Title
PhaseLift: A Novel Methodology for Phase Retrieval

Author
Voroninski, Vladislav

Publication Date
2013

Peer reviewed|Thesis/dissertation

PhaseLift: A novel framework for phase retrieval.

by

Vladislav Voroninski

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

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Strain, Chair
Professor Lawrence C. Evans
Professor Haiyan Huang

Spring 2013

PhaseLift: A novel framework for phase retrieval.

Abstract

PhaseLift: A novel framework for phase retrieval.

by

Vladislav Voroninski

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor John Strain, Chair

In many physical settings, it is difficult or impossible to measure the phase of a signal. The problem is then to recover a signal from intensity measurements only. This phase retrieval problem has challenged physicists, mathematicians and engineers for decades, being notoriously difficult to solve numerically. We propose a novel framework for phase retrieval, which recasts the problem as a low rank matrix recovery problem and provide theoretical guarantees and empirical demonstrations of its performance, as well as connections of our results to quantum mechanics and random matrix theory.

To Olympiada Mosunova

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors, John Strain and Emmanuel Candés, for helping me in the completion of this thesis. I am also grateful for the guidance of Lawrence C. Evans, Bernd Sturmfels and Thomas Strohmer.

Chapter 1

Introduction

1.1 Overview

In many applications, one would like to acquire information about an object but the physical nature of detectors make it difficult or impossible to measure the phase of the signal. Typically, detectors can often times record only the squared modulus of the Fresnel or Fraunhofer diffraction pattern of the radiation that is scattered from an object, leaving out desired structural information which comes from the phase of a signal. The problem then becomes to reconstruct an object from intensity measurements only. This phase retrieval problem has a long history of challenges, with a particularly important example in the practice of X-ray crystallography, in which one is faced with recovering a signal or image from the intensity measurements of its Fourier transform.[45, 52].

Phase retrieval problems are notoriously difficult to solve numerically. This thesis proposes, develops, analyses and tests a novel framework for phase retrieval, called PhaseLift, which is provably robust and numerically efficient.

In chapter 2, which is based on the paper "Phase retrieval from matrix completion", coauthored by Emmanuel Candes, Thomas Strohmer and Yonina Eldar [22], we explain the PhaseLift methodology, introduce some preliminary theoretical results and an empirical study of the effectiveness of PhaseLift. In chapter 3, which is based on the paper "PhaseLift: exact and stable quadratic recovery", co-authored by Emmanuel Candes and Thomas Strohmer [81], we show that PhaseLift exactly recovers a fixed signal with high probability from quadratic gaussian measurements and is furthermore stable with respect to noise. Chapter 4 makes a step in the direction of providing theoretical results for PhaseLift with more structured measurement ensembles and connects these results with Wright's conjecture from Quantum Mechanics.

1.2 The phase retrieval problem

Application and a historical perspective

Historically, one of the first applications of phase retrieval is X-ray crystallography [66, 42], an application that remains very important in the understanding of molecular structures. The phase

retrieval problem occurs in many other areas of imaging science such as diffraction imaging [2], optics [86], astronomical imaging [27], microscopy [65]. In particular, it is used in X-ray tomography, which has become an invaluable tool in biomedical imaging to generate quantitative 3D density maps of extended specimens on the nanoscale [29]. Other subjects where phase retrieval plays an important role are quantum mechanics [75, 26] and even makes an appearance in differential geometry [13]. Phase retrieval has seen more activity in recent years due to the desire to image individual molecules and other nano-particles, and the development of new imaging capabilities such as new X-ray synchrotron sources that provide extraordinary X-ray fluxes [67, 80, 14, 65, 29]. References and various instances of the phase retrieval problem as well as some theoretical and numerical techniques can be found in [45, 58, 52].

Statement of the problem

There are many ways in which one can pose the phase-retrieval problem, for instance depending upon assuming a continuous or discrete-space model for the signal. In this thesis, we consider finite length signals (one-dimensional or multi-dimensional) for simplicity, and because numerical algorithms ultimately operate with discrete data. While our claims apply only to the discrete version of the problem, our framework naturally extends to handle continuous objects as well.

Let $x \in \mathbb{C}^n$ be a signal and $z_i \in \mathbb{C}^n$ be a family of sensing vectors. Assume that measurements $b_i = |\langle x, z_i \rangle|^2$ are observed. Note that multiplying x by $e^{i\theta}$ for any $\theta \in \mathbb{R}$ does not change the measurements and thus we can hope to recover x only up to a multiplication by a constant phase factor of unit magnitude. From now on, if $x \in \mathbb{C}$ and $y \in \mathbb{C}^n$, we say that $x = y$ modulo phase if $\exists \quad \theta \in \mathbb{R}$ such that $x = e^{i\theta} y$. The problem of quadratic recovery is to recover x modulo phase from quadratic measurements $b_i = |\langle x, z_i \rangle|^2$.

Phase retrieval is the subset of instances of quadratic recovery which occur in practice, where the sensing vectors z_i are dictated by constraints of the physical setting. For instance, the Fourier transform of x is

$$\hat{x}[\omega] = \frac{1}{\sqrt{n}} \sum_{0 \le t < n} x[t] e^{-i2\pi\omega t/n}, \quad \omega \in \Omega. \tag{1.2.1}$$

Here, Ω is a grid of sampled frequencies, such as $\Omega = \{0, 1, \ldots, n-1\}$ so that the mapping above is the classical unitary discrete Fourier transform (DFT)[1]. Phase retrieval in the setting of X-ray crystallography then consists of finding x from the magnitude coefficients $|\hat{x}[\omega]|$, $\omega \in \Omega$. When Ω is the usual frequency grid as above and without further information about the unknown signal x, this problem is ill-posed since there are many different signals whose Fourier transforms have the same magnitude. Indeed, if x is a solution to the phase retrieval problem, then (1) cx for any scalar $c \in \mathbb{C}$ obeying $|c| = 1$ is also solution, (2) the "mirror function" or time-reversed signal $\bar{x}[-t \bmod n]$ where $t = 0, 1, \ldots, n-1$ is also solution, and (3) the shifted signal $x[t - a \bmod n]$ is also a solution. From a physical viewpoint these "trivial associates" of x are acceptable

[1] For later reference, we denote the Fourier transform operator by F and the inverse Fourier transform by F^{-1}.

ambiguities. But in general infinitely many solutions can be obtained from $\{|\hat{x}[\omega]|| : \omega \in \Omega\}$ beyond these trivial associates [79].

In practice, people have employed various assumptions on the signal, holographic or oversampling techniques to overcome this ambiguity. We will give a background on these approaches before introducing the PhaseLift methodology.

1.3 Main approaches to phase retrieval

While holographic techniques have been applied successfully in certain areas of optical imaging, they are generally difficult to implement in practice [1]. The development of algorithms for signal recovery from magnitude measurements is still a very active field of research. Existing methods for phase retrieval rely on various kinds of a priori information about the signal, such as positivity, atomicity, support constraints and real-valuedness [35, 34, 59, 25]. Direct methods [43] are limited in their applicability to small-scale problems due to their large computational complexity.

An approach proposed to alleviate the non-uniqueness problem is oversampling in the Fourier domain [48]. While oversampling provably offers no unicity advantage for one-dimensional signals, it does so for multidimensional signals, where it has been shown that twofold oversampling in each dimension almost always yields uniqueness for finitely supported, real-valued and non-negative signals [16, 44, 79]. In other words, a digital image of the form $x = \{x[t_1, t_2]\}$ with $0 \leq t_1 < n_1$ and $0 \leq t_2 < n_2$, whose Fourier transform is given by

$$\hat{x}[\omega_1, \omega_2] = \frac{1}{\sqrt{n_1 n_2}} \sum x[t_1, t_2] e^{-i2\pi(\omega_1 t_1/n_1 + \omega_2 t_2/n_2)}, \tag{1.3.1}$$

is almost always uniquely determined from the values of $|\hat{x}[\omega_1, \omega_2]|$ on the oversampled grid $\omega = (\omega_1, \omega_2) \in \Omega = \Omega_1 \times \Omega_2$ in which $\Omega_i = \{0, 1/2, 1, 3/2, \ldots, n_i + 1/2\}$. (In other words, if we think of $(1/n_1, 1/n_2)$ as some Nyquist frequency, then we would need to sample at a rate at least twice this Nyquist frequency.) This holds provided x has proper spatial support, is real valued and non-negative.

As pointed out in [58], these uniqueness results do not imply the existence of a stable recovery algorithm, or about the robustness and stability of commonly used reconstruction algorithms, a claim that we will back up with empirical evidence in chapter 1 of this thesis. In general, well-posedness does not immediately translate into numerical methods and as a result, the algorithmic and practical aspects of the phase retrieval problem (from noisy data) still pose significant challenges.

The most popular methods for phase retrieval from oversampled Fourier data are alternating projection algorithms proposed by Gerchberg and Saxton [37] and Fienup [35, 34]. These methods often require the exploitation of signal constraints and parameter selection to increase the likelihood of convergence to a correct solution [72, 59, 25, 61]. We describe below the simplest instance of a widely used alternating projection approach [70], which relies on support constraints in the spatial domain and oversampled measurements in the frequency domain. With T being a known subset containing the support of the signal x (supp$(x) \subset T$) and Fourier magnitude measurements $\{y[\omega]\}_{\omega \in \Omega}$ with $y[\omega] = |\hat{x}[\omega]|$, the method works as follows:

1. **I**nitialization: Choose an initial guess x_0 and set $z_0[\omega] = y[\omega]\frac{\hat{x}_0[\omega]}{|\hat{x}_0[\omega]|}$ for $\omega \in \Omega$.

2. **L**oop: For $k = 1, 2, \ldots$ inductively define

$$(1) \qquad x_k[t] = \begin{cases} (F^{-1}z_{k-1})[t] & \text{if } t \in T, \\ 0 & \text{else;} \end{cases}$$

$$(2) \qquad z_k[\omega] = y[\omega]\frac{\hat{x}_k[\omega]}{|\hat{x}_k[\omega]|} \qquad \text{for } \omega \in \Omega$$

   until convergence.

While this algorithm is simple to implement and amenable to additional constraints such as the positivity of x, no convergence guarantees are known and in fact the limit set of the iterations often depends non-trivially on the starting point. While projection algorithms onto convex sets are well understood [15, 41, 88, 6], the set $\{z : |\hat{z}[\omega]| = |\hat{x}[\omega]|\}$ is not convex and, therefore, the algorithm is not known to converge in general to the unique solution if there is one [53, 6, 58]. Some progress toward understanding the convergence behavior of certain alternating projection methods has been made in [57]. Good numerical results have been reported in certain oversampling settings, but they appear to be nevertheless somewhat problematic in light of our numerical experiments from Section 3.8. [60] points out that oversampling is not always practically feasible as certain experimental geometries allow only for sub-Nyquist sampling; an example is the Bragg sampling from periodic crystalline structures. Alternating projection algorithms may be more competitive when applied in the framework of multiple structured illuminations, as proposed in the PhaseLift methodology, instead of oversampling. Another direction of investigation is to utilize sparsity of the signal, see [60, 56, 87]. Here, the signal is known to have only a few non-zero coefficients, but the locations of the non-zero coefficients (that is, the support of the signal) are not known a priori.

In a different direction, a frame-theoretic approach to phase retrieval has been proposed in [3, 5], where the authors derive various necessary and sufficient conditions for the uniqueness of the solution, as well as various numerical algorithms. The practical applicability of these results is limited by the specific types of measurements required.

Chapter 2

The PhaseLift Methodology

2.1 PhaseLift – a novel methodology

This chapter develops a novel methodology for phase retrieval based on a rigorous and flexible numerical framework. Whereas most of the existing methods seek to overcome non-uniqueness by imposing additional constraints on the signal, PhaseLift employs different techniques. There are two main components to this approach.

- *Multiple structured illuminations.* We suggest collecting several diffraction patterns providing 'different views' of the sample, for instance, by modulating the light beam falling onto the sample or by placing a mask right after the sample, see Section 3.2 for details. Taking multiple diffraction patterns usually yields uniqueness as discussed in Section 2.4.

  The concept of using multiple measurements as an attempt to resolve the phase ambiguity for diffraction imaging was suggested in [68]. A variety of methods have been proposed to carry out these multiple measurements; various gratings and/or of masks, rotation of the axial position of the sample, and use of defocusing implemented by a spatial light modulator, see [1]. Other approaches include *ptychography*,where one records several diffraction patterns from overlapping areas of the sample, [77, 83].

- *Formulation of phase recovery as a matrix completion problem.* We suggest (1) lifting up the problem of recovering a vector from quadratic constraints into that of recovering a rank-one matrix from affine constraints, and (2) relaxing the combinatorial problem into a convex program. The price we pay for trading the nonconvex quadratic constraints into convex constraints is that we must deal with a highly underdetermined problem. However, recent advances in the areas of compressive sensing and matrix completion have shown that such convex approximations are often exact.

  Although PhaseLift is a new approach for phase retrieval, the idea of solving problems involving nonconvex quadratic constraints by semidefinite relaxations has a long history in optimization, see [10].

This chapter demonstrates that taken together, multiple coded illuminations and convex programming provide an effective approach to phase retrieval. Also, PhaseLift offers a principled way of dealing with noise, simplifying the use of various statistical noise models. This is important

because in practice, measurements are always noisy. In fact, this framework can be used to formulate a regularized maximum likelihood method. Lastly, the framework can also include a priori knowledge about the signal that can be formulated or relaxed as convex constraints.

2.2 Precedents

At the abstract level, the phase retrieval problem is that of finding $x \in \mathbb{C}^n$ obeying quadratic equations of the form $|\langle a_k, x\rangle|^2 = b_k$. Casting such quadratic constraints as affine constraints about the matrix variable $X = xx^\star$ has been widely used for finding good bounds on a number of quadratically constrained quadratic problems (QCQP). Indeed, solving the general case of a QCQP is known to be NP-hard since it includes the family of Boolean linear programs [10]. The approach usually consists in finding a relaxation of the QCQP using semidefinite programming (SDP). An important example of this strategy is Max Cut, an NP-hard problem in graph theory which can be formulated as a QCQP. In a celebrated paper, Goemans and Williamson introduced a relaxation [38] for this problem, which lifts a nonlinear, nonconvex problem to the space of symmetric matrices.

The idea of linearizing the phase retrieval problem by reformulating it as a problem of recovering a matrix from linear measurements can be found in [5]. While this reference also contains some intriguing numerical recovery algorithms, their practical relevance for most applications is limited by the fact that the proposed measurement matrices either require a very specific algebraic structure which does not seem to be compatible with the physical properties of diffraction, or the number of measurements is proportional to the square of the signal dimension, which is not feasible in most applications.

In terms of framework, the closest approach is the paper [24], in which the authors use a matrix completion approach for array imaging from intensity measurements. Although PhaseLift executes a similar relaxation, there are some differences. We present a "noise-aware" framework, which makes it possible to account for a variety of noise models in a systematic way. Moreover, our emphasis is on a novel combination of structured illuminations and convex programming.

2.3 Methodology

Structured illumination

Suppose $x = \{x[t]\}$ is the object of interest (t may be a one- or multi-dimensional index). We shall discuss illumination schemes collecting the diffraction pattern of the modulated object $w[t]x[t]$, where the waveforms or patterns $w[t]$ may be selected by the user. There are many ways in which this can be implemented in practice, and we discuss just a few of those.

- *Masking.* One possibility is to modify the phase front after the sample by inserting a mask or a phase plate, see [54] for example. A schematic layout is shown in Figure 2.1. In [49], the sample is scanned by shifting the phase plate as in ptychography (discussed below); the difference is that one scans the known phase plate rather than the object being imaged.

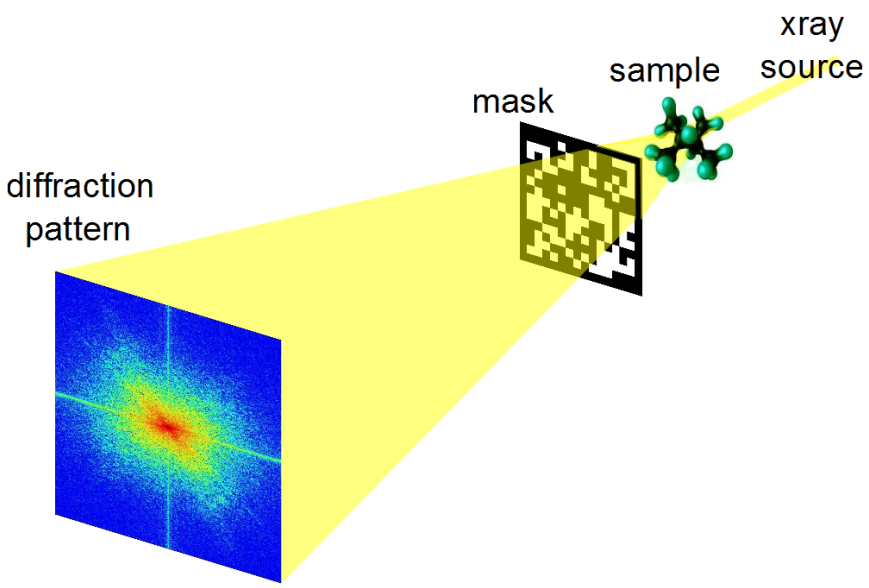

Figure 2.1: A typical setup for structured illuminations in diffraction imaging using a phase mask.

- *Optical grating.* Another standard approach would be to change the profile or modulate the illuminating beam, which can easily be accomplished by the use of optical gratings [55]. A simplified representation would look similar to the scheme depicted in Figure 2.1, with a grating instead of the mask (the grating could be placed before or after the sample).

- *Ptychography.* Here, one measures multiple diffraction patterns by scanning a finite illumination on an extended specimen [77, 83]. In this setup, it is common to maintain a substantial overlap between adjacent illumination positions.

- *Oblique illuminations.* One can use illuminating beams hitting the sample at user specified angle [31], see Figure 2.2 for a schematic illustration of this approach. One can also imagine having multiple simultaneous oblique illuminations.

As is clear, there is no shortage of options and one might prefer solutions which require generating as few diffraction patterns as possible for stable recovery.

Lifting

Suppose we have $x_0 \in \mathbb{C}^n$ or $\mathbb{C}^{n_1 \times n_2}$ (or some higher-dimensional version) about which we have quadratic measurements of the form

$$\mathbb{A}(x_0) = \{|\langle a_k, x_0 \rangle|^2 : k = 1, 2, \ldots, m\}. \tag{2.3.1}$$

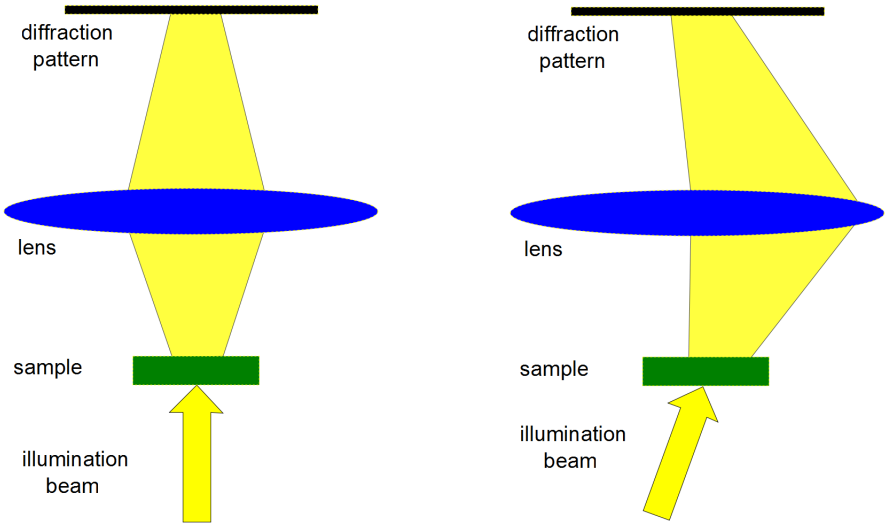

Figure 2.2: A typical setup for structured illuminations in diffraction imaging using oblique illuminations. The left image shows direct (on-axis) illumination and the right image corresponds to oblique (off-axis) illumination.

In the setting where we would collect the diffraction pattern of $w[t]x_0[t]$ as discussed earlier, then the waveform $a_k[t]$ can be written as

$$a_k[t] \propto w[t]e^{i2\pi \langle \omega_k, t \rangle}; \tag{2.3.2}$$

here, ω_k is a frequency value so that $a_k[t]$ is a patterned complex sinusoid. One can assume for convenience that the normalizing constant is such that a_k is unit normed, i.e. $\|a_k\|_2^2 = \sum_t |a_k[t]|^2 = 1$. Phase retrieval is then the feasibility problem

$$\begin{array}{ll} \text{find} & x \\ \text{obeying} & \mathbb{A}(x) = \mathbb{A}(x_0) := b. \end{array} \tag{2.3.3}$$

As is well known, quadratic measurements can be lifted up and interpreted as linear measurements about the rank-one matrix $X = xx^*$. Indeed,

$$|\langle a_k, x \rangle|^2 = \mathrm{Tr}(x^* a_k a_k^* x) = \mathrm{Tr}(a_k a_k^* x x^*) := \mathrm{Tr}(A_k X),$$

where A_k is the rank-one matrix $a_k a_k^*$. In what follows, we will let $\mathcal{A}$ be the linear operator mapping positive semidefinite matrices X into $\{\mathrm{Tr}(A_k X) : k = 1, \ldots, m\}$. Hence, the phase retrieval problem is equivalent to

$$\begin{array}{ll} \text{find} & X \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0 \\ & \mathrm{rank}(X) = 1 \end{array} \quad \Leftrightarrow \quad \begin{array}{ll} \text{minimize} & \mathrm{rank}(X) \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0. \end{array} \tag{2.3.4}$$

Upon solving the left-hand side of (3.2.4), we would factorize the rank-one solution X as xx^*, hence finding solutions to the phase-retrieval problem. Note that the equivalence between the left- and right-hand side of (3.2.4) is straightforward since by definition, $b = \mathbb{A}(x_0) = \mathcal{A}(x_0 x_0^*)$ and there exists a rank-one solution. Therefore, our problem is a rank-minimization problem over an affine slice of the positive semidefinite cone. As such, it falls in the realm of low-rank *matrix completion* or *matrix recovery*, a class of optimization problems that has gained tremendous attention in recent years, see e.g. [74, 19, 20]. Just as in matrix completion, the linear system $\mathcal{A}(X) = b$, with unknown in the positive semidefinite cone, is highly underdetermined. For instance suppose our signal x_0 has n complex unknowns. Then we may imagine collecting six diffraction patterns with n measurements for each (no oversampling). Thus $m = 6n$ whereas the dimension of the space of $n \times n$ Hermitian matrices over the reals is n^2, which is obviously much larger.

We are of course interested in low-rank solutions and this makes the search feasible. This also raises an important question: what is the minimal number of diffraction patterns needed to recover x, whatever x may be? Since each pattern yields n *real-valued* coefficients and x has n *complex-valued* unknowns, clearly at least two are needed. Further, in the context of quantum state tomography, Theorem II in [36] shows one needs at least $3n - 2$ intensity measurements to guarantee uniqueness, hence suggesting an absolute minimum of three diffraction patterns. Are three patterns sufficient? We partly address this question in Section 2.4.

Recovery via convex programming

The rank-minimization problem (3.2.4) is NP hard. We propose using the trace norm as a convex surrogate [8, 62] for the rank functional, giving the familiar SDP (and a crucial component of PhaseLift),

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(X) \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0; \end{array} \tag{2.3.5}$$

here and below $X \succeq 0$ means that X is Hermitian positive semidefinite. This problem is convex and there exists a wide array of numerical solvers including the popular Nesterov's accelerated first order method [69]. As far as the relationship between (3.2.4) and (3.2.5) is concerned, the matrix $\mathcal{A}$ in most diffraction imaging applications is not known to obey any of the conditions derived in the literature [19, 20, 74] that would guarantee a formal equivalence between the two programs. Nevertheless, the formulation (3.2.5) enjoys great empirical performance as demonstrated in Section 3.8. Furthermore, as shown in chapter 2, if measurement vectors a_k sampled independently and uniformly at random on the unit sphere, PhaseLift can recover x exactly (up to a global phase factor) with high probability, provided that the number of measurements is on the order of $n \log n$.

We mentioned earlier that measurements are typically noisy and that our formulation allows for a principled approach to deal with this issue for a variety of noise models. Suppose the measurement vector $\{b_k\}$ is sampled from a probability distribution $p(\cdot; \mu)$, where $\mu = \mathbb{A}(x_0)$ is the vector of noiseless values, $\mu_k = |\langle a_k, x_0 \rangle|^2$. Then a classical fitting approach simply consists of maximizing the likelihood,

$$\begin{array}{ll} \text{maximize} & p(b; \mu) \\ \text{subject to} & \mu = \mathbb{A}(x) \end{array} \tag{2.3.6}$$

with optimization variables μ and x. (A more concise description is to find x such that $p(b; \mathbb{A}(x))$ is maximum.) Using the lifting technique and the monotonicity of the logarithm, an equivalent formulation is

$$\begin{array}{ll} \text{minimize} & -\log(p(b;\mu)) \\ \text{subject to} & \mu = \mathcal{A}(X) \\ & X \succeq 0,\ \text{rank}(X) = 1. \end{array}$$

This is, of course, not tractable and our convex formulation suggests solving instead

$$\begin{array}{ll} \text{minimize} & -\log p(b;\mu) + \lambda \operatorname{Tr}(X) \\ \text{subject to} & \mu = \mathcal{A}(X) \\ & X \succeq 0 \end{array} \tag{2.3.7}$$

with optimization variables μ and X (in other words, find $X \succeq 0$ such that $-\log p(b; \mathcal{A}(X)) + \lambda \operatorname{Tr}(X)$ is minimum). Above, λ is a positive scalar and, hence, our approach is a penalized or regularized maximum likelihood method, which trades off between goodness and complexity of the fit. When the likelihood is log-concave, problem (2.3.7) is convex and solvable. We give two examples for concreteness:

- *Poisson data.* Suppose that $\{b_k\}$ is a sequence of independent samples from the Poisson distributions $\text{Poi}(\mu_k)$. The Poisson log-likelihood for independent samples has the form $\sum_k b_k \log \mu_k - \mu_k$ (up to an additive constant factor) and thus, our problem becomes

$$\begin{array}{ll} \text{minimize} & \sum_k [\mu_k - b_k \log \mu_k] + \lambda \operatorname{Tr}(X) \\ \text{subject to} & \mu = \mathcal{A}(X) \\ & X \succeq 0. \end{array}$$

- *Gaussian data.* Suppose that $\{b_k\}$ is a sequence of independent samples from the Gaussian distribution with mean μ_k and variance σ_k^2 (or is well approximated by Gaussian variables). Then our problem becomes

$$\begin{array}{ll} \text{minimize} & \sum_k \frac{1}{2\sigma_k^2} (b_k - \mu_k)^2 + \lambda \operatorname{Tr}(X) \\ \text{subject to} & \mu = \mathcal{A}(X) \\ & X \succeq 0. \end{array}$$

  If Σ is a diagonal matrix with diagonal elements σ_k^2, this can be written as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}[b - \mathcal{A}(X)]^* \Sigma^{-1} [b - \mathcal{A}(X)] + \lambda \operatorname{Tr}(X) \\ \text{subject to} & X \succeq 0. \end{array}$$

Both formulations are of course convex and in both cases, one recovers the noiseless trace minimization problem (3.2.5) as $\lambda \to 0^+$.

In addition, it is straightforward to include further constraints frequently discussed in the phase retrieval literature such as real-valuedness, positivity, atomicity and so on. Suppose the support of x is known to be included in a set T known a priori. Then we would add the linear constraint

$$X_{ij} = 0, \quad (i,j) \notin T \times T.$$

(Algorithmically, one would simply work with a reduced-size matrix.) Suppose we would like to enforce real-valuedness, then we simply assume that X is real valued and positive semidefinite. Finally positivity can be expressed as linear inequalities

$$X_{ij} \geq 0.$$

Of course, many other types of constraints can be incorporated in this framework, which provides appreciable flexibility.

PhaseLift with reweighting

The trace norm promotes low-rank solutions and this is why it is often used as a convex proxy for the rank. However, it is possible to further promote low-rank solutions by solving a sequence of weighted trace-norm problems, a technique which has been shown to provide even more accurate solutions [33, 21]. The reweighting scheme works like this: choose $\varepsilon > 0$; start with $W_0 = I$ and for $k = 0, 1, \ldots,$ inductively define X_k as the optimal solution to

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(W_k X) \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0 \end{array} \tag{2.3.8}$$

and update the 'weight matrix' as

$$W_{k+1} = (X_k + \varepsilon I)^{-1}.$$

The algorithm terminates on convergence or when the iteration count k attains a specified maximum number of iterations k_{max}. One can see that the first step of this procedure is precisely (3.2.5); after this initial step, the algorithm proceeds in solving a sequence of trace-norm problems in which the matrix weights W_k are roughly the inverse of the current guess.

As explained in the literature [33, 32], this reweighting scheme can be viewed as attempting to solve

$$\begin{array}{ll} \text{minimize} & f(X) = \log(\det(X + \varepsilon I)) \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0 \end{array} \tag{2.3.9}$$

by minimizing the tangent approximation to f at each iterate; that is to say, at step k, (2.3.8) is equivalent to minimizing $f(X_{k-1}) + \langle \nabla f(X_{k-1}), X - X_{k-1} \rangle$ over the feasible set. (As for the trace, the function $\log\det(X + \varepsilon I)$ serves as a surrogate for the rank functionial.) This can also be applied to noise-aware variants where one would simply replace the objective functional in (2.3.7) with

$$-\log p(b; \mu) + \lambda \operatorname{Tr}(W_k X),$$

at each step, and update W_k in exactly the same way as before.

2.4 Preliminary Theory

The PhaseLift framework poses two main theoretical questions:

1. When do multiple diffracted images imply unicity of the solution?
2. When does the convex heuristic succeed in recovering the unique solution to the phase-retrieval problem?

Developing comprehensive answers to these questions constitutes a whole research program, and we will address the second question under some measurement assumptions in the next chapter. In this chapter, we shall limit ourselves to introducing some theoretical results showing simple ways of designing diffraction patterns, which give unicity. Our focus is on getting uniqueness from a very limited number of diffraction patterns. For example, we shall demonstrate that in some cases three diffraction images are sufficient for perfect recovery. Thus, we give below partial answers to the first question and will begin to address the second in the next chapter.

A frequently discussed approach to retrieve phase information uses a technique from holography. Roughly speaking, the idea is to let the signal of interest x interfere with a known reference beam y. One typically measures $|x + y|^2$ and $|x - iy|^2$ and precise knowledge of y allows, in principle, to recover the amplitude and phase of x. Holographic techniques are hard to implement [1] in practice. Instead, we propose using a modulated version of the signal itself as a reference beam which in some cases may be easier to implement.

To discuss this idea, we need to introduce some notation. For a complex signal $z \in \mathbb{C}^n$, we let $|z|^2$ be the nonnegative real-valued n-dimensional vector containing the squared magnitudes of z. Suppose first that x is a one-dimensional signal $(x[0], x[1], \ldots, x[n-1])$ and that F_n is the $n \times n$ unitary DFT. In this section, we consider taking $3n$ real-valued measurements of the form

$$\mathbb{A}(x) = \{|F_n x|^2, |F_n(x + D^s x)|^2, |F_n(x - iD^s x)|^2\}, \tag{2.4.1}$$

where D is the modulation

$$D = \mathrm{diag}(\{e^{i2\pi t/n}\}_{0 \le t \le n-1}).$$

and s is a nonnegative integer. These measurements can be obtained by illuminating the sample with the three light fields 1, $1 + e^{i2\pi st/n}$ and $1 + e^{i2\pi(st/n - 1/4)}$. We show below that these $3n$ measurements are generally sufficient for perfect recovery.

Theorem 2.4.1 *Suppose the DFT of $x \in C^n$ does not vanish. Then x can be recovered up to global phase from the $3n$ real numbers $\mathbb{A}(x)$ (3.2.1) if and only if $\gcd(s, n) = 1$. In particular, assuming $\gcd(s, n) = 1$, if the trace-minimization program (3.2.5) or the iteratively reweighted algorithm return a rank-1 solution, then this solution is exact.*

Conversely, if the DFT vanishes at two frequency points k and k' obeying $k - k' \neq s \bmod n$, then recovery is not possible from the $3n$ real numbers (3.2.1).

The proof of this theorem is constructive and we give a simple algorithm that achieves perfect reconstruction. Further, one can use masks to scramble the Fourier transform as to make sure it

does not vanish. Suppose for instance that we collect

$$\mathbb{A}(Wx), \quad W = \text{diag}(\{z[t]\}_{0\le t\le n-1}).$$

where the $z[t]$'s are iid $\mathcal{N}(0,1)$. Then since the Fourier transform of $z[t]x[t]$ does not vanish with probability one, we have the following corollary.

Corollary 2.4.2 *Assume that* $\gcd(s,n) = 1$. *Then with probability one,* x *can be recovered up to global phase from the* $3n$ *real numbers* $\mathbb{A}(Wx)$ *where* W *is the diagonal matrix with Gaussian entries above.*

Of course, one could derive similar results by scrambling the Fourier transform with the aid of other types of masks, e.g. binary masks.

We now turn our attention to the situation in higher dimensions and will consider the 2D case (higher dimensions are treated in the same way). Here, we have a discrete signal $x[t_1,t_2] \in \mathbb{C}^{n_1\times n_2}$ about which we take the $3n_1n_2$ measurements

$$\{|\mathcal{F}_{n_1\times n_2}x|^2, |\mathcal{F}_{n_1\times n_2}(x+\mathcal{D}^s x)|^2, |\mathcal{F}_{n_1\times n_2}(x-i\mathcal{D}^s x)|^2\}, \quad s=(s_1,s_2); \tag{2.4.2}$$

$\mathcal{F}_{n_1\times n_2}$ is the 2D unitary Fourier transform defined by (1.3.1) in which the frequencies belong to the 2D grid $\{0,1,\dots,n_1-1\}\times\{0,1,\dots,n_2-1\}$, s is a pair of nonnegative integers and $\mathcal{D}^s$ is the modulation

$$[\mathcal{D}^s x][t_1,t_2] = e^{i2\pi s_1 t_1/n_1}\, e^{i2\pi s_2 t_2/n_1} x[t_1,t_2].$$

With these definitions, we have the following result:

Theorem 2.4.3 *Suppose the DFT of* $x \in C^{n_1\times n_2}$ *does not vanish. Then* x *can be recovered up to global phase from the* $3n_1n_2$ *real numbers* (2.4.2) *if and only if* $\gcd(s_1,n_1)=1$, $\gcd(s_2,n_2)=1$ *and* $\gcd(n_1,n_2)=1$. *Under these assumptions, if the trace-minimization program* (3.2.5) *or the iteratively reweighted algorithm return a rank-1 solution, then this solution is exact.*

Again, one can apply a random mask to turn this statement into a probabilistic statement holding either with probability one or with very large probability depending upon the mask that is used.

One can always choose s_1 and s_2 such that they be relatively prime to n_1 and n_2 respectively. The last condition may be less friendly but one can decide to pad one dimension with zeros to guarantee primality. This is equivalent to a slight oversampling of the DFT along one direction. An alternative is to take $5n_1n_2$ measurements in which we modulate the signal horizontally and then vertically; that is to say, we modulate with $s=(s_1,0)$ and then with $s=(0,s_2)$. These $5n_1n_2$ measurements guarantee recovery if s_1 is relatively prime to n_1 and s_2 is relatively prime to n_2 for all sizes n_1 and n_2, see Section 2.4 for details.

Proof of Theorem 2.4.1

Let $\hat{x} = (\hat{x}[0], \ldots, \hat{x}[n-1])$ be the DFT of x. Then knowledge of $\mathbb{A}(x)$ is equivalent to knowledge of

$$|\hat{x}[k]|^2, |\hat{x}[k] + \hat{x}[k-s]|^2, \text{ and } |\hat{x}[k] - i\hat{x}[k-s]|^2$$

for all $k \in \{0, 1, \ldots, n-1\}$ (above, $k - s$ is understood mod n). Write $\hat{x}[k] = |\hat{x}[k]|e^{i\phi[k]}$ so that $\phi[k]$ is the missing phase, and observe that

$$|\hat{x}[k] + \hat{x}[k-s]|^2 = |\hat{x}[k]|^2 + |\hat{x}[k-s]|^2 + 2|\hat{x}[k]||\hat{x}[k-s]|\text{Re}(e^{i(\phi[k-s]-\phi[k]}$$
$$|\hat{x}[k] - i\hat{x}[k-s]|^2 = |\hat{x}[k]|^2 + |\hat{x}[k-s]|^2 + 2|\hat{x}[k]||\hat{x}[k-s]|\text{Im}(e^{i(\phi[k-s]-\phi[k]}).$$

Hence, if $\hat{x}[k] \neq 0$ for all $k \in \{0, 1, \ldots, n-1\}$, our data gives us knowledge of all phase shifts of the form

$$\phi[k-s] - \phi[k], \quad k = 0, 1, \ldots, n-1.$$

We can, therefore, initialize $\phi[0]$ to be zero and then get the values of $\phi[-s]$, $\phi[-2s]$ and so on.

This process can be represented as a cycle in the group $\mathbb{Z}/n\mathbb{Z}$ as the sequence $(0, -s, -2s, \ldots)$. We would like this cycle to contain n unique elements, which is true if and only if the cyclic subgroup $(0, s, 2s, \ldots)$ has order n. This is equivalent to requiring $\gcd(s, n) = 1$. If this subgroup has a smaller order, then recovery is impossible since we finish the cycle before we have all the phases; the phases that we are able to recover do not enable us to determine any more phases without making further assumptions.

For the second part of the theorem, assume without loss of generality, that $s = -1$ and that $(k, k') = (0, k_0)$ $(1 < k_0 < n-1)$. For simplicity suppose these are the only zeros of the DFT. This creates two disjoint sets of frequency indices: those for which $0 < k < k_0$ and those for which $k_0 < k \leq n-1$. We are given no information about the phase difference between elements of these two subgroups, and hence recovery is not possible. This argument extends to situations where the DFT vanishes more often, in which case, we have even more indeterminacy.

Proof of Theorem 2.4.3

Let $\hat{x} = \{\hat{x}[k_1, k_2]\}$, where $(k_1, k_2) \in \{0, 1, \ldots n_1 - 1\} \times \{0, 1, \ldots, n_2 - 1\}$ be the DFT of x. Then we have knowledge of

$$|\hat{x}[k_1, k_2]|^2, |\hat{x}[k_1, k_2] + \hat{x}[k_1 - s_s, k_2 - s_2]|^2, \text{ and } |\hat{x}[k_1, k_2] - i\hat{x}[k_1 - s_1, k_2 - s_2]|^2$$

for all $(k_1 k_2)$. With the same notations as before, this gives knowledge of all phase shifts of the form

$$\phi[k_1 - s_1, k_2 - s_2] - \phi[k_1, k_2], \quad 0 \leq k_1 \leq n_1, 0 \leq k_2 \leq n_2 - 1.$$

Hence, we can initialize $\phi[0, 0]$ to be zero and then get the values of $\phi[-s_1, -s_2]$, $\phi[-2s_1, -2s_2]$ and so on. The argument is as before: we would like the cyclic subgroup $((0, 0), (s_1, s_2), (2s_1, 2s_2),$ ...) in $\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z}$ to have order $n_1 n_2$. Now the order of an element $(s_1, s_2) \in \mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z}$ is equal to

$$\text{lcm}(|s_1|, |s_2|) = \text{lcm}(n_1/\gcd(n_1, s_1), n_2/\gcd(n_2, s_2)),$$

where $|s_1|$ is the order of s_1 in $\mathbb{Z}/n_1\mathbb{Z}$ and likewise for $|s_2|$. Noting that $\text{lcm}(a,b) \leq ab$ and that equality is achieved if and only if $\gcd(a,b) = 1$, we must simultaneously have

$$\gcd(s_1, n_1) = 1, \quad \gcd(s_2, n_2) = 1 \quad \text{and } \gcd(n_1, n_2)) = 1$$

to have uniqueness.

Extensions

It is clear from our analysis that if we were to collect $|\mathcal{F}_{n_1\times n_2}x|^2$ together with

$$\{|\mathcal{F}_{n_1\times n_2}(x + \mathcal{D}^{s_k}x)|^2, |\mathcal{F}_{n_1\times n_2}(x - i\mathcal{D}^{s_k}x)|^2\}, \quad k = 1, \ldots, K,$$

so that one collects $(2K+1)n_1n_2$ measurements, then 2D recovery is possible if and only if $\{s_1, \ldots, s_K\}$ generates $\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z}$ (and the Fourier transform has no nonzero components). This can be understood by analyzing the generators of the group $\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z}$.

A simple instance consists in choosing one modulation pattern to be $(s_1, 0)$ and another to be $(0, s_2)$. If s_1 is relatively prime to n_1 and s_2 is relatively prime to n_2, these two modulations generate the whole group regardless of the relationship between n_1 and n_2. An algorithmic way to see this is as follows. Initialize $\phi(0,0)$. Then by using horizontal modulations, one recovers all phases of the form $\phi(k_1, 0)$. Further, by using vertical modulations (starting with $\phi(k_1, 0)$), one can recover all phases of the form $\phi(k_1, k_2)$ by moving upward.

2.5 Empirical Performance

This section introduces numerical simulations to illustrate and study the effectiveness of PhaseLift.

Numerical solvers

All numerical algorithms were implemented in Matlab using TFOCS [9] as well as modifications of TFOCS template files. TFOCS is a library of Matlab-files designed to facilitate the construction of first-order methods for a variety of convex optimization problems, which include those we consider.

In a nutshell, suppose we wish to solve the problem

$$\begin{array}{ll} \text{minimize} & g(X) := -\ell(b; \mathcal{A}(X)) + \lambda \operatorname{Tr}(X) \\ \text{subject to} & X \succeq 0 \end{array} \tag{2.5.1}$$

in which $\ell(b; \mathcal{A}(X))$ is a smooth and concave (in X) log-likelihood. Then a projected gradient method would start with an initial guess X_0, and inductively define

$$X_k = \mathcal{P}(X_{k-1} - t_k \nabla g(X_{k-1})), \tag{2.5.2}$$

where $\{t_k\}$ is a sequence of stepsize rules and $\mathcal{P}$ is the projection onto the positive semidefinite cone. (Various stepsize rules are typically considered including fixed stepsizes, backtracking line search, exact line search and so on.)

TFOCS implements a variety of accelerated first-order methods pioneered by Nesterov, see [69] and references therein. One variant [7] works as follows. Choose X_0, set $Y_0 = X_0$ and $\theta_0 = 1$, and inductively define

$$\begin{aligned} X_k &= \mathcal{P}(Y_{k-1} - t_k \nabla g(Y_{k-1})) \\ \theta_k &= 2\Big[1 + \sqrt{1 + 4/\theta_{k-1}^2}\Big]^{-1} \\ \beta_k &= \theta_k(\theta_{k-1}^{-1} - 1) \\ Y_k &= X_k + \beta_k(X_k - X_{k-1}), \end{aligned} \tag{2.5.3}$$

where $\{t_k\}$ is a sequence of stepsize rules as before. The sequence $\{\theta_k\}$ is usually referred to as a sequence of accelerated parameters, and $\{Y_k\}$ is an auxiliary sequence at which the gradient is to be evaluated. The advantage of this approach is that the computational work per iteration is as in the projected gradient method but the number of iterations needed to reach a certain accuracy is usually much lower [69]. TFOCS implements such iterations and others like it but with various improvements.

For large problems, e.g. images with a large number N of pixels, it is costly to hold the $N \times N$ optimization variable X in memory. To overcome this issue, our computational approach maintains a low-rank factorization of X. This is achieved by substituting the projection onto the semidefinite cone (the expensive step) with a proxy. Whereas $\mathcal{P}$ dumps the negative eigenvalues as in

$$\mathcal{P}(X) = \sum_i \max(\lambda_i, 0) u_i u_i^*,$$

where $\sum_i \lambda_i u_i u_i^*$ ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$) is any eigenvalue decomposition of X, our proxy only keeps the k largest eigenvalues in the expansion as in

$$\mathcal{P}_k(X) = \sum_{i \leq k} \max(\lambda_i, 0) u_i u_i^*. \tag{2.5.4}$$

For small values of k—we use k between 10 and 20—this can be efficiently computed since we only need to compute the top eigenvectors of a low-rank matrix at each step. Although this approximation gives good empirical results, convergence is no longer guaranteed. For a method like (2.5.2) or (2.5.3), the main computational cost of single iteration is dominated by computing (2.5.4) whose complexity is in turn governed by the costs of applying $\mathcal{A}$ and $\mathcal{A}^*$. By maintaining a low-rank factorization of X or Y, these costs are on the order of $k \times M \times n \log n$ for $x \in \mathbb{C}^n$, where M is the number of illuminations. Roughly, each iteration costs on the order of $k \times M$ FFTs.

Error measures

To measure performance, we will use the mean-square error (MSE). However, since a solution x_0 is only unique up to global phase, it makes no sense to compute the squared distance between x_0 and the recovery $\hat{x}_0$. Rather, we compute the distance to the solution space, i.e. we are interested in the relative MSE defined as

$$\min_{c:|c|=1} \frac{\|cx_0 - \hat{x}_0\|_2^2}{\|x_0\|_2^2}.$$

This is the definition we will adopt throughout the paper;[1] the Signal-to-Noise Ratio (SNR) of the measured data is defined as SNR $= 10 \log_{10} \|b - \tilde{b}\|_2^2 / \|b\|_2^2$, where $\tilde{b}$ denotes the noisy data.

Although our algorithm favors low-rank solutions, it is not guaranteed to find a rank-one solution. Therefore, if our optimal solution $\hat{X}_0$ does not have exactly rank one, we extract the rank-one approximation $\hat{x}_0 \hat{x}_0^*$ where $\hat{x}_0$ is an eigenvector associated with the largest eigenvalue. We use a scaling such that $\|\hat{x}_0\|_2^2 = \|x_0\|_2^2$. Note that the ℓ_2 norm of the true solution is generally known since by Parseval's theorem, the ℓ_2 norm of Fx_0 is equal to $\|x_0\|_2$. Hence, observing the diffraction pattern of the object x_0 reveals its squared ℓ_2 norm.

Alternating projections

For comparison, we will also apply an alternating projection algorithm in some of the experiments. To describe this algorithm, put $Ax := \{\langle a_j, x\rangle\}_{j=1}^m$ in which the a_j's are as in (2.3.2) so that $\mathbb{A}(x) = |Ax|^2$. In the setting of multiple illuminations, the alternating projection algorithm consists of the following steps: (1) choose an initial guess x_0; (2) compute $b_0 = Ax_0$ and for $k = 0, 1, \ldots,$

(i) adjust the modulus of b_k so that it fits the measurements b,

$$\tilde{b}_k[i] = b[i] \frac{b_k[i]}{|b_k[i]|}, \quad i = 1, \ldots, m.$$

(ii) Reproject $\tilde{b}_k$ onto the range of A,

$$x_{k+1} = \operatorname{argmin} \|Ax - \tilde{b}_k\|_2,$$
$$b_{k+1} = Ax_{k+1}.$$

Observe that we can incorporate appropriate additional information about x (such as positivity for example) via a suitable modification of the projection step (ii).

1-D simulations

Phase retrieval for one-dimensional signals arises in fiber optics [23, 47, 46], terahertz communications [51], speech recognition [73], as well as in the determination of concentration profiles and the detection of planar disorder in diffraction imaging [2, 84]. We evaluate PhaseLift for noiseless and noisy data using a variety of different 'illuminations' and test signals.

Noisefree measurements

In the first set of experiments we demonstrate the recovery of two very different signals from noiseless data. Both test signals are of length $n = 128$. The first signal, shown in Figure 2.3(a)) is a linear combination of a few sinusoids and represents a typical transfer function one might encounter in optics. The second signal is a complex signal, with independent Gaussian complex

[1] Alternatively, we could use $\|x_0 x_0^* - \hat{x}_0 \hat{x}_0^*\|_F / \|x_0 x_0^*\|_F$, which gives very similar values.

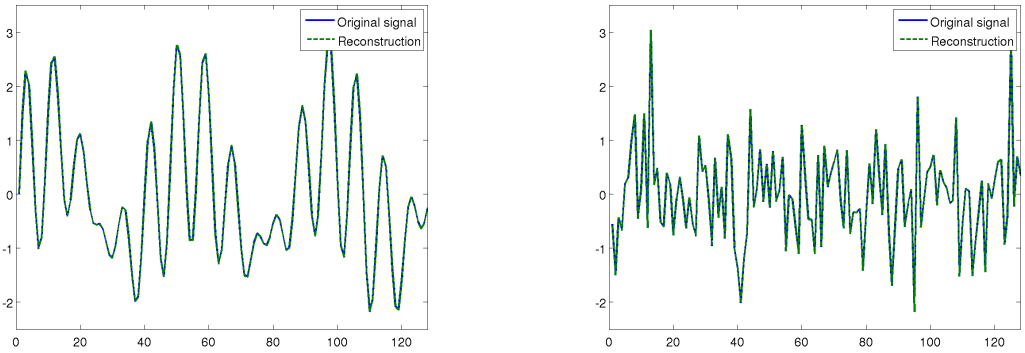

Figure 2.3: Two test signals and their reconstructions. The recovered signals are essentially indistinguishable from the originals. Left figure is smooth signal and its reconstruction. Right figure is random signal and its reconstruction.

entries (each entry is of the form $a + ib$ where a and b are independent $\mathcal{N}(0, 1)$ variables) so that the real and imaginary parts are independent white noise sequences; the real part of the signal is shown in Figure 2.3(b).

Four random binary masks are used to perform the structured illumination so that we measure $|Ax|^2$ in which

$$A = F \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix},$$

where each W_i is diagonal with either 0 or 1 on the diagonal, resulting in a total of 512 intensity measurements. We work with the objective functional $\frac{1}{2}\|b - \mathcal{A}(X)\|_2^2 + \lambda \operatorname{Tr}(X)$ and the constraint $X \succeq 0$ to recover the signal, in which we use a small value for λ such as 0.05 since we are dealing with noisefree data. We apply the reweighting scheme as discussed in Section 2.3. (To achieve perfect reconstruction, one would have to let $\lambda \to 0$ as the iteration count increases.) The algorithm is terminated when the relative *residual error* is less than a fixed tolerance, namely, $\|\mathcal{A}(\hat{x}_0\hat{x}_0^*) - b\|_2 \leq 10^{-6}\|b\|_2$, where $\hat{x}_0$ is the reconstructed signal just as before. The original and recovered signals are plotted in Figure 2.3(a) and (b). The MSE on a dB-scale (i.e., $10 \log_{10}$(MSE)) is 87.3dB in the first case and 90.5dB in the second.

We have repeated these experiments with the same test signals and the same algorithm, but using Gaussian masks instead of binary masks. In other words, the W_i's have Gaussian entries on the diagonal. It turns out that in this case, three illuminations—instead of four—were sufficient to obtain similar performance. Furthermore, we point out that no re-weighting was needed, when we used six or more Gaussian masks. Expressed differently, plain trace-norm minimization succeeds with $6n$ or more intensity measurements of this kind.

We also applied the alternating projection algorithm of Section 2.5, with random initial guess, to the examples above. When using three Gaussian masks (or the three operators I, F, FW related

to the quantum mechanical setting), alternating projections always failed. It never found the correct solution or even an approximation with a relative MSE less than 1. With four illuminations the alternating projections algorithm computed the correct solution in about 40% of the experiments, and returned a relative MSE larger than 1 in the other 60%. As we increase the number of masks, the behavior of alternating projections improved, it succeeded in about 99% of the experiments with eight Gaussian illuminations.

Noisy measurements

In the next set of experiments, we consider the case when the measurements are contaminated with Poisson noise. The test signal is again a complex random signal as above. Four, six, and eight illuminations with random binary masks are used. We add random Poisson noise to the measurements for five different SNR levels, ranging from about 16dB to about 52dB. Since the solution is known, we have calculated reconstructions for various values of the parameter λ balancing the negative log-likelihood and the trace norm, and report results for that λ giving the lowest MSE. We implemented this strategy via the standard Golden Section Search [50]. In practice one would have to find the best λ via a strategy like cross validation (CV) or generalized cross validation (GCV). For each SNR level we repeated the experiment ten times with different random noise and different binary masks.

Figure 2.4 shows the average relative MSE in dB (the values of $10\log_{10}(\text{rel. MSE})$ are plotted) versus the SNR. The error curves show clearly the desirable linear behavior between SNR and MSE with respect to the log-log scale. The performance degrades very gracefully with decreasing SNR. Furthermore, the difference of about 5dB between the error curve associated with four measurement and the error curve associated with eight measurements corresponds to an almost twofold error reduction, which is about as much improvement as one can hope to gain by doubling the number of measurements.

We repeat this experiment with determinstic masks as described in Section 2.4 (see (3.2.1)) instead of random masks. To achieve robustness vis a vis noise, three masks (as in Theorem 2.4.1) do not seem to suffice. We thus collect $7n$ measurements of the form $|F_n x|^2$, and then $\{|F_n(x + D^s x)|^2, |F_n(x - iD^s x)|^2$ with $s = 3, 5, 7$ as in (3.2.1). The recovery is very stable and the performance curve is shown in Figure 2.5. For comparison we also show the peformance curve corresponding to seven Gaussian random masks. Gaussian random masks yield better MSE in this example.

2-D simulations

We consider a stylized version of a setup one encounters in X-ray crystallography or diffraction imaging. The test image, shown in Figure 2.6(a) (magnitude), is a complex-valued image of size 256×256, whose pixel values correspond to the complex transmission coefficients of a collection of gold balls embedded in a medium.

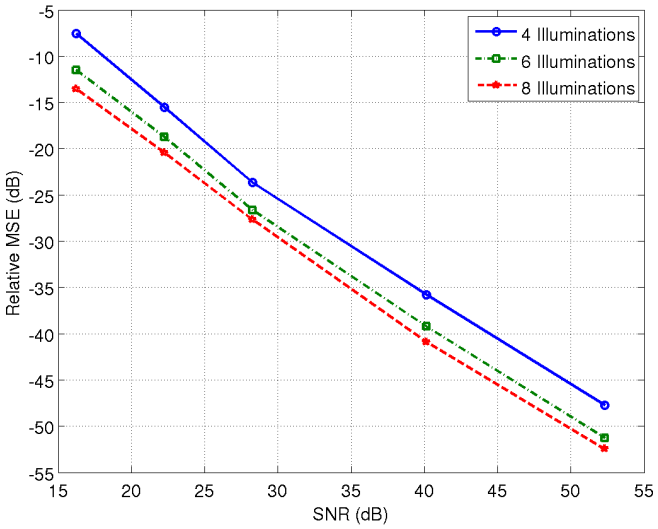

Figure 2.4: Relative MSE versus SNR on a dB-scale for different numbers of illuminations with binary masks. The linear relationship between SNR and MSE is apparent.

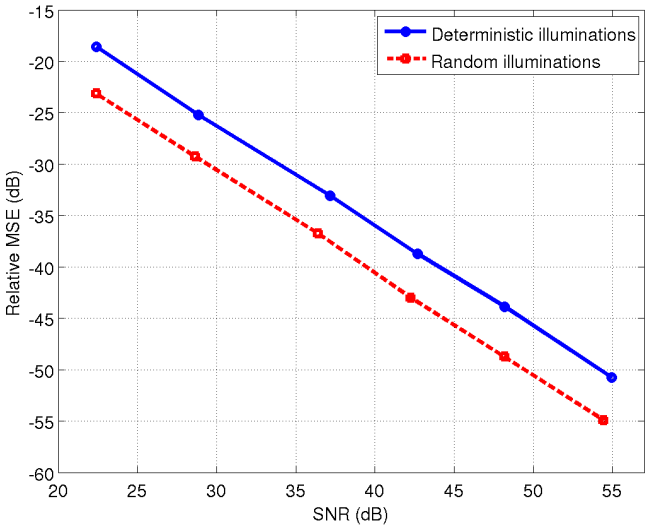

Figure 2.5: Relative MSE versus SNR on a dB-scale: seven illuminations with deterministic masks and with random masks.

Noisefree measurements

In the first experiment, we demonstrate the recovery of the image shown in Figure 2.6(a) from noiseless measurements. We consider two different types of illuminations. The first type uses Gaussian random masks in which the coefficients on the diagonal of W_k are independent real-valued standard normal variables. We use four illuminations, one being constant, i.e. $W_1 = I$, and the other three Gaussian. Again, we choose a small value of λ set to 0.05 in $\frac{1}{2}\|b - \mathcal{A}(X)\|_2^2 + \lambda \operatorname{Tr}(X)$ since we have no noise, and stop the reweighting iterations as soon as the residual error obeys $\|\mathcal{A}(\hat{x}_0\hat{x}_0^*) - b\|_2 \leq 10^{-4}\|b\|_2$. The reconstruction, shown in Figure 2.6(b), is visually indistinguishable from the original. Since the original image and the reconstruction are complex-valued, we only display the absolute value of each image throughout this and the next subsection.

Gaussian random masks may not be realizable in practice. Our second example uses simple random binary masks, where the entries are either 0 or 1 with equal probability. In this case, a larger number of illuminations as well as a larger number of reweighting steps are required to achieve a reconstruction of comparable quality. The result for eight binary illuminations is shown in Figure 2.6(c).

We repeated the experiment using as test signal an image with independent standard normal complex entries; that is, an entry is of the form $z_1 + iz_2$ where z_1 and z_2 are independent $\mathcal{N}(0, 1)$ variables. Here, four Gaussian masks were not sufficient, but we did achieve successful recovery with five Gaussian masks. We also applied the alternating projection algorithm to both test images. In the goldballs example, alternating projections succeeded both with four Gaussian masks and with eight binary masks. For the random image, however, alternating projections always failed when we used five Gaussian masks. As in the one-dimensional example, the performance of alternating projections improved as we increased the number of masks, eventually yielding consistent recovery of the correct image when we employed eight or more Gaussian masks.

Noisy measurements

In the second set of experiments we consider the same test image as before, but now with noisy measurements. In the first experiment the SNR is 20dB, in the second experiment the SNR is 60dB. We use 32 Gaussian random masks in each case. The resulting reconstructions are depicted in Figure 2.7(a) (20dB case) and Figure 2.7(b) (60dB case). The MSE in the 20dB case is 11.83dB. While the reconstructed image appears slightly more "fuzzy" than the original image, all features of the image are clearly visible. In the 60dB case the MSE is 47.96dB, and the reconstruction is virtually indistinguishable from the original image.

Multiple measurements via oversampling

Oversampling of two-dimensional signals is widely used to overcome the nonuniqueness of the phase retrieval problem. We now explore whether this approach is viable.

Here, we consider signals with real, non-negative values as test images, a case frequently considered in the literature, see e.g. [66, 64, 63]. These images are of size 128×128. We take noiseless measurements and apply PhaseLift as the alternating projection algorithm (also known as Fienup's Error Reduction Algorithm) [6, Section 4.A]. For each method, we terminate the iterations if the

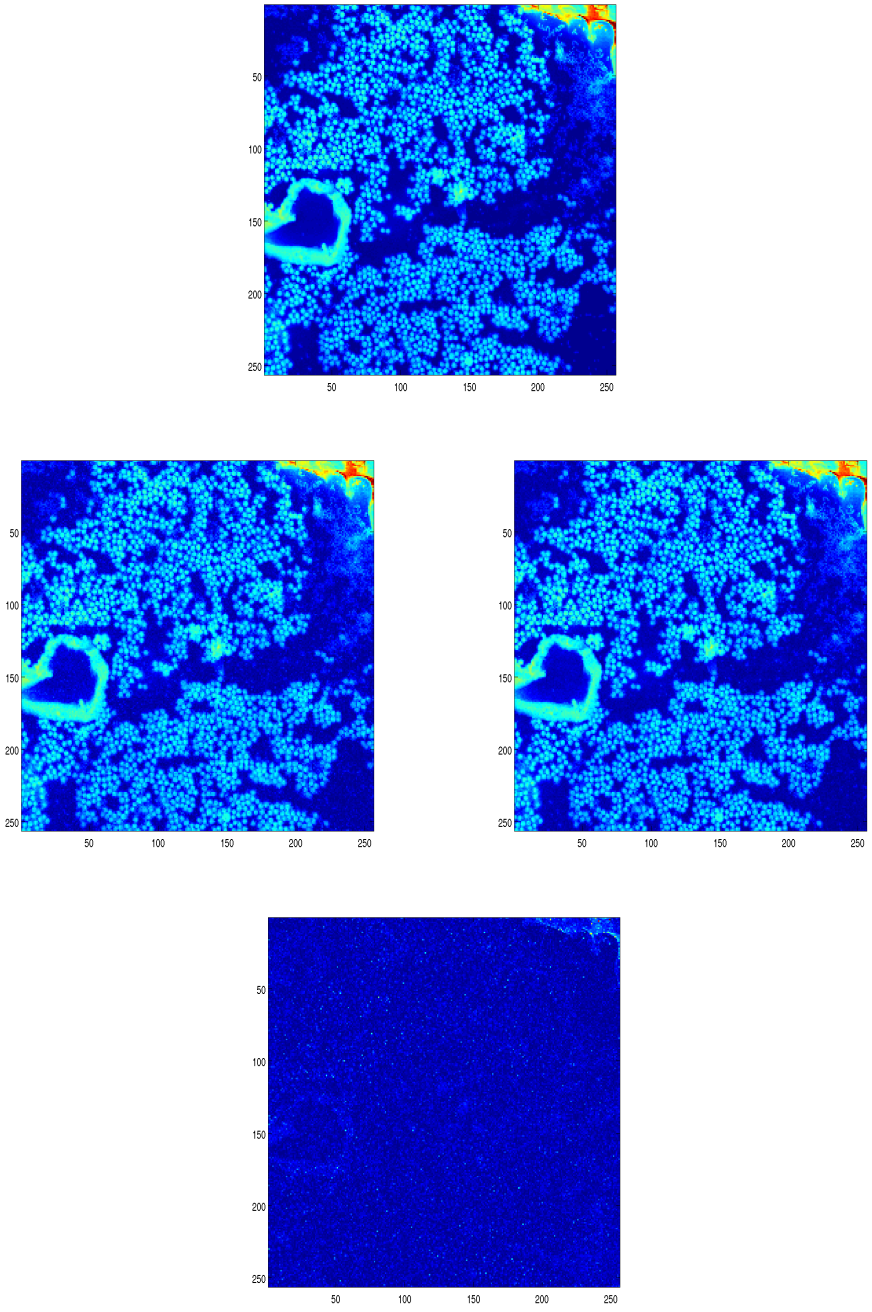

Figure 2.6: Original goldballs image and reconstructions via PhaseLift. Top: original image. Middle Left: reconstruction using 3 Gaussian masks. Middle Right: Reconstruction using 8 binary masks. Bottom: Error between Top and Middle Right.

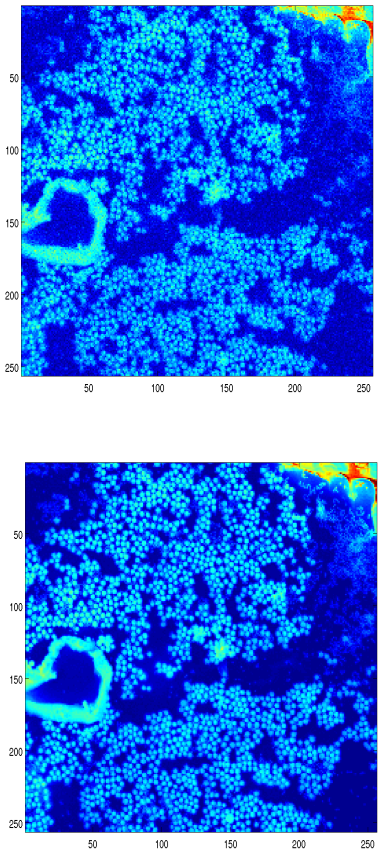

Figure 2.7: Reconstructions from noisy data via PhaseLift using 32 Gaussian random masks. Top: Low SNR. Bottom: High SNR.

relative residual error is less than 10^{-3} or if the relative error between two successive iterates is less than 10^{-6}. Since we assume that the support of the signal is known, there is no ambiguity of the solution with respect to translations. Moreover, the support is chosen to be non-symmetric around the origin, thus there is also no ambiguity with respect to reflections around the origin. Finally, since the signal is real valued and positive, there is no ambiguity with respect to global phase in this case.

- The simulations show that PhaseLift yields reconstructions that fit the measured data well, yielding a small relative residual error $\|\mathcal{A}(X) - y\|_2/\|y\|_2$, yet the reconstructions are far away from the true signal. This behavior is indicative of an ill-conditioned problem.
- The iterates of the alternating projection algorithm stagnated most of the time without converging to a solution. At other times it did yield reconstructions that fit the measured data well, but in either case the reconstruction was always very different from the true signal. Moreover, the reconstructions vary widely depending on the initial (random) guess.

Table 2.1 displays the results of PhaseLift as well as the alternating projection algorithm as described in [6, Section 4. A] (the other versions discussed in Section 4 of [6] yield comparable results). The setup is this: we oversample the signal in each dimension by a factor of r, where $r = 2, 3, 4, 5$. For each oversampling rate, we run ten experiments using a different test signal each time. The table shows the average residual errors over ten runs as well as the average relative MSE. The ill-posedness of the problem is evident from the disconnect between small residual error and large reconstruction error; that is to say, we fit the data very well and yet observe a large reconstruction error. *Thus, in stark contrast to what is widely believed, our simulations indicate that oversampling by itself is not a viable strategy for phase retrieval even for non-negative, real-valued images.*

Algorithm | Oversampling	2	3	4	5
$\|\mathcal{A}(X) - y\|_2/\|y\|_2$ (Alt.Proj.)	0.0650	0.0607	0.0541	0.0713
Relative MSE (Alt.Proj.)	0.6931	0.6882	0.6736	0.6878
$\|\mathcal{A}(X) - y\|_2/\|y\|_2$ (PhaseLift)	0.0051	0.0055	0.0056	0.0052
Relative MSE (PhaseLift)	0.4932	0.4893	0.4960	0.4981

Table 2.1: MSE obtained by alternating projections and by PhaseLift with reweighting from oversampled DFT measurements taken on 2D real-valued and positive test images. The alternating projection algorithm does not always find a signal consistent with the data as well as the support constraint. (After the projection step in the spatial domain, the current guess does not always match the measurement in Fourier space. After 'projection' in Fourier space, the signal is not the Fourier transform of a signal obeying the spatial constraints.) Our approach always finds signals matching measured data very well, and yet the reconstructions achieve a large reconstruction error. This indicates severe ill-posedness since we have several distinct solutions providing an excellent fit to the measured data.

2.6 Discussion

This chapter introduces a novel framework for phase retrieval, combining multiple illuminations with tools from convex optimization, which has been shown to work very well in practice and bears great potential. This work also calls for theory, improved algorithms and a physical implementation of these ideas. For now, we would like to bring up important open problems.

At the theoretical level, we need to understand for which families of physically implementable structured illuminations does the trace-norm heuristic succeed? How many diffraction patterns are provably sufficient for the PhaseLift convex programming approach to work? Also, we have empirically shown that our approach is robust to noise in the sense that the performance degrades very gracefully as the SNR decreases. Can this be made rigorous?

In the next chapter, it will be proven that for measurement vectors sampled independently and uniformly at random on the unit sphere, PhaseLift indeed reconstructs signals from noiseless measurements and is robust to noise, provided that the number of measurements is on the order of $n \log n$.

Chapter 3

PhaseLift for gaussian measuremenets

3.1 Overview

Suppose we wish to recover a signal $\boldsymbol{x} \in \mathbb{C}^n$ from m intensity measurements of the form $|\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle|^2$, $i = 1, 2, \ldots, m$; that is, from data in which phase information is missing. We prove that if the vectors $\boldsymbol{z}_i$ are sampled independently and uniformly at random on the unit sphere, then the signal $\boldsymbol{x}$ can be recovered exactly (up to a global phase factor) by solving a convenient semidefinite program—a trace-norm minimization problem; this holds with large probability provided that m is on the order of $n \log n$, and without any assumption about the signal whatsoever. This novel result demonstrates that in some instances, the combinatorial phase retrieval problem can be solved by convex programming techniques. Finally, we also prove that our methodology is robust to additive noise.

3.2 Introduction

Formally, suppose $\boldsymbol{x} \in \mathbb{C}^n$ is a discrete signal and that we are given information about the squared modulus of the inner product between the signal and some vectors $\boldsymbol{z}_i$, namely,

$$b_i = |\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle|^2, \quad i = 1, \ldots, m. \tag{3.2.1}$$

In truth, we would like to know $\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle$ and record both phase and magnitude information but can only record the magnitude; in other words, phase information is lost. In the classical example discussed above, the $\boldsymbol{z}_i$'s are complex exponentials at frequency ω_i so that one collects the squared modulus of the Fourier transform of $\boldsymbol{x}$. Of course, many other choices for the measurement vectors $\boldsymbol{z}_i$ are frequently discussed in the literature, see [36, 5] for instance.

We wish to recover $\boldsymbol{x}$ from the data vector $\boldsymbol{b}$, and suppose first that $\boldsymbol{x}$ is known to be real valued a priori. Then assuming that $\boldsymbol{x}$ is uniquely determined by $\boldsymbol{b}$ up to a global sign, the recovery may be cast as a combinatorial optimization problem: find a set of signs σ_i such that the solution to the linear equations $\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle = \sigma_i \sqrt{b_i}$, call it $\hat{\boldsymbol{x}}$, obeys $|\langle \hat{\boldsymbol{x}}, \boldsymbol{z}_i \rangle|^2 = b_i$. Clearly, there are 2^m choices for σ_i and only two choices of these signs yield $\boldsymbol{x}$ up to global phase. The complex case is harder yet, since resolving the phase ambiguities now consists of finding a collection σ_i of complex numbers,

each being on the unit circle. Formalizing matters, it has been shown that at least one version of the phase retrieval problem is NP-hard [78]. Thus, one of the major challenges in the field is to find conditions on m and $\boldsymbol{z}_i$ which guarantee efficient numerical recovery.

A frame-theoretic approach to signal recovery from magnitude measurements has been proposed in [3, 4, 5], where the authors derive various necessary and sufficient conditions for the uniqueness of the solution, as well as various polynomial-time numerical algorithms for very specific choices of $\boldsymbol{z}_i$. While theoretically quite appealing, the drawbacks are that the methods are (1) either algebraic in nature, thus severely limiting their stability in the presence of noise or slightly inexact data, or (2) the number m of measurements is on the order of n^2, which is much too large compared to the number of unknowns.

Here we follow a different route and establish that if the vectors $\boldsymbol{z}_i$ are independently and uniformly sampled on the unit sphere, then our signal can be recovered exactly from the magnitude measurements (3.2.1) by solving a simple convex program introduced below; this holds with high probability under the condition that the number of measurements is on the order of $n \log n$. Since there are n complex unknowns, we see that the number of samples is nearly minimal. To the best of our knowledge, this is the first result establishing that under appropriate conditions, the computationally challenging nonconvex problem of reconstructing a signal from magnitude measurements is formally equivalent to a convex program in the sense that they are guaranteed to have the same unique solution.

Finally, our methodology is robust with respect to noise in the measurements. That is, when the data are corrupted by a small amount of noise, we also prove that the recovery error is small.

Methodology

We introduce some notation that shall be used throughout to explain our methodology. Letting $\mathcal{A}$ be the linear transformation

$$\begin{array}{rcl} \mathcal{H}^{n\times n} & \to & \mathbb{R}^m \\ \boldsymbol{X} & \mapsto & \{\boldsymbol{z}_i^* \boldsymbol{X} \boldsymbol{z}_i\}_{1\le i\le m} \end{array} \tag{3.2.2}$$

which maps Hermitian matrices into real-valued vectors, one can express the data collection $b_i = |\langle \boldsymbol{x}, \boldsymbol{z}_i\rangle|^2$ as

$$\boldsymbol{b} = \mathcal{A}(\boldsymbol{x}\boldsymbol{x}^*). \tag{3.2.3}$$

For reference, the adjoint operator $\mathcal{A}^*$ maps real-valued inputs into Hermitian matrices, and is given by

$$\begin{array}{rcl} \mathbb{R}^m & \to & \mathcal{H}^{n\times n} \\ \boldsymbol{y} & \mapsto & \sum_i y_i \, \boldsymbol{z}_i \boldsymbol{z}_i^*. \end{array}$$

As observed in [24, 22] (see also [56]), the phase retrieval problem can be cast as the matrix recovery problem

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}) \\ \text{subject to} & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b} \\ & \boldsymbol{X} \succeq 0. \end{array} \tag{3.2.4}$$

Indeed, we know that a rank-one solution exists so the optimal $\boldsymbol{X}$ has rank at most one. We then factorize the solution as $\boldsymbol{x}\boldsymbol{x}^*$ in order to obtain solutions to the phase-retrieval problem. This gives

$\boldsymbol{x}$ up to multiplication by a unit-normed scalar. This is all we can hope for since if $\boldsymbol{x}$ is a solution to the phase retrieval problem, then $c\boldsymbol{x}$ for any scalar $c \in \mathbb{C}$ obeying $|c| = 1$ is also solution.[1]

Rank minimization is in general NP hard, and we propose, instead, solving a trace-norm relaxation. Although this is a fairly standard relaxation in control [8, 62], the idea of casting the phase retrieval problem as a trace-minimization problem over an affine slice of the positive semidefinite cone is very recent [24, 22]. Formally, we suggest solving

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(\boldsymbol{X}) \\ \text{subject to} & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b} \\ & \boldsymbol{X} \succeq 0. \end{array} \tag{3.2.5}$$

If the solution has rank one, we factorize it as above to recover our signal. This method which lifts up the problem of vector recovery from quadratic constraints into that of recovering a rank-one matrix from affine constraints via semidefinite programming is known under the name of *PhaseLift* [22].

The program (3.2.5) is a semidefinite program (SDP) in standard form, and there is a rapidly growing list of algorithms for solving problems of this kind as efficiently as possible. The crucial question is whether and under which conditions the combinatorially hard problem (3.2.4) and the convex problem (3.2.5) are formally equivalent.

Main result

In this chapter, we consider the simplest and perhaps most natural model of measurement vectors. In this statistical model, we simply assume that the vectors $\boldsymbol{z}_i$ are independently and uniformly distributed on the unit sphere of $\mathbb{C}^n$ or $\mathbb{R}^n$. To be concrete, we distinguish two models.

- *The real-valued model.* Here, the unknown signal $\boldsymbol{x}$ is real valued and the $\boldsymbol{z}_i$'s are independently sampled on the unit sphere of $\mathbb{R}^n$.
- *The complex-valued model.* The signal $\boldsymbol{x}$ is now complex valued and the $\boldsymbol{z}_i$'s are independently sampled on the unit sphere of $\mathbb{C}^n$.

Our main result is that the convex program recovers $\boldsymbol{x}$ exactly (up to global phase) provided the number m of magnitude measurements is on the order of $n \log n$.

Theorem 3.2.1 *Consider an arbitrary signal $\boldsymbol{x}$ in $\mathbb{R}^n$ or $\mathbb{C}^n$ and suppose that the number of measurements obeys $m \geq c_0\, n \log n$, where c_0 is a sufficiently large constant. Then in both the real and complex cases, the solution to the trace-minimization program is exact with high probability in the sense that* (3.2.5) *has a unique solution obeying*

$$\hat{\boldsymbol{X}} = \boldsymbol{x}\boldsymbol{x}^*. \tag{3.2.6}$$

This holds with probability at least $1 - 3e^{-\gamma \frac{m}{n}}$, where γ is a positive absolute constant.

[1] When the solution is unique up to multiplication by such a scalar, we shall say that unicity holds up to global phase.

Expressed differently, Theorem 3.2.1 establishes a rigorous equivalence between a class of phase retrieval problems and a class of semidefinite programs. Clearly, any phase retrieval algorithm, no matter how complicated or intractable, would need at least $2n$ quadratic measurements to recover a complex valued object $\boldsymbol{x} \in \mathbb{C}^n$. In fact recent results, compare Theorem II in [36], show that for complex-valued signals, one needs at least $3n - 2$ intensity measurements to guarantee uniqueness of the solution to (3.2.4). Further, Balan, Casazza and Edidin have shown that with probability 1, $4n - 2$ generic measurement vectors (which includes the case of random uniform vectors) suffice for uniqueness in the complex case [3].

Geometry

We find it remarkable that the only solution to (3.2.5) is $\hat{\boldsymbol{X}} = \boldsymbol{x}\boldsymbol{x}^*$. To see why this is perhaps unexpected, suppose for simplicity that the trace of the solution were known (we might be given some side information or just have additional measurements giving us this information) and equal to 1, say. In this case, the objective functional is of course constant over the feasible set, and our problem reduces to solving the feasibility problem

$$\begin{array}{ll} \text{find} & \boldsymbol{X} \\ \text{such that} & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b},\ \boldsymbol{X} \succeq 0 \end{array} \tag{3.2.7}$$

with again the assumption that knowledge of $\mathcal{A}(\boldsymbol{X})$ determines $\mathrm{Tr}(\boldsymbol{X})$ (equal to $\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^*) = \|\boldsymbol{x}\|_2 = 1$). In this context, our main theorem states that $\boldsymbol{x}\boldsymbol{x}^*$ is the unique feasible point. In other words, there is no other positive semidefinite matrix $\boldsymbol{X}$ in the affine space $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$. Naively, we would not expect this affine space of large dimension—it is of co-dimension about $n \log n$ and thus of dimension $n^2 - O(n \log n)$ in the complex case—to intersect the positive semidefinite cone in only one point. Indeed, counting degrees of freedom suggests that there are infinitely many candidates in the intersection. The reason why this is not the case, however, is precisely because there is a feasible solution with low rank. Indeed, the slice of the positive semidefinite cone $\{\boldsymbol{X} : \boldsymbol{X} \succeq 0\} \cap \{\mathrm{Tr}(\boldsymbol{X}) = 1\}$ is quite 'pointy' at $\boldsymbol{x}\boldsymbol{x}^*$ and it is, therefore, possible for the affine space $\{\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}\}$ to be tangent even though it is of very small codimension.

Figure 3.1 represents this geometry. In this example,

$$\boldsymbol{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \Longrightarrow \quad \boldsymbol{x}\boldsymbol{x}^* = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

and the affine space $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ is tangent to the positive semidefinite cone at the point $\boldsymbol{x}\boldsymbol{x}^*$.

Phase retrieval may be framed a problem in algebraic geometry since we are trying to find a solution to a set of polynomial equations. For instance, we prove that there is no other positive semidefinite matrix $\boldsymbol{X}$ in the affine space $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$, or equivalently, that a certain system of polynomial equations (a symmetric matrix is positive semidefinite if and only if the determinants of all the leading principal minors are nonnegative) only has one solution; this is a fact that general techniques from algebraic geometry appear to not detect.

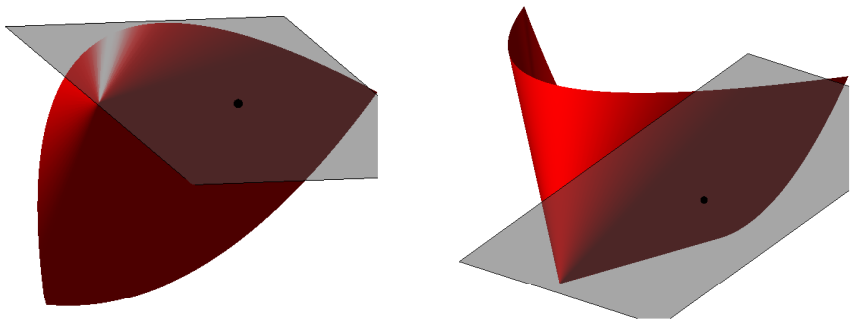

Figure 3.1: Representation of the affine space $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ (gray) and of the semidefinite cone $\begin{bmatrix} x & y \\ y & z \end{bmatrix} \succeq 0$ (red) which is a subset of $\mathbb{R}^3$. These two sets are drawn so that they are tangent to each other at the rank 1 matrix $\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ (black dot). Two views of the same 3D figure are provided for convenience.

Stability

In the real world, measurements are contaminated by noise. Using the frameworks developed in [18] and [40], it is possible to extend Theorem 3.2.1 to accommodate noisy measurements. One could consider a variety of noise models as discussed in [22] but we work here with a simple generic model in which we observe

$$b_i = |\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle|^2 + \nu_i, \tag{3.2.8}$$

where ν_i is a noise term with bounded ℓ_2 norm, $\|\boldsymbol{\nu}\|_2 \leq \varepsilon$. This model is nonstandard since the usual statistical linear model posits a relationship of the form $b_i = \langle \boldsymbol{x}, \boldsymbol{z}_i \rangle + \nu_i$ in which the mean response is a linear function of the unknown signal, not a quadratic function. Furthermore, we prefer studying (3.2.8) rather than the related model $b_i = |\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle| + \nu_i$ (the modulus is not squared) because in many applications of interest in optics and other areas of physics, one can measure squared magnitudes or intensities—not magnitudes.

We now consider the solution to

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(\boldsymbol{X}) \\ \text{subject to} & \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{b}\|_2 \leq \varepsilon \\ & \boldsymbol{X} \succeq 0. \end{array} \tag{3.2.9}$$

We do not claim that $\hat{\boldsymbol{X}}$ has low rank so we suggest estimating $\boldsymbol{x}$ by extracting the largest rank-1 component. Write $\hat{\boldsymbol{X}}$ as

$$\hat{\boldsymbol{X}} = \sum_{k=1}^{n} \hat{\lambda}_k \hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^*, \quad \hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_n \geq 0,$$

and set

$$\hat{\boldsymbol{x}} = \sqrt{\hat{\lambda}_1}\, \hat{\boldsymbol{u}}_1.$$

We prove the following estimate.

Theorem 3.2.2 *Fix $\boldsymbol{x} \in \mathbb{C}^n$ or $\mathbb{R}^n$ and assume the $\boldsymbol{z}_i$'s are uniformly sampled on the sphere of radius $\sqrt{n}$. Under the hypotheses of Theorem 3.2.1, the solution to* (3.2.9) *obeys ($\|\boldsymbol{X}\|_2$ is the Frobenius norm of $\boldsymbol{X}$)*

$$\|\hat{\boldsymbol{X}} - \boldsymbol{x}\boldsymbol{x}^*\|_2 \leq C_0\, \varepsilon \tag{3.2.10}$$

for some positive numerical constant C_0. We also have

$$\|\hat{\boldsymbol{x}} - e^{i\phi}\boldsymbol{x}\|_2 \leq C_0 \min(\|\boldsymbol{x}\|_2, \varepsilon/\|\boldsymbol{x}\|_2) \tag{3.2.11}$$

for some $\phi \in [0, 2\pi]$. Both these estimates hold with nearly the same probability as in the noiseless case.

Thus our approach also provides stable recovery in presence of noise. This important property is not shared by other reconstruction methods, which are of a more algebraic nature and rely on particular properties of the measurement vectors, such as the methods in [36, 3, 5], as well as the methods that appear implicitly in Theorem 3.1 and Theorem 3.3 of [22].

We note that one can further improve the accuracy of the solution $\hat{\boldsymbol{x}}$ by "debiasing" it. We replace $\hat{\boldsymbol{x}}$ by its rescaled version $s\hat{\boldsymbol{x}}$ where $s = \sqrt{\sum_{k=1}^{n} \hat{\lambda}_k}/\|\hat{\boldsymbol{x}}\|_2$. This corrects for the energy leakage occurring when $\hat{\boldsymbol{X}}$ is not exactly a rank-1 solution, which could cause the norm of $\hat{\boldsymbol{x}}$ to be smaller than that of the actual solution. Other corrections are of course possible.

Organization

The remainder of the chapter is organized as follows. Subsection 3.2 introduces some notation used throughout. In Section 3.3 we present the main architecture of the proof of Theorem 3.2.1, which comprises two key ingredients: approximate ℓ_1 isometries and approximate dual certificates. Section 3.4 is devoted to establishing approximate ℓ_1 isometries. In Section 3.5, we construct approximate dual certificates and complete the proof of Theorem 3.2.1 in the real-valued case. Section 3.6 shows how the proof for the real-valued case can be adapted to the complex-valued case. Section 3.7 is concerned with the proof of Theorem 3.2.2. Numerical simulations, illustrating our theoretical results, are presented in Section 3.8. We conclude the chapter with a short discussion in Section 3.9.

Notations

Here we introduce notations that shall be used throughout the chapter. Matrices and vectors are denoted in boldface (such as $\boldsymbol{X}$ or $\boldsymbol{x}$), while individual entries of a vector or matrix are denoted in normal font; e.g. the ith entry of $\boldsymbol{x}$ is x_i. For matrices, we define

$$\|\boldsymbol{X}\|_p = \Big[\sum_i \sigma_i^p(\boldsymbol{X})\Big]^{1/p},$$

(where $\sigma_i(\boldsymbol{X})$ denotes the ith singular value of $\boldsymbol{X}$), so that $\|\boldsymbol{X}\|_1$ is the nuclear norm, $\|\boldsymbol{X}\|_2$ is the Frobenius norm and $\|\boldsymbol{X}\|_\infty$ is the operator norm also denoted by $\|\boldsymbol{X}\|$. For vectors, $\|\boldsymbol{x}\|_p$ is the usual ℓ_p norm. We denote the $n-1$ dimensional sphere by S^{n-1}, i.e. the set $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$.

Next, we define $T_{\boldsymbol{x}}$ to be the set of symmetric matrices of the form

$$T_{\boldsymbol{x}} = \{\boldsymbol{X} = \boldsymbol{x}\boldsymbol{y}^* + \boldsymbol{y}\boldsymbol{x}^* : \boldsymbol{y} \in \mathbb{R}^n\} \tag{3.2.12}$$

and denote $T_{\boldsymbol{x}}^\perp$ by its orthogonal complement. Note that $\boldsymbol{X} \in T_{\boldsymbol{x}}^\perp$ if and only if both the column and row spaces of $\boldsymbol{X}$ are perpendicular to $\boldsymbol{x}$. Further, the operator $\mathcal{P}_{T_{\boldsymbol{x}}}$ is the orthogonal projector onto $T_{\boldsymbol{x}}$ and similarly for $\mathcal{P}_{T_{\boldsymbol{x}}^\perp}$. We shall almost always use $\boldsymbol{X}_{T_{\boldsymbol{x}}}$ as a shorthand for $\mathcal{P}_{T_{\boldsymbol{x}}}(\boldsymbol{X})$.

Finally, we will abuse language and say that a symmetric matrix $\boldsymbol{H}$ is feasible if and only if $\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{H}$ is feasible for our problem (3.2.5). This means that $\boldsymbol{H}$ obeys

$$\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{H} \succeq 0 \quad \text{and} \quad \mathcal{A}(\boldsymbol{H}) = 0. \tag{3.2.13}$$

3.3 Architecture of the Proof

In this section, we introduce the main architecture of the argument and defer the proofs of crucial intermediate results to later sections. We shall prove Theorem 3.2.1 in the real case first for ease of exposition. Then in Section 3.6, we shall explain how to modify the argument to the complex and more general case.

Suppose then that $\boldsymbol{x} \in \mathbb{R}^n$ and that the $\boldsymbol{z}_i$'s are sampled on the unit sphere. It is clear that we may assume without loss of generality that $\boldsymbol{x}$ is unit-normed. Further, since the uniform distribution on the unit sphere is rotationally invariant, it suffices to prove the theorem in the case where $\boldsymbol{x} = \boldsymbol{e}_1$. Indeed, we can write any unit vector $\boldsymbol{x}$ as $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{e}_1$ where $\boldsymbol{U}$ is orthogonal. Since

$$|\langle \boldsymbol{x}, \boldsymbol{z}_i\rangle|^2 = |\langle \boldsymbol{U}\boldsymbol{e}_1, \boldsymbol{z}_i\rangle|^2 = |\langle \boldsymbol{e}_1, \boldsymbol{U}^*\boldsymbol{z}_i\rangle|^2 =^d |\langle \boldsymbol{e}_1, \boldsymbol{z}_i\rangle|^2,$$

the problem is the same as that of finding $\boldsymbol{e}_1$. We henceforth assume that $\boldsymbol{x} = \boldsymbol{e}_1$.

Finally, the theorem can be equivalently stated in the case where the $\boldsymbol{z}_i$'s are i.i.d. copies of a white noise vector $\boldsymbol{z} \sim \mathcal{N}(0, I)$ with independent standard normals as components. Indeed, if $\boldsymbol{z}_i \sim \mathcal{N}(0, I)$,

$$|\langle \boldsymbol{x}, \boldsymbol{z}_i\rangle|^2 = b_i \quad \Longleftrightarrow \quad |\langle \boldsymbol{x}, \boldsymbol{u}_i\rangle|^2 = b_i/\|\boldsymbol{z}_i\|_2^2,$$

where $\boldsymbol{u}_i = \boldsymbol{z}_i/\|\boldsymbol{z}_i\|_2$ is uniformly sampled on the unit sphere. Since $\|\boldsymbol{z}_i\|_2$ does not vanish with probability one, establishing the theorem for Gaussian vectors establishes it for uniformly sampled vectors and vice versa. From now on, we assume $\boldsymbol{z}_i$ i.i.d. $\mathcal{N}(0, I)$.

Key lemma

The set $T := T_{\boldsymbol{e}_1}$ defined in (3.2.12) may be interpreted as the tangent space at $\boldsymbol{e}_1\boldsymbol{e}_1^*$ to the manifold of symmetric matrices of rank 1. Now standard duality arguments in semidefinite programming show that a sufficient (and nearly necessary) condition for $\boldsymbol{x}\boldsymbol{x}^*$ to be the unique solution to (3.2.5) is this:

- the restriction of $\mathcal{A}$ to T is injective ($\boldsymbol{X} \in T$ and $\mathcal{A}(X) = 0 \Rightarrow \boldsymbol{X} = 0$),
- and there exists a *dual certificate* $\boldsymbol{Y}$ in the range of $\mathcal{A}^*$ obeying[2]

$$\boldsymbol{Y}_T = \boldsymbol{e}_1\boldsymbol{e}_1^* \quad \text{and} \quad \boldsymbol{Y}_T^\perp \prec I_T^\perp. \tag{3.3.1}$$

The proof is straightforward and omitted. Our strategy to prove Theorem 3.2.1 hinges on the fact that a strengthening of the injectivity property allows to relax the properties of the dual certificate, as in the approach pioneered in [39] for matrix completion. We establish the crucial lemma below.

Lemma 3.3.1 *Suppose that the mapping $\mathcal{A}$ obeys the following two properties: for all positive semidefinite matrices X,*

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 < (1 + 1/9)\|\boldsymbol{X}\|_1; \tag{3.3.2}$$

and for all matrices $\boldsymbol{X} \in T$

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 > 0.94(1 - 1/9)\|\boldsymbol{X}\|. \tag{3.3.3}$$

Suppose further that there exists $\boldsymbol{Y}$ in the range of $\mathcal{A}^*$ obeying

$$\|\boldsymbol{Y}_T - \boldsymbol{e}_1\boldsymbol{e}_1^*\|_2 \le 1/3 \quad \textit{and} \quad \|\boldsymbol{Y}_T^\perp\| \le 1/2. \tag{3.3.4}$$

Then $\boldsymbol{e}_1\boldsymbol{e}_1^*$ is the unique minimizer to (3.2.5).

The first property (3.3.2) is reminiscent of the (one-sided) RIP property in the area of compressed sensing [17]. The difference is that it is expressed in the 1-norm rather than the 2-norm. Having said this, we note that RIP-1 properties have also been used in the compressed sensing literature, see [11] for example. We use this property instead of a property about $\|\mathcal{A}(\boldsymbol{X})\|_2$, because a RIP property in the 2-norm does not hold here because $\|\mathcal{A}(\boldsymbol{X})\|_2^2$, as we prove in the appendix, essentially because it involves fourth moments of Gaussian variables. The second property (3.3.3) is a form of local RIP-1 since it holds only for matrices in T.

We would like to emphasize that the bound for the dual certificate in (3.3.4) is loose in the sense that $\boldsymbol{Y}_T$ and $\boldsymbol{e}_1\boldsymbol{e}_1^*$ may not be that close, a fact which will play a crucial role in our proof. This is in stark contrast with the work of David Gross [39], which requires a very tight approximation.

Proof of Lemma 3.3.1

We need to show that there is no feasible $\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{H} \neq \boldsymbol{x}\boldsymbol{x}^*$ with $\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{H}) \le \mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^*)$. Consider then a feasible $\boldsymbol{H} \neq 0$ obeying $\mathrm{Tr}(\boldsymbol{H}) \le 0$, write

$$\boldsymbol{H} = \boldsymbol{H}_T + \boldsymbol{H}_T^\perp,$$

and observe that

$$0 = \|\mathcal{A}(\boldsymbol{H})\|_1 = \|\mathcal{A}(\boldsymbol{H}_T)\|_1 - \|\mathcal{A}(\boldsymbol{H}_T^\perp)\|_1. \tag{3.3.5}$$

[2]The notation $A \prec B$ means that $B - A$ is positive definite.

Now it is clear that $\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{H} \succeq 0 \Rightarrow \boldsymbol{H}_T^\perp \succeq 0$ and, therefore, (3.3.2) gives

$$m^{-1}\|\mathcal{A}(\boldsymbol{H}_T^\perp)\|_1 \leq (1+\delta)\operatorname{Tr}(\boldsymbol{H}_T^\perp)$$

for some $\delta < 1/9$. Also, $\operatorname{Tr}(\boldsymbol{H}_T) \leq -\operatorname{Tr}(\boldsymbol{H}_T^\perp) \leq 0$, which implies that $|\operatorname{Tr}(\boldsymbol{H}_T)| \geq \operatorname{Tr}(\boldsymbol{H}_T^\perp)$. We then show that the operator and Frobenius norms of $\boldsymbol{H}_T$ must nearly be the same.

Lemma 3.3.2 *Any feasible matrix $\boldsymbol{H}$ such that $\operatorname{Tr}(\boldsymbol{H}) \leq 0$ must obey*

$$\|\boldsymbol{H}_T\|_2 \leq \sqrt{\frac{17}{16}}\,\|\boldsymbol{H}_T\|.$$

Proof Since the matrix $\boldsymbol{H}_T$ has rank at most 2 and cannot be negative definite, it is of the form

$$-\lambda(\boldsymbol{u}_1\boldsymbol{u}_1^* - t\boldsymbol{u}_2\boldsymbol{u}_2^*),$$

where $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are orthonormal eigenvectors, $\lambda \geq 0$ and $t \in [0,1]$. We claim that we cannot have $t \geq 1/4$.[3] Suppose the contrary and fix $t \geq 1/4$. By (3.3.3), we know that

$$m^{-1}\,\|\mathcal{A}(\boldsymbol{H}_T)\|_1 \geq 0.94(1-\delta)\|\boldsymbol{H}_T\|.$$

Further, since

$$\|\boldsymbol{H}_T\| = \frac{|\operatorname{Tr}(\boldsymbol{H}_T)|}{1-t} \geq \frac{4}{3}|\operatorname{Tr}(\boldsymbol{H}_T)|$$

for $t \geq 1/4$, it holds that

$$0 \geq \frac{5}{4}(1-\delta)\,|\operatorname{Tr}(\boldsymbol{H}_T)| - (1+\delta)\operatorname{Tr}(\boldsymbol{H}_T^\perp).$$

The right-hand side above is positive if $\operatorname{Tr}(\boldsymbol{H}_T^\perp) < \frac{5}{4}\frac{(1-\delta)}{(1+\delta)}\,|\operatorname{Tr}(\boldsymbol{H}_T)|$, so that we may assume that

$$\operatorname{Tr}(\boldsymbol{H}_T^\perp) \geq \frac{5}{4}\frac{(1-\delta)}{(1+\delta)}\,|\operatorname{Tr}(\boldsymbol{H}_T)|\,.$$

Since, $|\operatorname{Tr}(\boldsymbol{H}_T)| \geq \operatorname{Tr}(\boldsymbol{H}_T^\perp)$, this gives

$$0 \geq \Big[\frac{5}{4}(1-\delta) - (1+\delta)\Big]\operatorname{Tr}(\boldsymbol{H}_T^\perp).$$

If $\delta < 1/9$, the only way this can happen is if $\operatorname{Tr}(\boldsymbol{H}_T^\perp) = 0 \Rightarrow \boldsymbol{H}_T^\perp = 0$. So we would have $\boldsymbol{H} = \boldsymbol{H}_T$ of rank 2 and $\mathcal{A}(\boldsymbol{H}_T) = 0$. Clearly, (3.3.3) implies that $\boldsymbol{H} = 0$.

Now that it is established that $t \leq 1/4$, the chain of inequalities follow from the relation between the eigenvalues of $\boldsymbol{H}_T$. ∎

[3] The choice of $1/4$ is somewhat arbitrary here.

To conclude the proof of Lemma 3.3.1, we show that the existence of an inexact dual certificate rules out the existence of matrices obeying the conditions of Lemma 3.3.2. From

$$0.94(1-\delta)\|\boldsymbol{H}_T\| \le m^{-1}\|\mathcal{A}(\boldsymbol{H}_T)\|_1 = m^{-1}\|\mathcal{A}(\boldsymbol{H}_T^\perp)\|_1 \le (1+\delta)\operatorname{Tr}(\boldsymbol{H}_T^\perp),$$

we conclude that

$$\operatorname{Tr}(\boldsymbol{H}_T^\perp) \ge 0.94\frac{1-\delta}{1+\delta}\|\boldsymbol{H}_T\| \ge 0.94\frac{1-\delta}{1+\delta}\sqrt{\frac{16}{17}}\|\boldsymbol{H}_T\|_2, \tag{3.3.6}$$

where we used Lemma 3.3.2. Next,

$$\begin{aligned} 0 \ge \operatorname{Tr}(\boldsymbol{H}_T) + \operatorname{Tr}(\boldsymbol{H}_T^\perp) &= \langle \boldsymbol{H}, \boldsymbol{e}_1\boldsymbol{e}_1^*\rangle + \operatorname{Tr}(\boldsymbol{H}_T^\perp) \\ &= \langle \boldsymbol{H}, \boldsymbol{e}_1\boldsymbol{e}_1^* - \boldsymbol{Y}\rangle + \langle \boldsymbol{H}, \boldsymbol{Y}\rangle + \operatorname{Tr}(\boldsymbol{H}_T^\perp) \\ &= \langle \boldsymbol{H}_T, \boldsymbol{e}_1\boldsymbol{e}_1^* - \boldsymbol{Y}_T\rangle - \langle \boldsymbol{H}_T^\perp, \boldsymbol{Y}_T^\perp\rangle + \operatorname{Tr}(\boldsymbol{H}_T^\perp) \\ &\ge \frac{1}{2}\operatorname{Tr}(\boldsymbol{H}_T^\perp) - \frac{1}{3}\|\boldsymbol{H}_T\|_2. \end{aligned}$$

The third line above follows from $\langle \boldsymbol{H}, \boldsymbol{Y}\rangle = 0$ and the fourth from Cauchy-Schwarz together with $|\langle \boldsymbol{H}_T^\perp, \boldsymbol{Y}_T^\perp\rangle| \le \frac{1}{2}\operatorname{Tr}(\boldsymbol{H}_T^\perp)$. Hence, it follows from (3.3.6) that

$$0 \ge \frac{1}{2}\Big(0.94\frac{1-\delta}{1+\delta}\sqrt{\frac{16}{17}} - \frac{2}{3}\Big)\|\boldsymbol{H}_T\|_2.$$

Since the numerical factor is positive for $\delta < 0.155$, the only way this can happen is if $\boldsymbol{H}_T = 0$. In turn, $\|\mathcal{A}(\boldsymbol{H}_T^\perp)\|_1 = 0 \ge (1-\delta)\operatorname{Tr}(\boldsymbol{H}_T^\perp)$ which gives $\boldsymbol{H}_T^\perp = 0$. This concludes the proof.

3.4 Approximate ℓ_1 Isometries

We have seen that in order to prove our main result, it suffices to show 1) that the measurement operator $\mathcal{A}$ enjoys approximate isometry properties (in an ℓ_1 sense) when acting on low-rank matrices and 2) that an inexact dual certificate exists. This section focuses on the former and establishes that both (3.3.2) and (3.3.3) hold with high probability. In fact, we shall prove stronger results than what is strictly required.

Lemma 3.4.1 *Fix any δ in $(0, 1/2)$ and assume $m \ge 20\delta^{-2}\, n$. Then for all unit vectors $\boldsymbol{u}$,*

$$(1-\delta) \le \frac{1}{m}\|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_1 \le (1+\delta) \tag{3.4.1}$$

on an event E_δ of probability at least $1 - 2e^{-m\varepsilon^2/2}$, where $\delta/4 = \varepsilon^2 + \varepsilon$. On the same event,

$$(1-\delta)\|\boldsymbol{X}\|_1 \le \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \le (1+\delta)\|\boldsymbol{X}\|_1$$

for all positive semidefinite matrices. The right inequality holds for all Hermitian matrices.

Proof This lemma has an easy proof. Let $\boldsymbol{Z}$ be the $m \times n$ matrix with $\boldsymbol{z}_i$'s as rows. Then

$$\|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_1 = \sum_i |\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle|^2 = \|\boldsymbol{Z}\boldsymbol{u}\|^2$$

so that

$$\sigma^2_{\min}(\boldsymbol{Z}) \le \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_1 \le \sigma^2_{\max}(\boldsymbol{Z}).$$

The claim is a consequence of well-known deviations bounds concerning the singular values of Gaussian random matrices [85], namely,

$$\begin{aligned}\mathbb{P}\left(\sigma_{\max}(\boldsymbol{Z}) > \sqrt{m} + \sqrt{n} + t\right) &\le e^{-t^2/2}\\ \mathbb{P}\left(\sigma_{\min}(\boldsymbol{Z}) < \sqrt{m} - \sqrt{n} - t\right) &\le e^{-t^2/2}.\end{aligned}$$

The conclusion follows from taking $m \ge \varepsilon^{-2} n$ and $t = \sqrt{m}\varepsilon$ (and from $\varepsilon^2 \ge \delta^2/20$ for $0 < \delta \le 1/2$). For the second part of the lemma, observe that $\boldsymbol{X} = \sum_j \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^*$ with nonnegative eigenvalues λ_j so that

$$\|\mathcal{A}(\boldsymbol{X})\|_1 = \sum_j \sum_i \lambda_j |\langle \boldsymbol{u}_j, \boldsymbol{z}_i\rangle|^2 = \sum_j \lambda_j \|\mathcal{A}(\boldsymbol{u}_j\boldsymbol{u}_j^*)\|_1.$$

The claim follows from (3.4.1). The last claim is a consequence of

$$\|\mathcal{A}(\boldsymbol{X})\|_1 \le \sum_j \sum_i |\lambda_j| |\langle \boldsymbol{u}_j, \boldsymbol{z}_i\rangle|^2$$

together with $\sum_j |\lambda_j| = \|\boldsymbol{X}\|_1$. ∎

Our next result is concerned with the mapping of rank-2 matrices.

Lemma 3.4.2 *Fix $\delta > 0$. Then there are positive numerical constants c_0 and γ_0 such that if $m \ge c_0 [\delta^{-2} \log \delta^{-1}] n$, $\mathcal{A}$ obeys the following property with probability at least $1 - 3e^{-\gamma_0 m \delta^2}$: for any symmetric rank-2 matrix $\boldsymbol{X}$,*

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \ge 0.94(1-\delta)\|\boldsymbol{X}\|. \tag{3.4.2}$$

Proof By homogeneity, it suffices to consider the case where $\|\boldsymbol{X}\| = 1$. Consider then a rank-2 matrix $\boldsymbol{X}$ with eigenvalue decomposition $X = \boldsymbol{u}_1\boldsymbol{u}_1^* - t\boldsymbol{u}_2\boldsymbol{u}_2^*$ with $t \in [-1, 1]$ and orthonormal $\boldsymbol{u}_i$'s. Note that for $t \le 0$, Lemma 3.4.1 already claims a tighter lower bound so it only suffices to consider $t \in [0, 1]$. We have

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 = \frac{1}{m}\sum_{i=1}^m \Big| |\langle \boldsymbol{u}_1, \boldsymbol{z}_i\rangle|^2 - t|\langle \boldsymbol{u}_2, \boldsymbol{z}_i\rangle|^2 \Big| = \frac{1}{m}\sum_i \xi_i,$$

where the ξ_i's are independent copies of the random variable

$$\xi = |Z_1^2 - tZ_2^2|$$

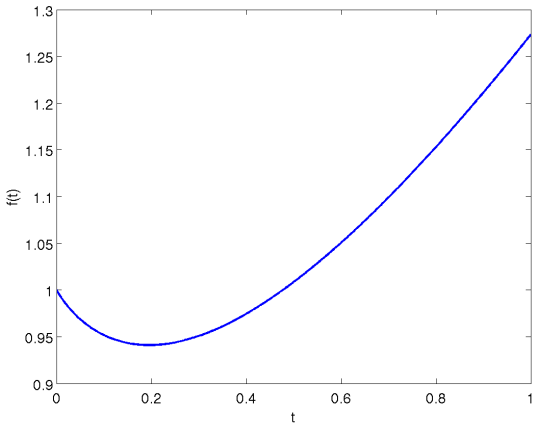

Figure 3.2: $f(t) = \mathbb{E}\,|Z_1^2 - tZ_2^2|$ as a function of t.

in which Z_1 and Z_2 are independent standard normal variables. This comes from the fact that $\langle \boldsymbol{u}_1, \boldsymbol{z}_i \rangle$ and $\langle \boldsymbol{u}_2, \boldsymbol{z}_i \rangle$ are independent standard normal. We calculate below that

$$\mathbb{E}\,\xi = f(t) = \frac{2}{\pi}\Big(2\sqrt{t} + (1-t)(\pi/2 - 2\arctan(\sqrt{t}))\Big). \tag{3.4.3}$$

The graph of this function is shown in Figure 3.2; we check that $f(t) \geq 0.94$ for all $t \in [0,1]$.

We now need a deviation bound concerning the fluctuation of $m^{-1}\sum_i \xi_i$ around its mean and this is achieved by classical Chernoff bounds. Note that $\xi \leq Z_1^2 + |t|Z_2^2$ is a sub-exponential variable and thus, $\|\xi\|_{\psi_1} := \sup_{p\geq 1} [\mathbb{E}\,|\xi|^p]^{1/p}$ is finite.[4]

Lemma 3.4.3 (Bernstein-type inequality [85]) *Let $X_1, \ldots, X_m$ be i.i.d. sub-exponential random variables. Then*

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}\,X_1\Big| \geq \varepsilon\Big) \leq 2\exp\Big[-c_0\, m \min\Big(\frac{\varepsilon^2}{\|X\|_{\psi_1}^2}, \frac{\varepsilon}{\|X\|_{\psi_1}}\Big)\Big]$$

in which c_0 is a positive numerical constant.

We have thus established that for a fixed X,

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \geq (0.94 - \varepsilon_0)\|\boldsymbol{X}\|$$

with probability at least $1 - 2e^{-\gamma_0 m\varepsilon_0^2}$ (provided $\varepsilon_0 \leq \|\xi\|_{\psi_1}$, which we assume).

To complete the argument, let $\mathcal{S}_\varepsilon$ be an ε net of the unit sphere, $\mathcal{T}_\varepsilon$ be an ε net of $[0,1]$, and set

$$\mathcal{N}_\varepsilon = \{\boldsymbol{X} = \boldsymbol{u}_1\boldsymbol{u}_1^* - t\boldsymbol{u}_2\boldsymbol{u}_2^* : (\boldsymbol{u}_1, \boldsymbol{u}_2, t) \in \mathcal{S}_\varepsilon \times \mathcal{S}_\varepsilon \times \mathcal{T}_\varepsilon\}.$$

Since $|\mathcal{S}_\varepsilon| \leq (3/\varepsilon)^n$, we have

$$|\mathcal{N}_\varepsilon| \leq (3/\varepsilon)^{2n+1}.$$

[4] It would be possible to compute a bound on this quantity but we will not pursue this at the moment.

Now for any $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{u}^* - t\boldsymbol{v}\boldsymbol{v}^*$, consider the approximation $\boldsymbol{X}_0 = \boldsymbol{u}_0\boldsymbol{u}_0^* - t_0\boldsymbol{v}_0\boldsymbol{v}_0^* \in \mathcal{N}_\varepsilon$, where $\|\boldsymbol{u}_0 - \boldsymbol{u}\|_2$, $\|\boldsymbol{v} - \boldsymbol{v}_0\|_2$ and $|t - t_0|$ are each at most ε. We claim that

$$\|\boldsymbol{X} - \boldsymbol{X}_0\|_1 \le 9\varepsilon, \tag{3.4.4}$$

and postpone the short proof. On the intersection of

$$E_1 = \{m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \le (1+\delta_1)\|X\|_1, \text{ for all } \boldsymbol{X}\}$$

with $E_2 := \{m^{-1}\|\mathcal{A}(\boldsymbol{X}_0)\|_1 \ge (0.94 - \varepsilon)\|\boldsymbol{X}_0\|, \text{ for all } \boldsymbol{X}_0 \in \mathcal{N}_\varepsilon\}$,

$$\begin{aligned} m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 &\ge \|\mathcal{A}(\boldsymbol{X}_0)\|_1 - \|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_0)\|_1 \\ &\ge (0.94 - \varepsilon)\|\boldsymbol{X}_0\| - 9(1+\delta_1)\varepsilon \\ &\ge (0.94 - \varepsilon)(\|\boldsymbol{X}\| - \|\boldsymbol{X}_0 - \boldsymbol{X}\|) - 9(1+\delta_1)\varepsilon \\ &\ge (0.94 - \varepsilon)(1 - 5\varepsilon) - 9(1+\delta_1)\varepsilon \\ &\ge 0.94 - (15 + 9\delta_1)\varepsilon, \end{aligned}$$

which is the desired bound by setting $0.94\delta = (15 + 9\delta_1)\varepsilon$. In conclusion, set $\delta_1 = 1/2$ and take $\varepsilon = 0.94\delta/20$. Then E_1 holds with probability at least $1 - O(e^{-\gamma_1 m\varepsilon^2})$ provided m obeys the condition of the theorem. Further, Lemma 3.4.2 states that E_2 holds with probability at least $1 - 2e^{-\gamma_2 m}$. This concludes the proof provided we check (3.4.4).

We begin with

$$\|\boldsymbol{X} - \boldsymbol{X}_0\|_1 \le \|\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{u}_0\boldsymbol{u}_0^*\|_1 + |t - t_0|\|\boldsymbol{v}\boldsymbol{v}^*\|_1 + |t_0|\|\boldsymbol{v}\boldsymbol{v}^* - \boldsymbol{v}_0\boldsymbol{v}_0^*\|_1.$$

Now

$$\|\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{u}_0\boldsymbol{u}_0^*\|_1 \le 2\|\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{u}_0\boldsymbol{u}_0^*\| \le 4\|\boldsymbol{u} - \boldsymbol{u}_0\|_2,$$

where the first inequality follows from the fact that $\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{u}_0\boldsymbol{u}_0^*$ is of rank at most 2, and the second follows from

$$\begin{aligned} \|\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{u}_0\boldsymbol{u}_0^*\| &= \sup_{\|\boldsymbol{x}\|_2=1} \left|\langle \boldsymbol{u}_0, \boldsymbol{x}\rangle^2 - \langle \boldsymbol{u}, \boldsymbol{x}\rangle^2\right| \\ &= \sup_{\|\boldsymbol{x}\|_2=1} \left|\langle \boldsymbol{u} - \boldsymbol{u}_0, \boldsymbol{x}\rangle\langle \boldsymbol{u} + \boldsymbol{u}_0, \boldsymbol{x}\rangle\right| \le \|\boldsymbol{u} - \boldsymbol{u}_0\|_2\|\boldsymbol{u} + \boldsymbol{u}_0\|_2 \le 2\|\boldsymbol{u} - \boldsymbol{u}_0\|_2. \end{aligned}$$

Similarly, $\|\boldsymbol{v}\boldsymbol{v}^* - \boldsymbol{v}_0\boldsymbol{v}_0^*\|_1 \le 4\varepsilon$ and this concludes the proof.[5] ∎

Lemma 3.4.4 *Let Z_1 and Z_2 be independent $\mathcal{N}(0,1)$ variables and $t \in [0,1]$. We have*

$$E|Z_1^2 - tZ_2^2| = f(t),$$

where $f(t)$ is given by (3.4.3)*.*

[5]The careful reader will remark that we have also used $\|\boldsymbol{X} - \boldsymbol{X}_0\| \le 5\varepsilon$, which also follows from our calculations.

Proof Set

$$\rho = \frac{1-t}{1+t} \text{ and } \cos\theta = \rho$$

in which $\theta \in [0, \pi/2]$. By using polar coordinates, we have

$$\begin{aligned}
\mathbb{E}\,|Z_1^2 - tZ_2^2| &= \frac{1}{2\pi}\int_0^\infty r^3 e^{-r^2/2}\,dr \int_0^{2\pi} |\cos^2\phi - t\sin^2\phi|\,d\phi \\
&= \frac{1}{\pi}\int_0^{2\pi} |\cos^2\phi - t\sin^2\phi|\,d\phi \\
&= \frac{2}{\pi}\int_0^{\pi} |\cos^2\phi - t\sin^2\phi|\,d\phi
\end{aligned}$$

Now using the identities $\cos^2\phi = (1+\cos 2\phi)/2$ and $\sin^2\phi = (1-\cos 2\phi)/2$, we have

$$\begin{aligned}
\mathbb{E}\,|Z_1^2 - tZ_2^2| &= \frac{1+t}{\pi}\int_0^{\pi} |\cos 2\phi + \rho|\,d\phi \\
&= \frac{1+t}{2\pi}\int_0^{2\pi} |\cos\phi + \rho|\,d\phi \\
&= \frac{1+t}{\pi}\int_0^{\pi} |\cos\phi + \rho|\,d\phi \\
&= \frac{1+t}{\pi}\int_0^{\pi} |\rho - \cos\phi|\,d\phi \\
&= \frac{1+t}{\pi}\Big[\int_0^{\theta} \cos\phi - \rho\,d\phi + \int_\theta^{\pi} \rho - \cos\phi\,d\phi\Big] \\
&= \frac{2}{\pi}(1+t)[\sin\theta + \rho(\pi/2 - \theta)].
\end{aligned}$$

We recognize (3.4.3). ∎

3.5 Dual Certificates

To prove our main theorem, it remains to show that one can construct an inexact dual certificate $\boldsymbol{Y}$ obeying the conditions of Lemma 3.3.1.

Preliminaries

The linear mapping $\mathcal{A}^*\mathcal{A}$ is of the form[6]

$$\mathcal{A}^*\mathcal{A} = \sum_{i=1}^m \boldsymbol{z}_i\boldsymbol{z}_i^* \otimes \boldsymbol{z}_i\boldsymbol{z}_i^*,$$

[6]For symmetric matrices, $\boldsymbol{A} \otimes \boldsymbol{B}$ is the linear mapping $\boldsymbol{H} \mapsto \langle \boldsymbol{A}, \boldsymbol{H}\rangle \boldsymbol{B}$.

which is another way to express that $\mathcal{A}^*\mathcal{A}(\boldsymbol{X}) = \sum_i \langle \boldsymbol{z}_i\boldsymbol{z}_i^*, \boldsymbol{X}\rangle \boldsymbol{z}_i\boldsymbol{z}_i^*$. Now observe the simple identity:

$$\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^* \otimes \boldsymbol{z}_i\boldsymbol{z}_i^*] = 2\mathcal{I} + \boldsymbol{I}_n \otimes \boldsymbol{I}_n := \mathcal{S}, \tag{3.5.1}$$

where $\mathcal{I}$ is the identity operator and $\boldsymbol{I}_n$ the n-dimensional identity matrix. Put differently, this means that for all $\boldsymbol{X}$,

$$\mathcal{S}(\boldsymbol{X}) = 2\boldsymbol{X} + \mathrm{Tr}(\boldsymbol{X})\boldsymbol{I}.$$

The proof is a simple calculation and omitted. It is also not hard to see that the mapping $\mathcal{S}$ is invertible and its inverse is given by

$$\mathcal{S}^{-1} = \frac{1}{2}\Big(\mathcal{I} - \frac{1}{n+2}\boldsymbol{I}_n \otimes \boldsymbol{I}_n\Big) \quad \Leftrightarrow \quad \mathcal{S}^{-1}(\boldsymbol{X}) = \frac{1}{2}\Big(\boldsymbol{X} - \frac{1}{n+2}\,\mathrm{Tr}(\boldsymbol{X})\boldsymbol{I}_n\Big).$$

We will use this object in the definition of our dual certificate.

Construction

For pedagogical reasons, we first introduce a possible candidate certificate defined by

$$\bar{\boldsymbol{Y}} := \frac{1}{m}\mathcal{A}^*\mathcal{A}\mathcal{S}^{-1}(\boldsymbol{e}_1\boldsymbol{e}_1^*). \tag{3.5.2}$$

Clearly, $\bar{\boldsymbol{Y}}$ is in the range of $\mathcal{A}^*$ as required. To justify this choice, the law of large numbers gives that in the limit of infinitely many samples,

$$\lim_{m\to\infty} \frac{1}{m}\sum_i (\boldsymbol{z}_i\boldsymbol{z}_i^* \otimes \boldsymbol{z}_i\boldsymbol{z}_i^*)\mathcal{S}^{-1}(\boldsymbol{e}_1\boldsymbol{e}_1^*) = \mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i^* \otimes \boldsymbol{z}_i\boldsymbol{z}_i^*)\mathcal{S}^{-1}(\boldsymbol{e}_1\boldsymbol{e}_1^*) = \boldsymbol{e}_1\boldsymbol{e}_1^*.$$

In other words, in the limit of large samples, we have a perfect certificate since $\bar{\boldsymbol{Y}}_T = \boldsymbol{e}_1\boldsymbol{e}_1^*$ and $\bar{\boldsymbol{Y}}_T^\perp = 0$. Our hope is that the sample average is sufficiently close to the population average so that one can check (3.3.4). In order to show that this is the case, it will be useful to think of $\bar{\boldsymbol{Y}}$ (3.5.2) as the random sum

$$\bar{\boldsymbol{Y}} = \frac{1}{m}\sum_i \boldsymbol{Y}_i,$$

where each matrix $\boldsymbol{Y}_i$ is an independent copy of the random matrix

$$\frac{1}{2}\Big[z_1^2 - \frac{1}{n+2}\|\boldsymbol{z}\|_2^2\Big]\boldsymbol{z}\boldsymbol{z}^*$$

in which $\boldsymbol{z} = (z_1, \ldots, z_n) \sim \mathcal{N}(0, I)$.

We would like to make an important point before continuing. We have seen that all we need from $\bar{\boldsymbol{Y}}$ is

$$\|\bar{\boldsymbol{Y}}_T - \boldsymbol{e}_1\boldsymbol{e}_1^*\|_2 \le 1/3$$

(and $\|\bar{\boldsymbol{Y}}_T^\perp\| \le 1/2$). This is in stark contrast with David Gross' approach [39] which requires a very small misfit, i.e. an error of at most $1/n^2$. In turn, this loose bound has an enormous implication: it eliminates the need for the golfing scheme and allows for the simple certificate candidate (3.5.2). In fact, our certificate can be seen as the first iteration of Gross' golfing scheme.

Truncation

For technical reasons, it is easier to work with a truncated version of $\bar{\boldsymbol{Y}}$ and our dual certificate is taken to be

$$\boldsymbol{Y} = \frac{1}{m}\sum_i \boldsymbol{Y}_i\, 1_{E_i}, \tag{3.5.3}$$

where the $\boldsymbol{Y}_i$'s are as before and 1_{E_i} are independent copies of 1_E with

$$E = \{|z_1| \le \sqrt{2\beta \log n}\} \cap \{\|\boldsymbol{z}\|_2 \le \sqrt{3n}\}.$$

We shall work with $\beta = 3$ so that $|z_1| \le \sqrt{6 \log n}$.

Lemma 3.5.1 *Let $\boldsymbol{Y}$ be as in* (3.5.3)*. Then*

$$\mathbb{P}\Big(\|\boldsymbol{Y}_T - \boldsymbol{e}_1\boldsymbol{e}_1^*\|_2 \ge \frac{1}{3}\Big) \le 2\exp\Big(-\gamma\frac{m}{n}\Big), \tag{3.5.4}$$

where $\gamma > 0$ is an absolute constant. This holds with the proviso that $m \ge c_1\, n$ for some numerical constant $c_1 > 0$, and that n is sufficiently large.

Lemma 3.5.2 *Let $\boldsymbol{Y}$ be as in* (3.5.3)*. Then*

$$\mathbb{P}\Big(\|\boldsymbol{Y}_T^{\perp}\| \ge \frac{1}{2}\Big) \le 4\exp\Big(-\gamma\frac{m}{\log n}\Big). \tag{3.5.5}$$

where $\gamma > 0$ is an absolute constant. This holds with the proviso that $m \ge c_1\, n \log n$ for some numerical constant $c_1 > 0$, and that n is sufficiently large.

$\boldsymbol{Y}$ on T and proof of Lemma 3.5.1

It is obvious that for any symmetric matrix $\boldsymbol{X} \in T$,

$$\|\boldsymbol{X}\|_2 \le \sqrt{2}\|\boldsymbol{X}\boldsymbol{e}_1\|_2$$

since only the first row and column are nonzero. We have

$$\boldsymbol{Y}_T\boldsymbol{e}_1 - \boldsymbol{e}_1 = \frac{1}{m}\sum_{i=1}^m \boldsymbol{y}_i 1_{E_i} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{e}_1\, 1_{E_i^c}, \tag{3.5.6}$$

where the $\boldsymbol{y}_i$'s are independent copies of the random vector

$$\boldsymbol{y} = \frac{1}{2}\Big[z_1^2 - \frac{1}{n+2}\|\boldsymbol{z}\|_2^2\Big] z_1\, \boldsymbol{z} - \boldsymbol{e}_1 := (\xi z_1)\, \boldsymbol{z} - \boldsymbol{e}_1. \tag{3.5.7}$$

We claim that

$$\Big\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{e}_1\, 1_{E_i^c}\Big\|_2 \le 1/9,$$

with probability at least $1 - 2e^{-\gamma m}$ for some $\gamma > 0$. This is a simple application of Bernstein's inequality. Set $\pi(\beta) = \mathbb{P}(E_i^c)$ and observe that

$$\pi(\beta) = \mathbb{P}(|z_1| \geq \sqrt{2\beta \log n}) + \mathbb{P}(\|\boldsymbol{z}\|_2^2 \geq 3n) \leq n^{-\beta} + e^{-\frac{n}{3}}. \tag{3.5.8}$$

The right-hand side follows from $\mathbb{P}(|z_1| \geq t) \leq e^{-t^2/2}$ which holds for $t \geq 1$ and from $\mathbb{P}(\|\boldsymbol{z}\|_2^2 \geq 3n) \leq e^{-n/3}$. In turn, this last bound follows from

$$\mathbb{P}(\|\boldsymbol{z}\|_2^2 - n \geq \sqrt{2n}t + t^2) \leq e^{-t^2/2}.$$

Returning to Bernstein, this gives

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^m 1_{E_i^c} - \pi(\beta)\Big| \geq t) \leq 2\exp\Big(-\frac{mt^2}{2\pi(\beta) + 2t/3}\Big).$$

Setting $t = 1/18$, $\beta = 3$ and taking n large enough so that $\pi(3) \leq 1/18$ proves the claim.

The main task is to bound the 2-norm of the sum $\sum_{i=1}^m \boldsymbol{y}_i 1_{E_i}$ and a convenient way to do this is via the vector Bernstein inequality.

Theorem 3.5.3 (Vector Bernstein inequality) *Let $\boldsymbol{x}_i$ be a sequence of independent random vectors and set $V \geq \sum_i \mathbb{E}\,\|\boldsymbol{x}_i\|_2^2$. Then for all $t \leq V/\max\|\boldsymbol{x}_i\|_2$, we have*

$$\mathbb{P}(\|\sum_i (\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\|_2 \geq \sqrt{V} + t) \leq e^{-t^2/4V}.$$

It is because this inequality requires bounded random vectors that we work with the truncation $\sum_{i=1}^m \boldsymbol{y}_i 1_{E_i}$.

Put $\bar{\boldsymbol{y}} = \boldsymbol{y}\, 1_E$. Since $\|\bar{\boldsymbol{y}}\|_2^2 \leq \|\boldsymbol{y}\|_2^2$, we first compute $\mathbb{E}\,\|\boldsymbol{y}\|_2^2$. We have

$$\|\boldsymbol{y}\|_2^2 = \|\boldsymbol{z}\|_2^2 z_1^2 \xi^2 - 2z_1^2\xi + 1, \qquad \xi = \frac{1}{2}\Big[z_1^2 - \frac{1}{n+2}\|\boldsymbol{z}\|_2^2\Big],$$

and a little bit of algebra yields

$$\|\boldsymbol{y}\|_2^2 = \frac{1}{4}z_1^6\|\boldsymbol{z}\|_2^2 - \frac{1}{2(n+2)}z_1^4\|\boldsymbol{z}\|_2^4 + \frac{1}{4(n+2)^2}z_1^2\|\boldsymbol{z}\|_2^6 - z_1^4 + \frac{1}{n+2}z_1^2\|\boldsymbol{z}\|_2^2 + 1.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{y}\|_2^2\right] &= \frac{1}{4}(15n+90) - \frac{1}{2(n+2)}(3n^2+30n+72) + \frac{1}{4(n+2)}(n+4)(n+6) - 1 \\ &\leq 4(n+4),\end{aligned} \tag{3.5.9}$$

where we have used the following identities

$$\begin{aligned}\mathbb{E}\left[z_1^2\|\boldsymbol{z}\|_2^2\right] &= n+2, \\ \mathbb{E}\left[z_1^2\|\boldsymbol{z}\|_2^6\right] &= (n+2)(n+4)(n+6), \\ \mathbb{E}\left[z_1^4\|\boldsymbol{z}\|_2^4\right] &= 3n^2+30n+72, \\ \mathbb{E}\left[z_1^6\|\boldsymbol{z}\|_2^2\right] &= 15n+90.\end{aligned}$$

Second, on the event of interest we have $|\xi| \leq \beta \log n$ (assuming $2\beta \log n \geq 3$), $|z_1| \leq \sqrt{2\beta \log n}$ and $\|\boldsymbol{z}\|_2 \leq \sqrt{3n}$ and, therefore,

$$\|\bar{\boldsymbol{y}}\|_2 \leq \sqrt{6n}\,(\beta \log n)^{3/2} + 1 \leq \sqrt{7n}(\beta \log n)^{3/2}$$

provided n is large enough.

Third, observe that by symmetry, all the entries of $\bar{\boldsymbol{y}}$ but the first have mean zero. Hence,

$$\|\,\mathbb{E}\,\bar{\boldsymbol{y}}\|_2 = |\,\mathbb{E}\,y_1 - \bar{y}_1| = |\,\mathbb{E}\,1_{E^c} y_1| \leq \sqrt{\mathbb{P}\,(E^c)}\,\sqrt{\mathbb{E}\,y_1^2}.$$

We have

$$y_1^2 = (\xi z_1^2 - 1)^2 = \frac{1}{4} z_1^8 - z_1^4 + \frac{1}{n+2}\|z\|_2^2 z_1^2 - \frac{1}{2(n+2)}\|z\|_2^2 z_1^6 + \frac{1}{4(n+2)^2}\|z\|_2^4 z_1^4 + 1$$

and using the identities above

$$\mathbb{E}\,y_1^2 = \frac{101}{4} - \frac{27n^2 + 210n + 288}{4(n+2)^2} \leq 22,$$

which gives

$$\|\,\mathbb{E}\,\bar{\boldsymbol{y}}\|_2 \leq \sqrt{22(n^{-\beta} + e^{-\frac{n}{3}})}.$$

Finally, with $V = 4m(n+4)$, Bernstein's inequality gives that for each

$$t \leq 4(n+4)/[\sqrt{7n}(\beta \log n)^{3/2}]$$

,

$$\|m^{-1}\sum_i (\bar{\boldsymbol{y}}_i - \mathbb{E}\,\bar{\boldsymbol{y}}_i)\|_2 \geq 2\sqrt{\frac{n+4}{m}} + t$$

with probability at most $\exp\!\big(-\frac{mt^2}{16(n+4)}\big)$. It follows that

$$\|m^{-1}\sum_i \bar{\boldsymbol{y}}_i\|_2 \geq \sqrt{22(n^{-\beta} + e^{-\frac{n}{3}})} + 2\sqrt{\frac{n+4}{m}} + t$$

with at most the same probability. Our result follows by taking $t = 1/6$, $\beta = 3$, $m \geq c_1 n$ where n and c_1 are sufficiently large such that

$$\sqrt{22(n^{-\beta} + e^{-\frac{n}{3}})} + 2\sqrt{\frac{n+4}{m}} + \frac{1}{6} \leq \frac{2}{9}.$$

$\boldsymbol{Y}$ on $T^\perp$ and proof of Lemma 3.5.2

We have

$$\boldsymbol{Y}_T^\perp = \frac{1}{m}\sum_i \boldsymbol{X}_i\, 1_{E_i},$$

where the $\boldsymbol{X}_i$'s are independent copies of the random matrix

$$\boldsymbol{X} = \frac{1}{2}\Big[z_1^2 - \frac{1}{n+2}\|\boldsymbol{z}\|_2^2\Big]\,\mathcal{P}_{T^\perp}(\boldsymbol{z}\boldsymbol{z}^T). \tag{3.5.10}$$

One natural way to bound the norm of this random sum is via the operator Bernstein's inequality. We develop a more customized approach, which gives sharper results.

Decompose $\boldsymbol{X}$ as

$$\boldsymbol{X} = \frac{1}{2}\Big[z_1^2 - 1\Big]\,\mathcal{P}_{T^\perp}(\boldsymbol{z}\boldsymbol{z}^T) + \frac{1}{2}\Big[1 - \frac{1}{n+2}\|\boldsymbol{z}\|_2^2\Big]\,\mathcal{P}_{T^\perp}(\boldsymbol{z}\boldsymbol{z}^T) := \boldsymbol{X}^{(0)} + \boldsymbol{X}^{(1)}.$$

Note that since z_1 and $\mathcal{P}_{T^\perp}(\boldsymbol{z}\boldsymbol{z}^T)$ are independent, we have $\mathbb{E}\,\boldsymbol{X}^{(0)} = 0$ and thus, $\mathbb{E}\,\boldsymbol{X}^{(1)} = 0$ since $\mathbb{E}\,\boldsymbol{X} = 0$. With $\bar{\boldsymbol{X}}_i^{(0)} = \boldsymbol{X}_i^{(0)} 1_{E_i}$ and similarly for $\bar{\boldsymbol{X}}_i^{(1)}$, it then suffices to show that

$$\Big\|\sum_i \bar{\boldsymbol{X}}_i^{(0)}\Big\| \le m/4 \quad \text{and} \quad \Big\|\sum_i \bar{\boldsymbol{X}}_i^{(1)}\Big\| \le m/4 \tag{3.5.11}$$

with large probability. Write the norm as

$$\Big\|\sum_i \bar{\boldsymbol{X}}_i^{(0)}\Big\| = \sup_{\boldsymbol{u}} \Big|\sum_i \langle \boldsymbol{u}, \bar{\boldsymbol{X}}_i^{(0)}\boldsymbol{u}\rangle\Big|,$$

where the supremum is over all unit vectors $\boldsymbol{u}$ that are orthogonal to $\boldsymbol{e}_1$. The strategy is now to find a bound on the right-hand side for each fixed $\boldsymbol{u}$ and apply a covering argument to control the supremum over the whole unit sphere. In order to do this, we shall make use of a classical large deviation result.

Theorem 3.5.4 (Bernstein inequality) *Let $\{X_i\}$ be a finite sequence of independent random variables. Suppose that there exist V_i and c such that for all X_i and all $k \ge 3$,*

$$\mathbb{E}\,|X_i|^k \le \frac{1}{2}k!V_i c_0^{k-2}.$$

Then for all $t \ge 0$,

$$\mathbb{P}\Big(\Big|\sum_i X_i - \mathbb{E}\,X_i\Big| \ge t\Big) \le 2\exp\Big(-\frac{t^2}{2\sum_i V_i + 2c_0 t}\Big). \tag{3.5.12}$$

For the first sum in (3.5.11), we write

$$\sum_i \langle \boldsymbol{u}, \bar{\boldsymbol{X}}_i^{(0)}\boldsymbol{u}\rangle = \sum_i \eta_i\, 1_{E_i},$$

where the η_i's are independent copies of

$$\eta = \frac{1}{2}\Big[z_1^2 - 1\Big]\langle \boldsymbol{z}, \boldsymbol{u}\rangle^2.$$

The point of the decomposition $\boldsymbol{X}^{(0)} + \boldsymbol{X}^{(1)}$ is that z_1 and $\langle \boldsymbol{z}, \boldsymbol{u}\rangle$ are independent since $\boldsymbol{u}$ is orthogonal to $\boldsymbol{e}_1$. We have $\mathbb{E}\,\eta = 0$ and for $k \geq 2$,

$$\mathbb{E}\,|\eta\, 1_E|^k \leq 2^{-k}\,\mathbb{E}\,|(z_1^2 - 1)\, 1_{\{z_1^2 \leq 2\beta \log n\}}|^k\ \mathbb{E}\,|\langle \boldsymbol{z}, \boldsymbol{u}\rangle|^{2k}.$$

First,

$$\mathbb{E}\,|(z_1^2 - 1)\, 1_{\{z_1^2 \leq 2\beta \log n\}}|^k \leq (2\beta \log n)^{k-2}\,\mathbb{E}(z_1^2 - 1)^2 = 2(2\beta \log n)^{k-2}.$$

Second, the moments of a chi-square variable with one degree of freedom are well known:

$$\mathbb{E}\,|\langle \boldsymbol{z}, \boldsymbol{u}\rangle|^{2k} = 1 \times 3 \times \ldots \times (2k-1) \leq 2^k k!$$

Hence we can apply Bernstein inequality with $V_i = 4, i = 1, \ldots, m$, and $c_0 = 2\beta \log n$ and, obtain

$$\mathbb{P}\Big(\Big|\sum_i \eta_i\, 1_{E_i} - \mathbb{E}[\eta_i\, 1_{E_i}]\Big| \geq mt\Big) \leq 2\exp\Big(-\frac{m}{4}\frac{t^2}{2 + \beta t \log n}\Big).$$

We now note that

$$|\,\mathbb{E}\,\eta_i 1_{E_i}| = |\,\mathbb{E}\,\eta_i 1_{E_i^c}| \leq \sqrt{\mathbb{P}(E_i^c)}\sqrt{\mathbb{E}\,\eta_i^2} = \sqrt{\frac{3\pi(\beta)}{2}}$$

which gives

$$\mathbb{P}\Big(m^{-1}\Big|\sum_i \eta_i\, 1_{E_i}\Big| \geq t + \sqrt{\frac{3\pi(\beta)}{2}}\Big) \leq 2\exp\Big(-\frac{m}{4}\frac{t^2}{2 + \beta t \log n}\Big).$$

For instance, take $t = 1/12$, $\beta = 3$, $m \geq c_1 n$ and n large enough to get

$$\mathbb{P}\Big(m^{-1}\Big|\sum_i \eta_i\, 1_{E_i}\Big| \geq 1/8\Big) \leq 2\exp\Big(-\gamma\frac{m}{\log n}\Big).$$

To derive a bound about $\|\bar{\boldsymbol{X}}^{(0)}\|$, we use (see Lemma 4 in [85])

$$\sup_u \Big|\langle \boldsymbol{u}, \bar{\boldsymbol{X}}^{(0)} u\rangle\Big| \leq 2 \sup_{\boldsymbol{u} \in \mathcal{N}_{1/4}} \Big|\langle \boldsymbol{u}, \bar{\boldsymbol{X}}^{(0)} \boldsymbol{u}\rangle\Big|,$$

where $\mathcal{N}_{1/4}$ is a $1/4$-net of the unit sphere $\{\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1, \boldsymbol{u} \perp \boldsymbol{e}_1\}$. Since $|\mathcal{N}_{1/4}| \leq 9^n$,

$$\mathbb{P}(m^{-1}\|\bar{\boldsymbol{X}}^{(0)}\| > 1/4) \leq \mathbb{P}\Big(m^{-1} \sup_{\boldsymbol{u} \in \mathcal{N}_{1/4}} \Big|\langle \boldsymbol{u}, \bar{\boldsymbol{X}}^{(0)} \boldsymbol{u}\rangle\Big| > 1/8\Big) \leq 9^n \times 2\exp\Big(-\gamma\frac{m}{\log n}\Big).$$

We deal with the second term in a similar way, and write

$$\sum_i \langle \boldsymbol{u}, \bar{\boldsymbol{X}}_i^{(1)} \boldsymbol{u} \rangle = \sum_i \eta_i \, 1_{E_i},$$

where the η_i's are now independent copies of

$$\eta = \frac{1}{2}\Big[1 - \frac{\|\boldsymbol{z}\|_2^2}{n+2}\Big] \langle \boldsymbol{z}, \boldsymbol{u} \rangle^2.$$

On E, $\|\boldsymbol{z}\|_2^2 \leq 3n$ and, therefore, $\mathbb{E}\,|\eta\, 1_E|^k \leq 2^k k!$. We can apply Bernstein's inequality with $c_0 = 2$ and $V = 8m$, which gives

$$\mathbb{P}\Big(\Big|\sum_i \eta_i \, 1_{E_i} - \mathbb{E}[\eta_i \, 1_{E_i}]\Big| \geq mt\Big) \leq 2\exp\Big(-\frac{m}{4}\frac{t^2}{4+t}\Big).$$

The remainder of the proof is identical to that above and is therefore omitted.

Proof of Theorem 3.2.1

We now assemble the various intermediate results to establish Theorem 3.2.1. As pointed out, Theorem 3.2.1 follows immediately from Lemma 3.3.1, which in turn hinges on the validity of the conditions stated in (3.3.2), (3.3.3), and (3.3.4).

Lemma 3.4.1 asserts that condition (3.3.2) holds with probability of failure at most p_1, where $p_1 = 2e^{-\gamma_1 m}$ and here and below, $\gamma_1, \ldots, \gamma_4$ are positive numerical constants. Similarly, Lemma 3.4.2 shows that condition (3.3.3) holds with probability of failure at most p_2, where $p_2 = 3e^{-\gamma_2 m}$. In both cases we need that $m > cn$ for an absolute constant $c > 0$.

Proceeding to the dual certificate in (3.3.4), we note that Lemma 3.5.1 establishes the first part of the dual certificate with a probability of failure at most p_3, where $p_3 = 3e^{-\gamma_3 m/n}$. The second part of the dual certificate in (3.3.4) is shown in Lemma 3.5.2 to hold with probability of failure at most p_4, where $p_4 = 4e^{-\gamma_4 \frac{m}{\log n}}$. In the former case we need $m > cn$ for an absolute constant $c > 0$ and in the latter $m > c' n \log n$.

Finally, the union bound gives that under the hypotheses of Theorem 3.2.1, exact recovery holds with probability at least $1 - 3e^{-\gamma m/n}$ for some $\gamma > 0$, as claimed.

3.6 The Complex Model

This section proves that Theorem 3.2.1 holds for the complex model as well. Not surprisingly, the main steps of the proof are the same as in the real case, but there are here and there some noteworthy differences. Instead of deriving the whole proof, we will carefully indicate the nontrivial changes that need to be carried out.

First, we can work with $\boldsymbol{x} = \boldsymbol{e}_1$ because of rotational invariance, and with independent complex valued Gaussian sequences $\boldsymbol{z}_i \sim \mathcal{CN}(0, I, 0)$. This means that the real and imaginary parts of $\boldsymbol{z}_i$ are independent white noise sequences with variance $1/2$.

The key Lemma 3.3.1 only requires a slight adjustment in the numerical constants. The reason for this is that while Lemma 3.4.1 does not require any modification, Lemma 3.4.2 changes slightly; in particular, the numerical constants are somewhat different. Here is the properly adjusted complex version.

Lemma 3.6.1 *Fix $\delta > 0$. Then there are positive numerical constants c_0 and γ_0 such that if $m \geq c_0 \left[\delta^{-2} \log \delta^{-1}\right] n$, $\mathcal{A}$ has the following property with probability at least $1 - 3e^{-\gamma_0 m \delta^2}$: for any Hermitian rank-2 matrix X,*

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \geq 2(\sqrt{2}-1)(1-\delta)\|\boldsymbol{X}\| \geq 0.828(1-\delta)\|\boldsymbol{X}\|. \tag{3.6.1}$$

The proof of this lemma follows essentially the proof of Lemma 3.4.2. The function $f(t)$ (cf. equation (3.4.3)) now takes the form

$$\mathbb{E}\,\xi = f(t) = \frac{1+t^2}{1+t}, \tag{3.6.2}$$

where $\xi = \left||Z_1|^2 - t|Z_2|^2\right|$, with Z_1 and Z_2 independent $\mathcal{CN}(0,1,0)$, as demonstrated in the following lemma.

Lemma 3.6.2 *Let Z_1 and Z_2 be independent $\mathcal{CN}(0,1,0)$ variables and $t \in [0,1]$. We have*

$$E||Z_1|^2 - t|Z_2|^2| = f(t),$$

where $f(t)$ is given by (3.6.2)*.*

Proof Set

$$\rho = \frac{1-t}{1+t} \text{ and } \cos\theta = \rho$$

in which $\theta \in [0, \pi/2]$. By using polar coordinates for the variables (x_1, y_1) associated with Z_1 and (x_2, y_2), associated with Z_2 we have

$$\begin{aligned}\mathbb{E}\,||Z_1|^2 - t|Z_2|^2| &= \frac{1}{2}\int_0^\infty \int_0^\infty |r_1^2 - t r_2^2| r_1 r_2 e^{-r_1^2/2} e^{-r_2^2/2}\, dr_1 dr_2 \\ &= \frac{1}{8}\int_0^\infty r^5 e^{-r^2/2}\, dr \int_0^{2\pi} |\sin\phi\cos\phi||\cos^2\phi - t\sin^2\phi|\, d\phi,\end{aligned}$$

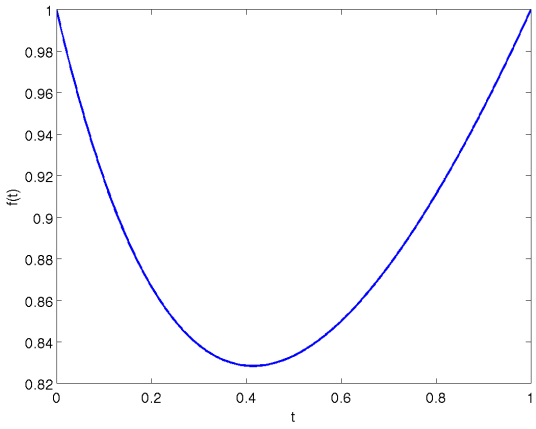

Figure 3.3: The function $f(t)$ in (3.6.2) as a function of t.

where we used polar coordinates again in variables (r_1, r_2). Now using the identities $\cos^2\phi = (1+\cos 2\phi)/2$, $\sin^2\phi = (1-\cos 2\phi)/2$ and $2\sin\phi\cos\phi = \sin 2\phi$ we have

$$\begin{aligned}
\mathbb{E}\,|Z_1^2 - tZ_2^2| &= \frac{1}{2}\int_0^\pi |\sin 2\phi||\cos 2\phi + \rho|\,d\phi \\
&= \frac{1}{2}\Big[\int_0^\theta \sin\phi(\cos\phi - \rho)\,d\phi + \int_\theta^\pi \sin\phi(\rho - \cos\phi)\,d\phi\Big] \\
&= \frac{1}{2}(1+t)[-\frac{1}{2}\cos 2\theta + 2\rho\cos\theta + \frac{1}{2}] \\
&= \frac{1}{2}(1+t)[\rho^2 + 1] \\
&= \frac{1+t^2}{1+t}
\end{aligned}$$

as claimed. ∎

The graph of $f(t)$ is shown in Figure 3.3. The minimum of this function on $[0,1]$ is $2(\sqrt{2}-1) > 0.828$. Furthermore, the covering argument in that proof has to be adapted; for example, unit spheres need to be replaced by complex unit spheres.

A consequence of this change in numerical values is that the numerical factors in Lemma 3.3.2 need to be adjusted.

Lemma 3.6.3 *Any feasible matrix $\boldsymbol{H}$ such that $\mathrm{Tr}(\boldsymbol{H}) \le 0$ must obey*

$$\|\boldsymbol{H}_T\|_2 \le \sqrt{\frac{5}{4}}\,\|\boldsymbol{H}_T\|.$$

Finally, with all of this in place, Lemma 3.3.1 becomes this:

Lemma 3.6.4 *Suppose that the mapping $\mathcal{A}$ obeys the following two properties: for some $\delta \leq 3/13$: 1) for all positive semidefinite matrices $\boldsymbol{X}$,*

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \leq (1+\delta)\|\boldsymbol{X}\|_1; \tag{3.6.3}$$

2) for all matrices $\boldsymbol{X} \in T$

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \geq 2(\sqrt{2}-1)(1-\delta)\|\boldsymbol{X}\| \geq 0.828(1-\delta)\|\boldsymbol{X}\|. \tag{3.6.4}$$

Suppose further that there exists Y in the range of $\mathcal{A}^*$ obeying

$$\|\boldsymbol{Y}_T - \boldsymbol{e}_1\boldsymbol{e}_1^*\|_2 \leq 1/5 \quad \textit{and} \quad \|\boldsymbol{Y}_T^{\perp}\| \leq 1/2. \tag{3.6.5}$$

Then $\boldsymbol{e}_1\boldsymbol{e}_1^*$ is the unique minimizer to (3.2.5).

We now turn our attention to the properties of the dual certificate we studied in Section 3.5. The first difference is that the expectation of $\mathcal{A}^*\mathcal{A}$ in (3.5.1) is different in the complex case. A simple calculation yields

$$\mathbb{E}\,\frac{1}{m}\mathcal{A}^*\mathcal{A} = \mathcal{I} + I_n \otimes I_n := \mathcal{S}.$$

This means that for all $\boldsymbol{X}$,

$$\mathcal{S}(\boldsymbol{X}) = \boldsymbol{X} + \mathrm{Tr}(\boldsymbol{X})\boldsymbol{I}. \tag{3.6.6}$$

We note that in this case

$$\mathcal{S}^{-1} = \mathcal{I} - \frac{1}{n+1}\boldsymbol{I}_n \otimes \boldsymbol{I}_n \quad \Leftrightarrow \quad \mathcal{S}^{-1}(\boldsymbol{X}) = \boldsymbol{X} - \frac{1}{n+1}\,\mathrm{Tr}(\boldsymbol{X})\boldsymbol{I}_n. \tag{3.6.7}$$

We of course use this new $\mathcal{S}^{-1}$ in the complex analog of the candidate certificate (3.5.3). A consequence is that in the proof of Lemma 3.5.1, for instance, (3.5.7) now takes the form

$$\boldsymbol{X} = \Big[|z_1|^2 - \frac{1}{n+1}\|\boldsymbol{z}\|_2^2\Big]\bar{z}_1\,\boldsymbol{z} - \boldsymbol{e}_1 := (\xi\bar{z}_1)\,\boldsymbol{z} - \boldsymbol{e}_1. \tag{3.6.8}$$

To bound the 2-norm of a sum of i.i.d. such random variables (as in Lemma 3.5.1), we employ the same Bernstein inequality for real vectors, using the fact that $\|\boldsymbol{z}\|_2 = \|(\Re(\boldsymbol{z}), \Im(\boldsymbol{z}))\|_2$ for any complex vector $\boldsymbol{z}$. Similarly (3.5.10) becomes

$$\boldsymbol{X} = \Big[|z_1|^2 - \frac{1}{n+1}\|\boldsymbol{z}\|_2^2\Big]\,\mathcal{P}_{T^\perp}(\boldsymbol{z}\boldsymbol{z}^*). \tag{3.6.9}$$

To bound the operator norm of a sum of i.i.d. such random matrices (as in Lemma 3.5.2), we again use a covering argument, this time working with chi-square variables with two degrees of freedom, since $|\langle \boldsymbol{z}, \boldsymbol{u}\rangle|^2$ is distributed as $\frac{1}{2}\chi^2(2)$. Since $|\langle \boldsymbol{z}, \boldsymbol{u}\rangle|^2$ are real random variables, we use the same version of the Bernstein inequality as in the real-valued case. The only difference is that the moments are now

$$\mathbb{E}\,|\langle \boldsymbol{z}, \boldsymbol{u}\rangle|^{2k} = 2^{-k} \times (2+0) \times (2+2) \times (2+4) \times \ldots \times (2+2k-2) = k!$$

3.7 Stability

This section proves the stability of our approach, namely, Theorem 3.2.2. Our proof parallels the argument of Candès and Plan for showing the stability of matrix completion [18] as well as that of Gross et al. in [40].

Just as before, we prove the theorem in the real case since the complex case is essentially the same. Further, we may still take $\boldsymbol{x} = \boldsymbol{e}_1$ without loss of generality. We shall prove stability when the $\boldsymbol{z}_i$'s are i.i.d. $\mathcal{N}(0, \boldsymbol{I}_n)$ and later explain how one can easily transfer a result for Gaussian vectors to a result for vectors sampled on the sphere. Under the assumptions of the theorem, the RIP-1-like properties, namely, Lemmas 3.4.1 and 3.4.2 hold with a numerical constant δ_1 we shall specify later. Under the same hypotheses, the dual certificate $\boldsymbol{Y}$ (3.5.2) obeys

$$\|\mathcal{P}_T(\boldsymbol{Y} - \boldsymbol{e}_1\boldsymbol{e}_1^*)\|_2 \le \gamma, \qquad \|\boldsymbol{Y}_{T^\perp}\| \le \frac{1}{2},$$

in which γ is a numerical constant also specified later.

Set $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^* = \boldsymbol{e}_1\boldsymbol{e}_1^*$ and write $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{H}$. We begin by recording two useful properties. First, since $\boldsymbol{X}$ is feasible for our optimization problem, we have

$$\operatorname{Tr}(\boldsymbol{X} + \boldsymbol{H}) \le \operatorname{Tr}(\boldsymbol{X}) \quad \Longleftrightarrow \quad \operatorname{Tr}(\boldsymbol{H}) \le 0. \tag{3.7.1}$$

Second, the triangle inequality gives

$$\|\mathcal{A}(\boldsymbol{H})\|_2 = \|\mathcal{A}(\hat{\boldsymbol{X}} - \boldsymbol{X})\|_2 \le \|\mathcal{A}(\hat{\boldsymbol{X}}) - \boldsymbol{b}\|_2 + \|\boldsymbol{b} - \mathcal{A}(\boldsymbol{X})\|_2 \le 2\varepsilon. \tag{3.7.2}$$

In the noiseless case, $\mathcal{A}(\boldsymbol{H}) = 0 \Longrightarrow \langle \boldsymbol{H}, \boldsymbol{Y}\rangle = 0$, by construction. In the noisy case, a third property is that $|\langle \boldsymbol{H}, \boldsymbol{Y}\rangle|$ is at most on the order of ε. Indeed,

$$m|\langle \boldsymbol{H}, \boldsymbol{Y}\rangle| = |\langle \mathcal{A}(\boldsymbol{H}), \mathcal{A}\mathcal{S}^{-1}(\boldsymbol{X})\rangle| \le \|\mathcal{A}(\boldsymbol{H})\|_\infty \|\mathcal{A}\mathcal{S}^{-1}(\boldsymbol{X})\|_1.$$

Since, $\|\mathcal{A}(\boldsymbol{H})\|_\infty \le \|\mathcal{A}(\boldsymbol{H})\|_2$ and

$$\|\mathcal{A}\mathcal{S}^{-1}(\boldsymbol{X})\|_1 \le m(1+\delta_1)\|\mathcal{S}^{-1}(\boldsymbol{X})\|_1 \le m(1+\delta_1),$$

we obtain

$$|\langle \boldsymbol{H}, \boldsymbol{Y}\rangle| \le 2\varepsilon(1+\delta_1). \tag{3.7.3}$$

We now reproduce the steps of the proof of Lemma 3.3.1, and obtain

$$0 \ge \operatorname{Tr}(\boldsymbol{H}_T) + \operatorname{Tr}(\boldsymbol{H}_T^\perp) \ge \frac{1}{2}\operatorname{Tr}(\boldsymbol{H}_T^\perp) - \gamma\|\boldsymbol{H}_T\|_2 - |\langle \boldsymbol{H}, \boldsymbol{Y}\rangle|,$$

which gives

$$\operatorname{Tr}(\boldsymbol{H}_T^\perp) \le 4\varepsilon(1+\delta_1) + 2\gamma\|\boldsymbol{H}_T\|_2 \le 4\varepsilon(1+\delta_1) + 2\sqrt{2}\gamma\|\boldsymbol{H}_T\|, \tag{3.7.4}$$

where we recall that $\boldsymbol{H}_T$ has rank at most 2. We also have

$$\begin{aligned} 0.94(1-\delta_1)\|\boldsymbol{H}_T\| \le m^{-1}\|\mathcal{A}(\boldsymbol{H}_T)\|_1 &\le m^{-1}\|\mathcal{A}(\boldsymbol{H})\|_1 + m^{-1}\|\mathcal{A}(\boldsymbol{H}_T^{\perp})\|_1 \\ &\le m^{-1/2}\|\mathcal{A}(\boldsymbol{H})\|_2 + (1+\delta_1)\operatorname{Tr}(\boldsymbol{H}_T^{\perp}) \qquad (3.7.5) \\ &\le 2m^{-1/2}\varepsilon + (1+\delta_1)\operatorname{Tr}(\boldsymbol{H}_{T^{\perp}}), \qquad (3.7.6) \end{aligned}$$

where the second inequality follows from the RIP-1 property together with the Cauchy-Schwarz inequality. Plugging this last bound into (3.7.4) gives

$$\operatorname{Tr}(\boldsymbol{H}_T^{\perp}) \le 4\varepsilon(1+\delta_1+\gamma\alpha m^{-1/2}) + \beta\gamma\operatorname{Tr}(\boldsymbol{H}_T^{\perp}),$$

where

$$\alpha = \frac{\sqrt{2}}{0.94(1-\delta_1)}, \quad \beta = 2\alpha(1+\delta_1).$$

Hence, when $\beta\gamma < 1$, we have

$$\operatorname{Tr}(\boldsymbol{H}_T^{\perp}) = \|\boldsymbol{H}_T^{\perp}\|_1 \le \frac{4(1+\delta_1+\gamma\alpha m^{-1/2})}{1-\beta\gamma}\varepsilon = c_1\,\varepsilon.$$

In addition, (3.7.6) then gives

$$\|\boldsymbol{H}_T\| \le \frac{2m^{-1/2} + (1+\delta_1)c_1}{0.94(1-\delta_1)}\varepsilon = c_2\,\varepsilon.$$

In conclusion,

$$\|\boldsymbol{H}\|_2 \le \|\boldsymbol{H}_T\|_2 + \|\boldsymbol{H}_T^{\perp}\|_2 \le \sqrt{2}\|\boldsymbol{H}_T\| + \|\boldsymbol{H}_T^{\perp}\|_1 \le (\sqrt{2}c_2 + c_1)\varepsilon = c_0\,\varepsilon,$$

and we also have $\|\boldsymbol{H}\| \le (c_2 + c_1)\varepsilon$.

It remains to show why the fact that $\hat{\boldsymbol{X}}$ is close to $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ in the Frobenius or operator norm produces a good estimate of $\boldsymbol{x}$ (recall that $\boldsymbol{x} = \boldsymbol{e}_1$). Set $\varepsilon_0 := \|\hat{\boldsymbol{X}} - \boldsymbol{X}\| \le c_0\,\varepsilon$. Below, $\hat{\lambda}_1 \ge 0$ is the largest eigenvalue of $\hat{\boldsymbol{X}} \succeq 0$, and $\hat{\boldsymbol{u}}_1$ the first eigenvector. Likewise, $\lambda_1 = 1$ is the top eigenvalue of $\boldsymbol{X} = \boldsymbol{e}_1\boldsymbol{e}_1^*$. Since $\operatorname{Tr}(\hat{\boldsymbol{X}}) \le \operatorname{Tr}(\boldsymbol{X})$,

$$\hat{\lambda}_1 \le \lambda_1.$$

In the other direction, we know from perturbation theory that

$$|\lambda_1 - \hat{\lambda}_1| \le \|\hat{\boldsymbol{X}} - \boldsymbol{X}\| = \varepsilon_0.$$

Assuming that $\varepsilon_0 < 1$, this gives $\hat{\lambda}_1 \in [1-\varepsilon_0, 1]$. The sin-$\theta$-Theorem [28] implies that

$$|\sin\theta| \le \frac{\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|}{|\hat{\lambda}_1|} \le \frac{\varepsilon_0}{1-\varepsilon_0},$$

where $0 \leq \theta \leq \pi/2$ is the angle between the spaces spanned by $\hat{\boldsymbol{u}}_1$ and $\boldsymbol{e}_1$. Writing

$$\hat{\boldsymbol{u}}_1 = \cos\theta \boldsymbol{e}_1 + \sin\theta \boldsymbol{e}_1^{\perp}$$

in which $\boldsymbol{e}_1^{\perp}$ is a unit vector orthogonal to $\boldsymbol{e}_1$, Pythagoras' relationship gives

$$\|\boldsymbol{e}_1 - \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{u}}_1\|_2^2 = (1 - \sqrt{\hat{\lambda}_1}\cos\theta)^2 + \hat{\lambda}_1 \sin^2\theta.$$

Since $\cos\theta = \sqrt{1-\sin^2\theta}$, we have

$$1 \geq \sqrt{\hat{\lambda}_1}\cos\theta \geq \sqrt{1 - \varepsilon_0 - \frac{\varepsilon_0^2}{1-\varepsilon_0}} \geq 1 - \varepsilon_0$$

for $\varepsilon_0 < 1/3$. Hence,

$$\|\boldsymbol{e}_1 - \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{u}}_1\|_2^2 \leq \varepsilon_0^2 + \frac{\varepsilon_0^2}{(1-\varepsilon_0)^2} \leq \frac{13}{4}\varepsilon_0^2$$

provided $\varepsilon_0 < 1/3$. Since we always have

$$\|\boldsymbol{e}_1 - \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{u}}_1\|_2 \leq \|\boldsymbol{e}_1\|_2 + \sqrt{\hat{\lambda}_1}\|\hat{\boldsymbol{u}}_1\|_2 \leq 2,$$

we have established

$$\|\boldsymbol{e}_1 - \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{u}}_1\|_2 \leq C_0 \min(\varepsilon, 1).$$

This holds for all values of ε_0 and proves the claim in the case where $\|\boldsymbol{x}\|_2 = 1$. The general case is obtained via a simple rescaling.

As mentioned above, we proved the theorem for Gaussian $\boldsymbol{z}_i$'s but it is clear that our results hold true for vectors sampled uniformly at random on the sphere of radius $\sqrt{n}$. The reason is that of course, $\|\boldsymbol{z}_i\|_2$ deviates very little from $\sqrt{n}$. Formally, set $\tilde{\boldsymbol{z}}_i = [\sqrt{n}/\|\boldsymbol{z}_i\|_2]\boldsymbol{z}_i$ so that these new vectors are independently and uniformly distributed on the sphere of radius $\sqrt{n}$. Then

$$\langle \boldsymbol{X}, \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}_i^*\rangle = \frac{n}{\|\boldsymbol{z}_i\|_2^2}\langle \boldsymbol{X}, \boldsymbol{z}_i\boldsymbol{z}_i^*\rangle,$$

and thus $\langle \boldsymbol{X}, \boldsymbol{z}_i\boldsymbol{z}_i^*\rangle$ is between $(1-\delta_2)\langle \boldsymbol{X}, \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}_i^*\rangle$ and $(1+\delta_2)\langle \boldsymbol{X}, \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}_i^*\rangle$ with very high probability. This holds uniformly over all Hermitian matrices. Thus if $\tilde{\mathcal{A}}(\boldsymbol{X}) = \{\tilde{\boldsymbol{z}}_i^* X \tilde{\boldsymbol{z}}_i\}_{1\leq i\leq m}$,

$$(1-\delta_2)\|\tilde{\mathcal{A}}(\boldsymbol{X})\|_q \leq \|\mathcal{A}(\boldsymbol{X})\|_q \leq (1+\delta_2)\|\tilde{\mathcal{A}}(\boldsymbol{X})\|_q$$

for any $1 \leq q \leq \infty$.

Now take $b_i = |\langle \boldsymbol{x}, \tilde{\boldsymbol{z}}_i\rangle|^2 + \nu_i$ and solve (3.2.9) to get $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \tilde{\boldsymbol{H}}$. Going through the same steps as above by using the relationships between $\mathcal{A}$ and $\tilde{\mathcal{A}}$ throughout, and by using the dual certificate $\boldsymbol{Y}$ associated with $\mathcal{A}$, we obtain

$$\|\tilde{\mathcal{A}}(\tilde{\boldsymbol{H}})\|_2 \leq 2\varepsilon, \qquad |\langle \tilde{\boldsymbol{H}}, \boldsymbol{Y}\rangle| \leq 2\varepsilon(1+\delta_1)(1+\delta_2),$$

and

$$\mathrm{Tr}(\tilde{\boldsymbol{H}}_{T^\perp}) \leq (1+\delta_2)c_1\varepsilon, \qquad \|\tilde{\boldsymbol{H}}_T\| \leq (1+\delta_2)c_2\varepsilon.$$

Therefore,

$$\|\tilde{\boldsymbol{H}}\|_2 \leq (1+\delta_2)(\sqrt{2}c_2 + c_1)\varepsilon.$$

The rest of the proof goes through just the same.

3.8 Numerical Simulations

In this section we illustrate our theoretical results with numerical simulations. In particular, we will demonstrate PhaseLift's robustness vis a vis additive noise.

We consider the setup in Section 3.2, where the measurements are contaminated with additive noise. The solution to (3.2.9) is computed using the following regularized nuclear-norm minimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{b}\|_2^2 + \lambda \operatorname{Tr}(\boldsymbol{X}) \\ \text{subject to} & \boldsymbol{X} \succeq 0. \end{array} \tag{3.8.1}$$

It follows from standard optimization theory [76] that (3.8.1) is equivalent to (3.2.9) for some value of λ. Hence, we use (3.8.1) to compute the solution of (3.2.9) by determining via a simple and efficient bisection search the largest value $\lambda(\varepsilon)$ such that $\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{b}\|_2 \leq \varepsilon$. The numerical algorithm to solve (3.8.1) was implemented in Matlab using TFOCS [9]. We then extract the largest rank-1 component as described in Section 3.2 to obtain an approximation $\hat{\boldsymbol{x}}$.

We will use the relative mean squared error (MSE) and the relative root mean squared error (RMS) to measure performance. However, since a solution is only unique up to global phase, it does not make sense to compute the distance between $\boldsymbol{x}$ and its approximation $\hat{\boldsymbol{x}}$. Instead we compute the distance modulo a global phase term and define the relative MSE between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ as

$$\min_{c:|c|=1} \frac{\|c\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2}{\|\boldsymbol{x}\|_2^2}.$$

The (relative) RMS is just the square root of the (relative) MSE.

In the first set of experiments, we investigate how the reconstruction algorithm performs as the noise level increases. The test signal is a complex-valued signal of length $n = 128$ with independent Gaussian complex entries (each entry is of the form $a + ib$ where a and b are independent $\mathcal{N}(0, 1)$ variables) so that the real and imaginary parts are independent white noise sequences. Obviously, the signal is arbitrary. We use $m = 6n$ measurement vectors sampled independently on the unit sphere $\mathbb{C}^n$.

We generate noisy data from both a Gaussian model and a Poisson model. In the Gaussian model, $b_i \sim \mathcal{N}(\mu_i, \sigma^2)$ where $\mu_i = |\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle|^2$ and σ is adjusted so that the total noise power is bounded by ε^2. In the Poisson model, $b_i \sim \text{Poi}(\mu_i)$ and the noise $b_i - \mu_i$ is rescaled to achieve a desired total power as above (we might do without this rescaling as well but have decided to work with a prescribed signal-to-noise ratio SNR for simplicity of exposition). We do this for five different SNR levels,[7] ranging from 5dB to 100dB. However, we point out that we do not make use of the noise statistics in our reconstruction algorithm[8], since our purpose is only to assume an upper bound on the total noise power, as in Theorem 3.2.2.

For each SNR level, we repeat the experiment ten times with different noise terms, different signals, and different random measurement vectors; we then record the average relative RMS over these ten experiments. Figure 3.4(a) shows the average relative MSE in dB (the values of

[7] The SNR of two signals $\boldsymbol{x}, \hat{\boldsymbol{x}}$ with respect to $\boldsymbol{x}$ is defined as $10 \log_{10} \|\boldsymbol{x}\|_2^2 / \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$. So we say that the SNR is 10dB if $10 \log_{10} \|\boldsymbol{x}\|_2^2 / \|\boldsymbol{\nu}\|_2^2 = 10$.

[8] We refer to [22] for efficient ways to incorporate statistical noise models into the reconstruction algorithm.

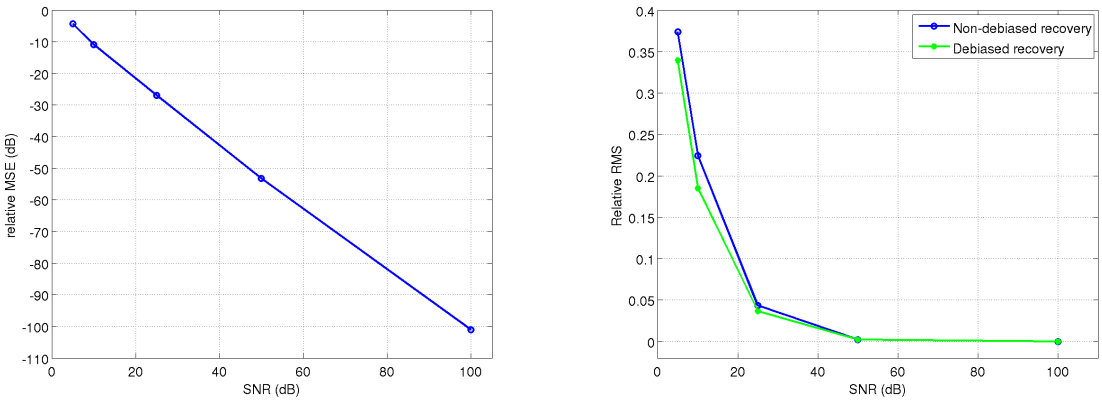

Figure 3.4: Performance of PhaseLift for Poisson noise. The stability of the algorithm is apparent as its performance degrades gracefully with decreasing SNR. (a) Relative MSE on a log-scale for the non-debiased recovery. (b) Relative RMS for the original and debiased recovery.

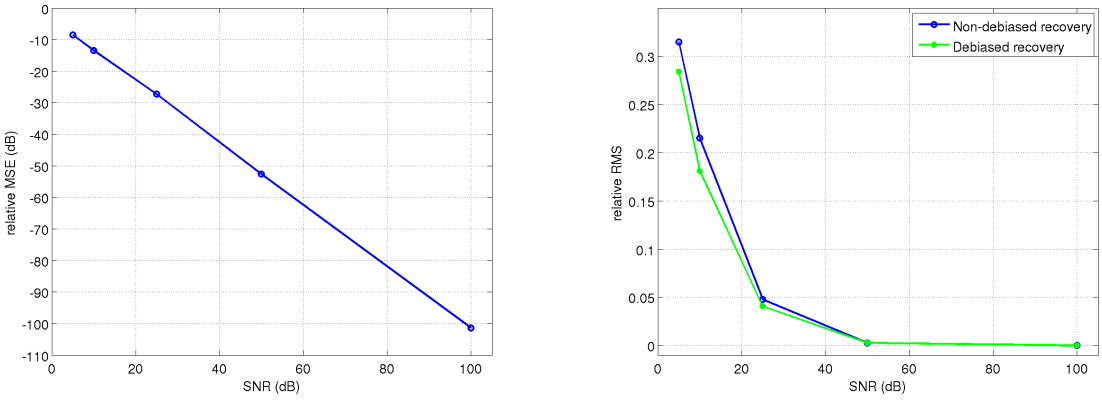

Figure 3.5: Performance of PhaseLift for Gaussian noise. (a) Relative MSE on a log-scale for the non-debiased recovery. (b) Relative RMS for the original and the debiased recovery.

$10 \log_{10}(\text{rel. MSE})$ are plotted) versus the SNR for Poisson noise. In each case, the performance degrades very gracefully with decreasing SNR, as predicted by Theorem 3.2.2. Debiasing as described at the end of Section 3.2 leads to a further improvement in the reconstruction for low SNR, as illustrated in Figure 3.4(b). The results for Gaussian noise are comparable, see Figure 3.5.

In the next experiment, we collect Poisson data about a complex-valued random signal just as above, and work with a fixed SNR set to 15dB. The number of measurements varies so that the oversampling rate m/n is between 5 and 22 (m is thus between $n \log n$ and $4.5n \log n$). We repeat the experiment ten times with different noise terms and different random measurement vectors for each oversampling rate; we then record the average relative RMS. Figure 3.6 shows the average

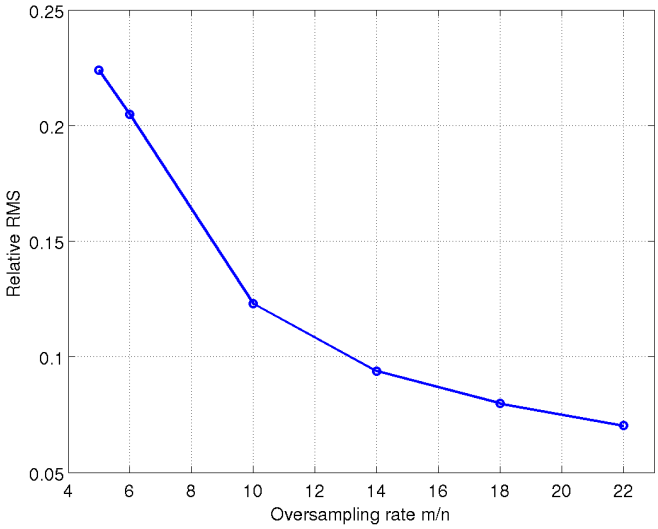

Figure 3.6: Oversampling rate versus relative RMS.

relative RMS of the solution to (3.2.5) versus the oversampling rate. We observe that the decrease in the RMS is inversely proportional to the number of measurements. For instance, the error reduces by a factor of two when we double the number of measurements. If instead we hold the standard deviation of the errors at a constant level, the mean squared error (MSE) reduces by a factor of about two when we double the number of measurements.

3.9 Discussion

In this chapter, we have shown that it is possible to recover a signal exactly (up to a global phase factor) from the knowledge of the magnitude of its inner products with a family of sensing vectors $\{\boldsymbol{z}_i\}$. The fact that on the order of $n \log n$ magnitude measurements $|\langle \boldsymbol{x}, \boldsymbol{z}_i \rangle|^2$ uniquely determine $\boldsymbol{x}$ is not surprising. The part we find unexpected, however, is that what appears to be a combinatorial problem is solved exactly by a convex program. Further, we have established the existence of a noise-aware recovery procedure—also based on a tractable convex program—which is robust vis a vis additive noise. To the best of our knowledge at the time of completion of this work, there are no other results about the recovery of an arbitrary signal from noisy quadratic data of this kind.

An appealing research direction is to study the recovery of a signal from other types of intensity measurements, and consider other families of sensing vectors. In particular, *structured* random families would be of great interest. The next and final chapter is dedicated to a step in this direction.

3.10 Appendix

We prove that the RIP in the 2-norm (and in any p-norm with $p > 1$) cannot hold for $\mathcal{A}$. We derive the claim for the real-valued setting, but the arguments can be easily extended to the complex-valued setting. Here and below, $|\boldsymbol{y}| = (|y_1|, \dots, |y_m|)$.

Consider an $m \times n$ matrix $\boldsymbol{A}$ with i.i.d. rows $\boldsymbol{z}_i =^d \mathcal{N}(0, I)$ and set $\mathcal{A}(\boldsymbol{X}) = \{\boldsymbol{z}_i^* \boldsymbol{X} \boldsymbol{z}_i\}_{i=1}^m$. Then for $\boldsymbol{x} \in \mathbb{R}^n$, $\mathcal{A}(\boldsymbol{x}\boldsymbol{x}^*) = |\boldsymbol{A}\boldsymbol{x}|^2$ and

$$\|\mathcal{A}(\boldsymbol{x}\boldsymbol{x}^*)\|_2 = \Big(\sum_{i=1}^m |\langle \boldsymbol{z}_i, \boldsymbol{x} \rangle|^4 \Big)^{1/2}.$$

Taking $\boldsymbol{x} = \boldsymbol{z}_1 / \|\boldsymbol{z}_1\|_2$, we get

$$\begin{aligned} \sup_{\boldsymbol{u} \in \mathcal{S}^{n-1}} \|\mathcal{A}(uu^*)\|_2 \geq \|\mathcal{A}(\boldsymbol{x}\boldsymbol{x}^*)\|_2 &= \left(\sum_{i=1}^m \left| \left\langle \boldsymbol{z}_i, \frac{\boldsymbol{z}_1}{\|\boldsymbol{z}_1\|_2} \right\rangle \right|^4 \right)^{1/2} \\ &\geq \left| \left\langle \boldsymbol{z}_1, \frac{\boldsymbol{z}_1}{\|\boldsymbol{z}_1\|_2} \right\rangle \right|^2 = \|\boldsymbol{z}_1\|_2^2 = \Omega(n), \end{aligned}$$

where the last equality holds with high probability.

Now, expand $\boldsymbol{A}$ into its singular value decomposition $\boldsymbol{A} = \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$ with $\sigma_1 \geq \sigma_2 \dots \geq \sigma_n$. As a consequence of well-known deviations bounds concerning the singular values of Gaussian random matrices [85], the inequalities

$$m(1-\delta) \leq \sigma_n^2 \leq \sigma_1^2 \leq m(1+\delta)$$

for some δ with $0 < \delta < 1$ hold with high probability provided that $m \geq Cn \log n$, where $C > 0$ is a suitable constant. All singular values of $\boldsymbol{A}$ are simple with probability 1 and thus $\boldsymbol{u}_n$, the singular vector corresponding to the smallest singular value, is well-defined and we can think of it as being distributed uniformly at random on the unit sphere. Therefore, with high probability

$$\|\boldsymbol{u}_n\|_\infty = \mathcal{O}(\frac{\sqrt{\log n}}{\sqrt{m}}).$$

This gives

$$\begin{aligned} \inf_{\boldsymbol{u} \in \mathcal{S}^{n-1}} \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_2 \leq \|\mathcal{A}(\boldsymbol{v}_n \boldsymbol{v}_n^*)\|_2 &= \||\boldsymbol{A}\boldsymbol{v}_n|^2\|_2 = \||\sigma_n \boldsymbol{u}_n|^2\|_2 \\ &= \sigma_n^2 \Big(\sum_{i=1}^m |\boldsymbol{u}_{ni}|^4 \Big)^{1/2} = \sigma_n^2 \mathcal{O}\left(\frac{\log n}{\sqrt{m}} \right) = \mathcal{O}(\sqrt{m} \log n) \end{aligned}$$

(also with high probability). This implies that

$$\frac{\sup_{\boldsymbol{u} \in \mathcal{S}^{n-1}} \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_2}{\inf_{\boldsymbol{u} \in \mathcal{S}^{n-1}} \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_2} = \Omega\left(\frac{n}{\sqrt{m} \log n} \right) \qquad \text{w.h.p.}$$

Therefore, unless we take m to be at least on the order of $n^2/\log^2 n$ (which is much too large to be of interest), the RIP-2 cannot hold. Similar arguments show that

$$\frac{\sup_{\boldsymbol{u}\in\mathcal{S}^{n-1}} \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_p}{\inf_{\boldsymbol{u}\in\mathcal{S}^{n-1}} \|\mathcal{A}(\boldsymbol{u}\boldsymbol{u}^*)\|_p} = \Omega\left(\frac{n}{m^{\frac{1}{p}}\log n}\right) \qquad \text{w.h.p.,}$$

and thus the RIP-p cannot hold for $p > 1$, unless m is at least on the order of $n^p/(\log n)^p$. Obviously, since the RIP does not hold for rank-1 matrices, it cannot hold for higher ranks.

Chapter 4

PhaseLift for unitary measurements

4.1 Overview

As proven above, PhaseLift recovers signals exactly with high probability when the measurement vectors z_i are iid gaussian and is furthermore provably stable with respect to measurement noise under the same assumptions. However, the gaussian measurement model is not know to be physically realizable. Therefore it is of great interest to prove exactness results for PhaseLift under more structured measurement assumptions, which will require far more technical proofs due to the lack of probabilistic independence between sensing vectors from structured random measurement ensembles. A step in this direction is to consider the z_i as rows of iid Haar distributed unitary matrices, which models a situation that occurs in quantum mechanics. In this chapter, we prove that PhaseLift succeeds with high probability under this measurement model as long as the number of measurements $m = O(n)$ and point out a corollary of the result which relates to Wright's conjecture from quantum mechanics.

Here we assume that measurements of the form $\{|U_k x|^2\}_{k=1}^r$ are available, where the U_i are sampled independently according to the Haar measure on $\mathbb{U}(n)$, the unitary group or $\mathbb{O}(n)$, the orthogonal group, and the total number of measurements is $m = rn$. Below, we will label the transpose of the row vectors of U_k as $u_i^{(k)}$ or enumerate them as $\{u_i\}_{i=1}^m$. As in the gaussian case, we may assume wlog that $x = e_1$, in this case by the unitary/orthogonal invariance of the Haar measure.

We proceed as in the previous chapter by showing that the measurement operator $\mathcal{A}$ in this setting obeys some nice properties with high probability. Namely, we need to verify that $\mathcal{A}$ satisfies the condition of the following lemma, which is a very slight modification of Lemma 3.6.4 achieved by noting that if $Y_T^\perp \prec 0$, then $\left\langle H_T^\perp, Y_T^\perp \right\rangle \leq 0$.

Lemma 4.1.1 *Suppose that the mapping $\mathcal{A}$ obeys the following two properties: for some $\delta \leq 3/13$:*

1) for all positive semidefinite matrices $\boldsymbol{X}$,

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \leq (1+\delta)\|\boldsymbol{X}\|_1; \tag{4.1.1}$$

2) for all matrices $\boldsymbol{X} \in T$

$$m^{-1}\|\mathcal{A}(\boldsymbol{X})\|_1 \geq 2(\sqrt{2}-1)(1-\delta)\|\boldsymbol{X}\| \geq 0.828(1-\delta)\|\boldsymbol{X}\|. \tag{4.1.2}$$

Suppose further that there exists Y in the range of $\mathcal{A}^*$ obeying

$$\|\boldsymbol{Y}_T - \boldsymbol{e}_1\boldsymbol{e}_1^*\|_2 \leq 1/5 \quad \textit{and} \quad Y_T^{\perp} \prec 0. \tag{4.1.3}$$

Then $\boldsymbol{e}_1\boldsymbol{e}_1^*$ is the unique minimizer to (3.2.5).

In particular, the RIP-1 property in this unitary case has implications related to Wright's conjecture. Furthermore, we adapt a trick in the construction of the dual certificate, used by [30] in the gaussian case, to reduce the number of necessary measurements from $O(n \log n)$ to $O(n)$. Establishing the above yields

Theorem 4.1.2 *Take $x \in \mathbb{C}^n$ and assume that measurements of the form $\{|U_k x|^2\}_{k=1}^r$ are available, where the U_i are sampled independently according to the Haar measure on $\mathbb{U}(n)$, the unitary group or $\mathbb{O}(n)$, the orthogonal group, so that the total number of measurements is $m = rn$. Then the PhaseLift algorithm succeeds in recovering x up to global phase with very high probability when $m = O(n)$.*

4.2 Restricted Isometry Property of type 1

In the sequel we will label the transpose of the row vectors of U_k as $u_i^{(k)}$ or enumerate them as $\{u_i\}_{i=1}^m$. As in the gaussian case, we may assume wlog that $x = e_1$, in this case by the unitary/orthogonal invariance of the Haar measure.

First, we aim to establish a RIP-1 property on rank-2 matrices for this class of measurements. Let $\mathcal{A}(X) = \{\sqrt{n(n+1)}\,\mathrm{Tr}(u_i u_i^* X)\}_{i=1}^m$, where $\mathcal{A}$ is a linear map from the Hermitian matrices. Let $X = x_1x_1^* - \lambda x_2 x_2^*$ be a rank-2 hermitian matrix in SVD form with $0 \leq \lambda \leq 1$. Then

$$\begin{aligned}\frac{1}{\sqrt{n(n+1)}}\mathcal{A}(x_1x_1^* - \lambda x_2x_2^*) &= \{|\langle u_i, x_1\rangle|^2 - \lambda|\langle u_i, x_2\rangle|^2\}_{i=1}^m \\ &=^d \{|\langle u_i, e_1\rangle|^2 - \lambda|\langle u_i, e_2\rangle|^2\}_{i=1}^m \\ &= \{|u_{i1}|^2 - \lambda|u_{i2}|^2\}_{i=1}^m\end{aligned}$$

where we used rotational invariance of Haar measure and the fact that there exist orthogonal or unitary transformations taking any real/complex orthobasis to another orthobasis and u_{ij} denotes the jth entry of the vector u_i.

As in Lemma 3.4.1, to establish $\frac{1}{m}\|\mathcal{A}(X)\|_1 \leq (1+\delta)\|X\|_1$ for all psd matrices, it is enough to consider X to be rank 1 psd. Taking any unit vector $x \in \mathbb{C}^n$, we have

$$\frac{1}{r}\frac{1}{\sqrt{n(n+1)}}\|\mathcal{A}(xx^*)\|_{l1} = \frac{1}{r}\sum_{i=1}^m |\langle u_m, x\rangle|^2 = 1$$

This implies that

$$\frac{1}{r}\frac{1}{\sqrt{n(n+1)}}\|\mathcal{A}(X)\|_1 \leq (1+\delta)\|X\|_1$$

for any $\delta > 0$ and for any psd X. Now since $\frac{m}{r\sqrt{n(n+1)}} = \sqrt{\frac{n}{n+1}}$, we can get the desired property with $\delta = \frac{3}{13}$.

To get the other part of RIP-1, we need to examine the quantity

$$\frac{1}{r}\frac{1}{\sqrt{n(n+1)}}\|\mathcal{A}(x_1x_1^* - \lambda x_2x_2^*)\|_{l1} =^d \frac{1}{r}\sum_{i=1}^{m}||u_{i1}|^2 - \lambda|u_{i2}|^2| = \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}||u_{i1}^{(k)}|^2 - \lambda|u_{i2}^{(k)}|^2|$$

and show that it is lower bounded by a multiple of the operator norm of $X = x_1x_1^* - \lambda x_2x_2^*$ whp. This sum may be expressed as a function of $2rn$ iid gaussian rvs. This function is not Lipschitz, so we will use a surrogate function that is Lipschitz in order to apply Talagrand's inequality [82] and then show that this introduces only a very small error.

We will treat the real and complex cases simultaneously. To be specific, one way to obtain the Haar measure on $\mathbb{O}(n)$ or $\mathbb{U}(n)$ is to perform Gram-Schmidt on the columns of a gaussian or complex gaussian matrix. Thus, we will consider the columns u_1 and u_2 of a Haar-distributed orthogonal matrix as the result of the Gram-Schmidt procedure on a pair of iid gaussian vectors ζ and z. Introduce the functions $v(x) = \frac{x}{\|x\|_2}$ and $t(x,y) = x - y\langle y, x\rangle$. Then if ζ, z are iid $\mathcal{N}(0,I)$ or $\mathcal{CN}(0,I,0)$, it can be verified that

$$(u_1, u_2) =^d (v(z), v(t(v(\zeta), v(z))))$$

We can now express the distribution of the quantity above as

$$\begin{aligned}\frac{1}{r}\sum_{i=1}^{m}||u_{i1}|^2 - \lambda|u_{i2}|^2| &=^d \frac{1}{r}\sum_{k=1}^{r}F(\zeta^{(k)}, z^{(k)}) \\ &= \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}||v(z^{(k)})_i|^2 - \lambda|v(t(v(\zeta^{(k)}), v(z^{(k)})))_i|^2|\end{aligned}$$

as a function of a $2rn$ component gaussian vector. The above function is not lipschitz and the issue occurs in two places: first, when we normalize the vectors ζ and z and then when we normalize the expression $t(v(\zeta), v(z))$. Let us introduce the surrogate functions $\tilde{v}_1(x) = \frac{x}{(\|x\|_2 \vee \sqrt{\frac{n}{c_1}})}$ where $c_1 >> 1$ and $\tilde{v}_2 = \frac{x}{(\|x\|_2 \vee \sqrt{c_2})}$ where $0 < c_2 < 1$. Now, the surrogate function

$$\frac{1}{r}\sum_{k=1}^{r}\tilde{F}(\zeta^{(k)}, z^{(k)}) = \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}||\tilde{v}_1(z^{(k)})_i|^2 - \lambda|\tilde{v}_2(t(\tilde{v}_1(\zeta^{(k)}), \tilde{v}_1(z^{(k)})))_i|^2|$$

is equal to the original with probability at least

$$1 - 2r\mathbb{P}\{\|\zeta\|_2^2 < \frac{n}{c_1}\} - r\mathbb{P}\{|\left\langle \frac{\zeta}{\|\zeta\|_2}, \frac{z}{\|z\|_2}\right\rangle|^2 > 1 - c_2\} \geq 1 - \mathrm{O}\left(re^{-\gamma n}\right)$$

where γ can be made arbitrarily large by taking c_1 large enough and c_2 small enough. It now remains to verify that the surrogate function is Lipschitz with a good enough constant and that it introduces only a small error.

Consider the function $g(x,y) = \sum_{i=1}^n ||x_i|^2 - \lambda|y_i|^2|$, with $x, y \in \mathbb{R}^n$ or $\mathbb{C}^n$ and assume $\|x_i\|_2 \leq 1, \|y_i\|_2 \leq 1$. We have

$$\begin{aligned}
|g(x_1,y_1) - g(x_2,y_2)| &\leq \sum_{i=1}^n \big| ||x_{1i}|^2 - \lambda|y_{1i}|^2| - ||x_{2i}|^2 - \lambda|y_{2i}|^2| \big| \\
&\leq \sum_{i=1}^n |(|x_{1i}|^2 - |x_{2i}|^2) - \lambda(|y_{1i}|^2 - |y_{2i}|^2)| \\
&\leq \sum_{i=1}^n |(|x_{1i}| + |x_{2i}|)(|x_{1i}| - |x_{2i}|)| + \lambda|(|y_{1i}| + |y_{2i}|)(|y_{1i}| - |y_{2i}|)| \\
&\leq \||x_1| + |x_2|\|_2 \|x_1 - x_2\|_2 + \lambda \||y_1| + |y_2|\|_2 \|y_1 - y_2\|_2 \\
&\leq 2\|x_1 - x_2\|_2 + 2\lambda \|y_1 - y_2\|_2
\end{aligned}$$

Now take

$$\begin{aligned}
&|\tilde{F}(\zeta_1,z_1) - \tilde{F}(\zeta_2,z_2)| \\
&= |g(\tilde{v}_1(z_1), \tilde{v}_2(t(\tilde{v}_1(\zeta_1), \tilde{v}_1(z_1)))) - g(\tilde{v}_1(z_2), \tilde{v}_2(t(\tilde{v}_1(\zeta_2), \tilde{v}_1(z_2))))| \\
&\leq 2\|\tilde{v}_1(z_1) - \tilde{v}_1(z_2)\|_2 + 2\lambda \|\tilde{v}_2(t(\tilde{v}_1(\zeta_1), \tilde{v}_1(z_1))) - \tilde{v}_2(t(\tilde{v}_1(\zeta_2), \tilde{v}_1(z_2)))\|_2 \\
&\leq 2\text{Lip}(\tilde{v}_1)\|z_1 - z_2\|_2 + 2\lambda \text{Lip}(\tilde{v}_2)\|t(\tilde{v}_1(\zeta_1), \tilde{v}_1(z_1)) - t(\tilde{v}_1(\zeta_2), \tilde{v}_1(z_2))\|_2 \\
&\leq 2\text{Lip}(\tilde{v}_1)\|z_1 - z_2\|_2 + 2\lambda \text{Lip}(\tilde{v}_2)\text{Lip}(t|_{\mathcal{B}(0,1)^2})\|(\tilde{v}_1(\zeta_1), \tilde{v}_1(z_1)) - (\tilde{v}_1(\zeta_2), \tilde{v}_1(z_2))\|_2 \\
&\leq 2\text{Lip}(\tilde{v}_1)\|z_1 - z_2\|_2 + 2\lambda \text{Lip}(\tilde{v}_2)\text{Lip}(t|_{\mathcal{B}(0,1)^2})\text{Lip}(\tilde{v}_1)\|(\zeta_1, z_1) - (\zeta_2, z_2)\|_2
\end{aligned}$$

One can verify that in either the real or complex case, when $\|x_i\|_2, \|y_i\|_2 \leq 1$, the function t satisfies

$$\|t(x_1,y_1) - t(x_2,y_2)\|_2 \leq 2\|(x_1,y_1) - (x_2,y_2)\|_2$$

For $x \in \mathbb{R}^n$, let $\tilde{v}(x) = \frac{x}{\|x\|_2 \vee c}$ for some positive constant c. Now, we have that $D(\frac{x}{\|x\|_2}) = \frac{1}{\|x\|_2^3}(\|x\|_2^2 I - xx^*)$ and hence $\|D(\frac{x}{\|x\|_2})\| \leq \frac{2}{\|x\|_2}$. Thus, $\|D(\tilde{v})\| \leq \frac{2}{c}$ on $\bar{\mathcal{B}}(0,c)^c$ so that $\text{Lip}(\tilde{v}|_U) \leq \frac{2}{c}$ for any open convex set $U \in \bar{\mathcal{B}}(0,c)^c$. Furthermore, we have $D(\tilde{v}) = \frac{1}{c} I$ on $\mathcal{B}(0,c)$ and thus $\|D(\tilde{v})\| \leq \frac{1}{c}$ on $\mathcal{B}(0,c)$ so that $\text{Lip}(\tilde{v}|_{\bar{\mathcal{B}}(0,c)}) \leq \frac{1}{c}$.

Take $x_i \in \bar{\mathcal{B}}(0,c)^c, i = 1,2$ such that the line connecting these two points intersects $\bar{\mathcal{B}}(0,c)$. Assume that the point(s) of intersection are z_1 and z_2 (with the line from x_1 to x_2 first hitting z_1 and then z_2). Then we have

$$\begin{aligned}\|\tilde{v}(x_1)-\tilde{v}(x_2)\|_2 &\leq \|\tilde{v}(x_1)-\tilde{v}(z_1)\|_2+\|\tilde{v}(z_1)-\tilde{v}(z_2)\|_2+\|\tilde{v}(z_2)-\tilde{v}(x_2)\|_2\\ &\leq \frac{2}{c}\|x_1-z_2\|_2+\frac{1}{c}\|z_1-z_2\|_2+\frac{2}{c}\|z_2-x_2\|_2\\ &\leq \frac{2}{c}\|x_1-x_2\|_2\end{aligned}$$

where for the sets U_i we take $\mathcal{B}(x_i, \|z_i - x_i\|_2)$. The other cases of arrangements of x_i are similarly proven. We conclude that $\text{Lip}(\tilde{v}) \leq \frac{2}{c}$. Note that this implies that in the complex case, we also have $\text{Lip}(\tilde{v}) \leq \frac{2}{c}$ with $\tilde{v}$ defined analogously. We have thus established that $\text{Lip}(\tilde{v}_1) \leq \frac{\sqrt{c_2}}{\sqrt{n}}$ and $\text{Lip}(\tilde{v}_2) \leq \frac{2}{\sqrt{c_1}}$. Using this information,

$$|\tilde{F}(\zeta_1,z_1)-\tilde{F}(\zeta_2,z_2)| \leq \frac{16\sqrt{\frac{c_1}{c_2}}}{\sqrt{n}}\|(\zeta_1,z_1)-(\zeta_2,z_2)\|_2$$

Finally, this implies

$$\text{Lip}(\frac{1}{r}\sum_{k=1}^{r}\tilde{F}(\zeta^{(k)},z^{(k)})) \leq \frac{1}{\sqrt{r}}\frac{16\sqrt{\frac{c_1}{c_2}}}{\sqrt{n}}$$

By Talagrand's inequality, we have

$$\mathbb{P}\{\left|\frac{1}{r}\sum_{k=1}^{r}\tilde{F}(\zeta^{(k)},z^{(k)})-\mathbb{E}\left[\frac{1}{r}\sum_{k=1}^{r}\tilde{F}(\zeta^{(k)},z^{(k)})\right]\right| \geq t\} \leq e^{-c(rn)t^2}$$

for a constant c which depends on c_i. Let

$$\tilde{G}=\frac{1}{r}\sum_{k=1}^{r}\tilde{F}(\zeta^{(k)},z^{(k)}), \qquad G=\frac{1}{r}\sum_{k=1}^{r}F(\zeta^{(k)},z^{(k)}).$$

G and $\tilde{G}$ are both bounded by 2 and disagree on a set of probability $\text{O}\left(re^{-\gamma n}\right)$, thus

$$\lim_{n\to\infty}\left|\mathbb{E}[\tilde{G}]-\mathbb{E}\left[G\right]\right| = 0.$$

So that if we fix t apriori, then for all n large enough

$$\begin{aligned}&\mathbb{P}\{|G-\mathbb{E}\left[G\right]| \geq t\}\\ &\leq \mathbb{P}\{\left|\tilde{G}-\mathbb{E}[\tilde{G}]\right| \geq t-\left|\tilde{G}-G\right|-\left|\mathbb{E}[\tilde{G}]-\mathbb{E}\left[G\right]\right|\}\\ &\leq \mathbb{P}\{G\neq\tilde{G}\}+\mathbb{P}\{\left|\tilde{G}-\mathbb{E}[\tilde{G}]\right| \geq t-\left|\mathbb{E}[\tilde{G}]-\mathbb{E}\left[G\right]\right|\}\\ &\leq \text{O}\left(re^{-\gamma n}\right)+e^{-c(rn)(t/2)^2}\end{aligned}$$

Therefore, we have established

$$\mathbb{P}\{\left|\frac{1}{r}\sum_{i=1}^{m}||u_{i1}|^2-\lambda|u_{i2}|^2|-\mathbb{E}\left[\frac{1}{r}\sum_{i=1}^{m}||u_{i1}|^2-\lambda|u_{i2}|^2|\right]\right|\geq t\}\leq e^{-c(rn)(t/2)^2}+\mathrm{O}\left(re^{-\gamma n}\right)$$

for constants c and γ which depend on c_i. To achieve an arbitrarily fast exponential rate, first select c_i so that γ is as large as needed, then fix r large enough.

We claim that $\mathbb{E}\left[\left||u_{i1}|^2-\lambda|u_{i2}|^2\right|\right]=\frac{1}{n}\frac{1+\lambda^2}{1+\lambda}$, which we compute below. We have from [4] that $(|u_{i1}|^2,\ldots,|u_{in-1}|^2)$ are uniformly distributed on $\{(x_1,\ldots x_{n-1});x_i\geq 0,\sum_{i=1}^{n-1}x_i\leq 1\}$. Thus

$$\begin{aligned}
&\mathbb{E}\left[\left||u_{i1}|^2-\lambda|u_{i2}|^2\right|\right]\frac{1}{(n-1)(n-2)}\\
&=(n-3)!\int_{\mathbb{R}^{n-1}}|x_1-\lambda x_2|\chi_{\{\sum_{i=1}^{n-1}x_i\leq 1,x_i\geq 0\}}dx_1\ldots dx_{n-1}\\
&=(n-3)!\int_{\mathbb{R}^2}|x_1-\lambda x_2|\chi_{\{x_1+x_2\leq 1,x_i\geq 0\}}\int_{\mathbb{R}^{n-3}}\chi_{\{x_3+\ldots x_{n-1}\leq 1-(x_1+x_2)\}}dx_3\ldots dx_{n-1}\\
&=\frac{(n-3)!}{(n-3)!}\int_{\mathbb{R}^2}|x_1-\lambda x_2|(1-(x_1+x_2))^{n-3}\chi_{\{x_1+x_2\leq 1,x_i\geq 0\}}dx_1dx_2\\
&=\int_0^1\int\left[\chi_{\{x_1\leq\lambda x_2\}}(\lambda x_2-x_1)+\chi_{\{x_1\geq\lambda x_2\}}(x_1-\lambda x_2)\right](1-(x_1+x_2))^{n-3}\chi_{\{0\leq x_2\leq 1-x_2\}}dx_1dx_2\\
&=\int_0^1\chi_{\{\lambda x_2\leq 1-x_2\}}\int_0^{\lambda x_2}(\lambda x_2-x_1)(1-(x_1+x_2))^{n-3}dx_1\\
&+\chi_{\{\lambda x_2\geq 1-x_2\}}\int_0^{1-x_2}(\lambda x_2-x_1)(1-(x_1+x_2))^{n-3}dx_1\\
&+\chi_{\{\lambda x_2\leq 1-x_2\}}\int_{\lambda x_2}^{1-x_2}(\lambda x_2-x_1)(1-(x_1+x_2))^{n-3}dx_1dx_2\\
&=\int_0^{\frac{1}{1+\lambda}}\int_0^{\lambda x_2}(\lambda x_2-x_1)(1-(x_1+x_2))^{n-3}dx_1+\int_{\lambda x_2}^{1-x_2}(x_1-\lambda x_1)(1-(x_1+x_2))^{n-3}dx_1dx_2\\
&+\int_{\frac{1}{1+\lambda}}^1\int_0^{1-x_2}(\lambda x_2-x_1)(1-(x_1+x_2))^{n-3}dx_2dx_2
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{\frac{1}{1+\lambda}} \lambda x_2 \left(\frac{-1}{n-2}(1-(x_1+x_2))^{n-2} \Big|_0^{\lambda x_2} \right) - \int_0^{\lambda x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 dx_2 \\
&+ \int_0^{\frac{1}{1+\lambda}} \int_{\lambda x_2}^{1-x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 - \lambda x_2 \left(\frac{-1}{n-2}(1-(x_1+x_2))^{n-2} \Big|_{\lambda x_2}^{1-x_2} \right) dx_2 \\
&+ \int_{\frac{1}{1+\lambda}}^1 \lambda x_2 \left(\frac{-1}{n-2}(1-(x_1+x_2))^{n-2} \Big|_0^{1-x_2} \right) - \int_0^{1-x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 dx_2 \\
&= \int_0^{\frac{1}{1+\lambda}} \lambda x_2 \left(\frac{-1}{n-2}(1-(1+\lambda)x_2))^{n-2} + \frac{1}{n-2}(1-x_2)^{n-2} \right) \\
&- \int_0^{\lambda x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 dx_2 \\
&+ \int_0^{\frac{1}{1+\lambda}} \int_{\lambda x_2}^{1-x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 - \lambda x_2 \left(\frac{1}{n-2}(1-(1+\lambda)x_2))^{n-2} \right) dx_2 \\
&+ \int_{\frac{1}{1+\lambda}}^1 \lambda x_2 \left(\frac{1}{n-2}(1-x_2)^{n-2} \right) dx_1 - \int_0^{1-x_2} x_1(1-(x_1+x_2))^{n-3} dx_1 dx_2 \\
&= \frac{\lambda}{(1+\lambda)^2} \frac{-1}{n-2} \int_0^1 x_2(1-x_2)^{n-2} dx_2 + \lambda \frac{1}{n-2} \int_0^{\frac{1}{1+\lambda}} x_2(1-x_2)^{n-2} dx_2 \\
&- \int_0^{\frac{1}{1+\lambda}} \frac{-1}{n-2} x_1(1-(x_1+x_2))^{n-2} \Big|_0^{\lambda x_2} - \frac{1}{(n-2)(n-1)}(1-(x_1+x_2))^{n-1} \Big|_0^{\lambda x_2} dx_2 \\
&+ \int_0^{\frac{1}{1+\lambda}} \frac{-1}{n-2} x_1(1-(x_1+x_2))^{n-2} \Big|_{\lambda x_2}^{1-x_2} - \frac{1}{(n-2)(n-1)}(1-(x_1+x_2))^{n-1} \Big|_{\lambda x_2}^{1-x_2} \\
&\frac{-\lambda}{(1+\lambda)^2} \frac{1}{n-2} \int_0^1 x_2(1-x_2)^{n-2} dx_2 + \frac{\lambda}{n-2} \int_{\frac{1}{1+\lambda}}^1 x_2(1-x_2)^{n-2} dx_2 \\
&- \int_{\frac{1}{1+\lambda}}^1 \frac{-1}{n-2} x_1(1-(x_1+x_2))^{n-2} \Big|_0^{1-x_2} - \frac{1}{(n-2)(n-1)}(1-(x_1+x_2))^{n-1} \Big|_0^{1-x_2} dx_2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\lambda}{(1+\lambda)^2}\frac{-1}{n-2}\int_0^1 x_2(1-x_2)^{n-2}dx_2 + \frac{\lambda}{n-2}\int_0^{\frac{1}{1+\lambda}} x_2(1-x)2)^{n-2}dx_2 \\
&- \int_0^{\frac{1}{1+\lambda}} \frac{-1}{n-2}\lambda x_2(1-(1+\lambda)x_2)^{n-2} - \frac{1}{(n-2)(n-1)}(1-(1+\lambda)x_2)^{n-1} \\
&+ \frac{1}{(n-2)(n-1)}(1-x_2)^{n-1}dx_2 + \int_0^{\frac{1}{1+\lambda}} \frac{1}{n-2}\lambda x_2(1-(1+\lambda)x_2)^{n-2} \\
&+ \frac{1}{(n-2)(n-1)}(1-(1+\lambda)x_2)^{n-1}dx_2 \\
&+ \frac{-\lambda}{(1+\lambda)^2}\frac{1}{n-2}\int_0^1 x_2(1-x_2)^{n-2}dx_2 + \frac{\lambda}{n-2}\int_{\frac{1}{1+\lambda}}^1 x_2(1-x_2)^{n-2}dx_2 \\
&- \int_{\frac{1}{1+\lambda}}^1 \frac{1}{(n-2)(n-1)}(1-x_2)^{n-1}dx_2 \\
&= \frac{\lambda}{(1+\lambda)^2)}\frac{-1}{n-2}\int_0^1 x(1-x)^{n-2}dx + \frac{\lambda}{n-2}\int_0^{\frac{1}{1+\lambda}} x(1-x)^{n-2}dx \\
&+ \frac{\lambda}{(1+\lambda)^2}\frac{1}{n-2}\int_0^1 x(1-x)^{n-2}dx + \frac{1}{1+\lambda}\frac{1}{(n-2)(n-2)}\int_0^1 (1-x)^{n-1}dx \\
&- \frac{1}{(n-2)(n-1)}\int_0^{\frac{1}{1+\lambda}} (1-x)^{n-1}dx + \frac{\lambda}{(1+\lambda)^2}\frac{1}{n-2}\int_0^1 x(1-x)^{n-2}dx \\
&+ \frac{1}{1+\lambda}\frac{1}{(n-2)(n-2)}\int_0^1 (1-x)^{n-1}dx - \frac{\lambda}{(1+\lambda)^2}\frac{1}{n-2}\int_0^1 x(1-x)^{n-2}dx \\
&+ \frac{\lambda}{n-2}\int_{\frac{1}{1+\lambda}}^1 x(1-x)^{n-2}dx - \frac{1}{(n-2)(n-1)}\int_{\frac{1}{1+\lambda}}^1 (1-x)^{n-1}dx \\
&= \frac{\lambda}{n-2}\int_0^1 x(1-x)^{n-2}dx + \left[2\frac{1}{1+\lambda} - 1\right]\frac{1}{(n-2)(n-1)}\int_0^1 (1-x)^{n-1}dx \\
&= \frac{1}{n(n-1)(n-2)}\left[\lambda + \frac{1-\lambda}{1+\lambda}\right] \\
&= \frac{1}{n(n-1)(n-2)}\frac{1+\lambda^2}{1+\lambda}
\end{aligned}
$$

Thus

$$\mathbb{E}\left[\frac{1}{r}\sum_{i=1}^m ||u_{i1}|^2 - \lambda|u_{i2}|^2|\right] = \frac{1+\lambda^2}{1+\lambda}$$

which, as in the complex gaussian case, achieves its minimum on $[0,1]$ of $2(\sqrt{2}-1) > 0.828$.

Implications related to Wright's conjecture

Using the same covering argument over rank-2 indefinite matrices as in Lemma 3.4.2, we obtain the RIP-1 property for unitary matrices. Since RIP-1 is stronger than injectivity of the measurements, this shows that there exists some integer r such that the measurements $|U_i x|_{i=1}^r$, where U_i are iid Haar distributed unitary matrices, are injective up to global phase with very high probability (Wright's conjecture is that there exist a set of 3 unitary operators which yield injective measurements). It would be interesting to see how small of an integer r can be achieved by probabilistic arguments, say by using more sophisticated concentration arguments.

4.3 Dual certification

We start with a useful property:

Moments of entries of a unitary matrix

Wlog, we shall further treat below the complex case only. We record some useful identities from [71]. Let u_{ij} be an entry of a $n \times n$ Haar distributed unitary matrix. Then

$$\mathbb{E}[|u_{ij}|^{2d}] = \frac{d!}{n(n+1)\dots(n+d-1)}$$

Which implies that $\mathbb{E}[|u_{ia}|^4] = \frac{2}{n(n+1)}$. Using the identity

$$\frac{1}{n} = \mathbb{E}[|u_{ia}|^2] = \mathbb{E}[|u_{ia}|^2(\sum_{b=1}^{n}|u_{ib}|^2)] = \mathbb{E}[|u_{ia}|^4] + (n-1)\,\mathbb{E}[|u_{ib}|^2|u_{ia}|^2]$$

we obtain, for $a \neq b$

$$\mathbb{E}[|u_{ia}|^2|u_{ib}|^2] = \frac{1}{n(n+1)}$$

Dual Certificates

With $\mathcal{A}$ as above, it can be verified that

$$\frac{1}{m}\,\mathbb{E}\,[\mathcal{A}^*\mathcal{A}] = I - \frac{1}{n+1} I \otimes I = \mathcal{S}$$

and we have $\mathcal{S}^{-1}(X) = X - \frac{1}{n+1}\operatorname{Tr}(X) I_n$. Thus, the regular construction of the dual certificate would be

$$\begin{aligned}\frac{1}{m}\mathcal{A}^*\mathcal{A}\mathcal{S}^{-1}(e_1e_1^*) &= \frac{1}{m}\sum_{i=1}^{m} n(n+1)u_iu_i^* \otimes u_iu_i^*(e_1e_1^* - \frac{1}{n+1}I_n) \\ &= \frac{n(n+1)}{m}\sum_{i=1}^{m}(|u_{i1}|^2 - \frac{1}{n+1})u_iu_i^* \\ &= \frac{n}{m}\sum_{i=1}^{m}((n+1)|u_{i1}|^2 - 1)u_iu_i^*\end{aligned}$$

Let $\psi_n = \mathbb{E}\left[(n)(n+1)(|u_{i1}| \wedge \frac{3}{\sqrt{n+1}})^4\right]$. ψ_n is slightly less than 2. Using a construction similar to that found in [5], we could then take the enhanced certificate to be

$$Y = \frac{1}{m}\sum_{i=1}^{m}(2n(n+1)(|u_{i1}| \wedge \frac{3}{\sqrt{n+1}})^2 - n(2\psi_n - 1))u_iu_i^*$$

We have then the expected value of this sum is 1 in the upper left corner, near to -1 on the rest of the diagonal and zero elsewhere. Furthermore, the contribution of the $|u_{i1}|$ term is capped to not be too large. We thus hope to acquire the same properties of the enhanced dual certificate as in the gaussian case.

Behavior of Y_T

Here we control the quantity $\|Y_T - e_1e_1^*\|_F$. We can re-write the certificate as

$$Y = \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}| \wedge \frac{3}{\sqrt{n+1}})^2 - (2\psi_n - 1))u_i^{(k)}u_i^{(k)*}$$

where $\{u_i^{(k)}\}_{i=1}^n$ are (indexed by k) iid Haar distributed on $\mathbb{U}_n$. To show that $\|Y_T - e_1e_1^*\|_F$ is small, it is enough to show that

$$\|\frac{1}{r}\sum_{k=1}^{r} x_k - e_1\|_2$$

is small, where

$$x_k =^d \sum_{i=1}^{n}(2(n+1)(|u_{i1}| \wedge \frac{3}{\sqrt{n+1}})^2 - (2\psi_n - 1))\bar{u}_{i1}u_i$$

We have

$$\begin{aligned}
\mathbb{E}\left[\|x_k\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}\left|(2(n+1)(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}})^2-(2\psi_n-1))\right|^2|u_{i1}|^2\right]\\
&= n\,\mathbb{E}\left[\left(4(n+1)^2\left(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}}\right)^4\right)|u_{i1}|^2\right]+\\
&\quad n\,\mathbb{E}\left[\left((2\psi_n-1)^2-4(n+1)(2\psi_n-1)\left(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}}\right)^2\right)|u_{i1}|^2\right]\\
&= 4n(n+1)^2\,\mathbb{E}\left[\left(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}}\right)^4|u_{i1}|^2\right]+(2\psi_n-1)^2\\
&\quad -4n(n+1)(2\psi_n-1)\,\mathbb{E}\left[\left(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}}\right)^2|u_{i1}|^2\right]\\
&\le 4n(n+1)^2\,\mathbb{E}\left[|u_{i1}|^6\right]+(2\psi_n-1)^2-4n(n+1)(2\psi_n-1)\,\mathbb{E}\left[\left(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}}\right)^4\right]\\
&= 4n(n+1)^2\frac{3!}{n(n+1)(n+2)}+4\psi_n^2-4\psi_n+1-4(2\psi_n-1)\psi_n\\
&= 24\frac{n+1}{n+2}+1-4\psi_n^2\le 24
\end{aligned}$$

Furthermore, we have

$$\|x_k\|_2=\left(\sum_{i=1}^{n}|(2(n+1)(|u_{i1}|\wedge\frac{3}{\sqrt{n+1}})^2-(2\psi_n-1))\bar{u}_{i1}|^2\right)^{1/2}\le\sqrt{21}$$

These facts allow us to apply the vector Bernstein inequality (Theorem 3.5.3) to get that $\|Y_T - e_1e_1^*\|_F$ is as small as necessary with probability at least $1-e^{-cr}$ for some constant c.

Behavior of $Y_{T^\perp}$

We would like to show that $Y_{T^\perp}\prec 0$ whp. It is enough to consider $\sup\{\langle x, Y_{T^\perp}x\rangle\,; x\in\mathbb{CS}^n, x_1=0\}$ and we aim to control this quantity via a covering argument. Using rotational invariance, we have

$$\begin{aligned}
\langle x, Y_{T^\perp}x\rangle =^d \langle e_2, Y_{T^\perp}e_2\rangle &= \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}|\wedge\frac{3}{\sqrt{n+1}})^2-(2\psi_n-1))|u_{i2}^{(k)}|^2\\
&= \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}|\wedge\frac{3}{\sqrt{n+1}})^2)|u_{i2}^{(k)}|^2-(2\psi_n-1)
\end{aligned}$$

A straightforward application of Talagrand's inequality fails here. Bernstein's inequality for weakly dependent variables also fails [12], so we will use an approach that involves conditioning and Talagrand's inequality. It suffices to show that

$$\frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n)|u_{i2}^{(k)}|^2$$

concentrates well about 0, where

$$\begin{aligned}\phi_n &= \mathbb{E}\left[\frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}| \wedge \frac{3}{\sqrt{n+1}})^2)|u_{i2}^{(k)}|^2\right] \\ &= \mathbb{E}\left[2n(n+1)(|u_{i1}| \wedge \frac{3}{\sqrt{n+1}})^2)|u_{i2}|^2\right] \leq 2\end{aligned}$$

we have,

$$\begin{aligned}&\frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|u_{i1}^{(k)}| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n)|u_{i2}^{(k)}|^2 \\ &=^d \frac{1}{r}\sum_{k=1}^{r} G(\zeta^{(k)}, z^{(k)}) \\ &= \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|v(z^{(k)})_i| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n)|v(t(v(\zeta^{(k)}), v(z^{(k)})))_i|^2\end{aligned}$$

and as before, we consider the surrogate function

$$\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)}) = \frac{1}{r}\sum_{k=1}^{r}\sum_{i=1}^{n}(2(n+1)(|\tilde{v}_1(z^{(k)})_i| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n)|\tilde{v}_2(t(\tilde{v}_1(\zeta^{(k)}), \tilde{v}_1(z^{(k)})))_i|^2$$

Now,

$$\begin{aligned}&\mathbb{P}\left(|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})| \geq t\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\chi_{\{|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})| \geq t\}}\middle|(z^{(1)}, \ldots, z^{(r)})\right]\right] \\ &= \mathbb{E}_z\left[\mathbb{P}_\zeta\left(\left|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})\right| \geq t\right)\right] \\ &\leq \mathbb{E}_z\left[\mathbb{P}_\zeta\left(\left|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)}) - \mathbb{E}_\zeta\left[\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})\right]\right| \geq t - \left|\mathbb{E}_\zeta\left[\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})\right]\right|\right)\right] \\ &\leq \mathbb{E}_z\left[\mathbb{P}_\zeta\left(\left|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)}) - f(\{z^{(i)}\}_{i=1}^{r})\right| \geq t - t_1\right)\chi_{\{|f(\{z^{(i)}\}_{i=1}^{r})| \leq t_1\}}\right] \\ &+ \mathbb{P}(|f(\{z^{(i)}\}_{i=1}^{r})| > t_1)\end{aligned}$$

where $f(\{z^{(i)}\}_{i=1}^r) = \mathbb{E}_\zeta \left[\frac{1}{r}\sum_{k=1}^r \tilde{G}(\zeta^{(k)}, z^{(k)})\right]$.

It now suffices to analyze the quantities $\text{Lip}_\zeta(\tilde{G}(\zeta, z))$ and $\mathbb{E}_\zeta\left[\tilde{G}(\zeta, z)\right]$ as functions of z. For $x \in \mathbb{R}^n$ or $\mathbb{C}^n$, $g(x) = \sum_{i=1}^n a_i|x_i|^2$ and $\|x_1\|_2 + \|x_2\|_2 \leq 2$, we have

$$|g(x_1) - g(x_2)| \leq 2\|a\|_\infty \|x_1 - x_2\|_2$$

Letting $a_i = \left(2(n+1)(|\tilde{v}_1(z)_i| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n\right)$ and noting $\|a\|_\infty \leq 20$

$$\begin{aligned}\left|\tilde{G}(\zeta_1, z) - \tilde{G}(\zeta_2, z)\right| &= |g(\tilde{v}_2(t(\tilde{v}_1(\zeta_1), \tilde{v}_1(z)))) - g(\tilde{v}_2(t(\tilde{v}_1(\zeta_2), \tilde{v}_1(z))))| \\ &\leq 2\|a\|_\infty \|\tilde{v}_2(t(\tilde{v}_1(\zeta_1), \tilde{v}_1(z))) - \tilde{v}_2(t(\tilde{v}_1(\zeta_2), \tilde{v}_1(z)))\|_2 \\ &\leq 40\text{Lip}(\tilde{v}_2)\text{Lip}(t|_{\mathcal{B}(0,1)^2})\text{Lip}(\tilde{v}_1)\|\zeta_1 - \zeta_2\|_2\end{aligned}$$

In conclusion

$$\text{Lip}_\zeta(\tilde{G}(\zeta, z)) \leq 8 * 40\frac{\sqrt{\frac{c_1}{c_2}}}{\sqrt{n}}$$

uniformly in z and thus

$$\text{Lip}(\frac{1}{r}\sum_{k=1}^r \tilde{G}(\zeta^{(k)}, z^{(k)})) \leq \frac{1}{\sqrt{r}} 8 * 40\frac{\sqrt{\frac{c_1}{c_2}}}{\sqrt{n}}$$

uniformly in $\left(z^{(1)}, \ldots, z^{(r)}\right)$. This gives that

$$\mathbb{P}_\zeta\left(\left|\frac{1}{r}\sum_{k=1}^r \tilde{G}(\zeta^{(k)}, z^{(k)}) - f(z^{(1)}, \ldots, z^{(r)})\right| \geq t)\right) \leq e^{-crnt^2}$$

for a constant c which depends on c_i but does not depend on z. Now we need to show that $f(z^{(1)}, \ldots, z^{(r)})$ concentrates well about its mean and that this mean is very small. We have

$$\begin{aligned}&f(z^{(1)}, \ldots, z^{(r)}) \\ &= \frac{1}{r}\sum_{k=1}^r \sum_{i=1}^n \mathbb{E}_\zeta\left[|\tilde{v}_2(t(\tilde{v}_1(\zeta^{(k)}), \tilde{v}_1(z^{(k)})))_i|^2\right]\left(2(n+1)(|\tilde{v}_1(z^{(k)})_i| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n\right)\end{aligned}$$

Let

$$h(z) = \{\mathbb{E}_\zeta\left[|\tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z)))_i|^2\right]\}_{i=1}^n.$$

and

$$p(z) = \{(2(n+1)(|\tilde{v}_1(z)_i| \wedge \frac{3}{\sqrt{n+1}})^2 - \phi_n)\}_{i=1}^n$$

First, using the following facts,

$$\begin{aligned}
&\mathbb{E}\left[\tilde{v}_1(\zeta)_i\right] = 0 \\
&\mathbb{E}\left[\tilde{v}_1(\zeta)_a \tilde{v}_1(\zeta)_b\right] = 0, a \neq b \\
&\mathbb{E}\left[|\tilde{v}_1(\zeta)_i|^2\right] \leq \frac{1}{n} \\
&\mathbb{E}\left[|\left\langle \tilde{v}_1(\zeta), y\right\rangle|^2\right] \leq \|y\|_2^2 \frac{1}{n} \\
&\mathbb{E}\left[2\tilde{v}_1(\zeta)_i \bar{y}_i \left\langle \tilde{v}_1(\zeta), y\right\rangle\right] = 0
\end{aligned}$$

for any $y \in \mathbb{C}^n$, we establish

$$\begin{aligned}
&\mathbb{E}_\zeta\left[|\tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z)))_i|^2\right] \\
&\leq \mathbb{E}_\zeta\left[\frac{1}{c_2}|(t(\tilde{v}_1(\zeta), \tilde{v}_1(z)))_i|^2\right] \\
&\leq \frac{1}{c_2}\mathbb{E}_\zeta\left[|\tilde{v}_1(\zeta)_i|^2 + |\tilde{v}_1(z)_i|^2 |\langle \tilde{v}_1(z), \tilde{v}_1(\zeta)\rangle|^2 - 2\Re(\tilde{v}_1(\zeta)_i \bar{\tilde{v}}_1(z)_i \left\langle \tilde{v}_1(\zeta), \tilde{v}_1(z)\right\rangle)\right] \\
&\leq \frac{1}{c_2}\left[\frac{1}{n} + |\tilde{v}_1(z)_i|^2 \|\tilde{v}_1(z)\|_2^2 \frac{1}{n}\right] \leq \frac{2}{c_2 n}
\end{aligned}$$

Thus, for any z, $\|h(z)\|_\infty \leq \frac{2}{c_2 n}$.

Now we shall compute $\text{Lip}(\sum_{i=1}^n h_i(z)p_i(z))$ directly:

$$\begin{aligned}
&\left|\sum_{i=1}^n h_i(z_1)p_i(z_1) - \sum_{i=1}^n h_i(z_2)p_i(z_2)\right| \\
&\leq \|h(z_1)\|_\infty \sum_{i=1}^n |p_i(z_1) - p_i(z_2)| + \|p(z_2)\|_\infty \sum_{i=1}^n |h(z_2) - h(z_2)| \\
&\leq 2(n+1)\|h(z_1)\|_\infty \sum_{i=1}^n \left|(|\tilde{v}_1(z_1)_i| \wedge \frac{3}{\sqrt{n+1}})^2 - (|\tilde{v}_1(z_2)_i| \wedge \frac{3}{\sqrt{n+1}})^2\right| + \\
&\|p(z_2)\|_\infty \mathbb{E}_\zeta\left[\sum_{i=1}^n \left||\tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z_1)))_i|^2 - |\tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z_2)))_i|^2\right|\right] \\
&\leq 2(n+1)\|h(z_1)\|_\infty 2(\||\tilde{v}_1(z_1)| \wedge \frac{3}{\sqrt{n+1}} - |\tilde{v}_1(z_2)| \wedge \frac{3}{\sqrt{n+1}}\|_2) + \\
&\|p(z_2)\|_\infty \mathbb{E}_\zeta\left[2\|\tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z_1))) - \tilde{v}_2(t(\tilde{v}_1(\zeta), \tilde{v}_1(z_2)))\|_2\right] \\
&\leq \left[2(n+1)\|h(z_1)\|_\infty 2\text{Lip}(\tilde{v}_1) + \|p(z_2)\|_\infty 2\text{Lip}(\tilde{v}_2)\text{Lip}(t|_{\mathcal{B}(0,1)^2})\text{Lip}(\tilde{v}_1)\right]\|z_1 - z_2\|_2 \\
&\leq \left(16\frac{n+1}{n}\frac{\sqrt{c_1}/c_2}{\sqrt{n}} + 320\frac{\sqrt{\frac{c_1}{c_2}}}{\sqrt{n}}\right)\|z_1 - z_2\|_2
\end{aligned}$$

Thus,

$$\text{Lip}(f(z^{(1)},\ldots,z^{(r)})) = \text{O}\left(\frac{\sqrt{c_1}}{c_2}\frac{1}{\sqrt{rn}}\right)$$

This will allow us to get the desired concentration of $f(z^{(1)},\ldots,z^{(r)})$ around its mean via Talagrand's inequality. Namely, we obtain

$$\mathbb{P}\left(\left|f(z^{(1)},\ldots,z^{(r)}) - \mathbb{E}\left[f(z^{(1)},\ldots,z^{(r)})\right]\right| \geq t\right) \leq e^{-crnt^2}$$

for a constant c which depends on c_i.

Let $F = \frac{1}{r}\sum_{k=1}^{r} G(\zeta^{(k)}, z^{(k)})$ and $\tilde{F} = \frac{1}{r}\sum_{k=1}^{r} \tilde{G}(\zeta^{(k)}, z^{(k)})$. Then $\mathbb{E}\left[f(z^{(1)},\ldots,z^{(r)})\right] = \mathbb{E}\left[\tilde{F}\right]$ and note $\mathbb{E}[F] = 0$. Since both F and $\tilde{F}$ are bounded and differ on a set of exponentially small probability, for any valid choice of c_i, $\lim_{n\to\infty}\mathbb{E}\left[\tilde{F}\right] = 0$ and so having fixed t apriori, for n large enough

$$\mathbb{P}\left(\left|f(z^{(1)},\ldots,z^{(r)})\right| \geq \frac{t}{2}\right) \leq e^{-crn(t/4)^2}$$

Taking $t_1 = \frac{t}{2}$, this implies

$$\mathbb{P}\left(|\frac{1}{r}\sum_{k=1}^{r}\tilde{G}(\zeta^{(k)}, z^{(k)})| \geq t\right) \leq e^{-crn(t-\frac{t}{2})^2} + e^{-crn(t/4)^2}$$

Now using that F and $\tilde{F}$ differ on a set of probability at most $\text{O}(re^{-\gamma n})$, we have

$$\begin{aligned}\mathbb{P}\left(|F| \geq t\right) &\leq \mathbb{P}\{|\tilde{F}| \geq t - |\tilde{F} - F|\} \\ &\leq \mathbb{P}\{F \neq \tilde{F}\} + \mathbb{P}\{|\tilde{F}| \geq t\} \\ &\leq \text{O}(re^{-\gamma n}) + e^{-crn(t/2)^2} + e^{-crn(t/4)^2}\end{aligned}$$

Therefore, we have established

$$\mathbb{P}\{\langle x, Y_{T^\perp} x\rangle \geq t + \phi_n - (2\psi_n - 1)\} \leq \text{O}(re^{-\gamma n}) + e^{-crn(t/2)^2} + e^{-crn(t/4)^2}$$

To get an arbitrarily fast exponential rate of concentration, fix γ to be as large as needed by choosing c_i appropriately, then fix r large enough. Note that $\phi_n \leq 2$ and ψ_n is very close to 2 so that $\phi_n - (2\psi_n - 1) \approx -1$. Choosing an appropriate t, we get that $Y_{T^\perp}$ is negative definite with high probability via the standard covering argument, which completes the proof of the main theorem.

Bibliography

[1] H. Duadi et. al. “Digital Holography and Phase Retrieval”. In: *Source: Holography, Research and Technologies*. Ed. by J. Rosen. InTech, 2011.

[2] O. Bunk et al. “Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels”. In: *Acta Cryst., Section A: Foundations of Crystallography* 63 (2007), pp. 306–314.

[3] R. Balan, P.G. Casazza, and D. Edidin. “On Signal Reconstruction without Noisy Phase”. In: *Appl. Comp. Harm. Anal.* 20 (2006), pp. 345–356.

[4] R. Balan et al. “Fast Algorithms for Signal Reconstruction without Phase”. In: *Wavelets XII*. Vol. 6701. Proc. SPIE. 2007, pp. 670111920–670111932.

[5] R. Balan et al. “Painless Reconstruction from Magnitudes of Frame Coefficients”. In: *J. Four. Anal. Appl.* 15 (2009), pp. 488–501.

[6] H.H. Bauschke, P.L. Combettes, and D.R. Luke. “Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization”. In: *J. Opt. Soc. Am. A* 19.7 (2002), pp. 1334–1345.

[7] A. Beck and M. Teboulle. “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems”. In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[8] C. Beck and R. D’Andrea. “Computational study and comparisons of LFT reducibility methods”. In: *Proceedings of the American Control Conference*. 1998, pp. 1013–1017.

[9] S. Becker, E.J. Candès, and M. Grant. *Templates for convex cone problems with applications to sparse signal recovery*. Tech. rep. Preprint available at http://tfocs.stanford.edu/tfocs/paper.shtml. Department of Statistics, Stanford University, 2010.

[10] A. Ben-Tal and A. S. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics, 2001.

[11] R. Berinde et al. “Combining geometry and combinatorics: A unified approach to sparse signal recovery”. In: *CoRR* (2008), pp. –1–1.

[12] “Bernstein’s Inequality”. In: *Wikipedia* ().

[13] G. Bianchi, F. Segala, and A. Volcic. “The solution of the covariogram problem for plane $\mathcal{C}^2_+$ convex bodies”. In: *J. Differential Geometry* 60 (2002), pp. 177–198.

[14] M.J. Bogan and et al. “Single Particle X-ray Diffractive Imaging”. In: *Nano Lett.* 8.1 (2008), pp. 310–316.

[15] L.M. Brègman. “The method of successive projection for finding a common point of convex sets”. In: *Soviet Math. Dokl.* 6 (1965), pp. 688–692.

[16] Y.M. Bruck and L.G. Sodin. “On the ambiguity of the image reconstruction problem”. In: *Opt. Comm.* 30 (1979), pp. 304–308.

[17] E. J. Candès and T. Tao. “Decoding by linear programming”. In: *IEEE Trans. on Information Theory* 51.2 (2005), pp. 4203–4215.

[18] E.J. Candès and Y. Plan. “Matrix completion with noise”. In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936.

[19] E.J. Candès and B. Recht. “Exact matrix completion via convex optimization”. In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[20] E.J. Candès and T. Tao. “The Power of Convex Relaxation: Near-Optimal Matrix Completion”. In: *IEEE Trans. Inform. Theory* 56.5 (2010), pp. 2053–2080.

[21] E.J. Candès, M.B. Wakin, and S.P. Boyd. “Enhancing Sparsity by Reweighted l_1 Minimization”. In: *J. Four. Anal. Appl.* 14 (2008), pp. 877–905.

[22] E.J. Candès et al. “Phase retrieval via matrix completion”. In: *Preprint* (2011).

[23] A. Carballar and M.A. Muriel. “Phase reconstruction from reflectivity in fiber Bragg gratings”. In: *J. Ligthwave Technol.* 15.8 (1997), pp. 1314–1322.

[24] A. Chai, M. Moscoso, and G. Papanicolaou. *Array imaging using intensity-only measurements*. Tech. rep. Stanford University, 2010.

[25] C.C. Chen et al. “Application of the optimization technique to noncrystalline X-Ray diffraction microscopy: guided hybrid input-output method (GHIO)”. In: *Phys. Rev. B.* 76 (2007), p. 064113.

[26] J.V. Corbett. “The Pauli problem, state reconstruction and quantum-real numbers”. In: *Rep. Math. Phys.* 57 (2006), pp. 53–68.

[27] J.C. Dainty and J.R. Fienup. “Phase retrieval and image reconstruction for astronomy”. In: *Image Recovery: Theory and Application*. Ed. by H. Stark. New York: Academic Press, 1987.

[28] C. Davis and W. M. Kahan. “The rotation of eigenvectors by a perturbation. III.” In: *SIAM J. Numer. Anal.* 7 (1970), pp. 1–46.

[29] M. Dierolf and et al. “Ptychographic X-ray computed tomography at the nanoscale”. In: *Nature* 467 (2010), pp. 436–440.

[30] Xiaodong Li Emmanuel Candes. “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns.” In: *Arxiv e-prints* (August 2012).

[31] A. Faridian et al. “Nanoscale imaging using deep ultraviolet digital holographic microscopy”. In: *Optics Express* 18.13 (2010), pp. 14159–14164.

[32] M. Fazel. “Matrix Rank Minimization with Applications”. PhD thesis. Stanford University, 2002.

[33] M. Fazel, H. Hindi, and S. Boyd. “Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices”. In: *Proc. Am. Control Conf.* 2003, pp. 2156–2162.

[34] J.R. Fienup. “Phase retrieval algorithms: A comparison”. In: *Applied Optics* 21.15 (1982), pp. 2758–2768.

[35] J.R. Fienup. “Reconstruction of an object from the modulus of its Fourier transform”. In: *Optics Letters* 3 (1978), pp. 27–29.

[36] J. Finkelstein. “Pure-state informationally complete and “really” complete measurements”. In: *Phys. Rev. A* 70 (2004), p. 052107.

[37] R.W. Gerchberg and W.O. Saxton. “A practical algorithm for the determination of phase from image and diffraction plane pictures”. In: *Optik* 35 (1972), pp. 237–246.

[38] M. X. Goemans and D. P. Williamson. “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. In: *J. ACM* 42 (6 1995), pp. 1115–1145.

[39] D. Gross. *Recovering low-rank matrices from few coefficients in any basis*. Available at http://arvix.org/abs/0910.1879. 2009.

[40] D. Gross et al. “Quantum-state tomography via compressed sensing”. In: *Physical Review Letters* 105.15 (2010).

[41] L. Gubin, B. Polyak, and E. Raik. “The method of projections for finding the common point of convex sets”. In: *USTSR Comput. Math. and Math. Phys.* 7 (1967), pp. 1–24.

[42] R.W. Harrison. “Phase problem in crystallography”. In: *J. Opt. Soc. Am. A* 10.5 (1993), pp. 1045–1055.

[43] H. Hauptman. “The Direct Methods of X-ray Crystallography”. In: *Science* 233.4760 (1986), pp. 178–183.

[44] M. Hayes. “The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform”. In: *IEEE Trans. Acoust., Speech, Signal Proc.* 30 (1982), pp. 140–154.

[45] N. Hurt. *Phase Retrieval and Zero Crossings*. Norwell, MA: Kluwer Academic Publishers, 1989.

[46] E. Ip et al. “Coherent detection in optical fiber systems”. In: *Optics Express* 16.2 (2008), pp. 753–791.

[47] Y. Ivankovski and D. Mendlovic. “High-rate long-distance fiber-optic communication based on advanced modulation techniques”. In: *Applied Optics* 38.26 (1999), pp. 5533–5540.

[48] H. N. Chapman J. Miao and D. Sayre. “.” In: *Microscopy and Microanalysis 3, supplement 2* (1997), pp. 1155–1156.

[49] I. Johnson et al. “Coherent Diffractive Imaging Using Phase Front Modifications”. In: *Phys. Rev. Lett.* 100.15 (2008), p. 155503.

[50] J. Kiefer. “Sequential minimax search for a maximum”. In: *Proceedings of the American Mathematical Society* 4.3 (1953), pp. 502–506.

[51] T. Kleine-Ostmann and T. Nagatsuma. “A Review on Terahertz Communications Research.” In: *ournal of Infrared, Millimeter, and Terahertz Waves* 32(2) (2011), pp. 143–171.

[52] M.V. Klibanov, P.E. Sacks, and A.V. Tikhonravov. “The phase retrieval problem”. In: *Inverse problems* 11 (1–28), p. 1995.

[53] A. Levi and H. Stark. “Restoration from phase and magnitude by generalized projections”. In: *Image Recovery: Theory and application*. Ed. by H. Stark. Acad. Press, 1987, pp. 277–320.

[54] Y.J Liu and et al. “Phase retrieval in x-ray imaging based on using structured illumination”. In: *Phys. Rev. A* 78 (2008), p. 023817.

[55] E.G. Loewen and E. Popov. *Diffraction Gratings and Applications*. Marcel Dekker, 1997.

[56] Y. Lu and M. Vetterli. “Sparse spectral factorization: Unicity and reconstruction algorithms.” In: *The 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP) Prague, Czech Republic.* (2011.).

[57] D.R. Luke. “Finding best approximation pairs relative to a convex and a prox-regular set in a Hilbert space.” In: *SIAM J. Optimiz., 19(2):7140–739, 2008.* (2008).

[58] D.R. Luke, J.V. Burke, and R.G. Lyon. “Optical Wavefront Reconstruction: Theory and Numerical Methods”. In: *SIAM Rev.* 44.2 (2002), pp. 169–224.

[59] S. Marchesini. “A unified evaluation of iterative projection algorithms for phase retrieval”. In: *Rev. Sci. Inst.* 78 (2007), pp. 011301 11–10.

[60] S. Marchesini. “Ab initio compressive phase retrieval”. In: 22.22 (22). Preprint, [arxiv:0809.2006], 2008., p. 22.

[61] S. Marchesini. “Phase retrieval and saddle-point optimization”. In: *J. Opt. Soc. Am. A* 24 (2007), pp. 3289–3296.

[62] M. Mesbahi and G. P. Papavassilopoulos. “On the Rank Minimization Problem Over a Positive Semidefinite Linear Matrix Inequality”. In: *IEEE Transactions on Automatic Control* 42.2 (1997), pp. 239–243.

[63] J. Miao, J. Kirz, and D. Sayre. “The oversampling phasing method”. In: *Acta Cryst.* 56:D13 (2000), pp. 12–15.

[64] J. Miao, D. Sayre, and H.N. Chapman. “Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects”. In: *J. Opt. Soc. Am. A* 15.6 (1998), pp. 1662–1669.

[65] J. Miao et al. “Extending X-Ray Crystallography to Allow the Imaging of Noncrystalline Materials, Cells and Single Protein Complexes”. In: *Annu. Rev. Phys. Chem.* 59 (2008), pp. 387–410.

[66] R.P. Millane. “Phase retrieval in crystallography and optics”. In: *J. Opt. Soc. Am. A.* 7 (1990), 394–411.

[67] R.P. Millane. “Recent advances in phase retrieval”. In: *Image Reconstruction from Incomplete Data IV*. Ed. by P.J. Bones, M.A. Fiddy, and R.P. Millane. Vol. 6316. Proc. SPIE. 2006, 63160E/1–11.

[68] D.L. Misell. “A method for the solution of the phase problem in electron microscopy”. In: *J. Phys. D: App. Phy.* 6.1 (1973), pp. L6–L9.

[69] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Vol. 87. Applied Optimization. Boston: Kluwer, 2004.

[70] J. von Neumann. “Functional Operators, Vol. II.” In: *Number 22 in Annals of Mathematics Studies. Princeton University Press* (1950).

[71] J. Novak. “Truncations of Random Unitary Matrices and Young Tableaux”. In: *The electronic journal of combinatorics* 14.R21 (2007).

[72] K.A. Nugent et al. “Unique phase recovery for nonperiodic objects”. In: *Phys. Rev. Lett.* 91 (2003), p. 203902.

[73] L. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. Signal Processing Series. Prentice Hall, 1993.

[74] B. Recht, M. Fazel, and P. Parrilo. “Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization”. In: *SIAM Review* (to appear).

[75] H. Reichenbach. *Philosophic Foundations of Quantum Mechanics*. Berkeley: University of California Press, 1944.

[76] R.T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

[77] J.M. Rodenburg. “Ptychography and related diffractive imaging methods”. In: *Advances in Imaging and Electron Physics, vol. 150* 150 (2008), pp. 87–184.

[78] H. Sahinoglou and S.D. Cabrera. “On phase retrieval of finite-length sequences using the initial time sample”. In: *IEEE Transactions on Circuits and Systems* 38.5 (1991), pp. 954–958.

[79] J.L.C. Sanz. “Mathematical Considerations for the Problem of Fourier Transform Phase Retrieval fromMagnitude”. In: *SIAM Journal on Applied Mathematics* 45.4 (1985), pp. 651–664.

[80] G. Scapin. “Structural Biology and Drug Discovery”. In: *Current Pharmaceutical Design* 12 (2006), pp. 2087–2097.

[81] E. Candès T. Strohmer and V. Voroninski. “PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming”. In: *Communications on Pure and Applied Mathematics* (2011).

[82] Terrence Tao. “Talagrands Concentration Inequality - Blog entry”. In: (2009).

[83] P. Thibault et al. “Probe retrieval in ptychographic coherent diffractive imaging”. In: *Ultramicroscopy* 109 (2009), pp. 338–343.

[84] D.P. Varn, G.S. Canright, and J.P. Crutchfield. “Discovering planar disorder in close-packed structures from x-ray diffraction: Beyond the fault mode”. In: *Phys. Rev. B.* 66.174110 (2002).

[85] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *Compressed Sensing: Theory and Applications*. Ed. by Yonina C. Eldar and Gitta Kutyniok. To Appear. Preprint available at `http://www-personal.umich.edu/~romanv/papers/papers.html`. Cambridge University Press, 2010.

[86] A. Walther. “The question of phase retrieval in optics”. In: *Opt. Acta* 10 (1963), pp. 41–49.

[87] A. Szameit Y. Shechtman Y.C. Eldar and M. Segev. “Sparsity Based Sub-Wavelength Imaging with Partially Incoherent Light Via Quadratic Compressed Sensing.” In: *Optics Express* 19(16): (2011), pp. 14807–14822.

[88] D.C. Youla. “Mathematical theory of image restoration by the method of convex projections”. In: *Image Recovery: Theory and application*. Ed. by H. Stark. Acad. Press, 1987, pp. 29–77.