1} \mathbb{E}[g^i]^{n_i(\mathbf{s})}.$$

Each cycle $\mathbf{u} \in [r]^{2p}$ with given shape $\mathbf{s}(\mathbf{u}) = \mathbf{s}$ is uniquely defined by specifying its $m(\mathbf{s})$ distinct vertices. Thus as long as $m(\mathbf{s}) \leq r$, there are precisely

$$r(r-1)\cdots(r-m(\mathbf{s})+1)$$

such cycles. But note that any even cycle of length $2p$ can visit at most $m(\mathbf{s}) \leq p+1$ distinct vertices, so the assumption $p < r$ implies the result. □

We can now complete the proof.

*Proof of Proposition 4.1.2.* Fix $p \in \mathbb{N}$ and let $r = \lceil \sigma^2 \rceil + p$. Then

$$(r-1)(r-2)\cdots(r-m(\mathbf{s})+1) \geq (\sigma^2 + p - m(\mathbf{s}) + 1)^{m(\mathbf{s})-1} \geq \sigma^{2(m(\mathbf{s})-1)}$$

for any $\mathbf{s} \in \mathcal{S}_{2p}$, where we have used that any even cycle of length $2p$ can visit at most $m(\mathbf{s}) \leq p+1$ distinct vertices. It remains to apply Lemmas 4.1.6 and 4.1.7. □

### 4.1.2 Extensions and adaptations

**Non-symmetric matrices**

Let $X$ be the $n \times m$ random rectangular matrix with $X_{ij} = g_{ij}b_{ij}$, where $\{g_{ij} : 1 \leq i \leq n, 1 \leq j \leq m\}$ are independent $N(0,1)$ random variables and $\{b_{ij} : 1 \leq i \leq n, 1 \leq j \leq m\}$ are given scalars. While this matrix is not symmetric, one can immediately obtain a bound on $\mathbb{E}\|X\|$ from Theorem 4.1.1 by applying the latter to the symmetric matrix

$$\tilde{X} = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix}.$$

Indeed, it is readily seen that $\|\tilde{X}\| = \|X\|$, so we obtain

$$\mathbb{E}\|X\| \leq (1+\varepsilon)\left\{2(\sigma_1 \vee \sigma_2) + \frac{6}{\sqrt{\log(1+\varepsilon)}}\,\sigma_* \sqrt{\log(n+m)}\right\}$$

for any $0 < \varepsilon \leq 1/2$ with

$$\sigma_1 := \max_i \sqrt{\sum_j b_{ij}^2}, \qquad \sigma_2 := \max_j \sqrt{\sum_i b_{ij}^2}, \qquad \sigma_* := \max_{ij} |b_{ij}|.$$

While this result is largely satisfactory, it does not lead to a sharp constant in the first term: it is known from asymptotic theory [110] that when $b_{ij} = 1$ for all $i, j$ we have $\mathbb{E}\|X\| \sim \sqrt{n} + \sqrt{m}$ as $n, m \to \infty$ with $n/m \to \gamma \in ]0, \infty[$, while the above bound can only give the weaker inequality $\mathbb{E}\|X\| \leq 2(1+o(1))(\sqrt{n} \vee \sqrt{m})$. The latter bound can therefore be off by as much as a factor 2.

We can regain the lost factor and also improve the logarithmic term by exploiting explicitly the bipartite structure of $\tilde{X}$ in the proof of Theorem 4.1.1. This leads to the following sharp analogue of Theorem 4.1.1 for rectangular random matrices.

**Theorem 4.1.8.** *Let $X$ be the $n \times m$ matrix with $X_{ij} = g_{ij} b_{ij}$. Then*

$$\mathbb{E}\|X\| \leq (1+\varepsilon)\left\{\sigma_1 + \sigma_2 + \frac{5}{\sqrt{\log(1+\varepsilon)}}\,\sigma_* \sqrt{\log(n \wedge m)}\right\}.$$

*for any* $0 < \varepsilon \leq 1/2$.

As the proof of this result closely follows the proof of Theorem 4.1.8, we will omit it and refer the reader to [42].

### Non–Gaussian entries

We have phrased Theorem 4.1.1 in terms of Gaussian random matrices for concreteness. However, note that the core argument of the proof of Theorem 4.1.1, the comparison principle of Proposition 4.1.2, did not depend at all on the Gaussian nature of the entries: it is only subsequently in Lemma 4.1.3 that we exploited the theory of Gaussian processes. The same observation applies to the proof of Theorem 4.1.8. As a consequence, we can develop various extensions of these results to more general distributions of the entries.

Let us begin by considering the case of subgaussian random variables.

**Corollary 4.1.9.** *Theorems 4.1.1 and 4.1.8 remain valid if the independent Gaussian random variables $g_{ij}$ are replaced by independent symmetric random variables $\xi_{ij}$ such that $\mathbb{E}[\xi_{ij}^{2p}] \le \mathbb{E}[g^{2p}]$ for every $p \in \mathbb{N}$ and $i, j$ ($g \sim N(0,1)$).*

*Proof.* As $\xi_{ij}$ are assumed to be symmetric, $\mathbb{E}[\xi_{ij}^{p}] = 0$ when $p$ is odd. It therefore follows readily by inspection of the proof that Proposition 4.1.2 (and its rectangular counterpart) remains valid under the present assumptions. □

Corollary 4.1.9 implies, for example, that the conclusions of Theorems 4.1.1 and 4.1.8 hold verbatim when $g_{ij}$ are replaced by independent Rademacher variables $\varepsilon_{ij}$, that is, $\mathbf{P}[\varepsilon_{ij} = \pm 1] = 1/2$ (see Section 4.1.5 below for more on such matrices). The moment assumption $\mathbb{E}[\xi_{ij}^{2p}] \le \mathbb{E}[g^{2p}]$ is somewhat unwieldy, however. In fact, we also establish the following Corollary (which is sharper for when the entries are bounded). For the sake of brevity, we refer the reader to [42] for a proof.

**Corollary 4.1.10.** *Let $X$ be the $n \times n$ symmetric random matrix with $X_{ij} = \xi_{ij} b_{ij}$, where $\{\xi_{ij} : i \ge j\}$ are independent symmetric random variables with unit variance and $\{b_{ij} : i \ge j\}$ are given scalars. Then we have for any $\alpha \ge 3$*

$$\mathbb{E}\|X\| \le e^{2/\alpha}\left\{2\sigma + 14\alpha \max_{ij} \|\xi_{ij} b_{ij}\|_{2\lceil \alpha \log n \rceil} \sqrt{\log n}\right\}.$$

### 4.1.3 Tail bounds

Given explicit bounds on the expectation $\mathbb{E}\|X\|$, we can readily obtain nonasymptotic tail inequalities for $\|X\|$ by applying standard concentration techniques. We record some useful results along these lines here.

**Corollary 4.1.11.** *Under the assumptions of Theorem 4.1.1, we have*

$$\mathbf{P}\left[\|X\| \geq (1+\varepsilon)\left\{2\sigma + \frac{6}{\sqrt{\log(1+\varepsilon)}}\,\sigma_*\sqrt{\log n}\right\} + t\right] \leq e^{-t^2/4\sigma_*^2}$$

*for any* $0 < \varepsilon \leq 1/2$ *and* $t \geq 0$. *In particular, for every* $0 < \varepsilon \leq 1/2$ *there exists a universal constant* $c_\varepsilon$ *such that for every* $t \geq 0$

$$\mathbf{P}[\|X\| \geq (1+\varepsilon)2\sigma + t] \leq ne^{-t^2/c_\varepsilon\sigma_*^2}.$$

*Proof.* As $\|X\| = \sup_v |\langle v, Xv\rangle|$ (the supremum is over the unit ball) and

$$\mathbb{E}[\langle v, Xv\rangle^2] = \sum_i b_{ii}^2 v_i^4 + 2\sum_{i\neq j} b_{ij}^2 v_i^2 v_j^2 \leq 2\sigma_*^2,$$

the first inequality follows from Gaussian concentration [53, Theorem 5.8] and Theorem 4.1.1. For the second inequality, note that we can estimate

$$\mathbf{P}[\|X\| \geq (1+\varepsilon)2\sigma + c_\varepsilon\sigma_* t] \leq \mathbf{P}[\|X\| \geq (1+\varepsilon)2\sigma + c'_\varepsilon\sigma_*\sqrt{\log n} + \sigma_* t] \leq e^{-t^2/4}$$

for $t \geq 2\sqrt{\log n}$ (with $c_\varepsilon, c'_\varepsilon$ chosen in the obvious manner), while

$$\mathbf{P}[\|X\| \geq (1+\varepsilon)2\sigma + c_\varepsilon\sigma_* t] \leq 1 \leq ne^{-t^2/4}$$

for $t \leq 2\sqrt{\log n}$. Combining these bounds completes the proof. □

Tail bounds on $\|X\|$ have appeared widely in the recent literature under the name “matrix concentration inequalities” (see [176, 220]). In the present setting, the corresponding result of this kind implies that for all $t \geq 0$

$$\mathbf{P}[\|X\| \geq t] \leq ne^{-t^2/8\sigma^2}.$$

The second inequality of Corollary 4.1.11 was stated for comparison with with this matrix concentration bound. Unlike the matrix concentration bound, Corollary 4.1.11 is essentially optimal in that it captures not only the correct mean, but also the correct tail behavior of $\|X\|$ [143, Corollary 3.2] (note that due to the factor $1 + \varepsilon$ in the leading term, we do not expect to see Tracy-Widom fluctuations at this scale).

**Remark 4.1.12.** *Integrating the tail bound obtained by the matrix concentration method yields the estimate $\mathbb{E}\|X\| \lesssim \sigma\sqrt{\log n}$. This method therefore yields an alternative proof of the noncommutative Khintchine bound that was discussed in the introduction. Combining this bound with concentration as in the proof of Corollary 4.1.11 already yields a better tail bound than the one obtained directly from the matrix concentration method. Nonetheless, it should be emphasized that the suboptimality of the above bound on the expected norm stems from the suboptimal tail behavior obtained by the matrix concentration method. Our sharp tail bounds help clarify the source of this inefficiency: the parameter $\sigma$ should only control the mean of $\|X\|$, while the fluctuations are controlled entirely by $\sigma_*$.*

The Gaussian concentration property used above is specific to Gaussian variables. However, there are many other situations where strong concentration results are available [53], and where similar results can be obtained. For example, if the Gaussian variables $g_{ij}$ are replaced by symmetric random variables $\xi_{ij}$ with $\|\xi_{ij}\|_\infty \leq 1$ (this captures in particular the case of Rademacher variables), Corollary 4.1.11 remains valid with slightly larger universal constants $c_\varepsilon, c'_\varepsilon$. This follows from the identical proof, up to the replacement of Gaussian concentration by a form of Talagrand’s concentration inequality [53, Theorem 6.10].

In the case of bounded entries, however, a more interesting question is whether it is possible to obtain tail bounds that capture the variance of the entries rather than their uniform norm (which is often much bigger than the variance), akin to the classical Bernstein inequality for sums of independent random variables. We presently develop a very useful result along these lines.

**Corollary 4.1.13.** *Let $X$ be an $n \times n$ symmetric matrix whose entries $X_{ij}$ are independent symmetric random variables. Then there exists for any $0 < \varepsilon \leq 1/2$ a universal constant $\tilde{c}_\varepsilon$ such that for every $t \geq 0$*

$$\mathbf{P}[\|X\| \geq (1+\varepsilon)2\sigma + t] \leq ne^{-t^2/\tilde{c}_\varepsilon \sigma_\infty^2},$$

*where we have defined*

$$\sigma := \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \qquad \sigma_\infty := \max_{ij} \|X_{ij}\|_\infty.$$

*Proof.* Let $X_{ij} = \tilde{X}_{ij}\mathbb{E}[X_{ij}^2]^{1/2}$, so that $\tilde{X}_{ij}$ have unit variance. Then

$$\mathbb{E}\|X\| \leq (1+\varepsilon)2\sigma + C_\varepsilon \sigma_\infty \sqrt{\log n}$$

for a suitable constant $C_\varepsilon$ by Corollary 4.1.10. On the other hand, a form of Talagrand's concentration inequality [53, Theorem 6.10] yields

$$\mathbf{P}[\|X\| \geq \mathbb{E}\|X\| + t] \leq e^{-t^2/c\sigma_\infty^2}$$

for all $t \geq 0$, where $c$ is a universal constant. The proof is completed by combining these bounds as in the proof of Corollary 4.1.11. □

Corollary 4.1.13 should be compared with the matrix Bernstein inequality in [220], which reads as follows in our setting (we omit the explicit constants):

$$\mathbf{P}[\|X\| \geq t] \leq n e^{-t^2/c(\sigma^2+\sigma_\infty t)}.$$

While this result looks quite different at first sight than Corollary 4.1.13, the latter yields strictly better tail behavior up to universal constants: indeed, note that

$$e^{-t^2/c^2\sigma_\infty^2} \leq e^{1-2t/c\sigma_\infty} \leq 3e^{-2t^2/c(\sigma^2+\sigma_\infty t)}$$

using $2x-1 \leq x^2$. The discrepancy between these results is readily explained. In our sharp bounds, the variance term $\sigma$ only appears in the mean of $\|X\|$ and not in the fluctuations: the latter only depend on the uniform parameter $\sigma_\infty$ and do not capture the variance.

**Corollary 4.1.14.** *Let $X$ be an $n \times n$ symmetric matrix whose entries $X_{ij}$ are independent centered random variables. Then there exists for any $0 < \varepsilon \leq 1/2$ a universal constant $\tilde{c}_\varepsilon$ such that for every $t \geq 0$*

$$\mathbf{P}[\|X\| \geq (1+\varepsilon)2\sqrt{2}\sigma + t] \leq n e^{-t^2/\tilde{c}_\varepsilon \sigma_\infty^2},$$

*where we have defined*

$$\sigma := \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \qquad \sigma_\infty := \max_{ij} \|X_{ij}\|_\infty.$$

Unfortunately, this results in an additional factor $\sqrt{2}$ in the leading term, which is suboptimal for Wigner matrices. We do not know whether it is possible, in general, to improve the constant when the entries are not symmetrically distributed.

### 4.1.4 Lower bounds

The main results of this section provide upper bounds on $\mathbb{E}\|X\|$. However, a trivial lower bound already suffices to establish the sharpness of our upper bounds in many cases of interest, at least for Gaussian variables.

**Lemma 4.1.15.** *In the setting of Theorem 4.1.1, we have*

$$\mathbb{E}\|X\| \gtrsim \sigma + \mathbb{E}\max_{ij} |b_{ij} g_{ij}|.$$

*Similarly, in the setting of Theorem 4.1.8*

$$\mathbb{E}\|X\| \gtrsim \sigma_1 + \sigma_2 + \mathbb{E}\max_{ij} |b_{ij} g_{ij}|.$$

*Proof.* Let us prove the second inequality; the first inequality follows in a completely analogous manner. As $\|X\| \geq \max_{ij} |X_{ij}|$, it is trivial that

$$\mathbb{E}\|X\| \geq \mathbb{E}\max_{ij} |X_{ij}| = \mathbb{E}\max_{ij} |b_{ij} g_{ij}|.$$

On the other hand, as $\|X\| \geq \max_i \|Xe_i\|$ ($\{e_i\}$ is the canonical basis in $\mathbb{R}^n$),

$$\mathbb{E}\|X\| \geq \max_i \mathbb{E}\|Xe_i\| \gtrsim \max_i \mathbb{E}[\|Xe_i\|^2]^{1/2} = \sigma_2.$$

Here we used the estimate

$$\mathbb{E}[\|Xe_i\|^2] = (\mathbb{E}\|Xe_i\|)^2 + \operatorname{Var}\|Xe_i\| \lesssim (\mathbb{E}\|Xe_i\|)^2,$$

where $\operatorname{Var}\|Xe_i\| \leq \max_j b_{ji}^2 \lesssim \max_j \mathbb{E}[b_{ji}|g_{ji}|]^2 \leq (\mathbb{E}\|Xe_i\|)^2$ by the Gaussian Poincaré inequality [53, Theorem 3.20]. Analogously, we obtain

$$\mathbb{E}\|X\| \geq \max_i \mathbb{E}\|X^* e_i\| \gtrsim \sigma_1.$$

Averaging these three lower bounds yields the conclusion. □

This simple bound shows that this section's main results are sharp as long there are enough large coefficients $b_{ij}$. This is the content of the following easy bound.

**Corollary 4.1.16.** *In the setting of Theorem 4.1.1, suppose that*

$$|\{ij : |b_{ij}| \geq c\sigma_*\}| \geq n^\alpha$$

*for some constants $c, \alpha > 0$. Then*

$$\mathbb{E}\|X\| \asymp \sigma + \sigma_* \sqrt{\log n},$$

*where the universal constant in the lower bound depends on $c, \alpha$ only. The analogous result holds in the setting of Theorem 4.1.8.*

*Proof.* Denote by $I$ the set of indices in the statement of the corollary. Then

$$\mathbb{E} \max_{ij} |b_{ij} g_{ij}| \geq \mathbb{E} \max_{ij \in I} |b_{ij} g_{ij}| \geq c\sigma_* \mathbb{E} \max_{ij \in I} |g_{ij}| \gtrsim \sigma_* \sqrt{\log |I|} \gtrsim \sigma_* \sqrt{\log n},$$

where we used a standard lower bound on the maximum of independent $N(0,1)$ random variables. The proof is concluded by applying Lemma 4.1.15. □

For example, it follows that this Section's main results are sharp as soon as every row of the matrix contains at least one large coefficient, that is, with magnitude of the same order as $\sigma_*$. This is the case is many natural examples of interest, and in these cases our results are optimal (up to the values of universal constants).

Of course, it quite possible that our bound is sharp even when the assumption of Corollary 4.1.16 fails: for example, in view of Lemma 4.1.15, our bound is sharp whenever $\sigma_* \sqrt{\log n} \lesssim \sigma$ regardless of any other feature of the problem.

**Remark 4.1.17.** *An intriguing observation that was made in [188] is that the trivial lower bound $\mathbb{E}\|X\| \geq \mathbb{E}\max_i \|Xe_i\|$ appears to be surprisingly sharp: we do not know of any example where the corresponding upper bound*

$$\mathbb{E}\|X\| \overset{?}{\lesssim} \mathbb{E}\max_i \|Xe_i\|$$

*fails. If such an inequality were to hold, the conclusion of Theorem 4.1.1 would follow easily using Gaussian concentration and a simple union bound. Recently, partial progress [225] was made towards understanding whether such an upper bound holds.*

**Remark 4.1.18.** *The conclusion of Corollary 4.1.16 relies heavily on the Gaussian nature of the entries. When the distributions of the entries are bounded, for example, it is possible that our bounds are no longer sharp. This issue will be discussed further in Section 4.1.5 below in the context of Rademacher matrices.*

### 4.1.5 Examples

#### Sparse random matrices

In the section, we consider the special case of Theorem 4.1.1 where the coefficients $b_{ij}$ can take the values zero or one only. This is in essence a sparse counterpart of Wigner matrices in which a subset of the entries has been set to zero. This rather general model covers many interesting random matrix ensembles, including the case of random band matrices where $b_{ij} = \mathbf{1}_{|i-j|\leq k}$ that has been of significant recent interest [137, 213, 47].

Let us fix a matrix $(b_{ij})$ of $\{0,1\}$-valued coefficients. We immediately compute

$$\sigma^2 = k, \qquad\qquad \sigma_* = 1,$$

where $k$ is the maximal number of nonzero entries in any row of the matrix of coefficients $(b_{ij})$. If we interpret $(b_{ij})$ as the adjacency matrix of a graph on $n$ points, then $k$ is simply the maximal degree of this graph. The following conclusion follows effortlessly from the results in this Section.

**Corollary 4.1.19.** *Let $X$ be the $n \times n$ symmetric random matrix with $X_{ij} = g_{ij} b_{ij}$, where $\{g_{ij}\}$ are independent $N(0,1)$ variables and $b_{ij} \in \{0,1\}$. Let $k$ be the maximal number of nonzero entries in a row of $(b_{ij})$. Then*

$$\mathbb{E}\|X\| \asymp \sqrt{k} + \sqrt{\log n},$$

*provided that every row of $(b_{ij})$ has at least one nonzero entry.*

### Rademacher matrices

We have seen that the results presented here provide sharp bounds in many cases on the norm of matrices with independent Gaussian entries. While our upper bounds continue to hold for subgaussian variables, this is not the case for the lower bounds in Section 4.1.4, and in this case we cannot expect our results to be sharp at the same level of generality. As a simple example, consider the case where $X$ is the diagonal matrix with i.i.d. entries on the diagonal. If the entries are Gaussian, then $\|X\| \gtrsim \sqrt{\log n}$, so that Theorem 4.1.1 is sharp. If the entries are bounded, however, then $\|X\| \lesssim 1$. On the other hand, the universality property of Wigner matrices shows that Theorem 4.1.1 is sharp in this case even when adapted to bounded random variables.

In view of these observations, it is natural to ask whether it is possible to obtain systematic improvement of this Section's main results that captures the size of the norm of random matrices with bounded entries. For concreteness, let us consider the case of Rademacher matrices $X_{ij} = \varepsilon_{ij} b_{ij}$, where $\{\varepsilon_{ij}\}$ are independent Rademacher (symmetric Bernoulli) random variables. In this setting, we can immediately obtain a trivial but useful improvement.

**Corollary 4.1.20.** *Let $X$ be the $n \times n$ symmetric random matrix with $X_{ij} = \varepsilon_{ij} b_{ij}$, where $\{\varepsilon_{ij}\}$ are independent Rademacher variables. Then*

$$\mathbb{E}\|X\| \lesssim (\sigma + \sigma_* \sqrt{\log n}) \wedge \|B\|,$$

*where $B := (|b_{ij}|)$ is the matrix of absolute values of the coefficients.*

*Proof.* In view of Corollary 4.1.9, it suffices to show that $\mathbb{E}\|X\| \leq \|B\|$. Note, however, that this inequality even holds pointwise: indeed,

$$\|X\| = \sup_v \sum_{ij} \varepsilon_{ij} b_{ij} v_i v_j \leq \sup_v \sum_{ij} |b_{ij} v_i v_j| = \|B\|,$$

where the supremum is taken over the unit ball in $\mathbb{R}^n$. □

Corollary 4.1.20 captures two reasons why a Rademacher matrix can have small norm: either it behaves like a Gaussian matrix with small norm; or its norm is uniformly bounded due to the boundedness of the matrix entries. This idea mirrors the basic ingredients in the general theory of Bernoulli processes [216, Chapter 5]. While simple, Corollary 4.1.20 captures at least the Wigner and diagonal examples considered above, albeit in a somewhat ad-hoc manner. We will presently show that a less trivial result can be easily derived from Corollary 4.1.20 as well.

The norm of Rademacher matrices was first investigated in a general setting by Seginer [199]. Using a delicate combinatorial method, he proves in this case that $\mathbb{E}\|X\| \lesssim \sigma \log^{1/4} n$. The assumption of Rademacher entries is essential: that such a bound cannot hold in the Gaussian case is immediate from the diagonal matrix example. Let us show that this result is an easy consequence of Corollary 4.1.20.

**Corollary 4.1.21.** *Let $X$ be the $n \times n$ symmetric random matrix with $X_{ij} = \varepsilon_{ij} b_{ij}$, where $\{\varepsilon_{ij}\}$ are independent Rademacher variables. Then*

$$\mathbb{E}\|X\| \lesssim \sigma \log^{1/4} n.$$

*Proof.* Fix $u > 0$. Let us split the matrix into two parts $X = X^+ + X^-$, where $X^+_{ij} = \varepsilon_{ij} b_{ij} \mathbf{1}_{|b_{ij}|>u}$ and $X^-_{ij} = \varepsilon_{ij} b_{ij} \mathbf{1}_{|b_{ij}|\le u}$. For $X^-$, we can estimate

$$\mathbb{E}\|X^-\| \lesssim \sigma + u\sqrt{\log n}.$$

On the other hand, we estimate for $X^+$ by the Gershgorin circle theorem

$$\mathbb{E}\|X^+\| \le \|(|b_{ij}|\mathbf{1}_{|b_{ij}|>u})\| \le \max_i \sum_j |b_{ij}|\mathbf{1}_{|b_{ij}|>u} \le \frac{\sigma^2}{u}.$$

We therefore obtain for any $u > 0$

$$\mathbb{E}\|X\| \lesssim \sigma + u\sqrt{\log n} + \frac{\sigma^2}{u}.$$

The proof is completed by optimizing over $u > 0$. □

Corollary 4.1.21 not only recovers Seginer's result with a much simpler proof, but also effectively explains why the mysterious term $\log^{1/4} n$ arises. More generally, the method of proof suggests how Corollary 4.1.20 can be used efficiently: we should attempt to split the matrix $X$ into two parts, such that one part is small by Theorem 4.1.1 and the other part is small uniformly. This idea also arises in a fundamental manner in the general theory of Bernoulli processes [216]. Unfortunately, it is generally not clear for a given matrix how to choose the best decomposition.

**Remark 4.1.22.** *In view of Corollary 4.1.19, one might hope that Corollary 4.1.20 (or a suitable adaptation of this bound) could yield sharp results in the general setting of sparse random matrices. The situation for Rademacher matrices turns out to be more delicate, however. To see this, let us consider two illuminating examples. In the following, let $k = \lceil \sqrt{\log n} \rceil$ and assume for simplicity that $n/k$ is integer.*

*First, consider the block-diagonal matrix $X$ of the form*

$$X = \begin{bmatrix} X_1 & & & & \\ & X_2 & & 0 & \\ & & \cdot & & \\ & 0 & & \cdot & \\ & & & & X_{n/k} \end{bmatrix},$$

*where each $X_i$ is a $k \times k$ symmetric matrix with independent Rademacher entries. Such matrices were considered by Seginer in [199], who shows by an elementary argument that $\mathbb{E}\|X\| \sim \sqrt{\log n}$. Thus Theorem 4.1.1 already yields a sharp result (and, in particular, the logarithmic term in Theorem 4.1.1 cannot be eliminated).*

*On the other hand, it was shown by Sodin [212] that if $X$ is the Rademacher matrix where the coefficient matrix $B$ is chosen to be a realization of the adjacency matrix of a random $k$-regular graph, then $\mathbb{E}\|X\| \sim \sqrt{k} \leq \log^{1/4} n$ with high probability. Thus in this case $\mathbb{E}\|X\| \sim \sigma$, and it appears that the logarithmic term in Theorem 4.1.1 is missing (evidently none of our bounds are sharp in this case).*

*Note, however, that in both these examples the parameters $\sigma, \sigma_*, \|B\|$ are identical: we have $\sigma = \sqrt{k}$, $\sigma_* = 1$, and $\|B\| = k$ (by the Perron-Frobenius theorem). In particular, there is no hope that the norm of sparse Rademacher matrices can be controlled using only the degree of the graph: the* structure *of the graph must come into play. It is an interesting open problem to understand precisely what aspect of this structure controls the norm of sparse Rademacher matrices. This question is closely* *connected to the study of random 2-lifts of graphs in combinatorics [49].*

## 4.2 Random Laplacian matrices

The largest eigenvalue of a matrix is always larger or equal than its largest diagonal entry. In this section (based on [29]), we show Theorem 4.2.1, which stats that for a large class of random Laplacian matrices, this bound is essentially tight: the largest eigenvalue is, up to lower order terms, often the size of the largest diagonal entry. Besides being a simple tool to obtain precise estimates on the largest eigenvalue of a large class of random Laplacian matrices, Theorem 4.2.1 played a crucial role in the analysis of semidefinite programming–based algorithms carried out in Chapter 3.

We use the term Laplacian matrix to refer to symmetric matrices whose rows and columns sum to zero. While oftentimes Laplacians are also thought of as being positive semidefinite, the matrices we will treat will not, in general, satisfy that property. Recall that given a symmetric matrix $X \in \mathbb{R}^{n\times n}$, we define the Laplacian $L_X$ of $X$ as

$$L_X = D_X - X,$$

where $D_X$ is the diagonal matrix whose diagonal entries are given by $(D_X)_{ii} = \sum_{j=1}^{n} X_{ij}$. Note that these are precisely the symmetric matrices $L$ for which $L\mathbf{1} = 0$. Laplacian matrices are particularly important in spectral graph theory [83] as the spectrum of the graph Laplacian matrix is known to contain important information about the graph (recall, for example Theorem 2.1.1), which has motivated its study for random graphs [95, 82, 61].

This section is concerned with random Laplacian matrices $L_X$ where the entries of the matrix $X$ are independent centered (but not identically distributed) random variables (so as to be able to use the machienary developped in Section 4.1). This section's main result is that, under mild and easily verifiable conditions, the largest eigenvalue of $L_X$ is, up to lower order terms, given by its largest diagonal entry. Our results will be of nonasymptotic nature (we refer the interested reader to [228] for a tutorial on nonasymptotic estimates in random matrix theory). As it was seen in Chapter 3, the largest diagonal value tends to be a quantity whose interpretation is intimately tied to the nature of the underlying problem.

We will illustrate the latter point further by turning to graph theory. It is well known that the spectrum of the Laplacian of a graph dictates whether or not the graph is connected. On the other hand, its diagonal is simply given by the degrees of the nodes of the graph. A relation between the spectrum of the Laplacian and its diagonal could then translate into a relation between degrees of nodes of a graph and its connectivity. In fact, such a relation is known to exist: *The phase transition for connectivity of Erdős-Rényi graphs* [2] *coincides with the one for the existence of isolated nodes.* While it is true that any graph with an isolated node (a node with degree zero) cannot be connected, the converse is far from true in general graphs, rendering this phenomenon particularly interesting. Indeed, for Section 4.2.2 we will use our main result to provide a simple and illustrative proof for this phenomenon.

### 4.2.1 Largest eigenvalue of Laplacian Matrices

We use this section to formulate precise versions of, and briefly discuss, the main results of this section.

**Theorem 4.2.1.** *[29]*

*Let $L$ be an $n \times n$ symmetric random Laplacian matrix (i.e. satisfying $L\mathbf{1} = 0$) with centered independent off-diagonal entries such that $\sum_{j\in[n]\setminus i}\mathbb{E}L_{ij}^2$ is equal for every $i$.*

[2] The Erdős-Rényi model for random graphs will be discussed in more detail in Section 4.2.2.

*Define $\sigma$ and $\sigma_\infty$ as*

$$\sigma^2 = \sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2 \quad \text{and} \quad \sigma_\infty^2 = \max_{i\neq j} \|L_{ij}\|_\infty^2 .$$

*If there exists $c > 0$ such that*

$$\sigma \geq c\,(\log n)^{\frac{1}{2}}\,\sigma_\infty, \tag{4.2}$$

*then there exists $c_1$, $C_1$, $\beta_1$, all positive and depending only on $c$, such that*

$$\lambda_{\max}(L) \leq \left(1 + \frac{C_1}{(\log n)^{\frac{1}{2}}}\right) \max_i L_{ii}$$

*with probability at least $1 - c_1 n^{-\beta_1}$.*

Even though we were not able to find a convincing application for which $\frac{\sigma}{\sigma_\infty}$ was asymptotically growing but slower than $\sqrt{\log n}$, we still include the theorem below for the sake of completeness.

**Theorem 4.2.2.** *[29]*

*Let $L$ be an $n \times n$ symmetric random Laplacian matrix (i.e. satisfying $L1 = 0$) with centered independent off-diagonal entries such that $\sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2$ is equal for every $i$.*

*Define $\sigma$ and $\sigma_\infty$ as*

$$\sigma^2 = \sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2 \quad \text{and} \quad \sigma_\infty^2 = \max_{i\neq j} \|L_{ij}\|_\infty^2 .$$

*If there exist $c$ and $\gamma > 0$ such that*

$$\sigma \geq c\,(\log n)^{\frac{1}{4}+\gamma}\,\sigma_\infty, \tag{4.3}$$

*then there exist $C_2$, $c_2$, $\epsilon$ and $\beta_2$, all positive and depending only on $c$ and $\gamma > 0$, such* *that*

$$\lambda_{\max}(L) \leq \left(1 + \frac{C_2}{(\log n)^{\epsilon}}\right) \max_i L_{ii},$$

*with probability at least* $1 - c_2 \exp\left[-\left(\log n\right)^{\beta_2}\right]$.

**Remark 4.2.3.** *In the theorems above, the condition that* $\sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2$ *is equal for every* $i$*, can be relaxed to the requirement that*

$$c'\sigma^2 \leq \sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2 \leq \sigma^2,$$

*for all* $i$*. This requires only simple adaptations to the proofs of these theorems.*

While we defer the proof of these theorems to Section 4.2.3, we briefly describe its idea. Corollary 4.1.14 estimates that

$$\|X\| \lesssim \sigma + \sigma_\infty \sqrt{\log n},$$

where $-X$ is the off-diagonal part of $L$. One the other hand, $L_{ii} = \sum_{j\in[n]\setminus i} X_{ij}$ has variance $\sigma^2$ and the Central Limit Theorem would suggest that $L_{ii}$ behave like independent gaussians of variance $\sigma^2$, which would mean that $\max_i L_{ii} \sim \sigma\sqrt{\log n}$ rendering the contribution of the off-diagonal entries (to the largest eigenvalue) negligible. However, several difficulties arise: the diagonal entries are not independent (as each pair shares a summand) and one needs to make sure that the central limit theorem behavior sets in (this is, in a way, ensured by requirements (4.2) and (4.3)). The proofs in Section 4.2.3 make many needed adaptations to this argument to make it rigorous.

### 4.2.2 Connectivity of Erdős-Rényi graphs

While the usefullness of these Theorems has already been demonstrated in Chapter 3, we now show how Theorem 4.2.1 can be used to provide a simple proof of the Erdős-Rényi connectivity phase transition. Besides illustrating further the utility of this result, it sheds light on an interesting connection between this phase transition and the ones established in Chapter 3, suggesting they are manifestations of one underlying phenomenon.

Recall that, for an integer $n$ and an edge probability parameter $0 \leq p \leq 1$, the Erdős-Rényi graph model [101] $\mathcal{G}(n,p)$ is a random graph on $n$ nodes where each one of the $\binom{n}{2}$ edges appears independently with probability $p$.

We are interested in understanding the probability that $G$, drawn according to $\mathcal{G}(n,p)$, is a connected graph. We will restrict our attention to the setting $p \leq \frac{1}{2}$. Let $L$ be the Laplacian of the random graph, given by $D - A$ where $A$ is its adjacency matrix and $D$ a diagonal matrix containing the degree of each node. Recall that $G$ connected is equivalent to $\lambda_2(L) > 0$[3].

It is clear that if $G$ has an isolated node then it cannot be connected. It is also known that for there not to be isolated nodes one needs the average degree of each node to be at least logarithmic [101]. For this reason we will focus on the regime

$$p = \frac{\rho \log n}{n},$$

for a constant $\rho$. It is easy to establish a phase transition on the degrees of the nodes of graphs drawn from $\mathcal{G}(n,p)$.

**Lemma 4.2.4.** *Let $n$ be a positive integer, $\rho$ a constant, and $p = \frac{\rho \log n}{n}$. Let $G$ be a random graph drawn from $\mathcal{G}(n,p)$, then for any constant $\Delta > 0$:*

1. *If $\rho > 1$ then, with high probability, $\min_{i \in [n]} \deg(i) \geq \frac{\Delta}{\sqrt{\log n}} \mathbb{E} \deg(i)$.*

[3] In fact, Theorem 2.1.1 is a refinement of this.

2. *If $\rho < 1$ then, with high probability,* $\min_{i\in[n]} \deg(i) = 0$. *That is, $G$ has at least one isolated node, thus being disconnected.*

Part (2) of the Lemma is a classical result [101], a particularly simple proof of it proceeds by applying the second moment method to the number of isolated nodes in $G$. For the sake of brevity we will skip those details, and focus on part (1). The main thing to note in part (1) of Lemma 4.2.4 is that the lower bound on minimum degree is asymptotically smaller than the average degree $\mathbb{E}\deg(i)$.

*Proof.* [of part (1) of Lemma 4.2.4]

Let $p = \frac{\rho\log n}{n}$ and $i$ denote a node of the graph, note that $\mathbb{E}\deg(i) = \frac{n-1}{n}\rho\log n$. We use Chernoff bound (see, for example, Lemma 2.3.3 in [99]) to establish, for any $0 < t < 1$,

$$\begin{aligned}
\mathbb{P}\left[\deg(i) < t\mathbb{E}\deg(i)\right] &\leq \left[\frac{\exp(-(1-t))}{t^t}\right]^{\mathbb{E}\deg(i)} \\
&= \left[\frac{\exp(-(1-t))}{t^t}\right]^{\frac{n-1}{n}\rho\log n} \\
&= \exp\left[-\left[1-t-t\log(1/t)\right]\frac{n-1}{n}\rho\log n\right].
\end{aligned}$$

Taking $t = \frac{\Delta}{\sqrt{\log n}}$ gives, for $n$ large enough (so that $t \leq 1$), that the probability that $\deg(i) < \frac{\Delta}{\sqrt{\log n}}\mathbb{E}\deg(i)$ is at most

$$\exp\left[-\left[1-\frac{\Delta}{\sqrt{\log n}}-\frac{\Delta}{\sqrt{\log n}}\log\left(\frac{\sqrt{\log n}}{\Delta}\right)\right]\frac{n-1}{n}\rho\log n\right],$$

which is easily seen to be $\exp\left[-\rho\log n + O(\sqrt{\log n}\log\log n)\right]$. A simple union bound over the $n$ vertices of $G$ gives

$$\mathbb{P}\left[\min_{i\in[n]}\deg(i) < \frac{\Delta}{\sqrt{\log n}}\mathbb{E}\deg(i)\right] \leq \exp\left[-(\rho-1)\log n + O(\sqrt{\log n}\log\log n)\right].$$

□

Using Theorem 4.2.1 we will show that, with high probability, as long as every node in $G$ is at least $\frac{\Delta}{\sqrt{\log n}}$ of the average degree, for a suitable $\Delta$, then $G$ is connected. This is made precise in the following Lemma.

**Lemma 4.2.5.** *Let $n \geq 2$ be an integer and $\varepsilon > 0$. Suppose that $\frac{\varepsilon \log n}{n} \leq p \leq \frac{1}{2}$ and $G$ a random graph drawn from $\mathcal{G}(n,p)$. There exists a constant $\Delta$ such that, with high probability, the following holds:*

*If*

$$\min_{i\in[n]} \deg(i) \geq \frac{\Delta}{\sqrt{\log n}} \mathbb{E} \deg(i),$$

*then $G$ is a connected graph (note that the right hand side does not depend on $i$).*

Before proving this Lemma, we note that Lemmas 4.2.4 and 4.2.5 immediately imply the well known phase transition phenomenon.

**Theorem 4.2.6.** *Let $n$ be a positive integer and $p = \frac{\rho \log n}{n}$.*

1. *If $\rho > 1$ then, with high probability, a random graph drawn from $\mathcal{G}(n,p)$ is connected.*

2. *If $\rho < 1$ then, with high probability, a random graph drawn from $\mathcal{G}(n,p)$ has at least one isolated node, thus being disconnected.*

While this phase transition is well understood, we find our proof through Lemmas 4.2.4 and 4.2.5 enlightening, as it provides a simple explanation of why the phase transition for disappearance of isolated nodes coincides with the one for connectivity. Moreover, it also emphasizes a connection with the optimality of the semidefinite relaxations in both $\mathbb{Z}_2$ Synchronization and the Stochastic Block Model that we established in Chatper 3.

*Proof.* [of Lemma 4.2.5]

Let $L$ be the graph Laplacian of $G$. Note that $\mathbb{E}(L) = npI - p11^T$, which means that

$$L = npI - p11^T - [-L + \mathbb{E}(L)]$$

Since $L\,\mathbf{1} = 0$, it is easy to see that $G$ is connected if and only if

$$\lambda_{\max}\left[-L + \mathbb{E}(L)\right] < np$$

We proceed by using Theorem 4.2.1 for

$$L = -L + \mathbb{E}(L).$$

The hypotheses of the Theorem are satisfied as the off-diagonal entries of $L$ are independent and

$$\sum_{j\in[n]\setminus i} \mathbb{E}L_{ij}^2 = (n-1)p(1-p) \geq \frac{np(1-p)}{2} \geq \frac{\varepsilon}{2}(1-p)^2\log n = \frac{\varepsilon}{2}\log n \max_{i\neq j}\|L_{ij}\|_\infty^2 .$$

This guarantees that there exists a constant $C_1$ such that, with high probability,

$$\lambda_{\max}\left[-L + \mathbb{E}(L)\right] \leq \left(1 + \frac{C_1}{\sqrt{\log n}}\right)\max_{i\in[n]}\left[-\deg(i) + (n-1)p\right] \tag{4.4}$$

where $\deg(i) = L_{ii}$ is the degree of node $i$. Equivalently,

$$\lambda_{\max}\left[-L + \mathbb{E}(L)\right] \quad\leq\quad np + \left(1 + \frac{C_1}{\sqrt{\log n}}\right)\left[-\min_{i\in[n]}\deg(i) + (n-1)p\right] - np$$

This means that, as long as (4.4) holds, then

$$\left(1 + \frac{C_1}{\sqrt{\log n}}\right)\left[-\min_{i\in[n]}\deg(i) + (n-1)p\right] - np < 0$$

implies the connectivity of $G$. Straighforward manipulations show that this conditions is equivalent to

$$\min_i \deg(i) > np\frac{C_1}{\sqrt{\log n}+C_1} - p,$$

which is implied by

$$\min_i \deg(i) \geq np\frac{C_1}{\sqrt{\log n}}. \tag{4.5}$$

The lemma follows by taking $\Delta = 2C_1$.

□

### 4.2.3 Proof of Theorems 4.2.1 and 4.2.2

We prove Theorems 4.2.1 and 4.2.2 through a few Lemmas. Let us define $X$ as the non-diagonal part of $-L$ and $y \in \mathbb{R}^n$ as $y = \operatorname{diag}(D_X)$, meaning that $y = \operatorname{diag}(L)$. Then $L = D_X - X$. We will separately lower bound $\max_i y_i$ and upper bound $\|X\|$. The upper bound on $\|X\|$ is obtained directly by Corollary 4.1.14. Before continuing with the proof let us recall the main idea: Corollary 4.1.14 gives that, with high probability,

$$\|X\| \lesssim \sigma + \sigma_\infty\sqrt{\log n},$$

where $X$ is the off-diagonal part of $-L$. One the other hand, $L_{ii} = \sum_{j\in[n]\setminus i} X_{ij}$ has variance $\sigma^2$. The Central Limit Theorem would thus suggest that $L_{ii}$ behave like a Gaussian of variance $\sigma^2$. Since different sums only share a single summand they are "almost" independent which by itself would suggest that $\max_i L_{ii} \sim \sigma\sqrt{\log n}$, which would imply the theorems. The proof that follows makes this argument precise.

We turn our attention to a lower bound on $\max_i y_i$. Recall that $y_i = \sum_{j=1}^n X_{ij}$. More specifically, we are looking for an upper bound on

$$\mathbb{P}\left[\max_i y_i < t\right],$$

for a suitable value of $t$. We note that, if the $y_i$'s were independent then this could be easily done via lower bounds on the upper tail of each $y_i$. Furthermore, if the random variable $y_i$ were gaussian, obtaining such lower bounds would be trivial. Unfortunately, the random variables in question are neither independent nor gaussian, forcing major adaptations to this argument. In fact, we will actually start by lower bounding

$$\mathbb{E} \max_{i \in [n]} y_i.$$

We will obtain such a bound via a comparison (using Jensen's inequality) with the maximum among certain independent random variables.

**Lemma 4.2.7.** *Let $\mathcal{I}$ and $\mathcal{J}$ be disjoint subsets of $[n]$. For $i \in \mathcal{I}$ define $z_i$ as*

$$z_i = \sum_{j \in \mathcal{J}} X_{ij}. \tag{4.6}$$

*Then*

$$\mathbb{E} \max_{i \in [n]} y_i \geq \mathbb{E} \max_{i \in \mathcal{I}} z_i.$$

*Proof.*

$$\mathbb{E} \max_{i \in [n]} y_i = \mathbb{E} \max_{i \in [n]} \sum_{j=1}^{n} X_{ij} \geq \mathbb{E} \max_{i \in \mathcal{I}} \sum_{j=1}^{n} X_{ij}.$$

Since $\mathcal{I} \cap \mathcal{J} = \emptyset$, $\{X_{ij}\}_{i \in \mathcal{I}, j \in \mathcal{J}} is independent from \{X_{ij}\}_{i \in \mathcal{I}, j \notin \mathcal{J}}$, and so Jensen's inequality gives

$$\mathbb{E} \max_{i \in \mathcal{I}} \sum_{j=1}^{n} X_{ij} \geq \mathbb{E} \max_{i \in \mathcal{I}} \left[ \sum_{j \in \mathcal{J}} X_{ij} + \sum_{j \notin \mathcal{J}} \mathbb{E} X_{ij} \right] = \mathbb{E} \max_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{ij} = \mathbb{E} \max_{i \in \mathcal{I}} z_i.$$

□

The following Lemma guarantees the existence of sets $\mathcal{I}$ and $\mathcal{J}$ with desired properties.

**Lemma 4.2.8.** *There exist $\mathcal{I}$ and $\mathcal{J}$ disjoint subsets of $[n]$ such that*

$$|\mathcal{I}| \geq \frac{1}{8}n,$$

*and, for every $i \in \mathcal{I}$,*

$$\mathbb{E}z_i^2 \geq \frac{1}{8}\sigma^2,$$

*where $z_i$ is defined, as in (4.6), to be $z_i = \sum_{j\in\mathcal{J}} X_{ij}$.*

*Proof.* Given the matrix $X$, we start by constructing a weighted graph on $n$ nodes such that $w_{ij} = \mathbb{E}X_{ij}^2$ (note that $w_{ii} = 0$, for al $i$). Let $(S, S^c)$ be a partition of the vertices of this graph, with $|S| \geq \frac{n}{2}$, that maximizes the cut

$$\sum_{i\in S,\, j\in S^c} w_{ij}.$$

It is easy to see that the maximum cut needs to be at least half of the total edge weights[4]. This readily implies

$$\sum_{i\in S,\, j\in S^c} w_{ij} \geq \frac{1}{2}\sum_{i<j} w_{ij} = \frac{1}{4}\sum_{i\in[n]}\sum_{j\in[n]} w_{ij} = \frac{1}{4}\sum_{i\in[n]}\sum_{j\in[n]} \mathbb{E}X_{ij}^2 = \frac{1}{4}n\sigma^2.$$

Consider $z_i$, for $i \in S$, defined as

$$z_i = \sum_{j\in S^c} X_{ij}.$$

We proceed by claiming that the set $\mathcal{I} \subset S$ of indices $i \in S$ for which

$$\mathbb{E}z_i^2 \geq \frac{1}{8}\sigma^2,$$

[4] One can build such a cut by consecutively selecting memberships for each node in a greedy fashion as to maximize the number of incident edges cut, see [197].

satisfies $|\mathcal{I}| \geq \frac{1}{8}n$. Thus, taking $\mathcal{J} = S^c$ would establish the Lemma.

To justify the claim, note that

$$\sum_{i \in S} \mathbb{E} z_i^2 = \sum_{i \in S,\, j \in S^c} w_{ij} \geq \frac{1}{4} n \sigma^2,$$

and

$$\sum_{i \in S} \mathbb{E} z_i^2 \leq |\mathcal{I}| \max_{i \in S} \mathbb{E} z_i^2 + (|S| - |\mathcal{I}|) \frac{1}{8} \sigma^2 \leq \left( |\mathcal{I}| + \frac{1}{8} |S| \right) \sigma^2 \leq \left( |\mathcal{I}| + \frac{1}{8} n \right) \sigma^2,$$

implying that $\left(|\mathcal{I}| + \frac{1}{8}n\right) \sigma^2 \geq \frac{1}{4} n \sigma^2$.

□

We now proceed by obtaining a lower bound for $\mathbb{E} \max_{i \in \mathcal{I}} z_i$, where $\mathcal{I}$ and $z_i$ are defined to satisfy the conditions in Lemma 4.2.8. We note that at this point the random variables $z_i$ are independent and each is a sum of independent random variables. We use Lemma 8.1 of [143] (for a fixed constant $\gamma = 1$) to obtain a lower bound on the upper tail of each $z_i$.

**Lemma 4.2.9.** [Lemma 8.1 of [143]] *In the setting described above, there exist two universal positive constants $K$ and $\varepsilon$ such that for every $t$ satisfying $t \geq K \frac{\sigma}{8}$ and $t \leq \varepsilon \frac{\sigma^2}{\sqrt{8}\sigma_\infty}$, we have (for every $i \in \mathcal{I}$ separately)*

$$\mathbb{P}\left[z_i > t\right] \geq \exp\left(-8 \frac{t^2}{\sigma^2}\right).$$

We are now ready to establish a lower bound on $\mathbb{E} \max_{i \in [n]} y_i$.

**Lemma 4.2.10.** *In the setting described above, there exist two universal positive constants $K$ and $\varepsilon$ such that for every $t$ satisfying $t \geq K \frac{\sigma}{8}$ and $t \leq \varepsilon \frac{\sigma^2}{\sqrt{8}\sigma_\infty}$, we have*

$$\mathbb{E} \max_{i \in [n]} y_i \geq t - (t + n\sigma_\infty) \exp\left(-\frac{n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right)$$

*Proof.* Let $K$ and $\varepsilon$ be the universal constants in Lemma 4.2.9 and $t$ such that $K\frac{\sigma}{8} \leq t \leq \varepsilon \frac{\sigma^2}{\sqrt{8}\sigma_\infty}$. Lemma 4.2.9 guarantees that, for any $i \in \mathcal{I}$,

$$\mathbb{P}\left[z_i > t\right] \geq \exp\left(-8\frac{t^2}{\sigma^2}\right).$$

Due to the independence of the random variables $z_i$, we have

$$\begin{aligned} \mathbb{P}\left[\max_{i\in\mathcal{I}} z_i \leq t\right] &= \prod_{i\in\mathcal{I}} \mathbb{P}\left[z_i \leq t\right] = \prod_{i\in\mathcal{I}} \left(1 - \mathbb{P}\left[z_i > t\right]\right) \\ &\leq \left(1 - \frac{1}{\exp\left(8\frac{t^2}{\sigma^2}\right)}\right)^{|\mathcal{I}|} \leq \left(1 - \frac{1}{\exp\left(8\frac{t^2}{\sigma^2}\right)}\right)^{n/8} \\ &\leq \exp\left(-\frac{n/8}{\exp\left(8\frac{t^2}{\sigma^2}\right)}\right) \end{aligned}$$

where the second to last inequality follows from the fact that $|\mathcal{I}| \geq \frac{1}{8}n$ and the last from the fact that $\left(1-\frac{1}{x}\right)^x \leq \exp(-1)$ for $x > 1$.

Since $\|X_{ij}\|_\infty \leq \sigma_\infty$ we have that, almost surely, $z_i \geq -(n-1)\sigma_\infty$. Thus,

$$\mathbb{E}\max_{i\in[n]} y_i \geq \mathbb{E}\max_{i\in\mathcal{I}} z_i \geq t\left[1 - \exp\left(-\frac{n/8}{\exp\left(8\frac{t^2}{\sigma^2}\right)}\right)\right] - (n-1)\sigma_\infty \exp\left(-\frac{n/8}{\exp\left(8\frac{t^2}{\sigma^2}\right)}\right),$$

which establishes the Lemma.

□

The last ingredient we need is a concentration result to control the lower tail of $\max_{i\in[n]} y_i$ by controling its fluctuations around $\mathbb{E}\max_{i\in[n]} y_i$. We make use of a result in [159].

**Lemma 4.2.11.** *In the setting described above, define $v$ as*

$$v = \mathbb{E}\left[\max_{i\in[n]} \sum_{j=1}^n \left(X_{ij} - X'_{ij}\right)^2\right], \tag{4.7}$$

*where $X'$ is an independent identically distributed copy of $X$.*

*Then, for any $x > 0$:*

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \mathbb{E}\left[\max_{i\in[n]} y_i\right] - x\right] \leq \exp\left(-\frac{x^2}{7(v+\sigma_\infty x)}\right).$$

*Proof.* This Lemma is a direct consequence of Theorem 12 in [159] by taking the independent random variables to be $Y_{(i,j)}$ such that $Y_{(i,j),t} = X_{ij}$ if $t = i$ and $Y_{(i,j),t} = 0$ otherwise. We note that there is a small typo (in the definition of the quantity $v$) in the Theorem as stated in [159]. □

At this point we need an upper bound on the quantity $v$ defined in (4.7). This is the purpose of the following Lemma.

**Lemma 4.2.12.** *In the setting above, let $X'$ is an independent identically distributed copy of $X$, then*

$$\mathbb{E}\left[\max_{i\in[n]} \sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right] \leq 9\sigma^2 + 90\sigma_\infty^2 \log n.$$

*Proof.* We apply a Rosenthal-type inequality from Theorem 8 of [52], for each $i \in [n]$ separately, and get, for any integer $p$ and $0 < \delta < 1$,

$$\begin{aligned} \left\|\sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right\|_p &\leq (1+\delta)\mathbb{E}\left[\sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right] + \frac{2p}{\delta}\left\|\max_{j\in[n]} \left(X_{ij} - X'_{ij}\right)^2\right\|_p \\ &\leq 2(1+\delta)\sigma^2 + \frac{8p}{\delta}\sigma_\infty^2. \end{aligned} \tag{4.8}$$

It is easy to see that

$$\mathbb{E}\left[\max_{i\in[n]} \sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right] \leq n^{\frac{1}{p}} \left\|\sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right\|_p.$$

Thus, taking $p = \lceil \alpha \log n \rceil$ for some $\alpha > 0$ gives

$$\begin{aligned} \mathbb{E}\left[\max_{i\in[n]} \sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right] &\leq n^{\frac{1}{\lceil \alpha \log n \rceil}} 2(1+\delta)\sigma^2 + n^{\frac{1}{\lceil \alpha \log n \rceil}} \frac{8\lceil \alpha \log n \rceil}{\delta} \sigma_\infty^2 \\ &\leq e^{\frac{1}{\alpha}} 2(1+\delta)\sigma^2 + e^{\frac{1}{\alpha}} \frac{8\lceil \alpha \log n \rceil}{\delta} \sigma_\infty^2. \end{aligned}$$

Taking, for example, $\delta = 0.5$ and $\alpha = 1$ gives

$$\mathbb{E}\left[\max_{i\in[n]} \sum_{j=1}^{n} \left(X_{ij} - X'_{ij}\right)^2\right] \leq 9\sigma^2 + 90\sigma_\infty^2 \log n.$$

□

We now collect all our bounds in a master Lemma.

**Lemma 4.2.13.** *In the setting described above, there exist universal constants $K > 0$ and $\varepsilon > 0$ such that, for any $t$ satisfying $K\frac{\sigma}{8} \leq t \leq \varepsilon \frac{\sigma^2}{\sqrt{8}\sigma_\infty}$, we have*

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{t}{2} - (t + n\sigma_\infty)\exp\left(\frac{-n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right)\right] \leq \exp\left(\frac{-t^2/10^4}{\sigma^2 + \sigma_\infty^2 \log n + \sigma_\infty t}\right)$$

*Proof.* Let $t > 0$ satisfy the hypothesis of the Lemma, and $x > 0$.

Recall that Lemma 4.2.11 gives

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \mathbb{E}\left[\max_{i\in[n]} y_i\right] - x\right] \leq \exp\left(-\frac{x^2}{7(v + \sigma_\infty x)}\right).$$

On the other hand, Lemma 4.2.10 and 4.2.12 control, respectively, $\mathbb{E}\left[\max_{i\in[n]} y_i\right]$ and $v$, giving

$$\mathbb{E}\left[\max_{i\in[n]} y_i\right] \geq t - (t + n\sigma_\infty)\exp\left(-\frac{n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right),$$

and

$$v \leq 9\sigma^2 + 90\sigma_\infty^2 \log n.$$

Combining all these bounds,

$$\begin{aligned}\mathbb{P}\left[\max_{i\in[n]} y_i \leq t - (t + n\sigma_\infty)\exp\left(-\frac{n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right) - x\right]&\\ \leq \exp\left(-\frac{x^2}{7(9\sigma^2 + 90\sigma_\infty^2 \log n + \sigma_\infty x)}\right).&\end{aligned}$$

Taking $x = t/2$ establishs the Lemma.

□

At this point, the proofs of Theorems 4.2.1 and 4.2.2 will consist essentially of applying Lemma 4.2.13 for appropriate values of $t$.

*Proof.* [of Theorem 4.2.1]

Let $\beta > 0$ be a constant to be defined later. Taking $t = \beta\sigma\sqrt{\log n}$ in Lemma 4.2.13 gives that, in the setting described above,

$$\begin{aligned}\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{\beta}{2}\sigma\sqrt{\log n} - \left(\beta\sigma\sqrt{\log n} + n\sigma_\infty\right)\exp\left(-n^{1-8\beta^2}\right)\right]&\\ \leq \exp\left(\frac{-\beta^2\sigma^2\log n/10^4}{\sigma^2 + \sigma_\infty^2\log n + \sigma_\infty(\beta\sigma\sqrt{\log n})}\right)&\\ = \exp\left(\frac{-\beta^2\log n/10^4}{1 + \left(\frac{\sigma_\infty}{\sigma}\right)^2\log n + \frac{\sigma_\infty}{\sigma}\beta\sqrt{\log n}}\right)&\\ = n^{-\left(\frac{\beta^2/10^4}{1+\left(\frac{\sigma_\infty}{\sigma}\right)^2\log n+\frac{\sigma_\infty}{\sigma}\beta\sqrt{\log n}}\right)},&\end{aligned}$$

provided that $K\frac{\sigma}{8} \leq \beta\sigma\sqrt{\log n} \leq \varepsilon\,\frac{\sigma^2}{\sqrt{8}\sigma_\infty}$, where $K$ and $\varepsilon$ are the universal constants in Lemma 4.2.13.

We start by noting that, if $0 < \beta < \frac{1}{\sqrt{8}}$ independent of $n$, then, for $n$ large enough (not depending on $\sigma$ or $\sigma_\infty$),

$$\left(\beta\sigma\sqrt{\log n} + n\sigma_\infty\right)\exp\left(-n^{1-8\beta^2}\right) \leq \frac{\beta}{6}\sigma\sqrt{\log n}.$$

Thus, provided that $\frac{K}{8\sqrt{\log n}} \leq \beta \leq \min\left\{\varepsilon\,\frac{\sigma}{\sqrt{8\log n}\sigma_\infty}, \frac{1}{3}\right\}$,

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{\beta}{3}\sigma\sqrt{\log n}\right] \leq n^{-\left(\frac{\beta^2/10^4}{1+(\frac{\sigma_\infty}{\sigma})^2\log n+\frac{\sigma_\infty}{\sigma}\beta\sqrt{\log n}}\right)}.$$

Let $c$ be the constant in the hypothesis of the theorem, then $\sigma > c\sqrt{\log n}\sigma_\infty$. Let $\beta = \min\left\{\frac{\varepsilon\, c}{\sqrt{8}}, \frac{1}{3}\right\}$. Clearly, for $n$ large enough,

$$\frac{K}{8\sqrt{\log n}} \leq \min\left\{\frac{\varepsilon\, c}{\sqrt{8}}, \frac{1}{3}\right\} \leq \min\left\{\varepsilon\,\frac{\sigma}{\sqrt{8\log n}\sigma_\infty}, \frac{1}{3}\right\},$$

and

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \min\left\{\frac{\varepsilon\, c}{6\sqrt{2}}, \frac{1}{9}\right\}\sigma\sqrt{\log n}\right] \leq n^{-\left(\frac{10^{-4}}{\max\left\{\frac{8}{\varepsilon^2 c^2},9\right\}+\max\left\{\frac{8}{\varepsilon^2},9c^2\right\}+\max\left\{\frac{\sqrt{8}}{\varepsilon},3c\right\}}\right)}.$$

This implies that there exist constants $c_1', C_1'$ and $\beta_1'$ such that

$$\mathbb{P}\left[\max_{i\in[n]} L_{ii} \leq C_1'\sigma\sqrt{\log n}\right] \leq c_1' n^{-\beta_1'}.$$

Recall that Corollary 4.1.14 ensures that, for a universal constant $c'$, and for every $u \geq 0$, by taking $t = u\sigma$,

$$\mathbf{P}[\|X\| > (3+u)\sigma] \leq ne^{-u^2\sigma^2/c'\sigma_\infty^2}. \tag{4.9}$$

It is easy to see that $ne^{-u^2\sigma^2/c'\sigma_\infty^2} \leq ne^{-u^2(\log n)c/c'} = n^{1-u^2c/c'}$. Taking $u = \sqrt{2c'/c}$

gives

$$\mathbf{P}\left[\|X\| > \left(3+\sqrt{2c'/c}\right)\sigma\right] \leq n^{-1}.$$

This means that, with probability at least $1 - c_1' n^{-\beta_1'} - n^{-1}$ we have

$$\|X\| < \left(3+\sqrt{2c'/c}\right)\sigma \leq \frac{3+\sqrt{2c'/c}}{C_1'\sqrt{\log n}} \max_{i\in[n]} L_{ii},$$

which, together with the fact that $\lambda_{\max}(L) \leq \|X\| + \max_{i\in[n]} L_{ii}$, establishes the theorem.

□

*Proof.* [of Theorem 4.2.2]

If $\sigma > \sqrt{\log n}\sigma_\infty$ then the result follows immediately from Theorem 4.2.1. For that reason we restrict our attention to the instances with $\sigma \leq \sqrt{\log n}\sigma_\infty$. We start by setting

$$t = 2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}. \tag{4.10}$$

Recall that there exist $c$ and $\gamma > 0$ such that $\sigma \geq c\,(\log n)^{\frac{1}{4}+\gamma}\sigma_\infty$, or equivalently

$$\frac{\sigma}{\sigma_\infty} \geq c\,(\log n)^{\frac{1}{4}+\gamma}.$$

This guarantees that, for $n$ large enough (not depending on $\sigma$ or $\sigma_\infty$), the conditions in Lemma 4.2.13 are satisfied. In fact,

$$\frac{K\sigma}{8} \leq 2\sigma\sqrt{c}\,(\log n)^{\frac{1}{4}+\frac{\gamma}{2}} \leq 2\sigma\sqrt{\frac{\sigma}{\sigma_\infty}}(\log n)^{\frac{1}{8}} \leq \frac{\varepsilon\,\sigma}{\sqrt{8}}\sqrt{\frac{\sigma}{\sigma_\infty}}\sqrt{c}\,(\log n)^{\frac{1}{8}+\frac{\gamma}{2}} \leq \frac{\varepsilon\,\sigma^2}{\sqrt{8}\sigma_\infty}.$$

Hence, Lemma 4.2.13 gives, for $t$ as in (4.10),

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{t}{2} - (t+n\sigma_\infty)\exp\left(\frac{-n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right)\right] \leq \exp\left(\frac{-t^2/10^4}{\sigma^2+\sigma_\infty^2\log n + \sigma_\infty t}\right).$$

We proceed by noting that, for $t = 2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}$ and $n$ large enough (not depending on $\sigma$ or $\sigma_\infty$),

$$(t + n\sigma_\infty)\exp\left(\frac{-n}{\exp\left(\frac{8t^2}{\sigma^2}\right)}\right) \leq \frac{t}{6}.$$

In fact, since $\sigma \leq \sigma_\infty\sqrt{\log n}$,

$$\exp\left(\frac{-n}{\exp\left(\frac{8\left(2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{1/2}(\log n)^{1/8}\right)^2}{\sigma^2}\right)}\right) \leq \exp\left(\frac{-n}{\exp\left(32(\log n)^{3/4}\right)}\right),$$

decreases faster than any polynomial.

Hence, since $t \geq 2\sigma\sqrt{c}\,(\log n)^{\frac{1}{4}+\frac{\gamma}{2}}$,

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{2}{3}\sigma\sqrt{c}\,(\log n)^{\frac{1}{4}+\frac{\gamma}{2}}\right] \leq \exp\left(\frac{-\left(2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}\right)^2/10^4}{\sigma^2 + \sigma_\infty^2\log n + \sigma_\infty 2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}}\right).$$

We proceed by noting that

$$\frac{\left(2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}\right)^2/10^4}{\sigma^2 + \sigma_\infty^2\log n + \sigma_\infty 2\sigma\left(\frac{\sigma}{\sigma_\infty}\right)^{\frac{1}{2}}(\log n)^{\frac{1}{8}}} = \frac{4(\log n)^{\frac{1}{4}}/10^4}{\frac{\sigma_\infty}{\sigma} + \left(\frac{\sigma_\infty}{\sigma}\right)^3\log n + 2\left(\frac{\sigma_\infty}{\sigma}\right)^{\frac{3}{2}}(\log n)^{\frac{1}{8}}}$$

Since $\frac{\sigma_\infty}{\sigma} \leq \frac{1}{c}(\log n)^{-\frac{1}{4}-\gamma}$, we have that, for $n$ large enough and a constant $c''$

$$\mathbb{P}\left[\max_{i\in[n]} y_i \leq \frac{2}{3}\sigma\sqrt{c}\,(\log n)^{\frac{1}{4}+\frac{\gamma}{2}}\right] \leq \exp\left(-c''(\log n)^{\gamma}\right).$$

At this point we upper bound $\|X\|$, as in the proof of Theorem 4.2.1. Recall, as in (4.9), for any $u > 0$,

$$\mathbf{P}[\|X\| > (3+u)\sigma] \leq ne^{-\frac{u^2\sigma^2}{c'\sigma_\infty^2}}.$$

Hence,

$$\mathbf{P}[\|X\| > (3+u)\sigma] \leq ne^{-\frac{u^2c^2}{c'}(\log(n))^{\frac{1}{2}+2\gamma}}.$$

Taking $u = (\log n)^{\frac{1}{4}}$ gives

$$\mathbf{P}[\|X\| > \left(3 + (\log n)^{\frac{1}{4}}\right)\sigma] \leq e^{-\frac{c^2}{c'}(\log(n))^{2\gamma}}.$$

The rest of the proof follows the final arguments in the proof of Theorem 4.2.1.

□

# Chapter 5

# Beyond exact recovery: Tightness of relaxations

## 5.1 Tightness for angular synchronization

In Chapter 3 we established exact recovery guarantees for semidefinite relaxations for a number of synchronization-type problems over $\mathbb{Z}_2$. Several guarantees of this type exist in different settings, to many a few: compressed sensing [65, 96, 223], matrix completion [70] and inverse problems [74, 20]. Usually this type of guarantees is similar in flavor to the ones derived in Chapter 3: there is a planted, or intended, solution and the random distribution on the input usually corresponds to noisy or otherwise corrupted measurements of the planted solution. The guarantees are then obtained by showing that the planted solution is, with high probability, the solution to a tractable relaxation.

However, unfortunately, there are many problems for which exact recovery is unrealistic. Either due to the amount of noise in the measurements or due to the topology of the space of solutions, one cannot expect to recover the original planted solution, even if armed with unlimited computational resources. For concreteness let us restrict the description to synchronization-type problems. Recall (Section 1.3.3) that the semidefinite programming-based relaxations are tractable relaxations for the MLE. If the MLE itself turns out not to correspond to the planted solution, than so cannot the relaxation, rendering the tools used in the examples above (and in Chapter 3) inadequate.

On the other hand, despite not coinciding with the planted solution, the MLE still enjoys many desirable statistical properties and our goal is still to compute it (although being, in general, non-tractable). Fortunately, it seems that in many instances these semidefinite relaxations tend to be tight (i.e. their solution coincides with the MLE) even in this setting, allowing us to compute the MLE efficiently [37, 30]. As, in many of these problems, the semidefinite relaxation being tight is equivalent to its optimal solution having the same rank as the planted solution, we refer to this remarkable phenomenon as *rank recovery*; it is the subject of this Chapter, which is mostly based on [30].

This behavior has been observed in many instances, including the multireference alignment problem by [32] (discussed in Section 6.1), the global registration problem [76], and camera motion estimation by [177, 178], to name a few. Yet, it seems there to be no theoretical understanding of this *rank recovery* phenomenon. We note that there has been work on understanding the rank of solutions of random SDPs by [19] but the results hold only under specific distributions and do not apply to these problems. It appears that the difficulty of analyzing rank recovery lies in the fact that, unlike in exact recovery, we cannot identify the exact form of the MLE, rendering dual certificate arguments very difficult to carry out.[1]

In this Section we will restrict our attention to the problem of Angular Synchronization (Section 1.2.1) and investigate the rank recovery phenomenon for this problems. The main contribution is a proof that, even though the angular synchro-

[1] In Section 7.2.1 we will discuss future directions of research related to these observations.

nization problem is NP-hard [240], in the face of Gaussian noise, the semidefinite relaxation for its MLE is indeed often tight, meaning that the MLE can be computed (and certified) in polynomial time. This remains true even for entry-wise noise levels growing to infinity as the size of the problem (the number of phases) grows to infinity.

Computing the MLE for angular synchronization is equivalent to solving a little Grothendieck problem over the complex numbers, similar to the ones described in Section 2.3. Our results show the surprising phenomenon that, in a randomized version of the Grothendieck problem, where there is a planted signal, there is no gap between the SDP relaxation and the original problem, with high probability.

The proposed result is qualitatively different from most tightness results available in the literature. Typical results establish either exact recovery of a planted signal [1, 130] (mostly in discrete settings), or exact recovery in the absence of noise, joint with stable (but not necessarily optimal) recovery when noise is present [67, 141, 66, 233, 93]. In contrast, here, we show optimal recovery even though exact recovery is not possible. In particular, Demanet and Jugnon showed stable recovery for angular synchronization via semidefinite programming, in an adversarial noise setting [93]. We complement that analysis by showing tightness in a non-adversarial setting, meaning that the actual MLE is computed.

Similarly to the exact recovery results established in Chapter 3, the proof relies on verifying that a certain candidate dual certificate is valid with high probability. The main difficulty comes from the fact that the dual certificate depends of the MLE, which does not coincide with the planted signal, and is a nontrivial function of the noise. We use necessary optimality conditions of the hard problem to both obtain an explicit expression for the candidate dual certificate, and to partly characterize the point whose optimality we aim to establish. This seems to be required since the MLE is not known in closed form.

In the context of sparse recovery, a result with similar flavor is support recovery guarantee [223], where the support of the estimated signal is shown to be contained in the support of the original signal. Due to the noise, exact recovery is also impossible in this setting. Other examples are recovery guarantee in the context of latent variable selection in graphical models [73] and sparse Principal Component Analysis [21].

Besides the relevance of angular synchronization in and of its own, we are confident this new insight will help uncover similar results in other applications where it has been observed that semidefinite relaxations can be tight even when the ground truth cannot be recovered, including severa synchronization-type problems [37].

The crux of our argument is to show that the SDP, with random data following a given distribution, admits a unique rank-one solution with high probability. We mention in passing that there are many other, deterministic results in the literature pertaining to the rank of solutions of SDP's. For example, Barvinok [45] and Pataki [179] both show that, in general, an SDP with only equality constraints admits a solution of rank at most (on the order of) the square root of the number of constraints. Furthermore, Sagnol [195] shows that under some conditions (that are not fulfilled in our case), certain SDP's related to packing problems always admit a rank-one solution. Sojoudi and Lavaei [214] study a class of SDP's on graphs which is related to ours and for which, under certain strong conditions on the topology of the graphs, the SDP's admit rank-one solutions—see also applications to power flow optimization [151].

### 5.1.1 The Angular Synchronization problem

We focus on the problem of *angular synchronization* [202, 41] presented in Section 1.2.1. Let us recall the setting: one wishes to estimate a collection of $n$ phases ($n \geq 2$) based on measurements of pairwise phase differences. We will restrict our analysis to the case where a measurement is available for every pair of nodes. More precisely, we let $z \in \mathbb{C}^n$ be an unknown, complex vector with unit modulus entries,

$|z_1| = \cdots = |z_n| = 1$, and we consider measurements of the form $C_{ij} = z_i\overline{z_j} + \varepsilon_{ij}$, where $\overline{z_j}$ denotes the complex conjugate of $z_j$ and $\varepsilon_{ij} \in \mathbb{C}$ is noise affecting the measurement. By symmetry, we define $C_{ji} = \overline{C_{ij}}$ and $C_{ii} = 1$, so that the matrix $C \in \mathbb{C}^{n\times n}$ whose entries are given by the $C_{ij}$'s is a Hermitian matrix. Further letting the noise $\varepsilon_{ij}$ be i.i.d. (complex) Gaussian variables for $i < j$, it follows directly that an MLE for the vector $z$ is any vector of phases $x \in \mathbb{C}^n$ minimizing $\sum_{i,j} |C_{ij}x_j - x_i|^2$. Equivalently, an MLE is a solution of the following quadratically constrained quadratic program (sometimes called the complex constant-modulus QP [153, Table 2] in the optimization literature, it also corresponds to a complex valued version of the Grothendieck problem treated in Section 2.3):

$$\max_{x\in\mathbb{C}^n} \; x^*Cx, \quad \text{subject to } |x_1| = \cdots = |x_n| = 1, \tag{5.1}$$

where $x^*$ denotes the conjugate transpose of $x$. Of course, this problem can only be solved up to a global phase, since only relative information is available. Indeed, given any solution $x$, all vectors of the form $xe^{i\theta}$ are equivalent solutions, for arbitrary phase $\theta$.

Similarly to the real case studied extensively in Section 2.3 and Chapter 3, solving (5.1) is, in general, an NP-hard problem [240, Prop. 3.5] and so we will consider a tractable semidefinite programming-based relaxation [240, 209, 153]. For any admissible $x$, the Hermitian matrix $X = xx^* \in \mathbb{C}^{n\times n}$ is Hermitian positive semidefinite, has unit diagonal entries and is of rank one. Conversely, any such $X$ may be written in the form $X = xx^*$ such that $x$ is admissible for (5.1). In this case, the cost function can be rewritten in linear form: $x^*Cx = \operatorname{Tr}(x^*Cx) = \operatorname{Tr}(CX)$. Dropping the rank constraint then yields the relaxation we set out to study:

$$\max_{X\in\mathbb{C}^{n\times n}} \; \operatorname{Tr}(CX), \quad \text{subject to } X_{ii} = 1, \ \forall_i \text{ and } X \succeq 0, \tag{5.2}$$

The techniques described in Section 2.3 can be easily adapted to this setting [209, 36, 115] asserting that the solution of (5.2) can be rounded to an approximate solution of (5.1), with a guaranteed approximation ratio.. But even better, when (5.2) admits an optimal solution $X$ of rank one, then no rounding is necessary: the leading eigenvector $x$ of $X = xx^*$ is a global optimum of (5.1), meaning we have solved the original problem exactly. Elucidating when the semidefinite program admits a solution of rank one, i.e., when the relaxation is *tight*, is the focus of the sequel.

As was mentioned earlier, in the presence of even the slightest noise, one cannot reasonably expect the true signal $z$ to be an optimal solution of (5.1) anymore (this can be formalized using Cramér-Rao bounds [55]). Nevertheless, we set out to show that (under some assumptions on the noise) solutions of (5.1) are close to $z$ and they can be computed via (5.2).

The proof follows, in spirit, that of the real case carried out in Section 3.1.1 but requires more sophisticated arguments because the solution of (5.1) is not known explicitly anymore. One effect of this is that the candidate dual certificate $Q$ will itself depend on the unknown solution of (5.1). With that in mind, the proof of the upcoming main lemma of this section (Lemma 5.1.5) follows this reasoning:

1. For small enough noise levels $\sigma$, any optimal solution $x$ of (5.1) is close to the sought signal $z$ (Lemmas 5.1.7 and 5.1.8).

2. Solutions $x$ are, a fortiori, local optimizers of (5.1), and hence satisfy first-order necessary optimality conditions. These take up the form $Qx = 0$, where $Q = \Re\{\mathrm{ddiag}(Cxx^*)\} - C$ depends smoothly on $x$ (see (5.9)). Note that $Q$ is a function of $x$, the MLE (which is not explicitly known).

3. Remarkably, this Hermitian matrix $Q$ can be used as a dual certificate for solutions of (5.2). Indeed, $X = xx^*$ is a solution of (5.2) if and only if $Qx = 0$ and $Q$ is positive semidefinite—these are the Karush-Kuhn-Tucker conditions

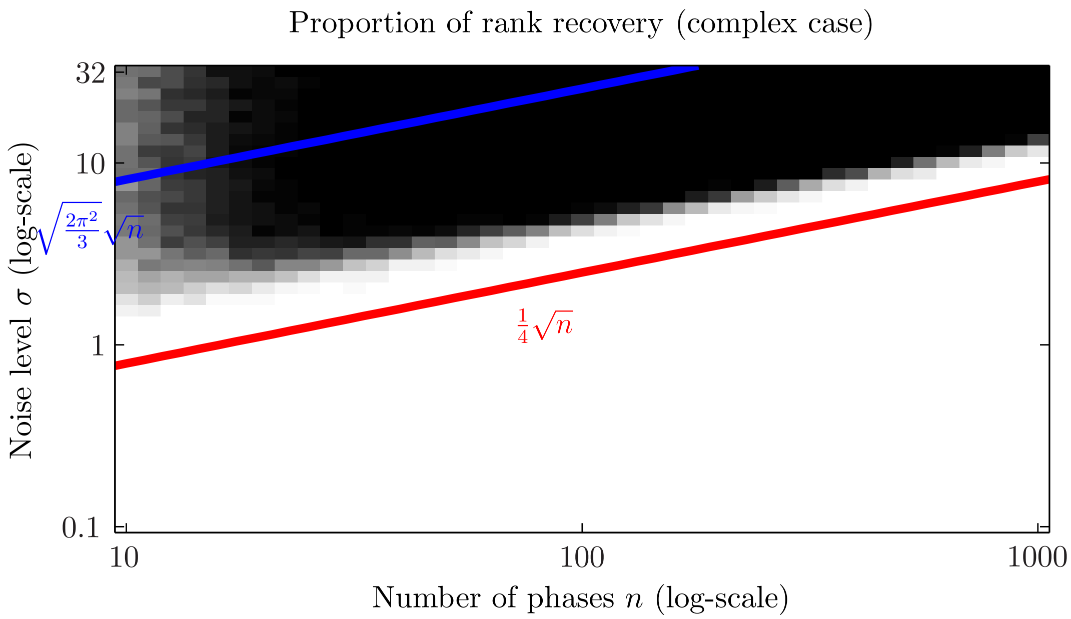

Figure 5.1: *Exact* recovery of the phases $z \in \{e^{i\theta} : \theta \in \mathbb{R}\}^n$ from the pairwise relative phase measurements $zz^* + \sigma W$ is hopeless as soon as $\sigma > 0$ [55]. Computing a maximum likelihood estimator (MLE) for $z$ is still interesting. In particular, below the red (bottom) line, with high probability, the MLE is closer to $z$ than the estimator with maximum error (Lemma 5.1.7). Computing the MLE is hard in general, but solving the semidefinite relaxation (5.2) is easy. When (5.2) has a rank-one solution, that solution coincides with the MLE. This figure shows, empirically, how frequently the (5.2) admits a unique rank-one solution: for each pair $(n, \sigma)$, 100 realizations of the noise are generated independently and thightness is checked. The frequency of success is coded by intensity (bright for 100% success, dark for 0% success). (5.2) is solved using a Riemannian optimization toolbox [54]. The (5.2) appears to be tight for remarkably large levels of noise. Theorem 5.1.2 partly explains this phenomenon, by showing that $\sigma$ can indeed grow unbounded while retaining rank recovery, albeit not at the rate witnessed here. We further note that, above the blue (top) line, no unbiased estimator for $z$ performs better than a random guess [55].

(KKT). Furthermore, that solution is unique if $\operatorname{rank}(Q) = n-1$ (Lemmas 5.1.9 and 5.1.10). Thus, it only remains to study the eigenvalues of $Q$.

4. In the absence of noise, $Q$ is a Laplacian for a complete graph with unit weights (up to a unitary transformation), so that its eigenvalues are 0 with multiplicity 1 and $n$ with multiplicity $n-1$. Then, $X = zz^*$ is always the unique solution of (5.2).

5. Adding small noise, because of the first point, the solution $x$ will move only by a small amount, and hence so will $Q$. Thus, the large eigenvalues should be controllable into remaining positive (Section 5.1.3).

6. The crucial fact follows: because of the way $Q$ is constructed (using first-order optimality conditions), the zero eigenvalue is "pinned down" (as long as $x$ is a local optimum of (5.1)). Indeed, both $x$ and $Q$ change as a result of adding noise, but the property $Qx = 0$ remains valid. Thus, there is no risk that the zero eigenvalue from the noiseless scenario would become negative when noise is added.

Following this road map, most of the work in the proof below consists in bounding how far away $x$ can be from $z$ (as a function of the noise level and structure) and in using that to control the large eigenvalues of $Q$.

**Remark 5.1.1** (The role of smoothness)**.** *The third point in the road map, namely the special role of $Q$, merits further comment. This way of identifying the dual certificate already appears explicitly in Journée et al. [134], who considered a different family of real, semidefinite programs which also admit a smooth geometry when the rank is constrained.*

*In essence, KKT conditions capture the idea that, at a (local) optimizer, there is no escape direction that—up to first order—both preserves feasibility and improves* *the objective function. The KKT conditions for* (5.2) *take up the classical form "if* $X$ *is optimal, then there exists a dual certificate* $Q$ *which satisfies such and such conditions." For the purpose of certifying a solution analytically, this is impractical, because there is no explicit formula stating which* $Q$ *to verify. Fortunately,* (5.2) *is* nondegenerate *[14] (meaning, roughly, that its underlying geometry is smooth). This leads to uniqueness of dual certificates, and hence suggests there may be an analytical expression for the dual candidate.*

*The convex problem* (5.2) *is a relaxation of* (5.1) *(up to the global phase). Hence, the KKT conditions for* (5.2) *at a rank-one solution* $xx^*$ *are the KKT conditions for* (5.1) *at* $x$ plus *additional conditions: the latter ensure none of the new directions are improving directions either. Because* (5.1) *is a smooth problem, the KKT conditions for* (5.1) *are explicit: if* $x$ *is a (local) optimizer of* (5.1)*, then* $\operatorname{grad} g(x) = -2Q(x)x = 0$ *(where* $g(x) = x^*Cx$ *and* $\operatorname{grad} g(x)$ *is its Riemannian gradient* (5.8)*).*

*This gives an explicit expression for a candidate dual* $Q$ *that satisfies part of the KKT conditions of* (5.2) *at* $xx^*$*. It then suffices to add the additional conditions of* (5.2) *(namely, that* $Q$ *be positive semidefinite) to obtain an explicit expression for the unique dual candidate.*

Our main theorem follows. In a nutshell, it guarantees that: under (complex) Wigner noise $W$, with high probability, solutions of (5.1) are close to $z$, and, assuming the noise level $\sigma$ is smaller than (on the order of) $n^{1/10}$, (5.2) admits a unique solution, it is of rank one and identifies the solution of (5.1) (unique, up to a global phase shift).

**Theorem 5.1.2.** *Let* $z \in \mathbb{C}^n$ *be a vector with unit modulus entries,* $W \in \mathbb{C}^{n\times n}$ *a Hermitian Gaussian Wigner matrix and let* $C = zz^* + \sigma W$*. Let* $x \in \mathbb{C}^n$ *be a global optimizer of* (5.1)*. With probability at least* $1 - 4n^{-\frac{1}{4}\log n + 2}$*, the following is true: The (unidentifiable) global phase of* $x$ *can be chosen such that* $x$ *is close to* $z$ *in the* *following two senses:*

$$\begin{aligned}\|x - z\|_\infty &\le 2\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}, \text{ and} \\ \|x - z\|_2^2 &\le 8\sigma n^{1/2}\min\left[1, 23\sigma\right].\end{aligned}$$

*Furthermore, there exists a universal constant $K > 0$ such that, if*

$$\sigma + \sigma^{2/5} \le K\frac{n^{1/10}}{\log(n)^{2/5}}, \tag{5.3}$$

*then the semidefinite program* (5.2)*, given by*

$$\max_{X\in\mathbb{C}^{n\times n}} \operatorname{Tr}(CX), \quad \text{subject to } \operatorname{diag}(X) = \mathbf{1} \text{ and } X \succeq 0,$$

*has, as its unique solution, the rank-one matrix $X = xx^*$.*

Notice that the numerical experiments (Figure 5.1) suggest it should be possible to allow $\sigma$ to grow at a rate of $\frac{n^{1/2}}{\operatorname{polylog}(n)}$ (as in the real case), but we were not able to establish that. Nevertheless, we do show that $\sigma$ can grow unbounded with $n$. To the best of our knowledge, this is the first result of this kind. We hope it might inspire similar results in other problems where the same phenomenon has been observed [37].

**Remark 5.1.3** (On the square-root rate)**.** *The relaxation* (5.2) *can be further relaxed by summarizing the constraints* $\operatorname{diag}(X) = \mathbf{1}$ *into the single constraint* $\operatorname{Tr}(X) = n$*. In so doing, the new relaxation always admits a rank-one solution $X = vv^*$ [179] such that $v$ is a dominant eigenvector of $C$, which corresponds to the spectral method described in Section 2.1. The data $C$ can be seen as a rank-one perturbation $zz^*$ of a random matrix $\sigma W$ [202]. For i.i.d. Gaussian noise, as soon as the operator norm of the noise matrix is smaller than (twice) the operator norm of the signal (equal to $n$), dominant eigenvectors of $C$ are expected to have better-than-random correlation with* *the signal $z$. Since the operator norm of such $W$'s grows as $n^{1/2}$, this explains why, even when $\sigma$ grows as $n^{1/2}$ itself, some signal is still present in the data.*

### 5.1.2 Main result

In this section we present our main technical result and show how it can be used to prove Theorem 5.1.2. Let us start by presenting a central definition in this section. Intuitively, this definition characterizes non-adversarial noise matrices $W$.

**Definition 5.1.4** ($z$-discordant matrix)**.** *Let $z \in \mathbb{C}^n$ be a vector with unit modulus entries. A matrix $W \in \mathbb{C}^{n\times n}$ is called $z$-*discordant *if it is Hermitian and satisfies all of the following:*

1. $\|W\|_{\mathrm{e},\infty} \leq \log(n)$,
2. $\|W\| \leq 3n^{1/2}$,
3. $\|Wz\|_\infty \leq n^{3/4}$,
4. $|z^*Wz| \leq n^{3/2}$.

Recall that $\|\cdot\|_{\mathrm{e},\infty}$ is the entry-wise infinity norm. The next lemma is the main technical contribution of this section. Note that it is a deterministic, non-asymptotic statement.

**Lemma 5.1.5.** *Let $z \in \mathbb{C}^n$ be a vector with unit modulus entries, let $W \in \mathbb{C}^{n\times n}$ be a Hermitian, $z$-discordant matrix (see Definition 5.1.4), and let $C = zz^* + \sigma W$. Let $x \in \mathbb{C}^n$ be a global optimizer of* (5.1)*. The (unidentifiable) global phase of $x$ can be chosen such that $x$ is close to $z$ in the following two senses:*

$$\|x - z\|_\infty \leq 2\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}, \text{ and}$$
$$\|x - z\|_2^2 \leq 8\sigma n^{1/2}\min\left[1, 23\sigma\right].$$

*Furthermore, there exists a universal constant $K > 0$ such that, if*

$$\sigma + \sigma^{2/5} \leq K \frac{n^{1/10}}{\log(n)^{2/5}}, \tag{5.4}$$

*then the semidefinite program* (5.2)*, given by*

$$\max_{X \in \mathbb{C}^{n\times n}} \operatorname{Tr}(CX), \quad \textit{subject to } \operatorname{diag}(X) = \mathbf{1} \ \textit{ and } X \succeq 0,$$

*has, as its unique solution, the rank-one matrix $X = xx^*$.*

We defer the proof of Lemma 5.1.5 to Section 5.1.3. The following proposition, whose proof we defer to [30] for the sake of brevity, shows how this lemma can be used to prove Theorem 5.1.2.

**Proposition 5.1.6.** *Let $z \in \mathbb{C}^n$ be a (deterministic) vector with unit modulus entries. Let $W \in \mathbb{C}^{n\times n}$ be a Hermitian standard Wigner matrix with zero diagonal, i.e. a random matrix with i.i.d. off-diagonal entries following a complex normal distribution and zeros on the diagonal. Thus, $W_{ii} = 0$, $W_{ij} = \overline{W_{ji}}$, $\mathbb{E}W_{ij} = 0$ and $\mathbb{E}|W_{ij}|^2 = 1$ (for $i \neq j$). Then, $W$ is $z$-discordant with probability at least $1 - 4n^{-\frac{1}{4}\log n + 2}$.*

The latter result is not surprising. Indeed, the definition of $z$-discordance requires two elements. Namely, (1) that $W$ be not too large (properties 1 and 2), and (2) that $W$ be not too aligned with $z$ (properties 3 and 4). For $W$ a Wigner matrix independent of $z$, those are indeed expected to hold.

The definition of $z$-discordance is not tightly adjusted to Wigner noise. As a result, it is expected that Lemma 5.1.5 will be applicable to show tightness of semidefinite relaxations for a larger span of noise models.

### 5.1.3 The proof

We now prove Lemma 5.1.5. See Section 5.1.1 for an outline of the proof. To ease the algebra involved in the proofs, and without loss of generality, we consider throughout that $z = \mathbf{1}$ (as it was done in Section 3.1.1). This corresponds to the deterministic change of variables $C \mapsto \mathrm{diag}(z)^* C \mathrm{diag}(z)$. Certainly, $W$ is $z$-discordant if and only if $\mathrm{diag}(z)^* W \mathrm{diag}(z)$ is $\mathbf{1}$-discordant.

**Global optimizers of** (5.1) **are close to** $z$

Let $x$ be any global optimizer of (5.1). Choosing the global phase of $x$ such that $\mathbf{1}^* x \geq 0$, we decompose $x$ as follows:

$$x = \mathbf{1} + \Delta, \tag{5.5}$$

where $\Delta \in \mathbb{C}^n$ should be thought of as an error term, as it represents how far a global optimizer of (5.1) is from the planted signal $\mathbf{1}$. This subsection focuses on bounding $\Delta$. We bound both its $\ell_2$ and $\ell_\infty$ norms.

The following easy $\ell_2$ bound is readily available:

$$\|\Delta\|_2^2 = \|x - \mathbf{1}\|_2^2 = x^*x + \mathbf{1}^*\mathbf{1} - 2\Re\{\mathbf{1}^* x\} = 2(n - \mathbf{1}^* x) \leq 2n. \tag{5.6}$$

The next lemma provides an improved bound when $\sigma \leq \frac{1}{4}n^{1/2}$.

**Lemma 5.1.7.** *If* $W$ *is* $\mathbf{1}$*-discordant, then*

$$\|\Delta\|_2^2 \leq 8\sigma n^{1/2}.$$

*Proof.* If $\sigma \geq \frac{1}{4}n^{1/2}$, the bound is trivial since $\|\Delta\|_2^2 \leq 2n \leq 8\sigma n^{1/2}$. We now prove the bound under the complementary assumption that $\sigma \leq \frac{1}{4}n^{1/2}$.

Since $x$ is a global maximizer of (5.1) it must, in particular, satisfy $x^*Cx \geq \mathbf{1}^* C\, \mathbf{1}$.

Hence,

$$x^*(\mathbf{1}\,\mathbf{1}^* + \sigma W)x \geq \mathbf{1}^*(\mathbf{1}\,\mathbf{1}^* + \sigma W)\,\mathbf{1},$$

or equivalently, $\sigma\left(x^*Wx - \mathbf{1}^*\,W\,\mathbf{1}\right) \geq \mathbf{1}^*\,\mathbf{1}\,\mathbf{1}^*\,\mathbf{1} - x^*\,\mathbf{1}\,\mathbf{1}^*\,x$. This readily implies that

$$\sigma\left(\|x\|_2^2\,\|W\| + |\,\mathbf{1}^*\,W\,\mathbf{1}\,|\right) \geq n^2 - |\,\mathbf{1}^*\,x|^2.$$

Hence, using $\|x\|_2^2 = n$ and $\mathbf{1}$-discordance of $W$ (more specifically, $\|W\| \leq 3n^{1/2}$ and $|\,\mathbf{1}^*\,W\,\mathbf{1}\,| \leq n^{3/2}$) we have

$$|\,\mathbf{1}^*\,x|^2 \geq n^2 - 3\sigma n^{3/2} - \sigma n^{3/2} = n^2 - 4\sigma n^{3/2}.$$

Since $\mathbf{1}^*\,x \geq 0$, under the assumption that $\sigma \leq \frac{1}{4}n^{1/2}$, we actually have:

$$\mathbf{1}^*\,x \geq \sqrt{n^2 - 4\sigma n^{3/2}}.$$

Combine the latter with the fact that $\|\Delta\|_2^2 = \|x - \mathbf{1}\,\|_2^2 = 2(n - \mathbf{1}^*\,x)$ to obtain

$$\|\Delta\|_2^2 \leq 2\left(n - \sqrt{n^2 - 4\sigma n^{3/2}}\right) \leq 8\sigma n^{1/2}.$$

The last inequality follows from $a - \sqrt{b} = \left(a - \sqrt{b}\right)\frac{a+\sqrt{b}}{a+\sqrt{b}} = \frac{a^2-b}{a+\sqrt{b}} \leq a - b/a$ for all $a > 0$ and $b \geq 0$ such that $a^2 \geq b$. □

The next lemma establishes a bound on the largest individual error, $\|\Delta\|_\infty$. This is informative for values of $n$ and $\sigma$ such that the bound is smaller than 2. Interestingly, for a fixed value of $\sigma$, the bound shows that increasing $n$ drives $\Delta$ to 0, uniformly.

**Lemma 5.1.8.** *If $W$ is $\mathbf{1}$-discordant, then*

$$\|\Delta\|_\infty \leq 2\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}.$$

*Proof.* We wish to upper bound, for all $i \in \{1, 2, \ldots, n\}$, the value of $|\Delta_i|$. Let $e_i \in \mathbb{R}^n$ denote the $i^{\text{th}}$ vector of the canonical basis (its $i^{\text{th}}$ entry is 1 whereas all other entries are zero). Consider $\hat{x} = x + (1 - x_i)e_i$, a feasible point of (5.1) obtained from the optimal $x$ by changing one of its entries to 1. Since $x$ is optimal, it performs at least as well as $\hat{x}$ according to the cost function of (5.1):

$$x^*Cx \geq \hat{x}^*C\hat{x} = x^*Cx + |1 - x_i|^2 C_{ii} + 2\Re\{(1 - \overline{x_i})e_i^*Cx\}.$$

Further develop the last term by isolating the diagonal term $C_{ii}$:

$$2\Re\{(1 - \overline{x_i})e_i^*Cx\} = 2C_{ii}\Re\{x_i - 1\} + 2\Re\Big\{(1 - \overline{x_i}) \sum_{j \neq i} C_{ij}x_j\Big\}.$$

Since $|1 - x_i|^2 C_{ii} = -2C_{ii}\Re\{x_i - 1\}$, combining the two equations above yields the following inequality:

$$\Re\Big\{(\overline{x_i} - 1) \sum_{j \neq i} (1 + \sigma W_{ij})x_j\Big\} \geq 0.$$

Injecting $x = \mathbf{1} + \Delta$ we get:

$$\Re\Big\{\overline{\Delta_i} \sum_{j \neq i} (1 + \sigma W_{ij})(1 + \Delta_j)\Big\} \geq 0.$$

Expand the product, remembering that $W_{ii} = 0$ by definition, to obtain:

$$\begin{aligned}(n-1)\Re\{\overline{\Delta_i}\} &\geq -\Re\Big\{\overline{\Delta_i} \sum_{j \neq i} \left(\sigma W_{ij} + \Delta_j + \sigma W_{ij}\Delta_j\right)\Big\} \\ &= |\Delta_i|^2 - \Re\Big\{\overline{\Delta_i} \sum_{j} \left(\sigma W_{ij} + \Delta_j + \sigma W_{ij}\Delta_j\right)\Big\}.\end{aligned}$$

At this point, recall we want to bound $|\Delta_i|$. Since

$$|\Delta_i|^2 = |x_i - 1|^2 = 2(1 - \Re\{x_i\}) = -2\Re\{\Delta_i\} = -2\Re\{\overline{\Delta_i}\},$$

the above inequality is equivalent to:

$$\begin{aligned}
|\Delta_i|^2 &\leq \frac{2}{n+1}\Re\Big\{\overline{\Delta_i}\sum_j (\sigma W_{ij} + \Delta_j + \sigma W_{ij}\Delta_j)\Big\} \\
&\leq \frac{2}{n+1}|\Delta_i|\left|\sum_j (\sigma W_{ij} + \Delta_j + \sigma W_{ij}\Delta_j)\right| \\
&\leq \frac{2}{n+1}|\Delta_i|\left(\sigma\, |e_i^* W\, \mathbf{1}| + |\, \mathbf{1}^*\, \Delta| + \sigma |e_i^* W \Delta|\right) \\
&\leq \frac{2}{n+1}|\Delta_i|\left(\sigma \|W\, \mathbf{1}\,\|_\infty + \frac{1}{2}\|\Delta\|_2^2 + \sigma\, \|W\|\, \|\Delta\|_2\right),
\end{aligned}$$

where we used the triangular inequality multiple times and the simple identity $\mathbf{1}^*\, \Delta = -\|\Delta\|_2^2/2$. The above inequality holds for all $1 \leq i \leq n$. We now leverage the **1**-discordance of $W$ (more precisely, $\|W\, \mathbf{1}\,\|_\infty \leq n^{3/4}$ and $\|W\| \leq 3n^{1/2}$) together with Lemma 5.1.7 to finally obtain:

$$\begin{aligned}
\|\Delta\|_\infty &\leq \frac{2}{n+1}\left(\sigma n^{3/4} + 4\sigma n^{1/2} + 6\sqrt{2}\,\sigma^{3/2} n^{3/4}\right) \\
&\leq 2\left(1 + 4n^{-1/4} + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4} \\
&\leq 2\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}.
\end{aligned}$$

This concludes the proof. □

As a side note, notice that, using the bound on $\|\Delta\|_\infty$, one obtains another bound on $\|\Delta\|_2^2$ as follows:

$$\|\Delta\|_2^2 \leq n\|\Delta\|_\infty^2 \leq 8\sigma n^{1/2}\left[\frac{(5 + 6\sqrt{2}\sigma^{1/2})^2}{2}\sigma\right]. \tag{5.7}$$

The factor in brackets is an increasing function of $\sigma$ that hits 1 for $\sigma \approx 0.0436$. Below that value, the above bound improves on Lemma 5.1.7 and the factor in brackets is bounded by $23\sigma$, thus yielding the bound as stated in Lemma 5.1.5. Nevertheless, in the remainder of the section, we use only Lemma 5.1.7 to bound $\|\Delta\|_2^2$. This is because we aim to allow $\sigma$ to grow with $n$, and Lemma 5.1.7 is sharper in that regime. The interest of the above bound is to show that, for small noise, the root mean squared error $\|\Delta\|_2/\sqrt{n}$ is at most on the order of $\sigma/n^{1/4}$.

### Optimality conditions for (5.2)

In Section 3.1.1 we described necessary conditions for a solution of the SDP to be an optimizer, via weak duality. Strong duality essentially states that, in many instances, these conditions (also called Karush-Kuhn-Tucker (KKT) conditions) are also sufficient. In fact, the global optimizers of the semidefinite program (5.2) can be characterized completely via the KKT conditions:

**Lemma 5.1.9.** *A Hermitian matrix $X \in \mathbb{C}^{n\times n}$ is a global optimizer of* (5.2) *if and only if there exists a Hermitian matrix $\hat{Q} \in \mathbb{C}^{n\times n}$ such that all of the following hold:*

1. $\operatorname{diag}(X) = \mathbf{1}$;
2. $X \succeq 0$;
3. $\hat{Q}X = 0$;
4. $\hat{Q} + C$ *is (real) diagonal; and*
5. $\hat{Q} \succeq 0$.

*If, furthermore,* $\operatorname{rank}(\hat{Q}) = n - 1$, *then $X$ has rank one and is the unique global optimizer of* (5.2).

*Proof.* These are the KKT conditions of (5.2) [194, Example 3.36]. Conditions 1 and 2 are primal feasibility, condition 3 is complementary slackness and conditions

4 and 5 encode dual feasibility. Since the identity matrix $I_n$ satisfies all equality constraints and is (strictly) positive definite, the so-called *Slater condition* is fulfilled. This ensures that the KKT conditions stated above are necessary and sufficient for global optimality [194, Theorem 3.34]. Slater's condition also holds for the dual. Indeed, let $\tilde{Q} = \alpha I - C$, where $\alpha \in \mathbb{R}$ is such that $\tilde{Q} \succ 0$ (such an $\alpha$ always exists); then $\tilde{Q} + C$ is indeed diagonal and $\tilde{Q}$ is strictly admissible for the dual. This allows to use results from [14]. Specifically, assuming $\operatorname{rank}(\hat{Q}) = n - 1$, Theorem 9 in that reference implies that $\hat{Q}$ is *dual nondegenerate.* Then, since $\hat{Q}$ is also optimal for the dual (by complementary slackness), Theorem 10 in that reference guarantees that the primal solution $X$ is unique. Since $X$ is nonzero and $\hat{Q}X = 0$, it must be that $\operatorname{rank}(X) = 1$. □

Certainly, if (5.2) admits a rank-one solution, it has to be of the form $X = xx^*$, with $x$ an optimal solution of the original problem (5.1). Based on this consideration, our proof of Lemma 5.1.5 goes as follows. We let $x$ denote a global optimizer of (5.1) and we consider $X = xx^*$ as a candidate solution for (5.2). Using the optimality of $x$ and assumptions on the noise, we then construct and verify a dual certificate matrix $Q$ as required per Lemma 5.1.9. In such proofs, one of the nontrivial parts is to guess an analytical form for $Q$ given a candidate solution $X$. We achieve this by inspecting the first-order optimality conditions of (5.1) (which $x$ necessarily satisfies). The main difficulty is then to show the suitability of the candidate $Q$, as it depends nonlinearly on the global optimum $x$, which itself is a complicated function of the noise $W$. Nevertheless, we show feasibility of $Q$ via a program of inequalities, relying heavily on the $\mathbf{1}$-discordance of the noise $W$ (see Definition 5.1.4).

### Construction of the dual certificate $Q$

Every global optimizer of the combinatorial problem (5.1) must, a fortiori, satisfy first-order necessary optimality conditions. We derive those now.

We endow the complex plane $\mathbb{C}$ with the Euclidean metric

$$\langle y_1, y_2 \rangle = \Re\{y_1^* y_2\}.$$

This is equivalent to viewing $\mathbb{C}$ as $\mathbb{R}^2$ with the canonical inner product, using the real and imaginary parts of a complex number as its first and second coordinates. Denote the complex circle by

$$\mathcal{S} = \{y \in \mathbb{C} : y^* y = 1\}.$$

The circle can be seen as a submanifold of $\mathbb{C}$, with tangent space at each $y$ given by (simply differentiating the constraint):

$$T_y\mathcal{S} = \{\dot{y} \in \mathbb{C} : \dot{y}^* y + y^* \dot{y} = 0\} = \{\dot{y} \in \mathbb{C} : \langle y, \dot{y} \rangle = 0\}.$$

Restricting the Euclidean inner product to each tangent space equips $\mathcal{S}$ with a Riemannian submanifold geometry. The search space of (5.1) is exactly $\mathcal{S}^n$, itself a Riemannian submanifold of $\mathbb{C}^n$ with the product geometry. Thus, problem (5.1) consists in maximizing a smooth function $g(x) = x^* C x$ over the smooth Riemannian manifold $\mathcal{S}^n$. Therefore, the first-order necessary optimality conditions for (5.1) (i.e., the KKT conditions) can be stated simply as $\operatorname{grad} g(x) = 0$, where $\operatorname{grad} g(x)$ is the Riemannian gradient of $g$ at $x \in \mathcal{S}^n$ [6]. This gradient is given by the orthogonal projection of the Euclidean (the classical) gradient of $g$ onto the tangent space of $\mathcal{S}^n$ at $x$ [6, eq. (3.37)]. The projector and the Euclidean gradient are given respectively by:

$$\begin{aligned} &\operatorname{Proj}_x \colon \mathbb{C}^n \to T_x\mathcal{S}^n \colon \dot{x} \mapsto \operatorname{Proj}_x \dot{x} = \dot{x} - \Re\{\operatorname{ddiag}(\dot{x}x^*)\}x, \\ &\nabla g(x) = 2Cx, \end{aligned}$$

where $\operatorname{ddiag} \colon \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ sets all off-diagonal entries of a matrix to zero. For $x$ a global optimizer of (5.1), it holds that

$$0 = \operatorname{grad} g(x) = \operatorname{Proj}_x \nabla g(x) = 2(C - \Re\{\operatorname{ddiag}(Cxx^*)\})x. \tag{5.8}$$

This suggests the following definitions:

$$X = xx^*, \qquad\qquad Q = \Re\{\operatorname{ddiag}(Cxx^*)\} - C. \tag{5.9}$$

Note that $Q$ is Hermitian and $Qx = 0$. Referring to the KKT conditions in Lemma 5.1.9, it follows immediately that $X$ is feasible for (5.2) (conditions 1 and 2); that $QX = (Qx)x^* = 0$ (condition 3); and that $Q + C$ is a diagonal matrix (condition 4). It thus only remains to show that $Q$ is also positive semidefinite and has rank $n - 1$. If such is the case, then $X$ is the unique global optimizer of (5.2), meaning that solving the latter solves (5.1). Note the special role of the first-order necessary optimality conditions: they guarantee complementary slackness, without requiring further work.

The following lemma further shows that $Q$ is the "right" candidate dual certificate. More precisely, for $x$ a critical point of (5.1), it is necessary and sufficient for $Q$ to be positive semidefinite in order for $X = xx^*$ to be optimal for (5.2).

**Lemma 5.1.10.** *$X$ (of any rank) is optimal for* (5.2) *if and only if it is feasible for* (5.2) *and $Q = \Re\{\operatorname{ddiag}(CX)\} - C$* (5.9) *is positive semidefinite and $QX = 0$. There exists no other dual certificate for $X$.*

*Proof.* The *if* part follows from Lemma 5.1.9. We show the *only if* part. Assume $X$ is optimal. Then, by Lemma 5.1.9, there exists $\hat{Q} \succeq 0$ which satisfies $\hat{Q}X = 0$ and $\hat{Q} + C = \hat{D}$, where $\hat{D}$ is diagonal. Thus, $CX = (\hat{D} - \hat{Q})X = \hat{D}X$ and $\Re\{\operatorname{ddiag}(CX)\} = \hat{D}$. Consequently, $Q = \hat{D} - C = \hat{Q}$. □

### A sufficient condition for $Q$ to be positive semidefinite with rank $n-1$

First, observe that the diagonal matrix $\mathrm{diag}(x)$ is a unitary matrix. Thus,

$$\mathcal{M} = \mathrm{diag}(x)^* Q \mathrm{diag}(x) = \Re\{\mathrm{ddiag}(Cxx^*)\} - \mathrm{diag}(x)^* C \mathrm{diag}(x) \tag{5.10}$$

is a Hermitian matrix whose spectrum is the same as that of $Q$. In particular, $Q$ and $\mathcal{M}$ share the same rank and they are simultaneously positive semidefinite, so that we now investigate $\mathcal{M}$.

Since $Qx = 0$, it follows that $\mathcal{M}\,\mathbf{1} = 0$ and $\mathcal{M}$ is positive semidefinite with rank $n-1$ if and only if $u^*\mathcal{M}u > 0$ for all $u \in \mathbb{C}^n$ such that $u \neq 0$ and $\mathbf{1}^*\,u = 0$. We set out to find sufficient conditions for the latter.

To this end, separate $u$ in its real and imaginary parts as $u = \alpha + i\beta$, with $\alpha, \beta \in \mathbb{R}^n$ satisfying $\mathbf{1}^\top\,\alpha = \mathbf{1}^\top\,\beta = 0$. The quadratic form expands as:

$$\begin{aligned} u^*\mathcal{M}u &= (\alpha^\top - i\beta^\top)\mathcal{M}(\alpha + i\beta) \\ &= \alpha^\top\mathcal{M}\alpha + \beta^\top\mathcal{M}\beta + i(\alpha^\top\mathcal{M}\beta - \beta^\top\mathcal{M}\alpha) \\ &= \alpha^\top\Re\{\mathcal{M}\}\alpha + \beta^\top\Re\{\mathcal{M}\}\beta - 2\alpha^\top\Im\{\mathcal{M}\}\beta. \end{aligned} \tag{5.11}$$

Let us inspect the last term more closely:

$$\begin{aligned} \Im\{\mathcal{M}\} &= -\Im\{\mathrm{diag}(x)^* C \mathrm{diag}(x)\} \\ &= -\Im\{\mathrm{diag}(x)^*\,\mathbf{1}\,\mathbf{1}^*\,\mathrm{diag}(x) + \sigma\mathrm{diag}(x)^* W \mathrm{diag}(x)\} \\ &= \Im\{xx^*\} - \sigma\Im\{\mathrm{diag}(x)^* W \mathrm{diag}(x)\}. \end{aligned}$$

At this step, we leverage the fact that, as per lemmas 5.1.7 and 5.1.8, if the noise level $\sigma$ is small enough, then $x$ is close to $z = \mathbf{1}$. We continue with the global phase convention $\mathbf{1}^*\,x \geq 0$ and the notation $x = \mathbf{1} + \Delta$ (5.5). Owing to $\alpha$ and $\beta$ having zero mean components, it follows that

$$\begin{aligned}\alpha^\top \Im\{xx^*\}\beta &= \alpha^\top \Im\{\mathbf{1}\,\mathbf{1}^* + \mathbf{1}\,\Delta^* + \Delta\,\mathbf{1}^* + \Delta\Delta^*\}\beta \\ &= \alpha^\top \Im\{\Delta\Delta^*\}\beta. \end{aligned} \tag{5.12}$$

Keeping in mind the intuition that $\Delta$ is an error term, we expect (5.12) to be small. We will make this precise later, and now turn our attention to the real part of $\mathcal{M}$, which turns out to be a Laplacian matrix. Indeed, this structure becomes apparent when the individual entries of the matrix are written out explicitly, starting from (5.10):

$$\Re\{\mathcal{M}\}_{ij} = \begin{cases} \Re\{(Cxx^*)_{ii} - C_{ii}\} = \sum_{\ell \neq i} \Re\{\bar{x}_i x_\ell C_{i\ell}\} & \text{if } i = j, \\ -\Re\{\bar{x}_i x_j C_{ij}\} & \text{if } i \neq j. \end{cases} \tag{5.13}$$

Recall that

$$L \colon \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n} \colon A \mapsto L_A = D_A - A \tag{5.14}$$

where $D_A = \mathrm{diag}(A\,\mathbf{1})$. Then,

$$\Re\{\mathcal{M}\} = L_{\Re\{\overline{C} \odot xx^*\}}, \tag{5.15}$$

where $\odot$ denotes the entry-wise (or Hadamard) product of matrices. Certainly, the symmetric matrix $\Re\{\mathcal{M}\}$ admits a zero eigenvalue associated to the all-ones vector, since for any $A$, $L_A\,\mathbf{1} = 0$. We define the *spectral gap* of the Laplacian $\Re\{\mathcal{M}\}$ as its smallest eigenvalue associated to an eigenvector orthogonal to the all-ones vector:

$$\lambda\left(\Re\{\mathcal{M}\}\right) = \min_{v\in\mathbb{R}^n, v\neq 0, \mathbf{1}^\top v = 0} \frac{v^\top \Re\{\mathcal{M}\} v}{v^\top v}. \tag{5.16}$$

Although this value could, in principle, be negative due to potential negative weights, the hope is that it will be positive and rather large (again, to be made precise later). Note that, if positive, $\lambda(L) = \lambda_2(L)$.

We now return to (5.11) and bound the expression:

$$\begin{aligned} u^*\mathcal{M}u &= \alpha^\top \Re\{\mathcal{M}\}\alpha + \beta^\top \Re\{\mathcal{M}\}\beta \\ &\qquad - 2\alpha^\top \Big( \Im\{\Delta\Delta^*\} - \sigma \Im\{\mathrm{diag}(x)^* W \mathrm{diag}(x)\} \Big) \beta \\ &\geq \left( \|\alpha\|_2^2 + \|\beta\|_2^2 \right) \lambda\left(\Re\{\mathcal{M}\}\right) \\ &\qquad - 2\|\alpha\|_2 \|\beta\|_2 \left( \|\Im\{\Delta\Delta^*\}\| + \sigma \left\|\Im\{\mathrm{diag}(x)^* W \mathrm{diag}(x)\}\right\| \right). \end{aligned}$$

For this inequality to lead to a guarantee of positivity of $u^*\mathcal{M}u$, it is certainly necessary to require $\lambda\left(\Re\{\mathcal{M}\}\right) > 0$. Using this and the fact[2] that $\|\Im\{A\}\| \leq \|A\|$, that the operator norm is invariant under unitary transformations and the simple inequality

$$0 \leq (\|\alpha\|_2 - \|\beta\|_2)^2 = \|\alpha\|_2^2 + \|\beta\|_2^2 - 2\|\alpha\|_2\|\beta\|_2,$$

it follows that

$$u^*\mathcal{M}u \geq 2\|\alpha\|_2\|\beta\|_2 \left( \lambda\left(\Re\{\mathcal{M}\}\right) - \|\Delta\|_2^2 - \sigma \|W\| \right).$$

Hence, a sufficient condition for $Q$ to be positive semidefinite with rank $n-1$ is:

$$\lambda\left(\Re\{\mathcal{M}\}\right) > \|\Delta\|_2^2 + \sigma \|W\|. \tag{5.17}$$

Let us now pause to reflect on condition (5.17) and to describe why it should hold. A bound on the operator norm of $W$ is readily available from **1**-discordance of $W$, and $\|\Delta\|_2^2$ is bounded by Lemma 5.1.7. Perhaps less obvious is why one would expect

[2] $\|A\|^2 = \max_{\substack{x\in\mathbb{C}^n \\ \|x\|=1}} \|Ax\|^2 \geq \max_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} \|Ax\|^2 = \max_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} \|\Re(A)x\|^2 + \|\Im(A)x\|^2 \geq \|\Im(A)\|^2.$

$\lambda\left(\Re\{\mathcal{M}\}\right)$ to be large. The intuition is that, for small enough noise, $x_i\bar{x}_j \approx z_i\bar{z}_j \approx C_{ij}$, so that $\Re\{\mathcal{M}\}$ is the Laplacian of a complete graph with large weights $\langle x_i\bar{x}_j, C_{ij}\rangle$. If this is the case, then it is known from graph theory that $\lambda\left(\Re\{\mathcal{M}\}\right)$ is large, because the underlying graph is well connected. The bound derived below on the spectral gap will, together with (5.17), reveal how large we may allow the noise level $\sigma$ to be.

### Bounding the spectral gap of $\Re\{\mathcal{M}\}$

This section is dedicated to lower bounding the spectral gap term (5.16). The right-hand side of (5.17) is on the order of $\sigma n^{1/2}$, so that showing that the spectral gap is at least on the order of $n - \mathcal{O}(\sigma n^{1/2})$ would yield an acceptable noise level of $\mathcal{O}(n^{1/2})$ for $\sigma$, as the numerical experiment suggests (Figure 5.1). Unfortunately, the bound we establish here is not as good, and thus constitutes the bottleneck in our analysis.[3]

**Lemma 5.1.11.** *If $W$ is $\mathbf{1}$-discordant, then*

$$\lambda\left(\Re\{\mathcal{M}\}\right) \geq n - \left[8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma n^{-1/4} + \left(6 + 40\sigma + 68\sigma^{3/2}\right)\log(n)\right]\sigma n^{3/4}.$$

*Proof.* Working from equation (5.15), we find that

$$\begin{aligned}
\Re\{\mathcal{M}\} &= L_{\Re\{\overline{C}\odot xx^*\}} \\
&= L_{\Re\{xx^*\}} + \sigma L_{\Re\left\{\overline{W}\odot(\mathbf{1}\,\mathbf{1}^* + xx^* - \mathbf{1}\,\mathbf{1}^*)\right\}} \\
&= L_{\Re\{xx^*\}} + \sigma L_{\Re\{W\}} + \sigma L_{\Re\left\{\overline{W}\odot(xx^* - \mathbf{1}\,\mathbf{1}^*)\right\}}.
\end{aligned}$$

[3] Even assuming (incorrectly) that $\Delta = 0$, so that $x = z = \mathbf{1}$, we would only get a spectral gap of $n - \mathcal{O}(\sigma n^{3/4})$ (because of the bound on $\|W\,\mathbf{1}\,\|_\infty$), yielding a final acceptable rate of $\sigma = \mathcal{O}(n^{1/4})$, which still falls short of the target rate $\mathcal{O}(n^{1/2})$ (all up to log factors).

Factor in the fact that for any $n \times n$ matrix $A$,

$$\begin{aligned} \|L_A\| &= \|\mathrm{diag}(A\,\mathbf{1}) - A\| \leq \|A\,\mathbf{1}\,\|_\infty + \|A\|\,, \text{ and} \\ \|A\| &\leq \|A\|_{\mathrm{F}} \leq n\,\|A\|_{\mathrm{e},\infty} \end{aligned}$$

to further obtain:

$$\begin{aligned} \lambda\left(\Re\{\mathcal{M}\}\right) &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \sigma\left(\left\|L_{\Re\{W\}}\right\| + \left\|L_{\Re\left\{\overline{W}\odot(xx^* - \mathbf{1}\,\mathbf{1}^*)\right\}}\right\|\right) \\ &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \sigma\left(\|W\,\mathbf{1}\,\|_\infty + \|W\|\right) \\ &\quad - \sigma\left(\left\|\left(\overline{W}\odot(xx^* - \mathbf{1}\,\mathbf{1}^*)\right)\mathbf{1}\right\|_\infty + n\left\|\overline{W}\odot(xx^* - \mathbf{1}\,\mathbf{1}^*)\right\|_{\mathrm{e},\infty}\right) \\ &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \sigma\left(\|W\,\mathbf{1}\,\|_\infty + \|W\| + 2n\,\|W\|_{\mathrm{e},\infty}\,\|xx^* - \mathbf{1}\,\mathbf{1}^*\|_{\mathrm{e},\infty}\right). \end{aligned}$$

We now rely on Lemma 5.1.8 to bound $\|xx^* - \mathbf{1}\,\mathbf{1}^*\|_{\mathrm{e},\infty}$. For all $1 \leq i, j \leq n$,

$$|x_i\overline{x_j} - 1| = |x_i - x_j| = |(x_i - 1) - (x_j - 1)| \leq |\Delta_i| + |\Delta_j| \leq 2\|\Delta\|_\infty.$$

Thus,

$$\|xx^* - \mathbf{1}\,\mathbf{1}^*\|_{\mathrm{e},\infty} \leq 2\|\Delta\|_\infty \leq 4\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}. \tag{5.18}$$

Combining the last equations with **1**-discordance of $W$ and the fact that for $n \geq 2$ we have $4 \leq 6\log(n)$, gives:

$$\begin{aligned} \lambda\left(\Re\{\mathcal{M}\}\right) &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \sigma\left(n^{3/4} + 3n^{1/2} + 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{3/4}\log(n)\right) \\ &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \left(4 + 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma\log(n)\right)\sigma n^{3/4} \\ &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \left(6 + 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma\right)\sigma n^{3/4}\log(n) \\ &\geq \lambda\left(L_{\Re\{xx^*\}}\right) - \left(6 + 40\sigma + 68\sigma^{3/2}\right)\sigma n^{3/4}\log(n). \end{aligned} \tag{5.19}$$

It remains to bound the dominating part of the spectral gap. To this effect, we use the fact that $\Re\left\{x_i\overline{x_j}\right\}$ is nonnegative when the noise level is low enough, so that (restricting ourselves to that regime) $\lambda\left(L_{\Re\{xx^*\}}\right)$ is the spectral gap of a complete graph with all weights strictly positive. That spectral gap must be at least as large as the smallest weight multiplied by the spectral gap of the complete graph with unit weights, namely, $\lambda\left(L_{\mathbf{1}\mathbf{1}^\top}\right) = n$. Formally, for all $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$ and $\mathbf{1}^\top v = 0$, by properties of Laplacian matrices it holds that

$$\begin{aligned} v^\top L_{\Re\{xx^*\}} v &= \sum_{i<j} \Re\{x_i\overline{x_j}\}(v_i - v_j)^2 \\ &\geq \min_{i,j} \Re\{x_i\overline{x_j}\} \sum_{i<j}(v_i - v_j)^2 \\ &= \min_{i,j} \Re\{x_i\overline{x_j}\}\ v^\top L_{\mathbf{1}\mathbf{1}^\top} v \\ &= n \min_{i,j} \Re\{x_i\overline{x_j}\}. \end{aligned}$$

Let us investigate the smallest weight. Recall that $|x_i - x_j| \leq 2\|\Delta\|_\infty$, $\Re\{x_i\overline{x_j}\} = 1 - \frac{1}{2}|x_i - x_j|^2$, and that $\|\Delta\|_\infty \leq 2\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)\sigma n^{-1/4}$ to get

$$\Re\{x_i\overline{x_j}\} \geq 1 - 2\|\Delta\|_\infty^2 \geq 1 - 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma^2 n^{-1/2}.$$

Hence,

$$\lambda\left(L_{\Re\{xx^*\}}\right) \geq n - 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma^2 n^{1/2}. \tag{5.20}$$

Merging the bounds (5.19) and (5.20) gives:

$$\begin{aligned} \lambda\left(\Re\{\mathcal{M}\}\right) &\geq n - 8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma^2 n^{1/2} - \left(6 + 40\sigma + 68\sigma^{3/2}\right)\sigma n^{3/4}\log(n) \\ &\geq n - \left[8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma n^{-1/4} + \left(6 + 40\sigma + 68\sigma^{3/2}\right)\log(n)\right]\sigma n^{3/4}. \end{aligned}$$

This establishes the lemma. □

### Concluding the proof

Recall that (5.2) is tight, in particular, if (5.17) holds:

$$\lambda\left(\Re\{\mathcal{M}\}\right) > \|\Delta\|_2^2 + \sigma\,\|W\|\,. \tag{5.21}$$

Still assuming **1**-discordance of $W$ (as it gives $\|W\| \leq 3n^{1/2}$) and collecting results from lemmas 5.1.7 and 5.1.11, we find that this condition is fulfilled in particular if

$$n - \left[8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma n^{-1/4} + \left(6 + 40\sigma + 68\sigma^{3/2}\right)\log(n)\right]\sigma n^{3/4} > 11\sigma n^{1/2}.$$

Reorder terms and divide through by $n^{3/4}$ to get the equivalent condition:

$$n^{1/4} > \left[\left(8\left(5 + 6\sqrt{2}\,\sigma^{1/2}\right)^2 \sigma + 11\right) n^{-1/4} + \left(6 + 40\sigma + 68\sigma^{3/2}\right)\log(n)\right]\sigma.$$

This can be written in the form (with some constant $c_1 > 0$):

$$n^{1/4} > f_1(\sigma) n^{-1/4} + f_2(\sigma)\log(n) + c_1\sigma^3 n^{-1/4},$$

where $f_1$ and $f_2$ are polynomials with nonnegative coefficients, lowest power $\sigma$ and highest power $\sigma^{5/2}$. Thus, their sum is upper-bounded by $c_2(\sigma + \sigma^{5/2})$, for some constant $c_2 > 0$. Hence, there exists a constant $c_3 > 0$ such that, if

$$c_1\sigma^3 n^{-1/4} < 0.99 n^{1/4}, \tag{5.22}$$

then

$$n^{1/4} > c_3(\sigma + \sigma^{5/2})\log(n)$$

is a sufficient condition. It is then easy to see that there exists a universal constant $K > 0$ such that

$$\sigma + \sigma^{2/5} \leq K \frac{n^{1/10}}{\log(n)^{2/5}}$$

is a sufficient condition for tightness of (5.2). This concludes the proof of Lemma 5.1.5.

# Chapter 6

# Multireference alignment and Cryo-Electron Microscopy

## 6.1 Alignment of signals

In this chapter we consider other instances of synchronization-type problems. This section, mostly based on [32], starts with an important and illustrative problem in signal processing, the multireference alignment problem (see Section 1.2.7): it consists of estimating an unknown signal $u$ from multiple noisy cyclically-shifted copies. More precisely, we are interested in the problem of estimating an unknown template vector $u \in \mathbb{R}^L$ from $n$ measurements $y_1, \dots, y_n$ of the form:

$$y_i = R_{l_i} u + \xi_i \in \mathbb{R}^L, \tag{6.1}$$

where $\xi_i \sim \mathcal{N}(0, \sigma^2 I_L)$ is gaussian white noise with variance $\sigma^2$, and $R_l$ denotes the index cyclic shift operator $(u_1, \dots, u_L) \mapsto (u_{1-l}, \dots, u_{L-l})$, represented as an $L \times L$ circulant permutation matrix.

The difficulty of this problem resides in the fact that both the template $u$ and the shifts $l_1, \dots, l_n \in \mathbb{Z}_L$ are unknown (moreover, no model is presumed a-priori for their distribution). If the shifts were known, one could easily estimate $u$ by unshifting the observations and averaging them. Motivated by this fact we will focus on the problem of estimating the shifts $l_1, \ldots, l_n$ (up to an undecidable global shift).

This problem has a vast list of applications. Alignment is directly used in structural biology [94, 218]; radar [241, 182]; crystalline simulations [215]; and image registration in a number of important contexts, such as in geology, medicine, and paleontology [97, 106].

Perhaps the most naïve approach to estimate the shifts in (6.1) would be to fix one of the observations, say $y_i$, as a reference template and align every other $y_j$ with it by the shift $\rho_{ij}$ minimizing their distance

$$\rho_{ij} = \operatorname{argmin}_{l \in \mathbb{Z}_L} \|R_l y_j - y_i\|_2. \tag{6.2}$$

While this solution works well at a high signal-to-noise ratio (SNR), it performs poorly at low SNR. Indeed, following the discussion in Section 1.3.1, we expect the recovered signal to be representative of $y_i$ than of $u$ (see Figure 1.3.1).

A more democratic approach would be to calculate pairwise relative shift estimates $\rho_{ij}$ for each pair and then attempt to recover the shifts $\{l_i\}$ by angular synchronization (recall Section 1.2.1), i.e., by attempting to minimize

$$\min_{l_1,\ldots,l_n \in \mathbb{Z}_L} \sum_{i,j=1}^{n} \left| e\left(\frac{l_i - l_j}{L}\right) - e\left(\frac{\rho_{ij}}{L}\right) \right|^2, \tag{6.3}$$

where $e(x) = e^{2\pi \imath x}$ denotes the classical Fourier basis function. Note how this formulation is of the form of (1.1) for $\mathcal{G} \cong \mathbb{Z}_L$ and

$$f_{ij}\left(l_i - l_j\right) = \left| e\left(\frac{l_i - l_j}{L}\right) - e\left(\frac{\rho_{ij}}{L}\right) \right|^2. \tag{6.4}$$

Moreover, note that (6.3) has the same optimizers as

$$\max_{l_1,\ldots,l_n\in\mathbb{Z}_L}\sum_{i,j=1}^{n} e\left(\frac{l_i-l_j}{L}\right)e\left(-\frac{\rho_{ij}}{L}\right), \tag{6.5}$$

which is linear in terms of the unidimensional representation of the cyclic group in $\mathbb{C}$. This suggests the use of the semidefinite program (1.9). Indeed, that approach coincides with the SDP-based relaxation for angular synchronization (see Section 1.2.1) described in Section 5.1. An alternative would be the spectral method described in Section 2.1.

The main shortcoming of this type of approach is that the only information it uses from the observations $\{y_i\}$ is the best relative shifts $\rho_{ij}$. This means that the performance of a given choice of $\{l_i\}$ can only be evaluated by comparing $l_i - l_j$ with $\rho_{ij}$ (in shift space) across pairs $(i,j)$. This does not take into account the cost associated with other possible relative shifts of $y_i$ and $y_j$. On the other hand, for a candidate solution $\{l_i\}$, relating $R_{-l_i}y_i$ and $R_{-l_i}y_j$ (in signal space) would take into account information about all possible shifts instead of just the best one (6.2). The quasi maximum likelihood estimator (Section 6.1.1) attempts to do exactly that by solving the minimization problem:

$$\min_{l_1,\ldots,l_n\in\mathbb{Z}_L}\sum_{i,j=1}^{n}\left\|R_{-l_i}y_i - R_{-l_j}y_j\right\|^2. \tag{6.6}$$

While this objective function can still be rewritten to be of the form of (1.1), by taking

$$f_{ij}\left(l_i-l_j\right) = \left\|y_i - R_{l_i-l_j}y_j\right\|^2, \tag{6.7}$$

it is no longer necessarily linear in terms of a unidimensional representation. Fortunately, as we will describe below, it is linear with respect to a $L$-dimensional representation, suggesting the use of a semidefinite program of the form of (1.9) with a matrix variable of size $nL \times nL$ rather than $n \times n$.

**Remark 6.1.1.** *This illustrates a crucial point regarding the semidefinite programming-based approach described in Section 1.3.3. While smaller dimension representations render smaller (and thus potentially more efficient) semidefinite programs, representations of larger dimension have the potential of allowing the use of more descriptive objective functions (for example,* (6.7) *vs* (6.4)*).*

### 6.1.1 Quasi Maximum Likelihood Estimator

The log likelihood function for model (6.1) is given by

$$\mathcal{L}(u, l_1, \ldots, l_n \,|\, y_1, \ldots, y_n) = \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma}\sum_{i\in[n]} \|R_{-l_i} y_i - u\|^2. \tag{6.8}$$

Maximizing $\mathcal{L}$ is equivalent to minimizing the sum of squared residuals $\sum_{i=1}^{n} \|R_{-l_i} y_i - u\|^2$. Fixing the $l_i$'s, the minimal value of $\mathcal{L}$ occurs at the average $u = \frac{1}{n}\sum_{i=1}^{n} R_{-l_i} y_i$. Making the tame assumption that $\|u\|^2$ is estimable (indeed the norm is shift-invariant) and thus fixed in (6.8), maximizing (6.8) is equivalent to maximizing the sum of the inner products $\left\langle R_{-l_i} y_i,\, R_{-l_j} y_j \right\rangle$ across all pairs $(i, j)$. Thus we consider the estimator

$$\operatorname{argmax}_{l_1,\ldots,l_n \in \mathbb{Z}_L} \sum_{i,j\in[n]} \langle R_{-l_i} y_i, R_{-l_j} y_j \rangle, \tag{6.9}$$

which coincides with the optimizers of (6.6). We refer to this estimator as the quasi-MLE.

It is not surprising that solving (6.9) is $\mathsf{NP}$-hard in general (the search space for this optimization problem has exponential size and is nonconvex). However, one could hope to approximate (6.9) up to some constant (in the spirit of the results established in Section 2.3). Unfortunately, the existence of an algorithm with such guarantees seems unlikely, as suggested by the theorem below.

**Theorem 6.1.2.** *Assuming no model on the observations $\{y_i\}$, it is* NP*-hard (under randomized reductions) to find a set of shifts approximating (6.9) within* $16/17+\varepsilon$ *of its optimum value. Furthermore, if the Unique-Games conjecture (Conjecture 1.2.2) is true, it is* NP*-hard to approximate (6.9) within any constant factor.*

We refer the reader to [32] for a proof of this theorem.

### 6.1.2 The semidefinite relaxation

Let us identify $R_l$ with the $L \times L$ permutation matrix that cyclicly permutes the entries fo a vector by $l_i$ coordinates:

$$R_l \begin{bmatrix} u_1 \\ \vdots \\ u_L \end{bmatrix} = \begin{bmatrix} u_{1-l} \\ \vdots \\ u_{L-l} \end{bmatrix}.$$

This corresponds to an $L$-dimensional representation of the cyclic group. Then, (6.9) can be rewritten:

$$\begin{aligned} \sum_{i,j\in[n]} \langle R_{-l_i} y_i, R_{-l_j} y_j \rangle &= \sum_{i,j\in[n]} \left(R_{-l_i} y_i\right)^T R_{-l_j} y_j \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[\left(R_{-l_i} y_i\right)^T R_{-l_j} y_j\right] \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[y_i^T R_{-l_i}^T R_{-l_j} y_j\right] \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[\left(y_i y_j^T\right)^T R_{l_i} R_{l_j}^T\right], \end{aligned}$$

confirming that (6.9) is indeed linear in this representation. Following the roadmap described in Section 1.3.3 we take

$$X = \begin{bmatrix} R_{l_1} \\ R_{l_2} \\ \vdots \\ R_{l_n} \end{bmatrix} \begin{bmatrix} R_{l_1}^T & R_{l_2}^T & \cdots & R_{l_n}^T \end{bmatrix} \in \mathbb{R}^{nL \times nL}. \tag{6.10}$$

We can then rewrite (6.9) as

$$\begin{aligned} \max \quad & \operatorname{Tr}(CX) \\ \text{s. t.} \quad & X_{ii} = I_{L\times L} \\ & X_{ij} \text{ is a circulant permutation matrix} \\ & X \succeq 0 \\ & \operatorname{rank}(X) \leq L, \end{aligned} \tag{6.11}$$

where $C$ is the rank 1 matrix given by

$$C = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} y_1^T & y_2^T & \cdots & y_n^T \end{bmatrix} \in \mathbb{R}^{nL \times nL}, \tag{6.12}$$

with blocks $C_{ij} = y_i y_j^T$.

The constraints $X_{ii} = I_{L\times L}$ and $\operatorname{rank}(X) \leq L$ imply that $\operatorname{rank}(X) = L$ and $X_{ij} \in O(L)$. Since the only doubly stochastic matrices in $O(L)$ are permutations, (6.11) can be rewritten as

$$\begin{aligned} \max \quad & \mathrm{Tr}(CX) \\ \text{s. t.} \quad & X_{ii} = I_{L\times L} \\ & X_{ij}\,\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & X \geq 0 \\ & X \succeq 0 \\ & \mathrm{rank}(X) \leq L. \end{aligned} \tag{6.13}$$

Removing the nonconvex rank constraint yields a semidefinite program, corresponding to (1.9),

$$\begin{aligned} \max \quad & \mathrm{Tr}(CX) \\ \text{s. t.} \quad & X_{ii} = I_{L\times L} \\ & X_{ij}\,\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & X \geq 0 \\ & X \succeq 0. \end{aligned} \tag{6.14}$$

**Another perspective**

The discrete optimization problem (6.9) may also be formulated using indicator variables as an integer programming problem

$$\mathrm{argmax}_{\{z_{ik}\}} \sum_{i,j=1}^{n} \sum_{k,l\in\mathbb{Z}_L} z_{ik} z_{jl} \langle R_{-k} y_i, R_{-l} y_j \rangle, \tag{6.15}$$

where $z_{ik} \in \{0,1\}$ and, for each $i$, $z_{ik} = 1$ for exactly one index $k$, corresponding to indicator variables $z_{ik} = 1_{\{l_i \equiv k\}}$. These requirements can be described with quadratic constraints (up to global sign, which cannot be fixed by quadratic constraints)

$$\begin{aligned} \sum_{k,l\in\mathbb{Z}_L} z_{ik}z_{jl} &= 1, \quad i,j\in[n] \\ z_{ik}z_{il} &= 0, \quad i\in[n],\ k\neq l\in\mathbb{Z}_L \\ z_{ik}z_{jl} &\geq 0, \quad i,j\in[n],\ k,l\in\mathbb{Z}_L. \end{aligned} \tag{6.16}$$

Since both the objective function (6.15) and the constraints (6.16) depend only on products of the form $z_{ik}z_{jl}$, the problem can be written linearly in terms of the Gram matrix $Z\in\mathbb{R}^{nL\times nL}$ with entries $Z_{ik;jl}=z_{ik}z_{jl}$, and can (after removing a rank constraint) be written in terms of a semidefinite program. In fact, this was the approach taken originally in [32], motivated by the semidefinite programming relaxation for the Unique-Games problem in [75]. After averaging the $L$ different solutions that this SDP has (corresponding to the global shift ambiguity), the two approaches are effectively equivalent. We refer the reader to [32] for a description of the problem using the alternative view.

### 6.1.3 Analysis of the SDP

If the SNR is high enough so that, between every pair of shifted measurements, the cross-correlation is maximal at the true offset, then it is not hard to show [32] that the relaxation (6.14) is exact. Unfortunately, this idealized scenario is unrealistic in practice. In fact, for realistic noise levels one does not expect even the MLE to coincide with the original shifts (if $\sigma > 0$, given enough signals, with high probability, there will be a noisy copy of $u$ that correlates less with the original vector $u$ than with a shift of it). Fortunately, similarly to the SDP relaxation investigated in Section 5.1, this SDP appears to be tight remarkably often (see Figure 7.1). However, an analysis of the type that was described in Section 5.1 appears to be more difficult in this setting and the understanding of this behavior remains an open problem — we will discuss this problem further in Section 7.2.1.

While no satisfactory tightness guarantee exists, we are able to give some stability guarantees to this relaxation. For simplicity, we will assume, without loss of generality, that all the ground truth shifts correspond to the zero shift. Hence $y_i = u + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma^2 I_L)$ i.i.d.. The ground truth solution for the SDP will then be a $X^\natural \in \mathbb{R}^{nL \times nL}$ with blocks $X^\natural_{ij} = I_{L\times L}$. We are interested in understanding whether the optimal solution to (6.14) resembles $X^\natural$. We will describe a guarantees characterized in terms of the gap between the correct offset and incorrect ones

$$\Delta = \|u\|^2 - \max_{l\neq 0}\langle u, R_l u\rangle, \tag{6.17}$$

and the noise level.

For any $X \in \mathbb{R}^{nL\times nL}$ lying in the SDP feasibility region, we can characterize the distance of $X$ from the desired solution $X^\natural$ by the differences

$$D_{ij} = \frac{1}{L}\sum_{k\neq l\in[L]} (X_{i,j})_{kl} = 1 - \frac{1}{L}\sum_{k\in[L]}(X_{i,j})_{kk} \in [0,1].$$

$D_{ij}$ is a measure of how much the solution of the SDP weighs shift preferences other than the ground truth. Note that $D_{ij}$ is always non-negative and, moreover, $D_{ij} = 0$, for all $i, j$, corresponds to $X = X^\natural$.

**Theorem 6.1.3.** *With probability* $1 - e^{-n+o(n)}$, *the solution to the SDP* (6.14) *satisfies*

$$\sum_{i,j} D_{ij} \leq \sigma\frac{(\|u\| + \sigma\sqrt{L})\cdot 12\log eL}{\Delta}\cdot n^2.$$

Theorem 6.1.3 indicates that at a sufficiently high SNR, (6.14) will produce a matrix $X$, of which each $L \times L$ block has most weight concentrated on its main diagonal. We refer the reader to [32] for a proof of Theorem 6.1.3. One of the main difficulties in the analysis of this problem is that noise model is on the vertices, which translates into dependent noise on the edges, rendering it qualitatively different from previous work analyzing similar SDPs in the context of the Unique-Games problem [23, 140, 139].

If the relaxation is not tight, a rounding step is needed: a procedure to produce a solution of the original problem (6.9) given a solution of (6.14). While it is not the main purpose of this section to discuss rounding procedures, we note that, if the optimal solution $X$ indeed places more weight on the correct shift (as predicted by Theorem 6.1.3), then even a naïve rounding scheme would likely interpret it as the correct shift being the optimal one. In fact, the numerical simulations in Figure 6.1.4 seem to suggest that this is case.

### 6.1.4 Numerical Results

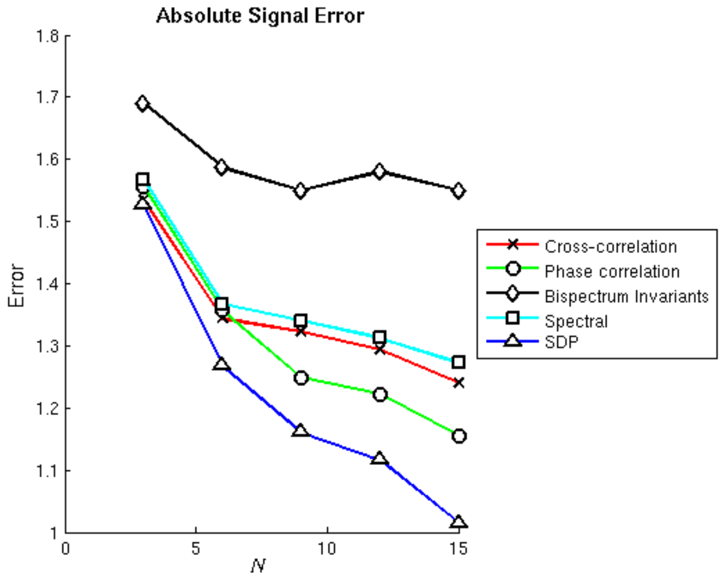

Figure 6.1: Averages of errors of several alignment methods across 500 draws of signal and noise. The parameters are $\sigma = 1$ and $L = 5$.

We implemented several baseline methods for multireference alignment, and plotted their average error performance across 500 draws of signal and noise in Figure 6.1.4. For each iteration, we chose a signal $u$ randomly from the distribution $\mathcal{N}(0, I_L)$, as well as $n$ i.i.d. noise vectors $\xi_i \sim \mathcal{N}(0, \sigma^2 I_L)$, and compared the performance of the SDP (6.14), together with a simple rounding procedure, with several other methods. Cross-correlation refers to the method of picking one of the observations as a reference to align the others, and Phase correlation to a normalized version of it [129]. The spectral method refers to a relaxation of the MLE to an eigenvector problem, and the bispectrum method is an invariant based method, we refer to [32] for a detailed description of these methods. These simulations suggest that the SDP (6.14) performs better than other benchmark techniques.

### 6.1.5 Computational complexity of the Semidefinite Program

Semidefinite programs can be solved, up to arbitrary accuracy in polynomial time [227]. However, their computational complexity still depend heavily on the number of variables and constraints. The SDP (6.14) has a matrix variable of size $nL \times nL$ and order of $n^2L^2$ constraints (the positivity constraints), which could render a naïve implementation of (6.14) impractical for large problem instances.

Fortunately, the SDP (6.14) has structure that can be exploited. Each $L \times L$ block $X_{ij}$ of $X$ is a circulant matrix. This means that they are simultaneously diagonalizable by the $L \times L$ discrete Fourier transform (DFT) matrix. This observation allows us us to find a unitary matrix $U$ (of size $nL \times nL$) so that $UXU^T$ is block-diagonal (with $L$ non-zero $n \times n$ diagonal blocks). After this transformation, the positive semidefinite constraint greatly simplifies, as it becomes decoupled on each block. We refer the reader to [32] for a more detailed discussion on this point.

**Remark 6.1.4.** *We note the general applicability of the idea briefly described above: If the group $\mathcal{G}$ is commutative, than we expect a representation of it to be simultaneously diagonalizable, meaning that the above speedup can be applied (replacing the DFT by another matrix) to the SDP* (1.9) *arising from synchronization on any compact abelian group. Also, if the compact $\mathcal{G}$ is not abelian, we still expect to be able to block* *diagonalize its representation, where the sizes of the blocks correspond to dimensions of the irreducible representations of $\mathcal{G}$. In fact, we believe this fact will play a crucial role in applying this approach to the problem of orientation estimation in Cryo-Electron Microscopy — which is a form of Synchronization over $SO(3)$ — and is discussed briefly in Section 6.3 but mostly deferred to a future publication.*

As another attempt to simplify the SDP, one might consider removing positivity constraints (since there are essentially $n^2L^2$ of them). Interestingly, this weaker SDP can be solved explicitly and is equivalent to the pairwise alignment method called phase correlation [129]. This method does not take into account information between all pairs of measurements, it essentially proceeds by selecting one of the measurements as a reference and align everything to it, not based on Cross-correlation but on a normalized version of it. This suggests that the full complexity of the SDP (6.14) is needed to obtain a good approximation to (6.6).

## 6.2 Clustering and graph multisection

In this section we briefly investigate an interesting connection between the signal alignment problem treated in Section 6.1 and the problem of clustering a graph or a point cloud in multiple ($L$) clusters. Countless versions of this problem exist. A particularly popular version of clustering a point cloud is $k$-means clustering [154, 149]. On the graph side, there are extensions of the Cheeger's inequality (Theorem 2.1.1) to provide guarantees of spectral clustering in this setting [144, 112]. There are also adaptations of some of the ideas presented in Section 3.2 to understand recovery in the Stochastic Block Model with multiple clusters [8, 4, 122].

For illustrative purposes we will consider the the min-sum k-clustering problem [239]. It will be clear that the ideas that follow can be adapted to other settings. Given $n$ points, $p_1, \ldots, p_n$, with pairwise distances $d_{ij}$ the min-sum k-clustering problem [239] consists in partioning the $n$ points in $L$ clusters $\mathcal{C}_1, \ldots, \mathcal{C}_L$ as to minimize

$$\sum_{k=1}^{L} \sum_{i,j \in \mathcal{C}_k} d_{ij}. \tag{6.18}$$

Note that if the points are in an euclidean space, $k$-means clustering can be described as minimizing a similar objective:

$$\sum_{k=1}^{L} \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} d_{ij},$$

where $d_{ij}$ is the squared of the $\ell_2$ distance between points $i$ and $j$. This objective is different due to the normalization factor depending on the cluster size.

### 6.2.1 The multireference alignment SDP for clustering

The idea to formulate (6.18) through the approach described in Section 6.1 (for the signal processing problem) is very simple: one can think of each point $p_i$ as a signal $y_i$ in $\mathbb{R}^L$ and think of a shift label as a cluster membership, the cost associated to the pair $i, j$ should then be $d_{ij}$ if the two signals are given the same shift and zero otherwise. This can be achieved by taking $C \in \mathbb{R}^{nL \times nL}$ to have $L \times L$ blocks $C_{ij} = -\frac{1}{L} d_{ij} I_{L \times L}$. In fact, by setting the cluster membership of $i$ to be $l_i$, it is easy to see that

$$\sum_{k=1}^{L} \sum_{i,j \in \mathcal{C}_k} d_{ij} = - \sum_{i,j \in [n]} \operatorname{Tr}\left[ C_{ij}^T R_{l_i} R_{l_j}^T \right].$$

This means that minimizing (6.18) is equivalent to (6.13) for this particular choice of $C$.

It is clear that this SDP has many optimal solutions. Given an optimal selection of cluster labelings, any permutation of these labels will yield a solution with the same objective. For that reason we can adapt the SDP to consider the average of such solutions. This is achieved by restricting each block $X_{ij}$ to be a linear combination of $I_{L\times L}$ and $\mathbf{1}\,\mathbf{1}^T$ (meaning that it is constant both on the diagonal and on the off-diagonal). Adding that constraint yields the following SDP.

$$\begin{array}{ll} \max & \operatorname{Tr}(CX) \\ \text{s. t.} & X_{ii} = I_{L\times L} \\ & X_{ij}\,\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & \left(X_{ij}\right)_{kk} = \left(X_{ij}\right)_{11} \\ & \left(X_{ij}\right)_{kl} = \left(X_{ij}\right)_{12}\,,\ \forall_{k\neq l} \\ & X \geq 0 \\ & X \succeq 0, \end{array} \tag{6.19}$$

Since the constraints in (6.19) imply

$$\left(X_{ij}\right)_{11} + (L-1)\left(X_{ij}\right)_{12} = 1,$$

(6.19) can be described completely in terms of the variables $\left(X_{ij}\right)_{11}$. For that reason we consider the matrix $Y \in \mathbb{R}^{n\times n}$ with entries $Y_{ij} = \left(X_{ij}\right)_{11}$. We can then rewrite (6.19) as

$$\begin{array}{ll} \max & \operatorname{Tr}\left(\tilde{C}Y\right) \\ \text{s. t.} & Y_{ii} = 1 \\ & Y \geq 0 \\ & Y^{(L)} \succeq 0, \end{array} \tag{6.20}$$

where $\tilde{C}_{ij} = -d_{ij}$ and $Y^{(L)}$ is the $nL \times nL$ matrix whose $n \times n$ diagonal blocks are equal to $Y$ and whose $n \times n$ non-diagonal blocks are equal to $\frac{11^T - Y}{L-1}$. For example,

$$Y^{(2)} = \begin{bmatrix} Y & 11^T - Y \\ 11^T - Y & Y \end{bmatrix} \quad \text{and} \quad Y^{(3)} = \begin{bmatrix} Y & \frac{11^T - X}{2} & \frac{11^T - Y}{2} \\ \frac{11^T - Y}{2} & Y & \frac{11^T - Y}{2} \\ \frac{11^T - Y}{2} & \frac{11^T - Y}{2} & Y \end{bmatrix}.$$

The following lemma gives a simpler characterization of the intriguing $Y^{(L)} \succeq 0$ constraint.

**Lemma 6.2.1.** *Let $Y$ be a symmetric matrix and $L \geq 2$ an integer. $Y^{(L)} \succeq 0$ if and only if $Y \succeq \frac{1}{L} \mathbf{1}\,\mathbf{1}^T$.*

Before proving Lemma 6.2.1, note that it implies we can succintly rewrite (6.20) as

$$\begin{aligned} \max \quad & \operatorname{Tr}\left(\tilde{C}Y\right) \\ \text{s. t.} \quad & Y_{ii} = 1 \\ & Y \geq 0 \\ & Y \succeq \tfrac{1}{L} \mathbf{1}\,\mathbf{1}^T. \end{aligned} \tag{6.21}$$

A simple change of variables $Z = \frac{L}{L-1} Y - \frac{1}{L-1} \mathbf{1}\,\mathbf{1}^T$, allows one to rewrite (6.21) as

$$\begin{aligned} \max \quad & \operatorname{Tr}\left(C'Z\right) - c' \\ \text{s. t.} \quad & Z_{ii} = 1 \\ & Z_{ij} \geq -\tfrac{1}{L-1} \\ & Z \succeq 0. \end{aligned} \tag{6.22}$$

for appropriate matrix $C'$ and constant $c'$. Remarkably, (6.22) coincides with the classical semidefinite relaxation for the `Max-k-Cut` problem [108].

*Proof.* [of Lemma 6.2.1]

Since, in this proof, we will be using $\mathbf{1}$ to refer to the all-ones vector in two different dimensions we will include a subscript denoting the dimension of the all-ones vector.

The matrix $Y^{(L)}$ is block circulant and so it can be block-diagonalizable by a block DFT matrix, $F_{L\times L}\otimes I_{n\times n}$, where $F_{L\times L}$ is the $L\times L$ (normalized) DFT matrix and $\otimes$ is the Kronecker product. In other words,

$$\left(F_{L\times L}\otimes I_{n\times n}\right)Y^{(L)}\left(F_{L\times L}\otimes I_{n\times n}\right)^T$$

is block diagonal. Furthermore, note that

$$Y^{(L)}=\left(\mathbf{1}_L\,\mathbf{1}_L^T\otimes\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-\mathbf{1}}\right)-\left(I_{L\times L}\otimes\left[Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\right]\right).$$

Also, It is easy to check that

$$\left(F_{L\times L}\otimes I_{n\times n}\right)\left(I_{L\times L}\otimes\left[Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\right]\right)\left(F_{L\times L}\otimes I_{n\times n}\right)^T=I_{L\times L}\otimes\left[Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\right],$$

and

$$\left(F_{L\times L}\otimes I_{n\times n}\right)\left(\mathbf{1}_L\,\mathbf{1}_L^T\otimes\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\right)\left(F_{L\times L}\otimes I_{n\times n}\right)^T=L\left(e_1e_1^T\otimes\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\right),$$

This means that $\left(F_{L\times L}\otimes I_{n\times n}\right)Y^{(L)}\left(F_{L\times L}\otimes I_{n\times n}\right)^T$ is a block diagonal matrix with the first block equal to $\mathcal{A}$ and all other diagonal blocks equal to $\mathcal{B}$ where $\mathcal{A}$ and $\mathcal{B}$ are given by

$$\mathcal{A}=Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}+L\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}=\mathbf{1}_n\,\mathbf{1}_n^T\ \text{ and }\mathcal{B}=Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}.$$

Thus, the condition $Y^{(L)}\succeq 0$ is equivalent to $Y-\frac{\mathbf{1}_n\,\mathbf{1}_n^T-Y}{L-1}\succeq 0$ which can be rewritten as,

$$Y-\frac{1}{L}\mathbf{1}_n\,\mathbf{1}_n^T\succeq 0.$$

□

We remark that exact recovery guarantees can be shown for similar SDPs for both $k$-means clustering [25] and recovery in the Stochastic Block Model with multiple clusters [8]. While we are not going to discuss these results here, we comment that such guarantees are obtained with arguments similar to those described in Chapter 3.

## 6.3 The viewing direction problem in Cryo-EM

We finish this chapter with a particularly challenging synchronization-type problem, the viewing direction estimation problem in Cryo-Electron Microscopy (Cryo-EM) [204], mentioned in Section 1.2.2. This section is meant as a brief preview for an upcoming publication of the author, Yutong Chen, and Amit Singer.

Cryo-EM is a technique used to determine the three-dimensional structured of biological macromolecules. The molecules are rapidly frozen in a thin layer of ice and imaged with an electron microscope, which gives two-dimensional projections (see Figure 6.2). One of the main difficulties with this imaging process is that these molecules are imaged at different unknown orientations in the sheet of ice and each molecule can only be imaged once (due to the destructive nature of the imaging process). More precisely, each measurement consists of a tomographic projection of a rotated (by an unknown rotation $R_i$) copy the molecule (as illustrated in Figure 6.2). The task is then to reconstruct the molecule density from many such measurement (see Figure 6.3) for a idealized density and measurement dataset). The problem of recovering the molecule density knowing the rotations fits in the framework of classical tomography for which effective methods exist, for this reason we will, once again, focus on determining the unknown rotations.

An added difficulty is the fact that the image process is extremely noisy. In fact, it is already non-trivial to distinguish whether a molecule was actually imaged or if the image consists only of noise (see Figure 6.4 for a real example of a measurement

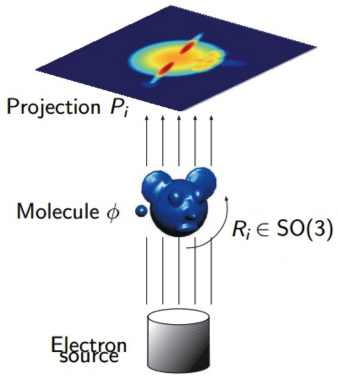

Figure 6.2: Illustration of the Cryo-EM imaging process: A molecule is imaged after being frozen at a random (unknown) rotation and a tomographic 2-dimensional projection is captured. Image courtesy of Amit Singer and Yoel Shkolnisky [204].

in this process). On the other hand, these datasets consist of many pictures which renders reconstruction possible.

The most direct way of describing this problem in the general framework is by considering the space of (say, bandlimited) molecule densities $\phi$ in $\mathbb{R}^3$ and $\mathcal{G} \cong SO(3)$ acting by rotating the molecule ($\phi(\cdot) \to \phi(g\cdot)$). In this case, this problem would be similar to multireference alignment in signal processing (Section 6.1) with the added difficulty that we do not measure a noisy copy of the molecule, but a noisy tomography projection.

Fortunately, the Fourier slice theorem, a central mathematical tool in the area of tomography, provides a way of comparing tomographic projections taken from different viewing directions. It states that the two-dimensional Fourier transform of a tomographic projection of a molecule density $\phi$ coincides with the restriction to a plane normal to the projection direction, a slice, of the three-dimensional Fourier transform of the density $\phi$. This means that we can formulate the problem as estimating the three-dimensional Fourier transforms of molecule density $\phi$ in $\mathbb{R}^3$, where the observation model now becomes: $SO(3)$ acts by rotating the molecule ($\phi(\cdot) \to \phi(g\cdot)$) and the measurement consists of a restriction to the equator slice.

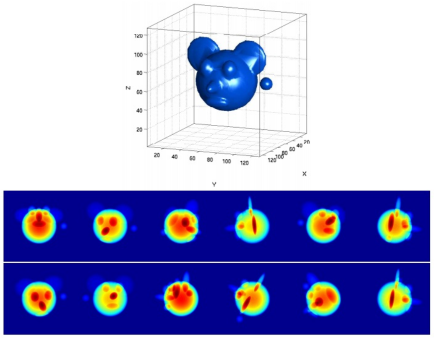

Figure 6.3: A illustration of a Cryo-EM dataset. Given tomographic projections of a molecule density taken at unknown rotations, we are interested in determining such rotations with the objective of reconstructing the molecule density. Image courtesy of Amit Singer and Yoel Shkolnisky [204].

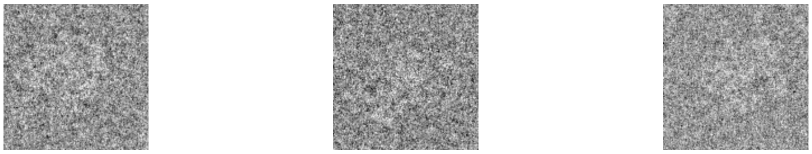

Figure 6.4: Sample images from the E. coli 50S ribosomal subunit, generously provided by Fred Sigworth at the Yale Medical School.

Any two two-dimensional slices have to intersect in a line (see Figure 6.5), meaning that, in the noiseless case, each pair of two-dimensional restriction must have a common line. Identifying these common lines and then, from them, estimating the rotations has been a strategy behind many successful recovery algorithms [226, 204, 207, 203].

Although having many extra difficulties, one can draw a parallel between this problem and the multireference alignment problem in signal processing (Section 6.1). A natural approach is to, for each pair of two-dimensional slices, estimate the common lines as the pair of lines that are the most similar. One can then attempt to find the rotations for each slice that are the most compatible with these pairs of common lines. In 2011, Singer and Shkolnisky [204] propose both a spectral method and a

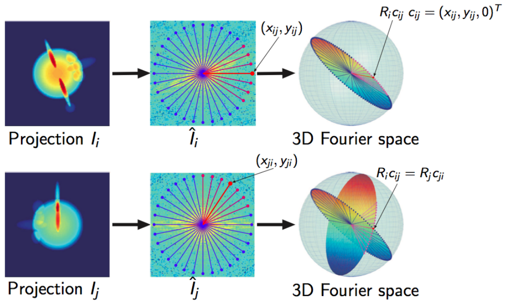

Figure 6.5: An illustration of the use of the Fourier slice theorem and the common lines approach on the viewing direction problem in Cryo-EM. Image courtesy of Amit Singer and Yoel Shkolnisky [204].

semidefinite programming–based method to find the rotations. The spectral method essentially corresponds to an instance of the general semidefinite program (1.9), considering the most popular representation of $SO(3)$, as $3 \times 3$ orthogonal matrices with positive determinant,

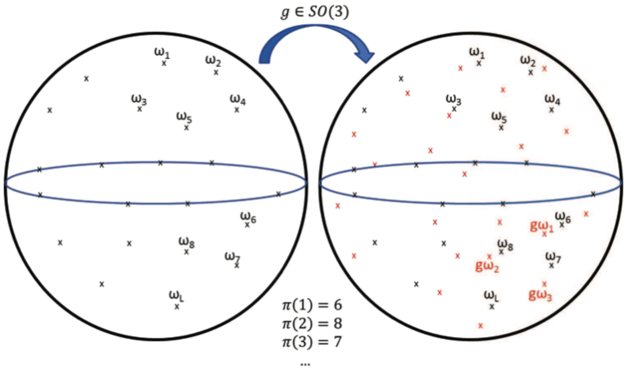

Figure 6.6: An illustration of the approximate representation of $SO(3)$ described by permutation on a discretization $\omega_1, \ldots, \omega_L$ of the sphere. For each element $g \in SO(3)$ we associate the permutation $\pi$ that $g$ approximately induces on the points of the discretization.

Similarly to the signal processing problem, this method has the shortcoming that it loses information, it only keeps the best pairwise common lines, ignoring how compatible different pairs of common lines would be. One can again formalize a quasi-MLE objective function that takes into account the likelihood of all pairs of common lines, not just the best ones. This objective function turns out to be approximated by a linear function if one takes a larger dimensional (approximate) representation of $SO(3)$. Given a discretization of the sphere, one can approximately represent an element of $g \in SO(3)$ by the permutation $\pi$ it (approximately) induces on the points (see Figure 6.6).

The idea is then to develop a semidefinite programming-based approach, inspired in (1.9). As suggested by Remark 6.1.4, representation theoretical properties of $SO(3)$ play a role at devising computationally efficient methods to solve the corresponding SDP. A detailed description (and analysis) of this approach will be subject of a future publication.

# Chapter 7

# Conclusions and Open Problems

## 7.1 Conclusions

Synchronization–type problems form a rich class of problems of great interest from both theoretical and practical viewpoints. In this thesis we only scratched the surface, as far as having a good understanding of these problems, their average hardness, and the effectiveness of different algorithmic approaches. We sincerely hope that this motivates further research towards a general understanding of this class of problems. In that note, we finish this thesis with a list of open problems, not necessarily related to the subject of this thesis, that are dear to the author.

## 7.2 Open Problems

### 7.2.1 Rank recovery for the multireference alignment SDP

Numerical simulations (see Figure 7.1 and [32, 37]) suggest that, below a certain noise level, the semidefinite program (6.14) is tight with high probability. However, an explanation of this phenomenon remains an open problem [37].

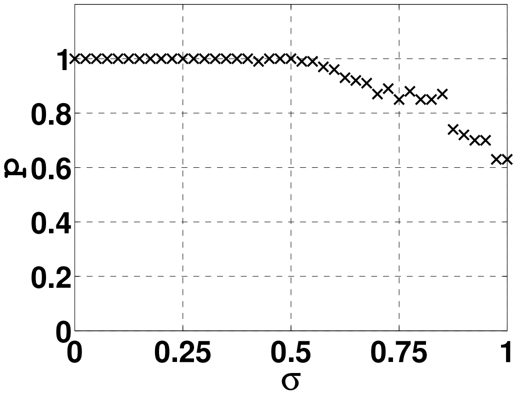

Figure 7.1: Fraction of trials (among 100) on which *rank recovery* was observed, for various values of noise $\sigma$. The plot corresponds to the multireference alignment problem treated in Section 6.1. It suggests that *rank recovery* happens with high probability, below certain noise levels [32, 37].

### 7.2.2 Sample complexity for multireference alignment

Another important question related to the signal processing problem described in Section 6.1 is to understand its sample complexity. Since the objective is to recover the underlying signal $u$, a larger number of observations $n$ should yield a better recovery (considering the model in (6.1)). The question can be formalized as: for a given value of $L$ and $\sigma$, how large does $n$ need to be in order to allow for a reasonably accurate recovery?

### 7.2.3 Deterministic RIP matrices

Let $K < M < N$ be positive integers and let $\delta > 0$. An $M \times N$ matrix $\Phi$ satisfies the $(K, \delta)$-*restricted isometry property* (RIP) if

$$(1-\delta)\|x\|^2 \leq \|\Phi x\|^2 \leq (1+\delta)\|x\|^2$$

whenever $x \in \mathbb{R}^N$ has at most $K$ nonzero entries (i.e. $x$ is a $K$-sparse vector). RIP matrices are important in signal processing, making possible the measurement and recovery of a sparse signal using significantly less measurements than the dimension of the signal [65, 96]. Random matrices are known to satisfy the RIP with high probability for several distributions for $K = O_\delta(M/\operatorname{polylog} N)$. However, matrices constructed randomly have a nonzero (albeit small) probability of failing to be RIP, and checking whether a given matrix satisfies this property is an NP-hard problem [43]. This has raised the interest in constructing explicit (deterministic) RIP matrices. Most deterministic constructions only achieve $K = O_\delta(\sqrt{M})$. The only construction so far to break this square root bottleneck is due to Bourgain, Dilworth, Ford, Konyagin and Kutzarova [57]; they construct a matrix satisfying RIP with $K = O_\delta(M^{1/2+\epsilon})$ for some $\epsilon > 0$ (see also [164]). Some effort has also been made in derandomizing the construction of RIP matrices [34] i.e. finding random constructions of RIP matrices using as few random bits as possible. In [35] it was conjectured that a certain submatrix of the DFT satisfies the $(K, \delta)$-RIP for some $\delta < \sqrt{2}-1$ and $K = O(M/\operatorname{polylog} N)$, this conjecture was later shown to be related to a conjecture in Number Theory [40].

### 7.2.4 Partial Fourier matrices satisfying the Restricted Isometry Property

Consider the random $M \times N$ matrix obtained by drawing rows uniformly with replacement from the $N \times N$ discrete Fourier transform matrix. If $M = \Omega_\delta(K \operatorname{polylog} N)$, then the resulting partial Fourier operator is known to satisfy the restricted isometry property with high probability, and this fact has been dubbed the *uniform uncertainty principle* [68]. A fundamental problem in compressed sensing is determining the smallest number $M$ of random rows necessary. To summarize the progress to date, Candès and Tao [68] first found that $M = \Omega_\delta(K \log^6 N)$ rows suffice, then Rudelson and Vershynin [192] proved $M = \Omega_\delta(K \log^4 N)$, and recently, Bourgain [56] achieved $M = \Omega_\delta(K \log^3 N)$; Nelson, Price and Wootters [172] also achieved $M = \Omega_\delta(K \log^3 N)$, but using a slightly different measurement matrix. As far as lower bounds, in [38] it was shown that $M = \Omega_\delta(K \log N)$ is necessary. This draws a constrast with random Gaussian matrices, where $M = \Omega_\delta(K \log(N/K))$ is known to suffice.

### 7.2.5 Monotonicity of average singular value

Recall Definition 2.3.2: For $d \geq 1$ and $G_{\mathbb{R}} \in \mathbb{R}^{d \times d}$ a gaussian random matrix (not necessarily symmetric) with i.i.d real valued entries $\mathcal{N}(0, d^{-1})$, $\alpha_{\mathbb{R}}(d)$ was defined as

$$\alpha_{\mathbb{R}}(d) := \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d} \sigma_j(G_{\mathbb{R}})\right],$$

where $\sigma_j(G)$ is the $j$th singular value of $G$.

Numerical simulations suggest that $\alpha_{\mathbb{R}}(d)$ is monotonically increasing with $d$ [36]. Perhaps more intriguing is the fact that the analogous quantity $\alpha_{\mathbb{C}}(d)$ over the complex numbers (where $G_{\mathbb{C}} \in \mathbb{C}^{d \times d}$ has complex values Gaussian entries — see [36]) appears to be monotonically decreasing. Establishing the monotonicity of either of these quantities remains an open problem [36] (see Conjecture 2.3.5).

### 7.2.6 Positive Principal Component Analysis

This problem is posed, by Andrea Montanari, in [166]. We briefly describe it here:

Given a symmetric matrix $W \in \mathbb{R}^{n \times n}$ the positive principal component analysis problem can be written as

$$\begin{array}{ll} \max & x^T W x \\ \text{s. t.} & \|x\| = 1 \\ & x \geq 0 \\ & x \in \mathbb{R}^n. \end{array} \tag{7.1}$$

In the flavor of the semidefinite relaxations considered in this thesis, (7.1) can be rewritten (for $X \in \mathbb{R}^{n\times n}$) as

$$\begin{array}{ll} \max & \mathrm{Tr}(WX) \\ \mathrm{s.\ t.} & \mathrm{Tr}(X) = 1 \\ & X \geq 0 \\ & X \succeq 0 \\ & \mathrm{rank}(X) = 1, \end{array}$$

and further relaxed to the semidefinite program

$$\begin{array}{ll} \max & \mathrm{Tr}(WX) \\ \mathrm{s.\ t.} & \mathrm{Tr}(X) = 1 \\ & X \geq 0 \\ & X \succeq 0. \end{array} \tag{7.2}$$

Similarly to the phenomenon investigated in Chapter 5, this relaxation appears to have a remarkable tendency to be tight. In fact, numerical simulations suggest that if $W$ is taken to be a Wigner matrix (symmetric with i.i.d. standard Gaussian entries), then the solution to (7.2) is rank 1 with high probability, but there is no explanation of this phenomenon. Note that this is qualitatively different from the examples studied in this thesis, as there isn't necessarily a planted solution in $W$.

### 7.2.7 Spectral norm of random matrices

Given $M$ symmetric matrices $A_1, \ldots, A_M \in \mathbb{R}^{n\times n}$, consider the random matrix

$$X = \sum_{k=1}^{M} g_k A_k, \tag{7.3}$$

where $g_k$ are i.i.d. standard Gaussian random variables. Both the noncommutative Khintchine inequality and the matrix concentration method (see [220, 221]) provide the following estimate for the spectral norm of $X$:

$$\mathbb{E}\|X\| \lesssim \sigma\sqrt{\log n} \qquad \text{with} \qquad \sigma^2 := \left\| \sum_{k=1}^{M} A_k^2 \right\|. \tag{7.4}$$

While (7.4) has proven to be of great use in many applications, the dimension factor is known not to always be tight. In fact, in Section 4.1 we described an improved bound (4.1) on the particular case where the entries of $X$ are independent.

It is an important problem to understand when the dimension dependence of the bound (7.4) can be improved. We refer the reader to [222] and references within for some progress in this direction.

# Bibliography

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *Network Science and Engineering, IEEE Transactions on*, 1(1):10–22, Jan 2014.

[2] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Linear inverse problems on Erdős-Rényi graphs: Information-theoretic limits and efficient recovery. *IEEE International Symposium on Information Theory (ISIT2014)*, 2014.

[3] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *Available online at arXiv:1405.3267v4 [cs.SI]*, 2014.

[4] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *Available online at arXiv:1503.00609 [math.PR]*, 2015.

[5] K. Rohe abd S. Chatterjee and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics.*

[6] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

[7] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 36–43, 2005.

[8] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. Multisection in the stochastic block model using semidefinite programming. *Submitted*, 2015.

[9] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *Twelfth IEEE International Conference on Computer Vision (ICCV 2009)*, Kyoto, Japan, September. IEEE.

[10] A. Agrawal, R. Raskar, and R. Chellappa. What is the range of surface reconstructions from a gradient field? In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 578–591. Springer Berlin Heidelberg, 2006.

[11] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *SIAM J. on Imaging Sci.*, 7(1):35–66, 2013.

[12] B. Alexeev, J. Cahili, and D. G. Mixon. Full spark frames. *Journal of Fourier Analysis and Applications*, 18(6):1167–1194, 2012.

[13] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5:13–51, 1993.

[14] F. Alizadeh, J.-P. Haeberly, and M.L. Overton. Complementarity and nondegeneracy in semidefinite programming. *Mathematical Programming*, 77(1):111–128, 1997.

[15] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986.

[16] N. Alon, K. Makarychev, Y. Makarychev, and A. Naor. Quadratic forms on graphs. *Invent. Math*, 163:486–493, 2005.

[17] N. Alon and V. Milman. Isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory*, 38:73–88, 1985.

[18] N. Alon and A. Naor. Approximating the cut-norm via Grothendieck's inequality. In *Proc. of the 36 th ACM STOC*, pages 72–80. ACM Press, 2004.

[19] D. Amelunxen and P. Bürgisser. Intrinsic volumes of symmetric cones and applications in convex programming. *Mathematical Programming*, pages 1–26, 2014.

[20] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. 2014.

[21] Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):pp. 2877–2921, 2009.

[22] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, New York, Melbourne, 2010.

[23] S. Arora, S. A. Khot, A. Kolla, D. Steurer, M. Tulsiani, and N. K. Vishnoi. Unique games on expanding constraint graphs are easy: extended abstract. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 21–28, New York, NY, USA, 2008. ACM.

[24] M.F. Atiyah and F. Hirzebruch. Quelques théorème de non-plongement pour les variétès differentiables. *Bull. Soc. Math. Fr.*, 87:383–396, 1959.

[25] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. Relax, no need to round: integrality of clustering formulations. *6th Innovations in Theoretical Computer Science (ITCS 2015)*, 2015.

[26] Z. D. Bai and Y. Q. Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *Ann. Probab.*, 16(4):1729–1741, 1988.

[27] R. Balan, B.G. Bodmann, P.G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *J. Fourier Anal. Appl. 15 (2009) 488–501*, 2009.

[28] R. Balan, P.G. Casazza, and D. Edidin. On signal reconstruction without phase. *Appl. Comput. Harmon. Anal. 20 (2006) 345–356*, 2006.

[29] A. S. Bandeira. Random Laplacian matrices and convex relaxations. *Available online at arXiv:1504.03987 [math.PR]*, 2015.

[30] A. S. Bandeira, N. Boumal, and A. Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Available online at arXiv:1411.3272 [math.OC]*, 2014.

[31] A. S. Bandeira, J. Cahil, D. G. Mixon, and A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis (ACHA)*, 37:106–125, 2014.

[32] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu. Multireference alignment using semidefinite programming. *5th Innovations in Theoretical Computer Science (ITCS 2014)*, 2014.

[33] A. S. Bandeira, Y. Chen, and D. G. Mixon. Phase retrieval from power spectra of masked signals. *Information and Inference: a Journal of the IMA*, 3:83–102, 2014.

[34] A. S. Bandeira, M. Fickus, D. G. Mixon, and J. Moreira. Derandomizing restricted isometries via the Legendre symbol. *Available online at arXiv:1406.4089 [math.CO]*, 2014.

[35] A. S. Bandeira, M. Fickus, D. G. Mixon, and P. Wong. The road to deterministic matrices with the restricted isometry property. *Journal of Fourier Analysis and Applications*, 19(6):1123–1149, 2013.

[36] A. S. Bandeira, C. Kennedy, and A. Singer. Approximating the little grothendieck problem over the orthogonal group. *Available online at arXiv:1308.5207 [cs.DS]*, 2013.

[37] A. S. Bandeira, Y. Khoo, and A. Singer. Open problem: Tightness of maximum likelihood semidefinite relaxations. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *JMLR W&CP*, pages 1265–1267, 2014.

[38] A. S. Bandeira, M. E. Lewis, and D. G. Mixon. Discrete uncertainty principles and sparse signal processing. *Available online at arXiv:1504.01014 [cs.IT]*, 2015.

[39] A. S. Bandeira and D. G. Mixon. Near-optimal sparse phase retrieval. *Proceedings of SPIE Volume 8858*, 2013.

[40] A. S. Bandeira, D. G. Mixon, and J. Moreira. A conditional construction of restricted isometries. *Available online at arXiv:1410.6457 [math.FA]*, 2014.

[41] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger inequality for the graph connection Laplacian. *SIAM J. Matrix Anal. Appl.*, 34(4):1611–1630, 2013.

[42] A. S. Bandeira and R. v. Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability, to appear*, 2015.

[43] A.S. Bandeira, E. Dobriban, D.G. Mixon, and W.F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Inform. Theory*, 59(6):3448–3450, 2013.

[44] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[45] A.I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(1):189–202, 1995.

[46] A. Ben-Tal and A. Nemirovski. On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty. *SIAM Journal on Optimization*, 12:811–833, 2002.

[47] F. Benaych-Georges and S. Péché. Largest eigenvalues and eigenvectors of band or sparse random matrices. *Electron. Commun. Probab.*, 19:4–9, 2014.

[48] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[49] Y. Bilu and N. Linial. Lifts, discrepancy and nearly optimal spectral gap. *Combinatorica*, 26(5):495–519, 2006.

[50] B.G. Bodmann and N. Hammen. Stable phase retrieval with low-redundancy frames. *Advances in Computational Mathematics*, pages 1–15, 2014.

[51] R. B. Boppana. Eigenvalues and graph bisection: An average-case analysis. *In 28th Annual Symposium on Foundations of Computer Science*, pages 280–285, 1987.

[52] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

[53] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, with a foreword by Michel Ledoux.

[54] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[55] N. Boumal, A. Singer, P.-A. Absil, and V. D. Blondel. Cramér-Rao bounds for synchronization of rotations. *Information and Inference: A journal of the IMA*, 3:1–39, 2014.

[56] J. Bourgain. An improved estimate in the restricted isometry problem. *Lect. Notes Math.*, 2116:65–70, 2014.

[57] J. Bourgain et al. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1), 2011.

[58] J. Briet, H. Buhrman, and B. Toner. A generalized Grothendieck inequality and nonlocal correlations that require high entanglement. *Communications in Mathematical Physics*, 305(3):827–843, 2011.

[59] J. Briet, F. M. O. Filho, and F. Vallentin. The positive semidefinite Grothendieck problem with rank constraint. In *Automata, Languages and Pro-*

*gramming*, volume 6198 of *Lecture Notes in Computer Science*, pages 31–42. Springer Berlin Heidelberg, 2010.

[60] J. Briet, F. M. O. Filho, and F. Vallentin. Grothendieck inequalities for semidefinite programs with rank constraint. *Theory of Computing*, 10:77–105, 2014.

[61] W. Bryc, A. Dembo, and T. Jiang. Spectral measure of large random Hankel, Markov and Toeplitz matrices. *The Annals of Probability*, 34(1):pp. 1–38, 2006.

[62] T. N. Bui, S. Chaudhuri, F. T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7:171–191, 1987.

[63] O. Bunk et al. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Cryst.*, A63:306–314, 2007.

[64] E. J. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[65] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52:489–509, 2006.

[66] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[67] E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[68] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[69] E.J. Candès, Y.C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[70] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[71] E. A. Carlen. Trace inequalities and quantum entropy: An introductory course. available at http://www.ueltschi.org/azschool/notes/ericcarlen.pdf, 2009.

[72] T. Carson and R. Impagliazzo. Hill-climbing finds random planted bisections. *Proc. 12th Symposium on Discrete Algorithms (SODA 01)*, pages 903–909, 2001.

[73] V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 08 2012.

[74] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[75] M. Charikar, K. Makarychev, and Y. Makarychev. Near-optimal algorithms for unique games. *Proceedings of the 38th ACM Symposium on Theory of Computing*, 2006.

[76] K. N. Chaudhury, Y. Khoo, and A. Singer. Global registration of multiple point clouds using semidefinite programming. *SIAM Journal on Optimization*, 25(1):126–185, 2015.

[77] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *Problems in analysis (Papers dedicated to Salomon Bochner, 1969), pp. 195–199. Princeton Univ. Press*, 1970.

[78] Y. Chen and A. J. Goldsmith. Information recovery from pairwise measurements. *IEEE International Symposium on Information Theory (ISIT2014)*, 2014.

[79] Y. Chen, Q.-X. Huang, and L. Guibas. Near-optimal joint object matching via convex relaxation. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[80] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 2012.

[81] F. Chung. Four proofs for the cheeger inequality and graph partition algorithms. *Fourth International Congress of Chinese Mathematicians, pp. 331–349*, 2010.

[82] F. Chung and L. Lu. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics).* American Mathematical Society, Boston, MA, USA, 2006.

[83] F. R. K. Chung. *Spectral Graph Theory.* AMS, 1997.

[84] J. Cohen. Is high-tech view of HIV too good to be true? *Science*, 341(6145):443–444, 2013.

[85] A. Conca, D. Edidin, M. Hering, and C. Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346 – 356, 2015.

[86] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Lecture Notes in Computer Science*, 1671:221–232, 1999.

[87] R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications.* Cambridge University Press, New York, NY, USA, 2011.

[88] M. Cucuringu. Synchronization over z2 and community detection in signed multiplex networks with constraints. *Journal of Complex Networks*, 2015.

[89] M. Cucuringu, Y. Lipman, and A. Singer. Sensor network localization by eigenvector synchronization over the euclidean group. *ACM Transactions on Sensor Networks*, 8(3):19:1–19:42, 2012.

[90] J.C. Dainty and J.R. Fienup. Phase retrieval and image reconstruction for astronomy. *In: H. Stark, ed., Image Recovery: Theory and Application, Academic Press, New York*, 1987.

[91] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook on the Geometry of Banach spaces*, volume 1, pages 317–366. Elsevier Science, 2001.

[92] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84, December 2011.

[93] L. Demanet and V. Jugnon. Convex recovery from interferometric measurements. *Available online at arXiv:1307.6864 [math.NA]*, 2013.

[94] R. Diamond. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Science*, 1(10):1279–1287, October 1992.

[95] X. Ding and T. Jiang. Spectral distribution of adjacency and Laplacian matrices of random graphs. *The Annals of Applied Probability*, 20(6):2086–2117, 2010.

[96] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[97] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. Wiley series in probability and statistics. Wiley, Chichester [u.a.], 1998.

[98] H. Duadi, O. Margalit, V. Mico, J.A. Rodrigo, T. Alieva, J. Garcia, and Z. Zalevsky. Digital holography and phase retrieval, in holography, research and technologies. *J. Rosen, ed., InTech; available online at* ***http://www.intechopen.com/books/holography-research-and-technologies/digital-holography-and-phase-retrieval***, 2011.

[99] R. Durrett. *Random Graph Dynamics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, New York, NY, USA, 2006.

[100] M. E. Dyer and A. M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451 – 489, 1989.

[101] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[102] K. Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, 1955.

[103] U. Feige and J. Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639 – 671, 2001.

[104] M. Fickus and D.G. Mixon. Numerically erasure-robust frames. *Linear Algebra Appl.*, 437:1394–1407, 2012.

[105] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Optics*, 21:2758–2769, 1982.

[106] H. Foroosh, J. B. Zerubia, and M. Berthod. Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, 11(3):188–200, 2002.

[107] J. Friedman. A proof of Alon's second eigenvalue conjecture and related problems. *Mem. Amer. Math. Soc.*, 195, 2008.

[108] A. Frieze and M. Jerrum. Improved approximation algorithms for MAX-k-CUT and MAX BISECTION. *Algorithmica*, 18(1):67–81, 1997.

[109] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

[110] S. Geman. A limit theorem for the norm of random matrices. *Ann. Probab.*, 8(2):252–261, 1980.

[111] R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik 35 (1972) 237–246*, 1972.

[112] S. O. Gharan and L. Trevisan. A higher–order cheeger's inequality. *Available online at arXiv:1107.2686 [cs.DS]*, 2011.

[113] A. Giridhar and P.R. Kumar. Distributed clock synchronization over wireless networks: Algorithms and analysis. In *Decision and Control, 2006 45th IEEE Conference on*, pages 4915–4920. IEEE, 2006.

[114] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefine programming. *Journal of the Association for Computing Machinery*, 42:1115–1145, 1995.

[115] M. X. Goemans and D. P. Williamson. Approximation algorithms for Max-3-Cut and other problems via complex semidefinite programming. *Journal of Computer and System Sciences*, 68(2):442–470, 2004.

[116] F. Gotze and A. Tikhomirov. On the rate of convergence to the Marchenko–Pastur distribution. *arXiv:1110.1284 [math.PR]*, 2011.

[117] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoustics Speech Signal Proc.*, 32:236–243, 1984.

[118] A. Grothendieck. Resume de la theorie metrique des produits tensoriels topologiques (french). *Reprint of Bol. Soc. Mat. Sao Paulo*, pages 1–79, 1996.

[119] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[120] Hagen and Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11:1074–1085, 1992.

[121] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *Available online at arXiv:1412.6156*, 2014.

[122] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *Available online at arXiv:1502.07738*, 2015.

[123] R.W. Harrison. Phase problem in crystallography. *J. Opt. Soc. Am. A 10 (1993) 1046–1055*, 1993.

[124] P. Harsha and A. Barth. Lecture 5: Derandomization (part ii). *http://www.tcs.tifr.res.in/ prahladh/teaching/05spring/lectures/lec5.pdf*, 2002.

[125] R. Hartley, J. Trumpf, Y. Dai, and H. Li. Rotation averaging. *International Journal of Computer Vision*, 103(3):267–305, 2013.

[126] T. Heinosaari, L. Mazzarella, and M. M. Wolf. Quantum tomography under prior information. *Communications in Mathematical Physics*, 318(2):355–374, 2013.

[127] R. Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45):18037–18041, 2013.

[128] N. J. Higham. Computing the polar decomposition – with applications. *SIAM J. Sci. Stat. Comput.*, 7:1160–1174, October 1986.

[129] J. L. Horner and P. D. Gianino. Phase-only matched filtering. *Appl. Opt.*, 23(6):812–816, 1984.

[130] Q.-X. Huang and L. Guibas. Consistent shape maps via semidefinite programming. *Computer Graphics Forum*, 32(5):177–186, 2013.

[131] C. Iaconis and I.A. Walmsley. Spectral phase interferometry for direct electric-field reconstruction of ultrashort optical pulses. *Opt. Lett.*, 23:792–794, 1998.

[132] M. Jerrum and G. B. Sorkin. The metropolis algorithm for graph bisection. *Discrete Applied Mathematics*, 82:155–175, 1998.

[133] X. Jiang, L.-H. LIM, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

[134] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[135] J. B. Keller. Closest unitary, orthogonal and hermitian operators to a given operator. *Mathematics Magazine*, 48(4):pp. 192–197, 1975.

[136] D. Keusters, H.-S. Tan, P. O'Shea, E. Zeek, R. Trebino, and W.S. Warren. Relative-phase ambiguities in measurements of ultrashort pulses with well-separated multiple frequency components. *J. Opt. Soc. Am. B*, 20:2226–2237, 2003.

[137] O. Khorunzhiy. Estimates for moments of random matrices with Gaussian elements. In *Séminaire de probabilités XLI*, volume 1934 of *Lecture Notes in Math.*, pages 51–92. Springer, Berlin, 2008.

[138] S. Khot. On the power of unique 2-prover 1-round games. *Thiry-fourth annual ACM symposium on Theory of computing*, 2002.

[139] A. Kolla, K. Makarychev, and Y. Makarychev. How to play unique games against a semi-random adversary. *FOCS*, 2011.

[140] A. Kolla and M. Tulsiani. Playing random and expanding unique games. *Unpublished.*

[141] P. Hand L. Demanet. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

[142] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282 (electronic), 2005.

[143] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991.

[144] J.R. Lee, S.O. Gharan, and L. Trevisan. Multi-way spectral partitioning and higher–order cheeger inequalities. *STOC '12 Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 2012.

[145] J. Lennard-Jones. On the determination of molecular fields. *Proc. R. Soc. Lond. A*, 106(738):463–477, 1924.

[146] O. Leveque. Random matrices and communication systems: Wishart random matrices: marginal eigenvalue distribution. *available at: http://ipg.epfl.ch/ leveque/Matrix/*, 2012.

[147] R.-C. Li. New perturbation bounds for the unitary polar factor. *SIAM J. Matrix Anal. Appl.*, 16(1):327–332, January 1995.

[148] G. Livan and P. Vivo. Moments of Wishart-Laguerre and Jacobi ensembles of random matrices: application to the quantum transport problem in chaotic cavities. *Acta Physica Polonica B*, 42:1081, 2011.

[149] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982.

[150] L. Lovasz. On the shannon capacity of a graph. *IEEE Trans. Inf. Theor.*, 25(1):1–7, 1979.

[151] S.H. Low. Convex relaxation of optimal power flow: a tutorial. In *Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium*, pages 1–15. IEEE, 2013.

[152] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8:261–277, 1988.

[153] Z. Luo, W. Ma, AMC So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *Signal Processing Magazine, IEEE*, 27(3):20–34, 2010.

[154] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, 1967.

[155] A. M. Maiden and J.M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109:1256–1262, 2009.

[156] M. Maila and J. Shi. A random walks view of spectral segmentation. *AI and STATISTICS (AISTATS)*, 2001.

[157] S. Marchesini. A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Inst.*, 78, 2007.

[158] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, pages 1–8, June 2007.

[159] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2), 2000.

[160] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703, New York, NY, USA, 2014. ACM.

[161] F. McSherry. Spectral partitioning of random graphs. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, pages 529–537, 2001.

[162] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem. 59 (2008) 387–410*, 2008.

[163] R.P. Millane. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A 7 (1990) 394–411.*, 1990.

[164] D. G. Mixon. Explicit matrices with the restricted isometry property: Breaking the square-root bottleneck. *available online at arXiv:1403.3427 [math.FA]*, 2014.

[165] D. Mondragon and V. Voroninski. Determination of all pure quantum states from a minimal number of observables. *arXiv:1306.1214*, 2013.

[166] A. Montanari. Principal component analysis with nonnegativity constraints. *http://sublinear.info/index.php?title=Open_Problems:62*, 2014.

[167] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. *Available online at arXiv:1407.1591v2 [math.PR]*, July 2014.

[168] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Available online at arXiv:1311.4115 [math.PR]*, January 2014.

[169] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *Probability Theory and Related Fields (to appear)*, 2014.

[170] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *Adv. Neural. Inf. Process. Syst.*, 18:955–962, 2005.

[171] A. Naor, O. Regev, and T. Vidick. Efficient rounding for the noncommutative Grothendieck inequality. In *Proceedings of the 45th annual ACM symposium on*

*Symposium on theory of computing*, STOC '13, pages 71–80, New York, NY, USA, 2013. ACM.

[172] J. Nelson, E. Price, and M. Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. *SODA*, pages 1515–1528, 2014.

[173] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Program.*, 109(2-3):283–317, 2007.

[174] Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9(1-3):141–160, 1998.

[175] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

[176] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010.

[177] O. Ozyesil, A. Singer, and R. Basri. Stable camera motion estimation using convex programming. *SIAM Journal on Imaging Sciences, to appear*, 2013.

[178] O. Ozyesil, A. Singer, and R. Basri. Robust camera location estimation by convex programming. *CVPR*, 2015.

[179] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

[180] G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge, 2003.

[181] G. Pisier. Grothendieck's theorem, past and present. *Bull. Amer. Math. Soc.*, 49:237–323, 2011.

[182] R. G. Pita, M. R. Zurera, P. J. Amores, and F. L. Ferreras. Using multilayer perceptrons to align high range resolution radar signals. In Wlodzislaw Duch, Janusz Kacprzyk, Erkki Oja, and Slawomir Zadrozny, editors, *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*, volume 3697 of *Lecture Notes in Computer Science*, pages 911–916. Springer Berlin Heidelberg, 2005.

[183] D. Puder. Expansion of random graphs: New proofs, new results. *Inventiones mathematicae*, pages 1–64, 2014.

[184] M. Püschel and J. Kovačević. Real, tight frames with maximal robustness to erasures. *Proc. Data Compr. Conf.*, pages 63–72, 2005.

[185] P. Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 245–254. ACM, 2008.

[186] O. Raz, N. Dudovich, and B. Nadler. Vectorial phase retrieval of 1-d signals. *IEEE Trans. Signal Process.*, 61:1632–1643, 2013.

[187] O. Raz et al. Vectorial phase retrieval for linear characterization of attosecond pulses. *Phys. Rev. Lett. 107 (2011) 133902/1–5*, 2011.

[188] S. Riemer and C. Schütt. On the expectation of the norm of random matrices with non-identically distributed entries. *Electron. J. Probab.*, 18, 2013.

[189] J. M. Rodenburg. Ptychography and related diffractive imaging methods. *Adv. Imag. Elect. Phys.*, 150:87–184, 2008.

[190] J.J. Rotman. *An introduction to the theory of groups.* Springer, 4 edition, 1994.

[191] J. Rubinstein and G. Wolansky. Reconstruction of optical surfaces from ray data. *Optical Review*, 8(4):281–283, 2001.

[192] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[193] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III*, pages 1576–1602. Hindustan Book Agency, New Delhi, 2010.

[194] A.P. Ruszczyński. *Nonlinear optimization.* Princeton University Press, 2006.

[195] G. Sagnol. A class of semidefinite programs with rank-one solutions. *Linear Algebra and its Applications*, 435(6):1446–1463, 2011.

[196] H. Sahinoglou and S.D. Cabrera. On phase retrieval of finite-length sequences using the initial time sample. *IEEE Trans. Circuits Syst. 38 (1991) 954–958*, 1991.

[197] S. Sahni and T. Gonzalez. P-complete approximation problems. *J. ACM*, 23(3):555–565, July 1976.

[198] P. H. Schonemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[199] Y. Seginer. The expected norm of random matrices. *Combin. Probab. Comput.*, 9(2):149–166, 2000.

[200] M. Shatsky, R. J. Hall, S. E. Brenner, and R. M. Glaeser. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of Structural Biology*, 166(1), 2009.

[201] J. Shen. On the singular values of gaussian random matrices. *Linear Algebra and its Applications*, 326(13):1 – 14, 2001.

[202] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20 – 36, 2011.

[203] A. Singer, R. R. Coifman, F. J. Sigworth, D. W. Chester, and Y. Shkolnisky. Detecting consistent common lines in cryo-em by voting. *Journal of Structural Biology*, pages 312–322.

[204] A. Singer and Y. Shkolnisky. Three-dimensional structure determination from common lines in Cryo-EM by eigenvectors and semidefinite programming. *SIAM J. Imaging Sciences*, 4(2):543–572, 2011.

[205] A. Singer and H.-T. Wu. Orientability and diffusion maps. *Appl. Comput. Harmon. Anal.*, 31(1):44–58, 2011.

[206] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.

[207] A. Singer, Z. Zhao, , Y. Shkolnisky, and R. Hadani. Viewing angle classification of Cryo-Electron Microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4:723–759, 2011.

[208] T. A. B. Snijders and K. Nowicki. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100, January 1997.

[209] A. So, J. Zhang, and Y. Ye. On approximating complex quadratic optimization problems via semidefinite programming relaxations. *Math. Program. Ser. B*, 110:93–110, 2007.

[210] A.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical Programming*, 130(1):125–151, 2011.

[211] A. M.-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.

[212] S. Sodin. The Tracy-Widom law for some sparse random matrices. *J. Stat. Phys.*, 136(5):834–841, 2009.

[213] S. Sodin. The spectral edge of some random band matrices. *Ann. of Math. (2)*, 172(3):2223–2251, 2010.

[214] S. Sojoudi and J. Lavaei. Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure. *SIAM Journal on Optimization*, 24(4):1746–1778, 2014.

[215] B. Sonday, A. Singer, and I. G. Kevrekidis. Noisy dynamic simulations in the presence of symmetry: Data alignment and model reduction. *Computers & Mathematics with Applications*, 65(10):1535 – 1557, 2013.

[216] M. Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems.

[217] T. Tao. *Topics in Random Matrix Theory.* Graduate studies in mathematics. American Mathematical Soc., 2012.

[218] D. L. Theobald and P. A. Steindel. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics*, 28(15):1972–1979, 2012.

[219] L. Trevisan. Max cut and the smallest eigenvalue. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, pages 263–272, New York, NY, USA, 2009. ACM.

[220] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[221] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning, to appear*, 2015.

[222] J. A. Tropp. Second-order matrix concentration inequalities. *In preparation*, 2015.

[223] J.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, March 2006.

[224] T. Tzeneva. Global alignment of multiple 3-d scans using eigevector synchronization. *Senior Thesis, Princeton University (supervised by S. Rusinkiewicz and A. Singer)*, 2011.

[225] R. van Handel. On the spectral norm of inhomogeneous random matrices. *Available online at arXiv:1502.05003 [math.PR]*, 2015.

[226] M. Van-Heel. Angular reconstitution: a posteriori assignment of projection directions for 3d reconstruction. *Ultramicroscopy*, 21:111–123, 1987.

[227] L. Vanderberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

[228] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press*, 2012.

[229] C. Vinzant. A small frame and a certificate of its injectivity. *Available online at arXiv:1502.04656 [math.FA]*, 2015.

[230] V. Vu. A simple svd algorithm for finding hidden partitions. *Available online at arXiv:1404.3918*, April 2014.

[231] I. Waldspurger, A. D'aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Math. Program.*, 149(1-2):47–81, 2015.

[232] A. Walther. The question of phase retrieval in optics. *Opt. Acta 10 (1963) 41–49*, 1963.

[233] L. Wang and A. Singer. Exact and stable recovery of rotations for robust synchronization. *Information and Inference*, 2(2):145–193, 2013.

[234] Y. Wang. Minimal frames for phase retrieval. *Workshop on Phaseless Reconstruction, U. Maryland, February 24*, 2013.

[235] Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, volume 166 of *International Series in Operations Research & Management Science*, pages 533–564. Springer US, 2012.

[236] Z. Wen, D. Goldfarb, and W. Yin. Alternating direction augmented lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3-4):203–230, 2010.

[237] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):pp. 325–327, 1958.

[238] D. P. Williamson and D. B. Shmoys. *The Design of Approximation Algorithms.* Cambridge University Press, 1 edition, 2011.

[239] D. Raz Y. Bartal, M. Charikar. Approximating min-sum k-clustering in metric spaces. *STOC'01*, 2001.

[240] S. Zhang and Y. Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.

[241] J. P. Zwart, R. van der Heiden, S. Gelsema, and F. Groen. Fast translation invariant classification of hrr range profiles in a zero phase representation. *Radar, Sonar and Navigation, IEE Proceedings*, 150(6):411–418, 2003.