Some Algorithms and Paradigms for Big Data

by

Yan Shuo Tan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mathematics)
in the University of Michigan
2018

Doctoral Committee:

Professor Anna Gilbert, Co-Chair
Professor Roman Vershynin, Co-Chair, University of California, Irvine
Professor Jinho Baik
Assistant Professor Laura Balzano
Professor Alexander Barvinok

Yan Shuo Tan

yanshuo@umich.edu

ORCID iD: 0000-0002-6670-9181

To my parents Chin Kong and Seok Lee, to my sister Yan An, and to Tiffany.

ACKNOWLEDGEMENTS

There are many people without whom this dissertation would have been impossible. First, I would like to thank my family in Singapore for their love and support, and for their understanding as four years of absence turned into nine, and as nine now turn into something longer and more indefinite.

I cannot be grateful enough to my adviser Roman Vershynin for his patience and guidance, especially in the early days when I was still struggling to find my feet.

I am also grateful to the various faculty, at Michigan and elsewhere, who took the time to offer me knowledge, wisdom, and also kindness. There are too many to name individually, but I would like to especially thank Laura Balzano, Anna Gilbert, Paul Kessenich, and Mahdi Soltanolkotabi.

Also, how can I forget my friends and colleagues who have made graduate school such a fun, enlightening, and energetic experience.

Lastly, I would like to thank Tiffany for constantly motivating me to be my best self, for sharing in my sorrow and joy, and for giving me my home away from home.

PREFACE

From self-driving cars to facial recognition to AlphaGo, the successes of big data have imprinted it upon the population imagination as a wellspring of technological wonder. Much less obvious to the public, but equally as important from an academic perspective, is the fact that big data has led to a growing synergy amongst the fields of statistics, computer science, and mathematics. In particular, many ideas from both pure and applied mathematics have proved useful in developing and understanding data analysis algorithms and algorithmic frameworks. This dissertation is entirely in this spirit. It has given me great joy to draw wield tools from high-dimensional probability, stochastic processes, convex geometry, and even some algebra to chisel out a modest niche in the growing edifice of mathematical data science.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The reality of big data poses both opportunities and challenges to modern researchers. Its key features – large sample sizes, high-dimensional feature spaces, and structural complexity – enforce new paradigms upon the creation of effective yet algorithmic efficient data analysis algorithms. In this dissertation, we illustrate a few paradigms through the analysis of three new algorithms. The first two algorithms consider the problem of phase retrieval, in which we seek to recover a signal from random rank-one quadratic measurements. We first show that an adaptation of the randomized Kaczmarz method provably exhibits linear convergence so long as our sample size is linear in the signal dimension. Next, we show that the standard SDP relaxation of sparse PCA yields an algorithm that does signal recovery for sparse, model-misspecified phase retrieval with a sample complexity that scales according to the square of the sparsity parameter. Finally, our third algorithm addresses the problem of Non-Gaussian Component Analysis, in which we are trying to identify the non-Gaussian marginals of a high-dimensional distribution. We prove that our algorithm exhibits polynomial time convergence with polynomial sample complexity.

CHAPTER 1

Introduction

1.1 Big Data

We live in the age of big data. As early as 2013, Cukier and Mayer-Schoenberger offered the following striking description for the size of our digital universe [30].

> In the third century bc, the Library of Alexandria was believed to house the sum of human knowledge. Today, there is enough information in the world to give every person alive 320 times as much of it as historians think was stored in Alexandria's entire collection – an estimated 1,200 exabytes' worth. If all this information were placed on CDs and they were stacked up, the CDs would form five separate piles that would all reach to the moon.

Since then, the sheer quantity of data that we possess has only gotten more ridiculous. Indeed, the digital universe continues to grow at an exponential rate, and is widely projected to double in size once every three years for the foreseeable future. This dizzying trend has captured the popular imagination, and many books have been written investigating its origins and consequences for society. It is not the place of this dissertation to add to this literature. Instead, we offer here a brief sketch of what other people have already said.

The first question to ask is: Where does all of this data comes from? At least some of it is the digitification of information that was previously stored in print or other analog media. Think, for instance, of Google's project to scan and render machine-readable all of the world's books. Similar to this is the migration of existing modes of communication and record-keeping to electronic forms – where written correspondence once took place through letters, they now occur via email. Both of these trends have resulted from the power, convenience and accessibility of personal computers, and, more recently, the growing ubiquity of all manner of digital devices. Indeed, according to Statista, it is projected that more than 36 percent of the world's population will own a smartphone in 2018 [101].

Yet, as “smart” devices increasingly penetrate our lives and integrate themselves into our lifestyles, their effect has not merely been to render old forms of data digital, but more consequentially, to create *new kinds* of data where there were none before. Take for instance the growing proportion of financial transactions that now take place using credit cards or other electronic payment methods. This allows transactions to be methodically recorded, allowing companies to create electronic profiles of their customers in order to pursue targeted marketing. The burgeoning use of social networks is another example. Again in this case, previously unrecorded information – a person’s social contacts, and her interactions with them – are recorded and “datafied”. Indeed, it seems that almost everyday, new types of data are being created and lending themselves to novel applications. An example from [30] is illustrative.

> Appreciating people’s posteriors is the art and science of Shigeomi Koshimizu, a professor at the Advanced Institute of Industrial Technology in Tokyo. Few would think that the way a person sits constitutes information, but it can. When a person is seated, the contours of the body, its posture, and its weight distribution can all be quantified and tabulated. Koshimizu and his team of engineers convert backsides into data by measuring the pressure they exert at 360 different points with sensors placed in a car seat and by indexing each point on a scale of zero to 256. The result is a digital code that is unique to each individual. In a trial, the system was able to distinguish among a handful of people with 98 percent accuracy. . . . Koshimizu’s plan is to adapt the technology as an anti-theft system for cars.

This explosion of data has presented enormous opportunities for researchers. From a statistical point of view, big data means more covariates or more samples or both, leading to better predictions when fitting traditional statistical models such as linear and logistic regression. Meanwhile, in computer science, a decades-long paradigm shift in artificial intelligence has reached maturity: instead of concocting algorithms directly for computers to perform certain tasks, more success can be attained by letting computers “learn” the algorithms themselves through applying learning algorithms to massive amounts of data. Here, the proliferation of data has been married with rapid advances in computing power to make data- and computation-intensive algorithms like deep learning feasible. The stunning success of deep learning has reverberated around academia as well as society as large. Amongst other things, it has enabled self-driving cars, facial recognition, automatic language translation, and AI for Go and other strategy games that can beat the very best human players.

1.2 A new mathematics of data

While the most visible success of big data has been its technological applications, it has also fertilized much mathematical research. First of all, the diversity of forms of data that we collect behooves us to study new statistical models that model different types of data better. Fitting these models then require new algorithms and strategies. The variety of ways in which data is collected and stored also lends itself to different algorithmic set-ups. For example, there has been much recent work on algorithms that work under the assumption that data is streaming, or that it is distributed across a number of different servers.

In addition, the sheer size of data has posed many interesting theoretical questions. Despite advances in computational power and resources available on ever smaller computers, there is sometimes more data than can be handled by traditional algorithms within a reasonable time frame. As such, there is a need for algorithms that have running times that are linear or even sublinear in their input parameters. In particular, this accounts for the heavy use of gradient descent and stochastic gradient descent in machine learning, and there is now a renewed emphasis on studying first order methods in optimization.

Another way to get around the computational bottleneck is by pre-processing the data to make it more tractable. Two of the most common methods of doing so are (1) data segmentation through clustering and (2) dimension reduction to reduce the number of covariates. Both of these areas continue to be rich topics of research. Furthermore, now that data is often no longer the only scarce resource, it is useful and important to investigate the trade off between the statistical and computational resources required to achieve a given performance criterion for a given inferential problem.

Thus far, we have discussed questions arising from having data that has both too many samples and too many covariates. In many situations, however, the problem with big data is not simply computational, but also statistical in the sense that we have too many covariates but *too few* samples. This is the case, for instance, with genomic data. When trying to predict what genes are predictive of a higher risk for cancer, a researcher could have, say, tens of thousands of candidate genes, but only a few hundred patients from which DNA samples were taken. Attempting to find the genes naively using linear regression is impossible. The problem, however, becomes feasible when we assert that the signal is *sparse* in the sense that only a few genes are predictive for cancer. Using such prior knowledge allows one to break sample complexity barriers, and there has been much progress in this direction over the last decade using ℓ_1 penalty techniques.

Researchers studying theoretical problems inspired by big data are scattered across many different departments. However, there is a growing sense in the community that the most rapid progress will come from combining expertise from statistics, computer science, and other mathematical domains. There is even a place for pure mathematics. Distributional assumptions and the stochastic nature of many big data algorithms mean that randomness is a central feature of the theoretical landscape, leading to heavy use of probabilistic tools. Indeed, scalar and matrix concentration inequalities coming from the field of high-dimensional probability are central to the analysis of many algorithms [113].

The usefulness of pure mathematics to big data is not limited to the field of probability. Theorems from convex geometry are central to analyzing sparse subspace clustering [98, 99]; Grothendieck's inequality from functional analysis yields sharp guarantees for a community detection algorithm [49]; concepts from Riemannian geometry and dynamic systems shed new light on accelerated optimization methods as well as optimization in non-convex settings [123, 71]; algebraic geometry can be used to prove results for learning Gaussian Mixture Models and for filling in missing data [9, 83]; tensor decomposition has emerged as a leading strategy for learning latent variable models [3]. These examples are just a slice of the growing synergy between pure mathematics and data science.

My own research, as presented in this dissertation, has been in this spirit. One of the algorithms we shall analyze was inspired by the theory of Fourier Analysis, while another is analyzed using stochastic process theory, and is partially inspired by the theory of Brownian Motion.

1.3 What this dissertation is about

In the last section, we saw how the field of mathematical data science has been developing in an exciting manner. It is again not the place of this dissertation to be a textbook, or even a survey of this emerging field. Instead, we will focus on a few new algorithms, each of which tackles a data science problem in a way that is representative of some broad paradigms for modern mathematical data science. In this section, we will serve a few small appetizers from each of these topics.

1.3.1 Phase retrieval and first order optimization methods

The first problem that we consider is phase retrieval. Mathematically, phase retrieval is the problem of solving systems of rank-1 quadratic equations in $\mathbb{R}^n$ or $\mathbb{C}^n$:

$$|\langle \mathbf{a}_i, \mathbf{x} \rangle|^2 = b_i^2, \qquad i = 1, 2, \ldots, N.$$

where $\mathbf{a}_i \in \mathbb{R}^n$ (or $\mathbb{C}^n$) are known sampling vectors, $b_i > 0$ are observed measurements, and $\mathbf{x} \in \mathbb{R}^n$ (or $\mathbb{C}^n$) is the decision variable. This problem is well motivated by practical concerns coming from optical imaging and has been a topic of study from at least the early 1980s [43, 96].

Early algorithms used in practice were based on alternating minimization, and hence had few theoretical guarantees. The first provably polynomial time algorithm, *PhaseLift*, was proposed by Candès et al in 2013 [22], and was based on semidefinite programming. While this was a theoretical breakthrough, the algorithm is not feasible for high-dimensional data. This is because the computational running time for even state-of-the-art SDP solvers does not scale well with the dimension n of the underlying vector space. Algorithms and their time complexity bounds are problem specific, so it is not possible to give a precise description of the running time required. Nonetheless, popular solvers based on interior point methods all have to perform multiple matrix factorizations at each step, each of which require $\Omega(n^3)$ basic operations. Considering that even 300 by 300 images are data points in a 90,000-dimensional space, it is clear that more efficient algorithms are required for data coming from real applications.

To address this issue, there has been a growing body of work on first order methods for phase retrieval applied to the natural objective functions associated with the phase retrieval problem. These algorithms are proved to work despite the non-convexity of these objectives, and yield rapid speed-ups over *PhaseLift*. In Chapter 3 of this dissertation, we will discuss and prove a guarantee for a stochastic gradient scheme. We will prove that provided we start with an initial estimate $\mathbf{x}_0$ that is within constant distance of the true signal vector $\mathbf{x}_*$, then for any error tolerance ϵ, $O(n^2/\epsilon)$ basic operations are sufficient to obtain an estimate $\hat{\mathbf{x}}_*$ that is within distance ϵ of $\mathbf{x}_*$. This is provided that we are given $N = \Omega(n)$ independent Gaussian measurements. A valid initial estimate $\mathbf{x}_0$ can be provably obtained using a spectral initialization method, but numerical experiments show that the algorithm works from arbitrary initializations.

As mentioned in the previous section, first order methods have come to fore in recent years because of the large number of covariates of modern data. In this new regime, second order methods like Newton's method are too expensive, thereby removing many of the traditional tools in the optimization toolkit. Often, stochastic schemes such as subsampling can also lead to rapid speed-ups, and in this way, our algorithm for phase retrieval is representative of many successful algorithms for big data.

1.3.2 Sparsity and ℓ_1 penalties

Sparsity is a major theme that runs through much of modern data science. This is the case first of all because sparsity is a feature of many modern data sets and data analysis problems. For instance, although we now have thousands or even millions of covariates in regression problems, usually only a tiny fraction of them are predictive of the response variable. In signal processing, we often have sparse signals in high dimensional vector spaces and wish to linearly compress them into a vector space of much smaller dimension. In some set ups, naturally occurring signals that are not themselves sparse, such as natural images, become sparse when we use a carefully chosen basis or dictionary.

In addition, sparsity can be thought of as a statistical resource. By assuming that our regression vector is small, we vastly reduce the search space for the regression problem. This allows us to be able to estimate the regression vector with a number of samples that is much smaller than the number of covariates. For instance, if we know that the regression vector is s-sparse, then it is easy to prove that an exhaustive search over all subsets of s coordinates will allow us to estimate the regression vector accurately, so long as we have $\Omega(s \log n)$ samples. On the other hand, such an exhaustive search is not computationally feasible, so this alone does not yield a scheme for making use of sparsity.

Fortunately, there is actually such a computationally feasible scheme. The idea is to relax the sparsity constraint, which is combinatorial, to an ℓ_1-norm constraint, which is geometric. Moreover, since ℓ_1-constraint is convex, the theory of convex optimization tells us that it can be incorporated into the linear regression problem in a computationally feasible manner. Finally, we need to be sure that this relaxation is tight, i.e. that the solution to the ℓ_1-constrained problem remains the solution to the original sparse regression problem. This turns out to be true when our data matrices satisfy the "restricted isometry property", which holds with high probability for data that satisfy reasonable distributional assumptions, and when the number of samples is again of the order of $\Omega(s \log n)$.

This remarkable sequence of ideas was first discovered by Candès, Tao and Romberg [23], and applies also to the signal processing setting we mentioned earlier: one can compress an s-sparse signal $\mathbf{s}$ by applying a known random projection $\mathbf{A}$. The original signal can then be recovered from the compressed vector $\mathbf{x} := \mathbf{A}\mathbf{s}$ using the linear programming method described above. Their work helped to found the modern field of compressed sensing, which continues to be a highly active today.

It is natural to extend the theoretical framework of sparsity and ℓ_1-regularized optimization to phase retrieval. The phase retrieval model differs from that of linear regression only in the sense that linear measurements are replaced with quadratic measurements. Indeed both models are instances of single index models, the study of which have a rich history in statistics. Furthermore, the signal vectors that arise in phase retrieval are often sparse [96]. As such, it is unsurprising that there is already a body of work on the problem of sparse phase retrieval. For a brief overview of the literature, see Chapter 4.

1.3.3 Model misspecification in sparse phase retrieval

When trying to model real world data with parametric statistical models, it is also important to account for the possibility that the model might be misspecified, or in other words, that the true distribution does not lie in the parametric family that we have assumed. Whenever this is likely, we need to have algorithms that are robust, i.e. that the estimate produced by our algorithm continue to have good predictive power, however this is quantified.

In the case of linear regression, a common way in which model misspecification happens is that instead of having linear measurements, we receive measurements of the form

$$b_i := f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle), \qquad i = 1, 2, \ldots, N.$$

Here, f is an unknown, possibly random, link function. Indeed, real data is never precisely linear. Nonetheless, naive linear regression continues to work well as an algorithm for estimating $\mathbf{x}_*$ (up to norm), and researchers have for some time been using *Lasso* and other sparse linear regression algorithms for data in which the response variable is clearly non-linear, as is the case when it is discrete.

Plan and Vershynin were able to justify this practice theoretically in their work on the *non-linear Lasso* in 2016 [87]. They showed that, assuming that the measurement vectors $\mathbf{a}_i$'s are independent Gaussians, then Lasso continues to work with the same sample complexity guarantee so long as the link function f satisfies some regularity properties, and such that it is "positively correlated" with a true linear function.

Again, it is natural to extend this framework to the setting of phase retrieval. Here, we are concerned with having measurements that are not precisely quadratic. Note that the earlier analysis for Lasso does not apply to this setting because our unknown link functions should still be close in some sense to the square function, which is easily shown to be "uncorrelated" with linear functions.

In order to overcome this, we propose combining the "lifting" procedure of PhaseLift [22] with the "correlation maximization" algorithm that Plan and Vershynin proposed to solve the problem of 1-bit compressed sensing [85]. It turns out that the resulting algorithm is essentially the convex relaxation of sparse PCA proposed some years earlier by d'Aspremont et al. [31]. We are able that our algorithm has sample complexity $O(s^2 \log n)$, where s is again the sparsity parameter. This matches the performances of other algorithms that have been proposed for sparse phase retrieval. Nonetheless, it is not information theoretic optimal, and it is an open question whether the optimal rate can be achieved. We will pursue this discussion in much more detail in Chapter 4.

1.3.4 Learning with moments

One way to fit a parametric model is to use the method of moments. This method has a long and venerable history, having been first proposed and used by no less than Karl Pearson in 1894. Pearson had a set of data comprising the ratio of forehead to body length for 1000 crabs, and believed the crabs to have come from two different species rather than one. As such, he postulated that the forehead to body length ratio measurements could be modeled by the sum of two Gaussian components, which he then fit by matching the first 6 moments of the model with the empirical moments that he computed from the data [76].

More generally, the method of moments proceeds as follows: we estimate the parameter $\boldsymbol{\theta}$ to be the value such that the moments of the distribution $\mu_{\boldsymbol{\theta}}$ are "close" to the empirical moments computed from sample data, which in the basic setting comprises independent copies of a random variable drawn from the true distribution. Here, "closeness' means different things in different contexts. While Pearson first proposed the method to study distributions on $\mathbb{R}$, it can also be applied to distributions on $\mathbb{R}^n$. It is important to note that the moments of multivariate distributions are not scalars but tensors, and this adds an additional layer of complexity.

Recently, the method has been successfully applied to provide polynomial time algorithms for learning various latent variable models [54, 3], including topic models, hidden Markov models, and high-dimensional Gaussian mixture models. There are two key ideas that underpin these algorithms. First, the low-order moment tensors of the distributions have low-rank decompositions, i.e. each of them can be written as the sum of a small number of pure tensors). The individual pure tensors summands the yield information about the model parameter vector $\boldsymbol{\theta}$. Second, there is a robust polynomial time method for finding these low-rank decompositions [3].

Another variant of the method of moments also yield a polynomial time algorithm for solving the problem of Independent Component Analysis (ICA). ICA is a semi-parametric model that has applications to blind source separation. In this model, the signal is a random vector $\mathbf{s}$ in $\mathbb{R}^n$ with independent non-Gaussian entries, and the observations made by the observer are of the form $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is an unknown n by n mixing matrix. The goal of the problem is to learn the mixing matrix $\mathbf{A}$. The algorithm, introduced by Frieze et al. [44] and further studied by Arora et al. [4] is an iterative algorithm based on local search.

It exploits the fact that the columns of $\mathbf{A}$ are the local optima for the 4-th order moment tensor, i.e. for the function $f\colon S^{n-1} \to \mathbb{R}$ defined by

$$f(\mathbf{v}) := \mathbb{E}\{\langle \mathbf{v}, \mathbf{x} \rangle^4\}.$$

In summary, we see that the method of moments is a useful tool for many learning problems.

1.3.5 Dimension reduction through linear projections

The high-dimensional nature of many modern data sets make many otherwise useful data analysis algorithms inefficient, especially those whose running time scales as a high degree polynomial in the dimension of the ambient vector space. On the other hand, the *intrinsic* dimension of the data is usually a lot smaller than its ambient dimension. For instance, it is often the case the signal in the data is localized to a low-dimensional subspace. In such a situation, one would like to do dimension reduction. We would like to project the data to its “true” subspace, and run our algorithms on the projected data points instead, thereby leading to potentially massive speedups.

There are many methods for dimension reduction in the literature. We will mention two of the most basic here, both of which involve linear projections. Principal Component Analysis (PCA) involves projecting the data points to a subspace of maximal variance. Practically, one forms the sample covariance matrix of the data, computes its eigendecomposition, and then projects the data to the subspace spanned by the top k eigenvectors, where k is an algorithmic parameter that is supplied using prior knowledge or through model selection. The motivation for PCA is the assumption that the directions with more variance have more “explanatory value”. This would be true, for example, if the data arises from points lying on a subspace perturbed by a small amount of orthogonal noise.

Random projections have also turned out to be very useful. The celebrated Johnson-Lindestrauss lemma tells us that the pairwise ℓ_2 distances between N points are preserved under a random projection to a vector space of dimension $\Omega(\log N)$. In this instance, by random, we mean that the target subspace is drawn uniformly from the relevant Grassmannian. Indeed, random projections tend to preserve the “structure” of “data” so long as the target space has large enough dimension.

One way to make this precise is in the setting of structured regression. Suppose we are given the prior information that a signal $\mathbf{x}_*$ lies in a compact set $K \subset \mathbb{R}^n$. Then $\mathbf{x}_*$ can be estimated from a random projection $\mathbf{P}\mathbf{x}_*$ so long as the target subspace has dimension larger than a constant multiple of $w(K)^2$. Here $w(K)$ denotes the Gaussian width of the set

K, and we see that square corresponds to the “statistical dimension” of the set [115]. This has connections to the “M^* bound” and related theory in geometric functional analysis.

Apart from preserving information while simultaneously allowing us to work in a lower-dimensional space, random projections are also cheap to compute. This is because tall matrices with independent Gaussian entries are approximate isometries whose column spans are drawn uniformly from the Grassmannian. Such matrices are easy to construct using pseudorandom number generators. This fact allows us to have speedups in computing approximate matrix factorizations via random projections, thereby contributing to much of the success of randomized numerical linear algebra [50].

1.3.6 NGCA and reweighted PCA

In the problem of Non-Gaussian Component Analysis (NGCA), we assume that we have data points in a low dimensional subspace E that are perturbed by orthogonal Gaussian noise. The goal is to estimate this structured subspace E. If the noise is a lot smaller than the variance of the points within E, we can solve this problem using PCA. If this is either unknown or not the case, then new assumptions and ideas are needed.

A reasonable assumption to make is that the data points have non-Gaussian marginals in the directions that lie in E. In this case, we can use moment information to determine the non-Gaussian directions, thereby finding E. Indeed, this is what Vempala and Xiao proposed in 2011 [112]. Their idea was to adapt the local search algorithm proposed for solving ICA that we have discussed in a previous section [44]. Recall that this algorithm is able to recover the columns of the mixing matrix **A** as the local optima of the fourth moment tensor. In the NGCA case, we no longer assume that there are *independent* non-Gaussian directions, so a much more delicate analysis is required. Furthermore, we no longer assume that the non-Gaussian marginals differ from a Gaussian in the fourth moment. As such, higher moment tensors have to be considered.

Although Vempala and Xiao’s algorithm comes with a polynomial running time and sample complexity guarantee, it is fairly complicated and requires delicate parameter tuning. Furthermore, it is rather computationally inefficient. In Chapter 6, we will show that NGCA can also be solved by the far easier algorithm of reweighted PCA. To run this algorithm, we first place the data points in isotropic position. Next, we attach to each data point $\mathbf{X}_i$ the weight $\exp(-\alpha\|\mathbf{X}_i\|_2^2)$, where α is a parameter that can either be chosen with prior knowledge, or found through the running of the algorithm. We next do PCA on the reweighted sample, and then extract non-Gaussian directions as the eigenvectors to outlier eigenvalues.

Performing PCA with a reweighted sample has been applied successfully in several other contexts [18, 48]. In our case, the reason why it works is because of a new characterization of multi-dimensional Gaussian distributions: we are able to tell whether a random vector $\mathbf{X}$ is a standard Gaussian by inspecting the moments of its norm $\|\mathbf{X}\|_2$ and those of its dot product with an independent copy $\langle \mathbf{X}, \mathbf{X}' \rangle$. This moment information can be extracted from the reweighted sample covariance matrix, as well as from an auxiliary matrix that has be used in adversarial situations.

We will develop the characterization theorem further in Chapter 5. The theory turns out to also be useful for proving several theorems on energy minimization for distributions on the sphere.

1.4 Notes

Many of the results in this dissertation have been the result of joint work with my adviser, Roman Vershynin. Each of the following chapters is based on work that is available as a preprint or published paper. See [107, 108, 105, 106].

CHAPTER 2

High-Dimensional Probability

2.1 What is high-dimensional probability?

High-dimensional probability is the study of random objects defined over $\mathbb{R}^n$ or $\mathbb{C}^n$, where n is a large but fixed number. Such objects include vectors, matrices, tensors, and graphs. Despite its fundamental importance today, the reader may notice that there aren't many textbooks on high-dimensional probability, the reason being that the field is still very young, both in content and in name. Traditionally, probability theorists were interested in asymptotic results such as Central Limit Laws or in stochastic process theory. Many of the ideas, techniques and theorems in what we now call high-dimensional probability were instead developed to answer questions in geometric functional analysis, and later, non-asymptotic random matrix theory.

In recent years, researchers studying algorithms either possessing internal randomization or handling data with distributional assumptions naturally found themselves faced with questions about high-dimensional random objects. Sometimes, these questions could be answered with simple concentration inequalities such as Chernoff's inequality, but often, more sophisticated results are called for. As such, results from high-dimensional probability have begun to garner more attention, and in a synergistic manner, the field has begun to take on more independent research interest, emerging out of the shadow of geometric functional analysis into a more cohesive whole.

In this chapter, we collate some results from high-dimensional probability that will be used in the rest of this dissertation. These are collated from [70], [110], [113], [74], as well as several other sources that will be mentioned where appropriate.

2.2 ψ_α random variables

Definition 2.2.1 (Orlicz norms)**.** Let $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ be a convex, increasing function with $\psi(0) = 0$. Define the *Orlicz norm* of a random variable X with respect to ψ as

$$\|X\|_\alpha := \inf\{\lambda > 0 : \ \mathbb{E}\{\psi(|X|/\lambda)\} \leq 1\}.$$

Equipped with this norm, the space of random variables with finite norm forms a Banach space, called an *Orlicz space*.

We are especially interested in the Orcliz spaces corresponding to ψ_α for $\alpha > 0$. These are defined as follows. When $\alpha \geq 1$, we set $\psi_\alpha(x) := \exp(x^\alpha) - 1$. When $0 < \alpha \leq 1$, this function is no longer convex, so we convexify it by fiat, setting $\psi_\alpha(x) := \exp(x^\alpha) - 1$ for $x \geq x(\alpha)$ large enough, and taking ψ_α to be linear on $[0, x(\alpha)]$. If some random variable X has a finite ψ_α norm $\|X\|_{\psi_\alpha}$, we say that it is a ψ_α *random variable*.

Readers may already be familiar with ψ_2 and ψ_1 Orcliz spaces. These correspond to *subgaussian* and *subexponential* random variables respectively (see [113] for more details). For these two classes of random variables, we have the well-known Hoeffding's and Bernstein's inequalities.

Proposition 2.2.2 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_m$ be independent, centered, subgaussian random variables. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^m X_i\right|\right\} \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^m \|X_i\|_{\psi_2}^2}\right),$$

where $c > 0$ is an absolute constant.

Proposition 2.2.3 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_m$ be independent, centered, subexponential random variables. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^m X_i\right|\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^m \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right),$$

where $c > 0$ is an absolute constant.

In subsequent chapters, however, it will be useful for us to consider Orlicz spaces in full generality. This is because we will need to work with $\psi_{1/2}$ random variables, for which many of the standard concentration inequalities do not hold. Nonetheless, we still have the following.

Proposition 2.2.4 (Characterization of $\psi_{1/2}$ RVs)**.** *Let X be a real-valued random variable. Then the following properties are equivalent. The parameters $C_i > 0$ appearing in these properties differ from each other by at most an absolute contant factor.*

1. *The tails of X satisfy*

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp\Big(-\sqrt{t/C_1}\Big).$$

2. *The moments of X satisfy*

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq C_2 p^2.$$

3. *The $\psi_{1/2}$ norm of X satisfies*

$$\|X\|_{\psi_{1/2}} \leq C_3.$$

Proof. Same as in the case of ψ_1 and ψ_2. See [113]. □

We have the following further properties.

Proposition 2.2.5 (Products, Lemma 8.5 in [74])**.** *Let X and Y be ψ_α random variables for some $\alpha > 0$. Then XY is a $\psi_{\alpha/2}$ random variable with $\psi_{\alpha/2}$ norm satisfying*

$$\|XY\|_{\psi_{\alpha/2}} \leq C_\alpha \|X\|_{\psi_\alpha} \|Y\|_{\psi_\alpha}.$$

Here, C_α is an absolute constant depending only on α.

Proposition 2.2.6 (Centering)**.** *Let X be a ψ_α random variable for some $\alpha > 0$. Then*

$$\|X - \mathbb{E}X\|_{\psi_\alpha} \leq 2\|X\|_{\psi_\alpha}.$$

Proof. We have $\|X - \mathbb{E}X\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha} + \|\mathbb{E}X\|_{\psi_\alpha}$. Now check the definition of the norm to verify that $\|\mathbb{E}X\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha}$. □

Proposition 2.2.7 (Sums, Theorem 6.21 in [70])**.** *Let $0 < \alpha \leq 1$, and let $X_1, \dots, X_m$ be a sequence of independent, centered ψ_α random variables. Then*

$$\left\|\sum_{i=1}^m X_i\right\|_{\psi_\alpha} \leq C_\alpha \left(\mathbb{E}\left|\sum_{i=1}^m X_i\right| + \left\|\max_{1 \leq i \leq m} |X_i|\right\|_{\psi_\alpha}\right).$$

Here, C_α is an absolute constant depending only on α.

Proposition 2.2.8 (Maxima, Lemma 2.2.2 in [110])**.** *Let $0 < \alpha \leq 1$, and let $X_1, \ldots, X_m$ be independent, centered ψ_α random variables. Then*

$$\left\| \max_{1 \leq i \leq m} |X_i| \right\|_{\psi_\alpha} \leq C_\alpha \psi_\alpha^{-1}(m) \max_{1 \leq i \leq m} \|X_i\|_{\psi_\alpha}.$$

Here, C_α is an absolute constant depending only on α.

Proposition 2.2.9 (Bernstein-type inequality for $\psi_{1/2}$ RVs)**.** *Let $X_1, \ldots, X_m$ be an independent, centered $\psi_{1/2}$ random variables. There is an absolute contant C such that $S_m := \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i$ is a $\psi_{1/2}$ random variable with $\psi_{1/2}$ norm satisfying*

$$\|S_m\|_{\psi_{1/2}} \leq C \max_{1 \leq i \leq m} \|X_i\|_{\psi_{1/2}}.$$

In particular, if $\max_{1 \leq i \leq m} \|X_i\|_{\psi_{1/2}}$ is bounded above by a constant, for every $t \geq 0$, we have

$$\mathbb{P}\{|S_m| \geq t\} \leq 2 \exp(-\sqrt{t/C}).$$

Proof. This follows more or less immediately from the last two propositions. First, notice that

$$\mathbb{E} \left| \sum_{i=1}^m X_i \right| \leq \left(\mathbb{E} \left| \sum_{i=1}^m X_i \right|^2 \right)^{1/2} \leq C\sqrt{m} \max_{1 \leq i \leq m} \|X_i\|_{\psi_{1/2}}$$

Here, the first inequality is an application of Jensen's inequality, and the second uses the moment bound in Proposition 2.2.4. Next, we compute $\psi_\alpha^{-1}(m) = (\log(m+1))^2$, and use Proposition 2.2.8, we get

$$\left\| \max_{1 \leq i \leq m} |X_i| \right\|_{\psi_\alpha} \leq C(\log m)^2 \max_{1 \leq i \leq m} \|X_i\|_{\psi_\alpha}.$$

Finally, plug these two bounds into the inequality given by Proposition 2.2.7, and note that $\log(m+1)/\sqrt{m} \leq 5$. This completes the proof of the first statement. The tail bound follows from Proposition 2.2.4. □

2.3 Subgaussian random vectors and random matrices

We say that a random vector $\mathbf{X}$ in $\mathbb{R}^n$ is *subgaussian* if all one-dimensional marginals of $\mathbf{X}$ are subgaussian. Furthermore, if all these marginals have subgaussian norm bounded by a constant K, we abuse notation slightly and say that $\mathbf{X}$ has subgaussian norm $\|\mathbf{X}\|_{\psi_2}$ bounded above by K.

Theorem 2.3.1 (Concentration of norm for general sub-Gaussian vectors)**.** *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$, with $\|\boldsymbol{X}\|_{\psi_2} \leq K$. There is a universal constant C such that for each positive integer $r > 0$, the moments of $\|\boldsymbol{X}\|_2$ and $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ satisfy*

$$(\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^{1/r} \leq CK(\sqrt{n} + \sqrt{r}) \tag{2.1}$$

$$(\mathbb{E}\{|\langle \boldsymbol{X}, \boldsymbol{X}' \rangle|^r\})^{1/2r} \leq CK(\sqrt{n} + \sqrt{r}). \tag{2.2}$$

Proof. The second bound follows from the first, since by Cauchy-Schwarz,

$$(\mathbb{E}\{|\langle \mathbf{X}, \mathbf{X}' \rangle|^r\})^{1/2r} \leq (\mathbb{E}\{\|\mathbf{X}\|_2^r \|\mathbf{X}'\|_2^r\})^{1/2r} = (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^{1/r}$$

To prove (2.1), pick a $\frac{1}{2}$-net $\mathcal{N}$ on S^{n-1}. A volumetric argument shows that one may pick $\mathcal{N}$ to have size no more than 5^n (see [114]). We then have

$$\|\mathbf{X}\|_2 = \sup_{\mathbf{v} \in S^{n-1}} \langle \mathbf{X}, \mathbf{v} \rangle \leq 2 \sup_{\mathbf{v} \in \mathcal{N}} \langle \mathbf{X}, \mathbf{v} \rangle.$$

By definition, there is a universal constant c such that for any fixed unit vector $v \in S^{d-1}$, $\mathbb{P}\{\langle \mathbf{X}, \mathbf{v} \rangle > t\} \leq 2\exp\left(-\frac{ct^2}{K^2}\right)$. Taking a union bound over the net thus gives

$$\mathbb{P}\{\|\mathbf{X}\|_2 > 2t\} \leq 2\exp\left(n\log 5 - \frac{ct^2}{K^2}\right). \tag{2.3}$$

Next, we integrate out the tail bound (2.3) to obtain bounds for the moments. Observe that if $\frac{ct^2}{2K^2} \geq n\log 5$, we have $n\log 5 - \frac{ct^2}{K^2} \leq -\frac{ct^2}{2K^2}$. This condition on t is equivalent to $t \geq CK\sqrt{n}$, so we have

$$\mathbb{P}\{\|\mathbf{X}\|_2 > 2t\} \leq \begin{cases} 1 & t < CK\sqrt{n} \\ 2\exp\left(-\frac{ct^2}{K^2}\right) & t \geq CK\sqrt{n} \end{cases} \tag{2.4}$$

For any positive integer r, we integrate this bound to get

$$\begin{aligned} \mathbb{E}\{\|\mathbf{X}\|_2^r\} &= \int_0^\infty rt^{r-1}\mathbb{P}\{\|\mathbf{X}\| > t\}dt \\ &\leq \int_0^{CK\sqrt{n}} rt^{r-1}dt + \int_{CK\sqrt{n}}^\infty 2rt^{r-1}\exp\left(-\frac{ct^2}{K^2}\right)dt \\ &\leq C^rK^rn^{r/2} + C^rK^rr\int_0^\infty t^{r/2-1}e^{-t}dt. \end{aligned}$$

The integral in the last line is the gamma function, so in short, we have shown that

$$\mathbb{E}\{\|\mathbf{X}\|_2^r\} \leq C^r K^r (n^{r/2} + \Gamma(r/2+1)). \tag{2.5}$$

Taking r-th roots of both sides and using Hölder, together with the fact that $\Gamma(x)^{1/x} \lesssim x$, gives (2.1). □

Lemma 2.3.2 (Covariance estimation for sub-Gaussian random vectors)**.** *Let $\boldsymbol{X}$ be a centered sub-Gaussian random vector in $\mathbb{R}^n$ with covariance matrix $\boldsymbol{\Sigma}$ and sub-Gaussian norm satisfying $\|\boldsymbol{X}\|_{\psi_2} \leq K$ for some $K \geq 1$. Let $\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i\boldsymbol{X}_i^T$ denote the sample covariance matrix from N independent samples. Then there is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, we have $\mathbb{P}\left\{\left\|\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}\right\| > \epsilon\right\} \leq \delta$ so long as $N \geq CK^2(n + \log(1/\delta))\epsilon^{-2}$.*

Proof. This is essentially Theorem 5.39 in [114]. □

Lemma 2.3.3 (Moments of spherical marginals)**.** *Let $\boldsymbol{\theta}$ be uniformly distributed on the sphere S^{n-1}. Then for any unit vector $\mathbf{v} \in S^{n-1}$ and any positive integer k, we have*

$$\mathbb{E}\{\langle\boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\} = \frac{1 \cdot 3 \cdots (2k-1)}{n \cdot (n+2) \cdots (n+2k-2)} \tag{2.6}$$

Proof. There are several ways to prove this identity. We shall prove this by computing Gaussian integrals. Let g and $\mathbf{g}_n$ denote standard Gaussians in 1 dimension and n dimensions respectively. Then using the radial symmetry of $\mathbf{g}$, we have

$$\mathbb{E}\{g^{2k}\} = \mathbb{E}\{\langle\mathbf{g}_n, \mathbf{v}\rangle^{2k}\} = \mathbb{E}\{\langle\|\mathbf{g}_n\|_2\boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\} = \mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\}\mathbb{E}\{\langle\boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\}.$$

Rearranging gives

$$\mathbb{E}\{\langle\boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\} = \frac{\mathbb{E}\{g^{2k}\}}{\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\}}.$$

We then compute

$$\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\} = \frac{\omega_n}{(2\pi n)^{n/2}}\int_0^\infty r^{2k} r^{n-1} e^{-r^2/2} dr, \tag{2.7}$$

where ω_n is the volume of the sphere S^{n-1}. It is well known that

$$\omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)},$$

while we also have

$$\int_0^\infty r^{2k} r^{n-1} e^{-r^2/2} dr = 2^{n/2+k-1}\Gamma(n/2+k).$$

Substituting these back into (2.7) gives

$$\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\} = 2^k \frac{\Gamma(n/2+k)}{\Gamma(n/2)} = n \cdot (n+2) \cdots (n+2k-2). \tag{2.8}$$

This yields the denominator in (2.6). A similar calculation for $\mathbb{E}\{g^{2k}\}$ yields the numerator.

□

2.4 Chaining

Many concentration inequalities for random vectors and random matrices make use of net arguments. For example, consider Lemma 2.3.2 for bounding the operator norm of a random matrix. To prove this, one makes use of the fact that the operator norm of an n by n matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\| = \sup_{v \in S^{n-1}} \|\mathbf{A}\mathbf{v}\|. \tag{2.9}$$

The net argument is to approximate the supremum over the sphere by a maximum over an ϵ-*net*, that is, a collection of points $\mathcal{N}$ for which every other point on the sphere is ϵ-close to a point in $\mathcal{N}$. The error is then controlled using continuity properties of the ℓ_2 norm. In this case, as in many others, such an argument produces optimal results. However, this is not always the case.

To see why, it is insightful to view (2.9) as saying that $\|\mathbf{A}\|$ is the supremum of a random process $(X_\mathbf{v})$ indexed by $\mathbf{v}$ over the index set S^{n-1}. This supremum is bounded by considering the process increments $X_\mathbf{v} - X_\mathbf{u}$ at a scale of $\|\mathbf{v} - \mathbf{u}\| \approx \epsilon$, where ϵ is the parameter of the net that we are using. In the matrix case, the choice of $\epsilon$was not too important, but on many occasions it is. Worse still is the situation where, because of the non-uniformly of process increments, there is not a single choice of ϵ that works best, . In this case, it is helpful to try to consider all scales simultaneously. One way to address this is using the idea of generic chaining, which was first invented by Talagrand [70, 104]. We shall use a variant of his approach which is appropriate for our purposes. This approach was developed by Dirksen in [38].

Let (T, d) be a metric space. A sequence $\mathcal{T} = (T_k)_{k \in \mathbb{Z}_+}$ of subsets of T is called *admissible* if $|T_0| = 1$, and $|T_k| \leq 2^{2^k}$ for all $k \geq 1$. For any $0 < \alpha < \infty$, we define the γ_α

functional of (T, d) to be

$$\gamma_\alpha(T, d) := \inf_{\mathcal{T}} \sup_{t \in T} \sum_{k=0}^{\infty} 2^{k/\alpha} d(t, T_k). \tag{2.10}$$

Let d_1 and d_2 be two metrics on T. We say that a process (Y_t) has *mixed tail increments* with respect to (d_1, d_2) if there are constants c and C such that for all $s, t \in T$, we have the bound

$$\mathbb{P}(|Y_s - Y_t| \geq c(\sqrt{u} d_2(s, t) + u d_1(s, t))) \leq C e^{-u}. \tag{2.11}$$

Remark 2.4.1. In [38], processes with mixed tail increments are defined as above but with the further restriction that $c = 1$ and $C = 2$. This is not necessary for the result that we need (Lemma 2.4.2) to hold. The indeterminacy of c and C gets absorbed into the final constant in the bound.

Lemma 2.4.2 (Mixed tail processes, Theorem 5 in [38])**.** *If $(Y_t)_{t \in T}$ has mixed tail increments, then there is a constant C such that for any $u \geq 1$, with probability at least $1 - e^{-u}$,*

$$\sup_{t \in T} |Y_t - Y_{t_0}| \leq C(\gamma_2(T, d_2) + \gamma_1(T, d_1) + \sqrt{u}\,\text{diam}(T, d_2) + u\,\text{diam}(T, d_1)).$$

At first glance, the γ_2 and γ_1 quantities seem mysterious and intractable. We will show however, that they can be bounded by more familiar quantities that are easily computable in our situation. First, given a set T with metric d, we define the *covering number* of T at scale u to bethe smallest number of radius u balls needed to cover T. We denote this quantity by $N(T, d, u)$. Interchanging the supremum and the sum in (2.10), and then doing a few usual tricks yields the famous Dudley inequality.

Lemma 2.4.3 (Dudley's inequality)**.** *For each $\alpha > 0$, there is an absolute constant C_α for which for any metric space (T, d), one has*

$$\gamma_\alpha(T, d) \leq C_\alpha \int_0^\infty (\log N(T, d, u))^{1/\alpha} du. \tag{2.12}$$

2.5 Growth functions and VC dimension

Unlike most of the topics that we have discussed thus far, growth functions and VC dimension emerged not out of geometric functional analysis, but instead as part of an attempt to proof uniform limit laws in statistics. The theory first started with Vapnik and Chervonenkis's foundational paper [111], and has since grown into an indispensable tool for theoretical machine learning, and in particular PAC theory. We state here some definitions and standard results that will be required in 3.5. We refer the interested reader to [91] for a more in-depth exposition on these topics.

Let $\mathcal{X}$ be a set and $\mathcal{C}$ be a family of subsets of $\mathcal{X}$. For a given set $C \in \mathcal{C}$, we slightly abuse notation and identify it with its indicator function $1_C \colon \mathcal{X} \to \{0,1\}$. The *growth function* $\Pi_{\mathcal{C}} \colon \mathbb{N} \to \mathbb{R}$ of $\mathcal{C}$ is defined via

$$\Pi_{\mathcal{C}}(m) := \max_{x_1,\ldots,x_m \in \mathcal{X}} |\{(C(x_1), C(x_2), \ldots, C(x_m)) \,:\, C \in \mathcal{C}\}|.$$

Meanwhile, the *VC dimension* of $\mathcal{C}$ is defined to be the largest integer m for which $\Pi_{\mathcal{C}}(m) = 2^m$. These two concepts are fundamental to statistical learning theory. The key connection between them is given by the Sauer-Shelah lemma.

Lemma 2.5.1 (Sauer-Shelah, Corollary 3.3 in [91])**.** *Let $\mathcal{C}$ be a collection of subsets of VC dimension d. Then for all $m \geq d$, have*

$$\Pi_{\mathcal{C}}(m) \leq \left(\frac{em}{d}\right)^d.$$

The reason why we are interested in the growth function of a family of subsets $\mathcal{C}$ is because we have the following guarantee for the uniform convergence for the empirical measures of sets belonging to $\mathcal{C}$.

Proposition 2.5.2 (Uniform deviation, Theorem 2 in [111])**.** *Let $\mathcal{C}$ be a family of subsets of a set $\mathcal{X}$. Let μ be a probability measure on $\mathcal{X}$, and let $\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ be the empirical measure obtained from m independent copies of a random variable X with distribution μ. For every u such that $m \geq 2/u^2$, the following deviation inequality holds:*

$$\mathbb{P}(\sup_{C \in \mathcal{C}} |\hat{\mu}_m(C) - \sigma(C)| \geq u) \leq 4\Pi_{\mathcal{C}}(2m)\exp(-mu^2/16). \tag{2.13}$$

We now state and prove two simple claims.

Claim 2.5.3. *Let $\mathcal{C}$ be the collection of all hemispheres in S^{n-1}. Then the VC dimension of $\mathcal{C}$ is bounded from above by $n+1$.*

Proof. It is a standard fact from statistical learning theory [91] that the VC dimension of half-spaces in $\mathbb{R}^n$ is $n+1$. Since S^{n-1} is a subset of $\mathbb{R}^n$, the claim follows by the definition of VC dimension. □

Claim 2.5.4. *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of functions from a set $\mathcal{X}$ to $\{0, 1\}$. Using $\triangle$ to denote symmetric difference, we define*

$$\mathcal{C} \triangle \mathcal{D} := \{C \triangle D \mid C \in \mathcal{C}, D \in \mathcal{D}\}. \tag{2.14}$$

Then the growth function $\Pi_{\mathcal{C}\triangle\mathcal{D}}$ of $\mathcal{C}\triangle\mathcal{D}$ satisfies $\Pi_{\mathcal{C}\triangle\mathcal{D}}(m) \leq \Pi_{\mathcal{C}}(m) \cdot \Pi_{\mathcal{D}}(m)$ for all $m \in \mathbb{Z}_+$.

Proof. Fix m, and points $x_1, \ldots, x_m \in \mathcal{X}$. Then every possible configuration $(f(x_1), f(x_2), \ldots, f(x_m))$ arising from some $f \in \mathcal{C}\triangle\mathcal{D}$ is the point-wise symmetric difference

$$(f(x_1), f(x_2), \ldots, f(x_m)) = (C(x_1), C(x_2), \ldots, C(x_m))\triangle(D(x_1), D(x_2), \ldots, D(x_m))$$

of configurations arising from some $C \in \mathcal{C}$ and $D \in \mathcal{D}$. By the definition of growth functions, there are at most $\Pi_{\mathcal{C}}(m) \cdot \Pi_{\mathcal{D}}(m)$ pairs of these configurations, from which the bound follows. □

Remark 2.5.5. There is an extensive literature on how to bound the VC dimension of concept classes that arise from finite intersections or unions of those from a known collection of concept classes, each of which has bounded VC dimension. We won't require this much sophistication here, and refer the reader to [17] for more details.

CHAPTER 3

Phase Retrieval and the Randomized Kaczmarz Method

3.1 Introduction

The mathematical phase retrieval problem is that of solving a system of quadratic equations

$$|\langle \mathbf{a}_i, \mathbf{z}\rangle^2| = b_i^2, \qquad i = 1, 2, \ldots, m \tag{3.1}$$

where $\mathbf{a}_i \in \mathbb{R}^n$ (or $\mathbb{C}^n$) are known sampling vectors, $b_i > 0$ are observed measurements, and $z \in \mathbb{R}^n$ (or $\mathbb{C}^n$) is the decision variable. The solution to the problem, $\mathbf{x}_*$, is called the signal vector. It is customary to use terminology from signal processing in the phase retrieval, since the mathematical problem is inspired by practical applications to do with signal recovery.

One such application is Coded Diffraction Imaging (CDI). In this procedure, a small two-dimensional object is illuminated by a coherent wave, and its far field diffraction intensity pattern is observed. The intensity function so derived corresponds roughly to the squared magnitude of the 2D Fourier transform of the object's transmittance function. The problem of recovering the transmittance function then fits into the framework of our mathematical model (3.1). $\mathbf{x}_*$ is now the discretized transmittance function, while the $\mathbf{a}_i$'s are 2D DFT vectors. Generally, the high frequency spectrum of light waves makes it impossible for optical detection devices to measure their phase. Having to recovery a signal from the amplitudes of its Fourier transform is thus a central feature of optical imaging. Phase retrieval also arises naturally in many other settings in science and engineering, including electron microscopy, crystallography, and astronomy.

In line with the development of optical imaging, researchers have proposed and studied algorithms to address phase retrieval since at least the 1970s. The first algorithms, such as those proposed by Gerchberg-Saxton and Fienup were based on alternating projections

Loss function	Name	Papers
$f(\mathbf{z}) = \sum_{i=1}^{m} (\lvert\langle \mathbf{a}_i, \mathbf{z}\rangle\rvert^2 - b_i^2)^2$	Squared loss for intensities	[20, 103]
$f(\mathbf{z}) = \sum_{i=1}^{m} (\lvert\langle \mathbf{a}_i, \mathbf{z}\rangle\rvert - b_i)^2$	Squared loss for amplitudes	[119, 126]
$f(\mathbf{z}) = \sum_{i=1}^{m} \big\lvert\lvert\langle \mathbf{a}_i, \mathbf{z}\rangle\rvert^2 - b_i^2\big\rvert$	ℓ_1 loss for intensities	[42, 40, 35]

Table 3.1: Non-convex loss functions for phase retrieval

[43, 96]. These were shown to exhibit empirical convergence to a global minimum in the noise-free oversampled setting, but were not robust to noise, and had few theoretical guarantees.

Over the last half a decade, there has been great interest in constructing and analyzing algorithms with provable guarantees given certain classes of sampling vector sets. One line of research involves "lifting" the quadratic system to a linear system, which is then solved using convex relaxation (*PhaseLift*) [22]. A second method is to formulate and solve a linear program in the natural parameter space using an anchor vector (*PhaseMax*) [47, 6, 52]. Although both of these methods can be proved to have near optimal sample efficiency, the most empirically successful approach has been to directly optimize various naturally-formulated non-convex loss functions, the most notable of which are displayed in Table 3.1.

These loss functions enjoy nice properties which make them amenable to various optimization schemes [103, 42]. Those with provable guarantees include the prox-linear method of [40], and various gradient descent methods [20, 26, 119, 126, 35]. Some of these methods also involve adaptive measurement pruning to enhance performance.

In 2015, Wei [121] proposed adapting a family of randomized Kaczmarz methods for solving the phase retrieval problem. He was able to show using numerical experiments that these methods perform comparably with state-of-the-art Wirtinger flow (gradient descent) methods when the sampling vectors are real or complex Gaussian, or when they follow the coded diffraction pattern (CDP) model [20]. He also showed that randomized Kaczmarz methods outperform Wirtinger flow when the sampling vectors are the concatenation of a few unitary bases. Unfortunately, [121] was not able to provide adequate theoretical justification for the convergence of these methods (see Theorem 2.6 in [121]).

In this chapter, we attempt to bridge this gap by showing that the basic randomized Kaczmarz scheme used in conjunction with truncated spectral initialization achieves *linear convergence* to the solution with high probability, whenever the sampling vectors are drawn uniformly from the sphere[1] S^{n-1} and the number of measurements m is larger than a constant times the dimension n.

It is also interesting to note that the basic randomized Kaczmarz scheme is exactly *stochastic gradient descent* for the Amplitude Flow objective, which suggests that other gradient descent schemes can also be accelerated using stochasticity.

3.1.1 Randomized Kaczmarz for solving linear systems

The Kaczmarz method is a fast iterative method for solving systems of overdetermined linear equations that works by iteratively satisfying one equation at a time. In 2009, Strohmer and Vershynin [102] were able to give a provable guarantee on its rate of convergence, provided that the equation to be satisfied at each step is selected using a prescribed randomized scheme.

Suppose our system to be solved is given by

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{3.2}$$

where $\mathbf{A}$ is an m by n matrix. Denoting the rows of A by $\mathbf{a}_1^T, \dots, \mathbf{a}_m^T$, we can write (3.2) as the system of linear equations

$$\langle \mathbf{a}_i, \mathbf{x} \rangle = b_i, \quad i = 1, \dots, m.$$

The solution set of each equation is a hyperplane. The randomized Kaczmarz method is a simple iterative algorithm in which we *project the running approximation onto the hyperplane of a randomly chosen equation*. More formally, at each step k we randomly choose an index $r(k)$ from $[m]$ such that the probability that $r(k) = i$ is proportional to $\|\mathbf{a}_i\|_2^2$, and update the running approximation as follows:

$$\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{r(k)} - \langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle}{\|\mathbf{a}_{r(k)}\|_2^2} \mathbf{a}_{r(k)}.$$

Strohmer and Vershynin [102] were able to prove the following theorem:

Theorem 3.1.1 (Linear convergence for linear systems). *Let $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|_F / \sigma_{\min}(\boldsymbol{A})$. Then for any initialization $\boldsymbol{x}_0$ to the equation* (3.2)*, the estimates given to us by randomized Kaczmarz satisfy*

$$\mathbb{E}\|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2^2 \leq \left(1 - \kappa(A)^{-2}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2.$$

[1] This is essentially equivalent to being real Gaussian because of the concentration of norm phenomenon in high dimensions. Also, one may normalize vectors easily.

Note that if A has bounded condition number, then $\kappa(A) \asymp \sqrt{n}$.

3.1.2 Randomized Kaczmarz for phase retrieval

In the phase retrieval problem (3.1), each equation

$$|\langle \mathbf{a}_i, \mathbf{x}_* \rangle| = b_i$$

defines two hyperplanes, one corresponding to each of $\pm x$. A natural adaptation of the randomized Kaczmarz update for this situation is then to project the running approximation to the *closer hyperplane*. We restrict to the case where each measurement vector $\mathbf{a}_i$ has unit norm, so that in equations, this is given by

$$\mathbf{x}_k := \mathbf{x}_{k-1} + \eta_k \mathbf{a}_{r(k)}, \tag{3.3}$$

where

$$\eta_k = \text{sign}(\langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle) b_{r(k)} - \langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle.$$

In order to obtain a convergence guarantee for this algorithm, we need to choose $\mathbf{x}_0$ so that it is close enough to the signal vector $\mathbf{x}_*$. This is unlike the case for linear systems where we could start with an arbitrary initial estimate $\mathbf{x}_0 \in \mathbb{R}^n$, but the requirement is par for the course for phase retrieval algorithms. Unsurprisingly, there is a rich literature on how to obtain such estimates [22, 26, 126, 119]. The best methods are able to obtain a good initial estimate using $O(n)$ samples.

3.1.3 Main results

The main result of this chapter guarantees the linear convergence of the randomized Kaczmarz algorithm for phase retrieval for random measurements $\mathbf{a}_i$ that are drawn independently and uniformly from the unit sphere.

Theorem 3.1.2 (Convergence guarantee for algorithm)**.** *Fix $\epsilon > 0$, $0 < \delta_1 \leq 1/2$, and $0 < \delta, \delta_2 \leq 1$. There are absolute constants $C, c > 0$ such that if*

$$m \geq C(n \log(m/n) + \log(1/\delta)),$$

then with probability at least $1 - \delta$, m sampling vectors selected uniformly and independently from the unit sphere S^{n-1} form a set such that the following holds: Let $x \in \mathbb{R}^n$ be a signal vector and let $\boldsymbol{x}_0$ be an initial estimate satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq c\sqrt{\delta_1}\|\boldsymbol{x}_*\|_2$. Then for *any $\epsilon > 0$, if*

$$K \geq 2(\log(1/\epsilon) + \log(2/\delta_2))n,$$

then the K-th step randomized Kaczmarz estimate $\boldsymbol{x}_K$ satisfies $\|\boldsymbol{x}_K - \boldsymbol{x}_*\|_2^2 \leq \epsilon \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2$ with probability at least $1 - \delta_1 - \delta_2$.

Comparing this result with Theorem 3.1.1, we observe two key differences. First, there are now two sources of randomness: one is in the creation of the measurements $\mathbf{a}_i$, and the other is in the selection of the equation at every iteration of the algorithm. The theorem gives a guarantee that holds with high probability over both sources of randomness. Theorem 3.1.2 also requires an initial estimate $\mathbf{x}_0$. This is not hard to obtain. Indeed, using the *truncated spectral initialization* method of [26], we may obtain such an estimate with high probability given $m \gtrsim n$. For more details, see Proposition 3.7.1.

The proof of this theorem is more nontrivial than the Strohmer-Vershynin analysis of randomized Kaczmarz algorithm for linear systems [102]. We break down the argument in smaller steps, each of which may be of independent interest to researchers in this field.

First, we generalize the Kaczmarz update formula (3.3) and define what it means to take a randomized Kaczmarz step with respect to any probability measure on the sphere S^{n-1}: we choose a measurement vector at each step according to this measure. Using a simple geometric argument, we then provide a bound for the expected decrement in distance to the solution set in a single step, where the quality of the bound is given in terms of the properties of the measure we are using for the Kaczmarz update (Lemma 3.2.1).

Performing the generalized Kaczmarz update with respect to the uniform measure on the sphere corresponds to running the algorithm with unlimited measurements. We utilize the symmetry of the uniform measure to compute an explicit formula for the bound on the stepwise expected decrement in distance. This decrement is geometric whenever we make the update from a point making an angle of less than $\pi/8$ with the true solution, so we obtain linear convergence conditioned on no iterates escaping from the "basin of linear convergence". We are able to bound the probability of this bad event using a supermartingale inequality (Theorem 3.3.1).

Next, we abstract out the property of the uniform measure that allows us to obtain local linear convergence. We call this property the *anti-concentration on wedges* property, calling it ACW for short. Using this convenient definition, we can easily generalize our previous proofs for the uniform measure to show that all ACW measures give rise to randomized Kaczmarz update schemes with local linear convergence (Theorem 3.4.3).

The usual Kaczmarz update corresponds running the generalized Kaczmarz update with respect to $\mu_{\mathbf{A}} := \frac{1}{m}\sum_{i=1} \delta_{\mathbf{a}_i}$. We are able to prove that when the $\mathbf{a}_i$'s are selected uniformly and independently from the sphere, then $\mu_{\mathbf{A}}$ satisfies the ACW condition with high probability, so long as $m \gtrsim n$ (Theorem 3.5.7). The proof of this fact uses VC theory and a *chaining argument*, together with metric entropy estimates.

Finally, we are able to put everything together to prove a guarantee for the full algorithm in Section 3.6. In that section, we also discuss the failure probabilities δ, δ_1 and δ_2, and how they can be controlled.

3.1.4 Notes

This chapter is adapted from the paper "Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees" [107]. During the preparation of that manuscript, we became aware of independent simultaneous work done by Jeong and Güntürk. They also studied the randomized Kaczmarz method adapted to phase retrieval, and obtained almost the same result that we did (see [61] and Theorem 1.1 therein). In order to prove their guarantee, they use a stopping time argument similar to ours, but replace the ACW condition with a stronger condition called *admissibility*. They prove that measurement systems comprising vectors drawn independently and uniformly from the sphere satisfy this property with high probability, and the main tools they use in their proof are hyperplane tessellations and a net argument together with Lipschitz relaxation of indicator functions.

After submitting the first version of the manuscript, we also became aware of independent work done by Zhang, Zhou, Liang, and Chi [126]. Their work examines stochastic schemes in more generality (see Section 3 in their paper), and they claim to prove linear convergence for both the randomized Kaczmarz method as well as what they called *Incremental Reshaped Wirtinger Flow*. However, they only prove that the distance to the solution decreases in expectation under a single Kaczmarz update (an analogue of our Lemma 3.2.1 specialized to real Gaussian measurements). As we will see in this chapter, this bound cannot be naively iterated.

3.2 Computations for a single step

In this section, we will compute what happens in expectation for a single update step of the randomized Kaczmarz method. It will be convenient to generalize our sampling scheme slightly as follows. When we work with a fixed matrix A, we may view our selection of a random row $\mathbf{a}_{r(k)}$ as drawing a random vector according to the measure $\mu_{\mathbf{A}} := \frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{a}_i}$. We need not restrict ourselves to sums of Dirac delta functions. For any probability measure

μ on the sphere S^{n-1}, we define the random map $\mathbf{P} = \mathbf{P}_\mu$ on vectors $z \in \mathbb{R}^n$ by setting

$$\mathbf{Pz} := \mathbf{z} + \eta \mathbf{a}, \tag{3.4}$$

where

$$\eta = \text{sign}(\langle \mathbf{a}, \mathbf{z} \rangle) |\langle \mathbf{a}, \mathbf{x}_* \rangle| - \langle \mathbf{a}, \mathbf{z} \rangle \quad \text{and} \quad \mathbf{a} \sim \mu. \tag{3.5}$$

Note that as before, $\mathbf{x}_*$ is a fixed vector in $\mathbb{R}^n$ (think of $\mathbf{x}_*$ as the actual solution of the phase retrieval problem). We call $\mathbf{P}_\mu$ the *generalized Kaczmarz projection with respect to* μ. Using this update rule over independent realizations of $\mathbf{P}$, $\mathbf{P}_1, \mathbf{P}_2, \ldots$, together with an initial estimate $\mathbf{x}_0$, gives rise to a *generalized randomized Kaczmarz algorithm* for finding $\mathbf{x}_*$: set the k-th step estimate to be

$$\mathbf{x}_k := \mathbf{P}_k \mathbf{P}_{k-1} \cdots \mathbf{P}_1 \mathbf{x}_0. \tag{3.6}$$

Fix a vector $\mathbf{z} \in \mathbb{R}^n$ that is closer to $\mathbf{x}_*$ than to $-\mathbf{x}_*$, i.e. so that $\langle \mathbf{x}_*, \mathbf{z} \rangle > 0$, and suppose that we are trying to find $\mathbf{x}_*$. Examining the formula in (3.5), we see that $\mathbf{P}$ projects $\mathbf{z}$ onto the right hyperplane (i.e., the one passing through $\mathbf{x}_*$ instead of the one passing through $-\mathbf{x}_*$) if and only if $\langle \mathbf{a}, \mathbf{z} \rangle$ and $\langle \mathbf{a}, \mathbf{x}_* \rangle$ have the same sign. In other words, this occurs if and only if the random vector a does *not* fall into the region of the sphere defined by

$$W_{\mathbf{x}_*, \mathbf{z}} := \{ \mathbf{v} \in S^{n-1} \mid \text{sign}(\langle \mathbf{v}, \mathbf{x}_* \rangle) \neq \text{sign}(\langle \mathbf{v}, \mathbf{z} \rangle) \}. \tag{3.7}$$

This is the region lying between the two hemispheres with normal vectors $\mathbf{x}_*$ and $\mathbf{z}$. We call such a region a *spherical wedge*, since in three dimensions it has the shape depicted in Figure 3.1.

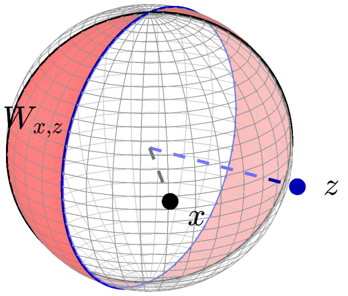

Figure 3.1: Geometry of $W_{\mathbf{x}_*, \mathbf{z}}$

When $\mathbf{a} \notin W_{\mathbf{x}_*, \mathbf{z}}$, we can use the Pythagorean theorem to write

$$\|\mathbf{z} - \mathbf{x}_*\|_2^2 = \|\mathbf{Pz} - \mathbf{x}_*\|_2^2 + \langle \mathbf{z} - \mathbf{x}_*, \mathbf{a} \rangle^2. \tag{3.8}$$

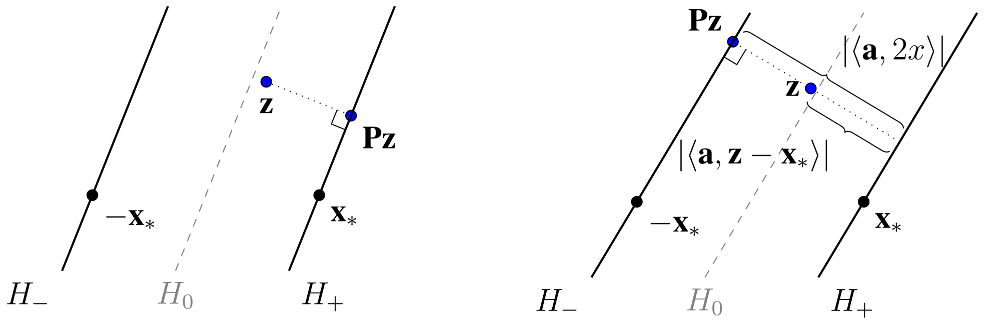

Figure 3.2: Orientation of $\mathbf{x}_*$, $\mathbf{z}$, and $\mathbf{Pz}$ when $a \in W_{\mathbf{x}_*,\mathbf{z}}$ and when $a \notin W_{\mathbf{x}_*,\mathbf{z}}$. H_+ and H_- denote respectively the hyperplanes defined by the equations $\langle \mathbf{y}, \mathbf{a}\rangle = b$ and $\langle \mathbf{y}, \mathbf{a}\rangle = -b$. H_0 denotes the hyperplane defined by the equation $\langle \mathbf{y}, \mathbf{a}\rangle = 0$. The left diagram demonstrates the situation when $a \in W_{\mathbf{x}_*,\mathbf{z}}$, thereby justifying (3.8). The right diagram demonstrates the situation when $a \notin W_{\mathbf{x}_*,\mathbf{z}}$, thereby justifying (3.11).

Rearranging gives

$$\|\mathbf{Pz} - \mathbf{x}_*\|_2^2 = \|\mathbf{z} - \mathbf{x}_*\|_2^2(1 - \langle \tilde{\mathbf{z}}, \mathbf{a}\rangle^2), \tag{3.9}$$

where $\tilde{\mathbf{z}} = (\mathbf{z} - \mathbf{x}_*)/\|\mathbf{z} - \mathbf{x}_*\|_2$.

In the complement of this event, we get

$$\mathbf{Pz} = \mathbf{z} + \langle \mathbf{a}, (-\mathbf{x}_*) - \mathbf{z}\rangle \mathbf{a} = \mathbf{z} - \langle \mathbf{a}, \mathbf{z} - \mathbf{x}_*\rangle + \langle \mathbf{a}, -2\mathbf{x}_*\rangle,$$

and using orthogonality,

$$\|\mathbf{Pz} - \mathbf{x}_*\|_2^2 = \|\mathbf{z} - \mathbf{x}_*\|_2^2 - \langle \mathbf{a}, \mathbf{z} - \mathbf{x}_*\rangle^2 + \langle \mathbf{a}, 2\mathbf{x}_*\rangle^2. \tag{3.10}$$

Since $\mathbf{z}$ gets projected to the hyperplane containing $-\mathbf{x}_*$, it may move further away from $\mathbf{x}_*$. However, we can bound how far away it can move. Because $\langle \mathbf{a}, \mathbf{x}_*\rangle$ has the opposite sign as $\langle \mathbf{a}, \mathbf{z}\rangle$, we have

$$|\langle \mathbf{a}, \mathbf{z} + \mathbf{x}_*\rangle| < |\langle \mathbf{a}, \mathbf{z} - \mathbf{x}_*\rangle|,$$

and so

$$|\langle \mathbf{a}, 2\mathbf{x}_*\rangle| = |\langle \mathbf{a}, (\mathbf{z} - \mathbf{x}_*) - (\mathbf{z} + \mathbf{x}_*)\rangle| < 2|\langle \mathbf{a}, \mathbf{z} - \mathbf{x}_*\rangle|.$$

Substituting this into (3.10), we get the bound

$$\|\mathbf{Pz} - \mathbf{x}_*\|_2^2 \leq \|\mathbf{z} - \mathbf{x}_*\|_2^2 + 3\langle \mathbf{a}, \mathbf{z} - \mathbf{x}_*\rangle^2 = \|\mathbf{z} - \mathbf{x}_*\|_2^2(1 + 3\langle \tilde{\mathbf{z}}, \mathbf{a}\rangle^2), \tag{3.11}$$

where $\tilde{\mathbf{z}}$ is as before.

We can combine (3.9) and (3.11) into a single inequality by writing

$$\begin{aligned}\|\mathbf{P}\mathbf{z}-\mathbf{x}_*\|_2^2 &\le \|\mathbf{z}-\mathbf{x}_*\|_2^2(1-\langle\tilde{\mathbf{z}},\mathbf{a}\rangle^2)1_{W^c_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}) + \|\mathbf{z}-\mathbf{x}_*\|_2^2(1+3\langle\tilde{\mathbf{z}},\mathbf{a}\rangle^2)1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a})\\ &= \|\mathbf{z}-\mathbf{x}_*\|_2^2(1-(1-4\cdot 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}))\langle\tilde{\mathbf{z}},\mathbf{a}\rangle^2)\\ &= \|\mathbf{z}-\mathbf{x}_*\|_2^2(1-\langle\tilde{\mathbf{z}},(1-4\cdot 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}))\mathbf{a}\mathbf{a}^T\tilde{\mathbf{z}}\rangle).\end{aligned}$$

Taking expectations, we can remove the role that $\tilde{\mathbf{z}}$ plays by bounding this as follows.

$$\begin{aligned}&\mathbb{E}[\|\mathbf{z}-\mathbf{x}_*\|_2^2(1-\langle\tilde{\mathbf{z}},(1-4\cdot 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}))\mathbf{a}\mathbf{a}^T\tilde{\mathbf{z}}\rangle)]\\ &\qquad\qquad = \|\mathbf{z}-\mathbf{x}_*\|_2^2(1-\langle\tilde{\mathbf{z}},\mathbb{E}[(1-4\cdot 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}))\mathbf{a}\mathbf{a}^T]\tilde{\mathbf{z}}\rangle)\\ &\qquad\qquad \le \|\mathbf{z}-\mathbf{x}_*\|_2^2\big[1-\lambda_{\min}(\mathbb{E}\mathbf{a}\mathbf{a}^T-4\mathbb{E}\mathbf{a}\mathbf{a}^T 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a}))\big].\end{aligned}$$

We may thus summarize what we have obtained in the following lemma.

Lemma 3.2.1 (Expected decrement)**.** *Fix vectors $\boldsymbol{x}_*,\boldsymbol{z}\in\mathbb{R}^n$, a probability measure μ on S^{n-1}, and let $P=\boldsymbol{P}_\mu$, $W_{\boldsymbol{x}_*,\boldsymbol{z}}$ be defined as in* (3.4) *and* (3.7) *respectively. Then*

$$\mathbb{E}\|\boldsymbol{P}\boldsymbol{z}-\boldsymbol{x}_*\|_2^2 \le \big[1-\lambda_{\min}(\mathbb{E}\boldsymbol{a}\boldsymbol{a}^T-4\mathbb{E}\boldsymbol{a}\boldsymbol{a}^T 1_{W_{\boldsymbol{x}_*,\boldsymbol{z}}}(\boldsymbol{a}))\big]\|\boldsymbol{z}-\boldsymbol{x}_*\|_2^2.$$

Let us next compute what happens for $\mu=\sigma$, the uniform measure on the sphere. It is easy to see that $\mathbb{E}\mathbf{a}\mathbf{a}^T=\frac{1}{n}\mathbf{I}_n$, so it remains to compute $\mathbb{E}\mathbf{a}\mathbf{a}^T 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a})$. To do this, we make a convenient choice of coordinates: Let θ be the angle between $\mathbf{z}$ and $\mathbf{x}_*$. We assume that both points lie in the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$, the first two basis vectors, and that the angle between $\mathbf{z}$ and $\mathbf{x}_*$ is bisected by $\mathbf{e}_1$, as illustrated in Figure 3.3.

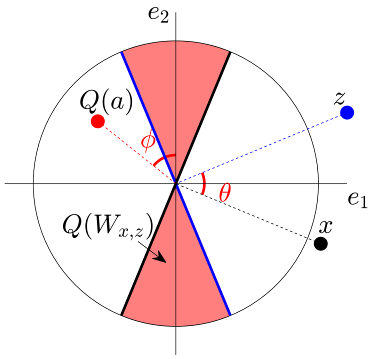

Figure 3.3: Choice of coordinates

For convenience, denote $\mathbf{M}:=\mathbb{E}\mathbf{a}\mathbf{a}^T 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a})$. Let $\mathbf{Q}$ denote the orthogonal projection operator onto the span of $\mathbf{e}_1$ and $\mathbf{e}_2$. Then $\mathbf{Q}(W_{\mathbf{x}_*,\mathbf{z}})$ is the union of two sectors of angle θ, which are respectively bisected by $\mathbf{e}_2$ and $-\mathbf{e}_2$. Recall that all coordinate projections of the uniform random vector a are uncorrelated. It is clear that from the symmetry in Figure 3.3 that they remain uncorrelated even when conditioning on the event that $\mathbf{a} \in W_{\mathbf{x}_*,\mathbf{z}}$. As such, $\mathbf{M}$ is a diagonal matrix.

Let ϕ denote the anti-clockwise angle of $\mathbf{Qa}$ from $\mathbf{e}_2$ (see Figure 3.3). We may write

$$\langle \mathbf{a}, \mathbf{e}_1 \rangle^2 = \|\mathbf{Qa}\|_2^2 \langle \mathbf{Qa}/\|\mathbf{Qa}\|_2, \mathbf{e}_1 \rangle^2 = \|\mathbf{Qa}\|_2^2 \sin^2 \phi.$$

Note that the magnitude and direction of $\mathbf{Qa}$ are independent, and $a \in W_{\mathbf{x}_*,\mathbf{z}}$ if either ϕ or $\phi - \pi$ lies between $-\theta/2$ and $\theta/2$. We therefore have

$$\mathbf{M}_{11} = \mathbb{E}[\langle \mathbf{a}, \mathbf{e}_1 \rangle^2 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a})] = \mathbb{E}[\|\mathbf{Qa}\|_2^2 \mathbb{E} \sin^2 \phi 1_{(-\theta/2,\theta/2)}(\phi \text{ or } \phi - \pi)].$$

By a standard calculation using symmetry, we have $\mathbb{E}\|\mathbf{Qa}\|_2^2 = 2/n$. Since ϕ is distributed uniformly on the circle, we can compute

$$\mathbb{E} \sin^2 \phi 1_{(-\theta/2,\theta/2)}(\phi \text{ or } \phi - \pi) = \frac{1}{\pi} \int_{-\theta/2}^{\theta/2} \sin^2 t dt = \frac{1}{\pi} \int_{-\theta/2}^{\theta/2} \frac{1 - \cos(2t)}{2} dt = \frac{\theta - \sin \theta}{2\pi}.$$

As such, we have $\mathbf{M}_{11} = (\theta - \sin \theta)/n\pi$, and by a similar calculation, $\mathbf{M}_{22} = (\theta + \sin \theta)/n\pi$. Meanwhile, for $i \geq 3$ we have

$$\begin{aligned} \mathbf{M}_{ii} &= \frac{\text{Tr}(\mathbf{M}) - \mathbf{M}_{11} - \mathbf{M}_{22}}{n-2} \\ &= \frac{\mathbb{E}[\|(\mathbf{I} - \mathbf{Q})\mathbf{a}\|_2^2 1_{W_{\mathbf{x}_*,\mathbf{z}}}(\mathbf{a})]}{n-2} \\ &= \frac{\mathbb{E}\|(\mathbf{I} - \mathbf{Q})\mathbf{a}\|_2^2 \mathbb{E} 1_{(-\theta/2,\theta/2)}(\phi \text{ or } \phi - \pi)}{n-2} \\ &= \frac{(n-2)/n \cdot \theta/\pi}{n} = \frac{\theta}{n\pi}. \end{aligned}$$

This implies that

$$\lambda_{\max}(\mathbf{M}_\theta) = \frac{\theta + \sin \theta}{n\pi}. \tag{3.12}$$

We have now completed proving the following lemma.

Lemma 3.2.2 (Expected decrement for uniform measure)**.** *Fix vectors $x, z \in \mathbb{R}^n$ such that $\langle \boldsymbol{z}, \boldsymbol{x}_* \rangle > 0$, and let $\boldsymbol{P} = \boldsymbol{P}_\sigma$ denote the generalized Kaczmarz projection with respect to σ, the uniform measure on the sphere. Let θ be the angle between $\mathbf{z}$ and $\boldsymbol{x}_*$. Then*

$$\mathbb{E}\|\boldsymbol{Pz} - \boldsymbol{x}_*\|_2^2 \leq \left[1 - \frac{1 - 4(\theta + \sin \theta)/\pi}{n}\right] \|\boldsymbol{z} - \boldsymbol{x}_*\|_2^2.$$

Remark 3.2.3. By being more careful, one may compute an *exact* formula for the expected decrement rather than a bound as is the case in previous lemma. This is not necessary for our purposes and does not give better guarantees in our analysis, so the computation is omitted.

3.3 Local linear convergence using unlimited uniform measurements

In this section, we will show that if we start with an initial estimate that is close enough to the ground truth $\mathbf{x}_*$, then repeatedly applying generalized Kaczmarz projections with respect to the uniform measure σ gives linear convergence in expectation. This is exactly the situation we would be in if we were to run randomized Kaczmarz given an unlimited supply of independent sampling vector $\mathbf{a}_1, \mathbf{a}_2, \ldots$ drawn uniformly from the sphere.

We would like to imitate the proof for linear convergence of randomized Kaczmarz for linear systems (Theorem 3.1.1) given in [102]. We denote by $\mathbf{X}_k$ the estimate after k steps, using capital letters to emphasize the fact that it is a random variable. If we know that $\mathbf{X}_k$ takes the value $\mathbf{x}_k \in \mathbb{R}^n$, and the angle θ_k that $\mathbf{z}$ makes with $\mathbf{x}_k$ is smaller than $\pi/8$, then, Lemma 3.2.2 tells us

$$\mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 \mid \mathbf{X}_k = \mathbf{x}_k] \leq (1 - \alpha_\sigma/n)\|\mathbf{x}_k - \mathbf{x}_*\|_2^2, \tag{3.13}$$

where $\alpha_\sigma := 1/2 - 4\sin(\pi/8)/\pi > 0$.

The proof for Theorem 3.1.1 proceeds by unconditioning and iterating a bound similar to (3.13). Unfortunately, our bound depends on $\mathbf{x}_k$ being in a specific region in $\mathbb{R}^n$ and does not hold arbitrarily. Nonetheless, by using some basic concepts from stochastic process theory, we may derive a *conditional* linear convergence estimate. The details are as follows.

For each k, let $\mathcal{F}_k$ denote the σ-algebra generated by $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$, where $\mathbf{a}_k$ is the sampling vector used in step k. Let $B \subset \mathbb{R}^n$ be the region comprising all points making an angle less than or equal to $\pi/8$ with $\mathbf{x}_*$. This is our *basin of linear convergence*. Let us assume a fixed initial estimate $\mathbf{x}_0 \in B$. Now define a stopping time τ via

$$\tau := \min\{k : \mathbf{X}_k \notin B\}. \tag{3.14}$$

For each k, and $\mathbf{x}_k \in B$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k+1} \mid \mathbf{X}_k = \mathbf{x}_k] &\leq \mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k} \mid \mathbf{X}_k = \mathbf{x}_k] \\ &= \mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k} \mid \mathbf{X}_k = \mathbf{x}_k, \mathcal{F}_k] \\ &= \mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 \mid \mathbf{X}_k = \mathbf{x}_k, \mathcal{F}_k] 1_{\tau > k} \\ &\leq (1 - \alpha_\sigma/n)\|\mathbf{x}_k - \mathbf{x}_*\|_2^2 1_{\tau > k}.\end{aligned}$$

Here, the first inequality follows from the inclusion $\{\tau > k+1\} \subset \{\tau > k\}$, the first equality statement from the Markov nature of the process $(\mathbf{X}_k)$, the second equality statement from the fact that τ is a stopping time, while the second inequality is simply (3.13). Taking expectations with respect to $\mathbf{X}_k$ then gives

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k+1}] &= \mathbb{E}[\mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k+1} \mid \mathbf{X}_k]] \\ &\leq (1 - \alpha_\sigma/n)\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 1_{\tau > k}].\end{aligned}$$

By induction, we therefore obtain

$$\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 1_{\tau > k}] \leq (1 - \alpha_\sigma/n)^k \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2.$$

We have thus proven the first part of the following convergence theorem.

Theorem 3.3.1 (Linear convergence from unlimited measurements)**.** *Let $\boldsymbol{x}_*$ be a vector in $\mathbb{R}^n$, let $\delta > 0$, and let $\boldsymbol{x}_0$ be an initial estimate to $\boldsymbol{x}_*$ such that $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \delta\|\boldsymbol{x}_*\|_2$. Suppose that our measurements $a_1, a_2, \ldots$ are fully independent random vectors distributed uniformly on the sphere S^{n-1}. Let $\boldsymbol{X}_k$ be the estimate given by the randomized Kaczmarz update formula* (3.6) *at step k, and let τ be the stopping time defined via* (3.14)*. Then for every $k \in \mathbb{Z}_+$,*

$$\mathbb{E}[\|\boldsymbol{X}_k - \boldsymbol{x}_*\|_2^2 1_{\tau = \infty}] \leq (1 - \alpha_\sigma/n)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2, \tag{3.15}$$

where $\alpha_\sigma = 1/2 - 4\sin(\pi/8)/\pi > 0$. Furthermore, $\mathbb{P}(\tau < \infty) \leq (\delta/\sin(\pi/8))^2$.

Proof. In order to prove the second statement, we combine a stopping time argument with a supermartingale maximal inequality. Set $Y_k := \|\mathbf{X}_{\tau \wedge k} - \mathbf{x}_*\|_2^2$. We claim that Y_k is a supermartingale. To see this, we break up its conditional expectation as follows:

$$\begin{aligned}\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] &= \mathbb{E}[\|\mathbf{X}_{\tau \wedge (k+1)} - \mathbf{x}_*\|_2^2 1_{\tau \leq k} \mid \mathcal{F}_k] + \mathbb{E}[\|\mathbf{X}_{\tau \wedge (k+1)} - \mathbf{x}_*\|_2^2 1_{\tau > k} \mid \mathcal{F}_k] \\ &= \mathbb{E}[\|\mathbf{X}_{\tau \wedge k} - \mathbf{x}_*\|_2^2 1_{\tau \leq k} \mid \mathcal{F}_k] + \mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau > k} \mid \mathcal{F}_k].\end{aligned}$$

Since $\|\mathbf{X}_{\tau\wedge k} - \mathbf{x}_*\|_2^2$ is measurable with respect to $\mathcal{F}_k$, we get

$$\mathbb{E}[\|\mathbf{X}_{\tau\wedge k} - \mathbf{x}_*\|_2^2 1_{\tau\leq k} \mid \mathcal{F}_k] = \|\mathbf{X}_{\tau\wedge k} - \mathbf{x}_*\|_2^2 1_{\tau\leq k} = Y_k 1_{\tau\leq k}.$$

Meanwhile, on the event $\tau > k$, we have $\mathbf{X}_k \in B$, so we may use (3.13) to obtain

$$\mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 1_{\tau>k} \mid \mathcal{F}_k] = \mathbb{E}[\|\mathbf{X}_{k+1} - \mathbf{x}_*\|_2^2 \mid \mathcal{F}_k] 1_{\tau>k} \leq (1-\alpha_\sigma/n)\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 1_{\tau>k}.$$

Next, notice that

$$\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 1_{\tau>k} = \|\mathbf{X}_{\tau\wedge k} - \mathbf{x}_*\|_2^2 1_{\tau>k} = Y_k 1_{\tau>k}.$$

Combining these calculations gives

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] \leq Y_k 1_{\tau\leq k} + (1-\alpha_\sigma/n) Y_k 1_{\tau>k} \leq Y_k.$$

Now define a second stopping time T to be the earliest time k such that $\|\mathbf{X}_k - \mathbf{x}_*\|_2 \geq \sin(\pi/8)\cdot\|\mathbf{x}_*\|_2$. A simple geometric argument tells us that $T \leq \tau$, and that T also satisfies

$$T = \inf\{k \mid Y_k \geq \sin^2(\pi/8)\|\mathbf{x}_*\|_2^2\}.$$

As such, we have

$$\mathbb{P}(\tau < \infty) \leq \mathbb{P}(T < \infty) = \mathbb{P}\Big(\sup_{1\leq k<\infty} Y_k \geq \sin^2(\pi/8)\|\mathbf{x}_*\|_2^2\Big).$$

Since (Y_k) is a non-negative supermartingale, we may apply the supermartingale maximal inequality to obtain a bound on the right hand side:

$$\mathbb{P}\Big(\sup_{1\leq k<\infty} Y_k \geq \sin^2(\pi/8)\|\mathbf{x}_*\|_2^2\Big) \leq \frac{\mathbb{E}Y_0}{\sin^2(\pi/8)\|\mathbf{x}_*\|_2^2} \leq (\delta/\sin(\pi/8))^2.$$

This completes the proof of the theorem. □

Corollary 3.3.2. *Fix $\epsilon > 0$, $0 < \delta_1 \leq 1/2$, and $0 < \delta_2 \leq 1$. In the setting of Theorem 3.3.1, suppose that $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \sqrt{\delta_1}\sin(\pi/8)\|\boldsymbol{x}_*\|_2$. Then with probability at least $1-\delta_1-\delta_2$, if $k \geq (\log(2/\epsilon) + \log(1/\delta_2))n/\alpha_\sigma$ then $\|\boldsymbol{X}_k - \boldsymbol{x}_*\|_2^2 \leq \epsilon\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2$.*

Proof. First observe that

$$\mathbb{P}(\tau < \infty) \leq \left(\frac{\sqrt{\delta_1}\sin(\pi/8)}{\sin(\pi/8)}\right)^2 = \delta_1 \leq 1/2.$$

Next, since

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 1_{\tau=\infty}] &= \mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \mid \tau = \infty]\mathbb{P}(\tau = \infty) + 0 \cdot \mathbb{P}(\tau < \infty) \\ &\geq \frac{1}{2}\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \mid \tau = \infty],\end{aligned}$$

applying Theorem 3.3.1 gives

$$\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \mid \tau = \infty] \leq 2(1 - \alpha_\sigma/n)^k \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2.$$

Applying Markov's inequality then gives

$$\begin{aligned}\mathbb{P}(\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 > \epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2 \mid \tau = \infty) &\leq \frac{\mathbb{E}[\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \mid \tau = \infty]}{\epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2} \\ &\leq \frac{2(1 - \alpha_\sigma/n)^k}{\epsilon}.\end{aligned}$$

Plugging our choice of k into this last bound shows that it is in turn bounded by δ_2. We therefore have

$$\begin{aligned}\mathbb{P}(\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \leq \epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2) &= \mathbb{P}(\|\mathbf{X}_k - \mathbf{x}_*\|_2^2 \leq \epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2 \mid \tau = \infty)\mathbb{P}(\tau = \infty) \\ &\geq (1 - \delta_2)(1 - \delta_1) \\ &\geq 1 - \delta_1 - \delta_2\end{aligned}$$

as we wanted. □

3.4 Local linear convergence for $\mathrm{ACW}(\theta, \alpha)$ measures

We would like to extend the analysis in the previous section to the setting where we only have access to finitely many uniform measurements, i.e. when we are back in the situation of (3.1). When we sample uniformly from the rows of A, this can be seen as running the generalized randomized Kaczmarz algorithm using the measure $\mu_{\mathbf{A}} = \frac{1}{m}\sum_{i=1}^m \delta_{\mathbf{a}_i}$ as opposed to $\mu = \sigma$.

If we retrace our steps, we will see that the key property of the uniform measure σ that we used was that if $W \subset S^{n-1}$ is a wedge[2] of angle θ, then we could make $\lambda_{\max}(\mathbb{E}_\sigma \mathbf{a}\mathbf{a}^T 1_W(\mathbf{a}))$ arbitrarily small by taking θ small enough (see equation (3.12)). We

[2] Recall that a wedge of angle θ is the region of the sphere between two hemispheres with normal vectors making an angle of θ.

do not actually need such a strong statement. It suffices for there to be an absolute constant α such that

$$\lambda_{\min}(\mathbb{E}\mathbf{a}\mathbf{a}^T - 4\mathbb{E}\mathbf{a}\mathbf{a}^T 1_W(\mathbf{a})) \geq \frac{\alpha}{n} \tag{3.16}$$

holds for θ small enough.

Definition 3.4.1 (Anti-concentration)**.** If a probability measure μ on S^{n-1} satisfies (3.16) for all wedges W of angle less than θ, we say that it is *anti-concentrated on wedges of angle θ at level α*, or for short, that it satisfies the ACW(θ, α) condition.

Abusing notation, we say that a measurement matrix A is ACW(θ, α) if the uniform measure on its rows is ACW(θ, α). Plugging in this definition into Lemma 3.2.1, we immediately get the following statement.

Lemma 3.4.2 (Expected decrement for ACW measure)**.** *Let μ be a probability measure on the sphere S^{n-1} satisfying the $ACW(\theta, \alpha)$ condition for some $\alpha > 0$ and some acute angle $\theta > 0$. Let $\boldsymbol{P} = \boldsymbol{P}_\mu$ denote the generalized Kaczmarz projection with respect to μ. Then for any $\boldsymbol{x}_*, \boldsymbol{z} \in \mathbb{R}^n$ such that the angle between them is less than θ, we have*

$$\mathbb{E}\|\boldsymbol{P}\boldsymbol{z} - \boldsymbol{x}_*\|_2^2 \leq (1 - \alpha/n)\|\boldsymbol{z} - \boldsymbol{x}_*\|_2^2. \tag{3.17}$$

We may now imitate the arguments in the previous section to obtain a guarantee for local linear convergence for the generalized randomized Kaczmarz algorithm using such a measure μ.

Theorem 3.4.3 (Linear convergence for ACW measure)**.** *Suppose μ is an* ACW(θ, α) *measure. Let $\boldsymbol{x}_*$ be a vector in $\mathbb{R}^n$, let $\delta > 0$, and let $\boldsymbol{x}_0$ be an initial estimate to $\boldsymbol{x}_*$ such that $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \delta\|\boldsymbol{x}_*\|_2$. Let $\boldsymbol{X}_k$ denote the k-th step of the generalized randomized Kaczmarz method with respect to the measure μ, defined as in* (3.6)*. Let Ω be the event that for every $k \in \mathbb{Z}_+$, $\boldsymbol{X}_k$ makes an angle less than θ with $\boldsymbol{x}_*$. Then for every $k \in \mathbb{Z}_+$,*

$$\mathbb{E}[\|\boldsymbol{X}_k - \boldsymbol{x}_*\|_2^2 1_\Omega] \leq (1 - \alpha/n)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2. \tag{3.18}$$

Furthermore, $\mathbb{P}(\Omega^c) \leq (\delta/\sin\theta)^2$.

Proof. We repeat the proof of Theorem 3.3.1. Let $B_\mu \subset S^{n-1}$ be the region on the sphere comprising all points making an angle less than or equal to $\pi/8$ with $\mathbf{x}_*$. Define stopping times τ_μ and T_μ as the earliest times that $\mathbf{X}_k \notin B_\mu$ and $\|\mathbf{X}_k - \mathbf{x}_0\|_2 \geq \sin(\theta)\|\mathbf{x}_*\|_2$ respectively. Again, $Y_k := \mathbf{X}_{k\wedge\tau_\mu}$ is a supermartingale, so we may use the supermartingale inequality to bound the probability of Ω^c. Conditioned on the event Ω, we may iterate the bound given by Lemma 3.4.2 to obtain (3.18). □

Corollary 3.4.4. *Fix $\epsilon > 0$, $0 < \delta_1 \leq 1/2$, and $0 < \delta_2 \leq 1$. In the setting of Theorem 3.4.3, suppose that $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \sqrt{\delta_1}\sin(\theta)\|\boldsymbol{x}_*\|_2$. Then with probability at least $1 - \delta_1 - \delta_2$, if $k \geq (\log(2/\epsilon) + \log(1/\delta_2))n/\alpha$ then $\|\boldsymbol{X}_k - \boldsymbol{x}_*\|_2^2 \leq \epsilon\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2$.*

3.5 $\mathrm{ACW}(\theta, \alpha)$ condition for finitely many uniform measurements

Following the theory in the previous section, we see that to prove linear convergence from finitely many uniform measurements, it suffices to show that the measurement matrix A is $\mathrm{ACW}(\theta, \alpha)$ for some θ and α.

For a *fixed* wedge W, we can easily achieve (3.16) by using a standard matrix concentration theorem. By taking a union bound, we can guarantee that it holds over exponentially many wedges with high probability. However, the function $W \mapsto \lambda_{\max}(\mathbb{E}\mathbf{a}\mathbf{a}^T 1_W(\mathbf{a}))$ is not Lipschitz with respect to any natural parametrization of wedges in S^{n-1}, so a naive net argument fails. To get around this, we use VC theory, metric entropy, and a chaining theorem from [38].

First, we will use the theory of VC dimension and growth functions to argue that all wedges contain approximately the right fraction of points. This is the content of the next lemma.

Lemma 3.5.1 (Uniform concentration of empirical measure over wedges)**.** *Fix an acute angle $\theta > 0$. Let $\mathcal{W}_\theta$ denote the collection of all wedges of S^{n-1} of angle less than θ. Suppose A is an m by n matrix with rows $\boldsymbol{a}_i$ that are independent uniform random vectors on S^{n-1}, and let $\mu_{\boldsymbol{A}} = \frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{a}_i}$. Then if $m \geq (4\pi/\theta)^2(2n\log(2em/n) + \log(2/\delta))$, with probability at least $1 - \delta$, we have*

$$\sup_{W \in \mathcal{W}} \mu_{\boldsymbol{A}}(W) \leq 2\theta/\pi.$$

Proof. Using VC theory (Proposition 2.5.2), we have

$$\mathbb{P}(\sup_{W \in \mathcal{W}} |\mu_{\mathbf{A}}(W) - \sigma(W)| \geq u) \leq 4\Pi_{\mathcal{W}_\theta}(2m)\exp(-mu^2/16) \tag{3.19}$$

whenever $m \geq 2/u^2$. Let $\mathcal{S}$ be the collection of all sectors of any angle, and let $\mathcal{H}$ denote the collection of all hemispheres. By Claim 2.5.3 and the Sauer-Shelah lemma (Lemma 2.5.1) relating VC dimension to growth functions, we have $\Pi_{\mathcal{H}}(2m) \leq (2em/n)^n$.

Next, notice that using the notation in (2.14), we have $\mathcal{W} = \mathcal{H} \triangle \mathcal{H}$. As such, we may apply Claim 2.5.4 to get

$$\Pi_{\mathcal{W}}(2m) \leq (2em/n)^{2n}.$$

We now plug this bound into the right hand side of (3.19), set $u = \theta/\pi$, and simplify to get

$$\mathbb{P}(\sup_{W \in \mathcal{W}} |\mu_{\mathbf{A}}(W) - \sigma(W)| \geq \theta/\pi) \leq 4\exp(2n\log(2em/n) - m(\theta/\pi)^2/16).$$

Our assumption implies that $m \geq 2/(\theta/\pi)^2$ so the bound holds, and also that the bound is less than δ. Finally, since $\mathcal{W}_\theta \subset \mathcal{W}$, on the complement of this event, any $W \in \mathcal{W}_\theta$ satisfies

$$\mu_{\mathbf{A}}(W) \leq \sigma(W) + \theta/\pi \leq 2\theta/\pi$$

as we wanted. □

For every wedge $W \in \mathcal{W}_\theta$, we may associate the configuration vector

$$\mathbf{s}_{W,\mathbf{A}} := (1_W(\mathbf{a}_1), 1_W(\mathbf{a}_2), \ldots, 1_W(\mathbf{a}_m)).$$

We can write

$$\lambda_{\max}(\mathbb{E}_{\mu_{\mathbf{A}}}\mathbf{a}\mathbf{a}^T 1_W(\mathbf{a})) = \frac{1}{m}\lambda_{\max}(\mathbf{A}^T\mathbf{S}_{W,\mathbf{A}}\mathbf{A}), \tag{3.20}$$

where $\mathbf{S}_{W,A} = \text{diag}(\mathbf{s}_{W,\mathbf{A}})$. $\mathbf{S}_{W,\mathbf{A}}$ is thus a selector matrix, and if we condition on the good event given to us by the previous theorem, it selects at most a $2\theta/\pi$ fraction of the rows of A. This means that $\mathbf{s}_{W,\mathbf{A}} \in \mathcal{S}_{2\theta/\pi}$, where we define

$$\mathcal{S}_\tau := \{\mathbf{d} \in \{0,1\}^m \mid \langle \mathbf{d}, \mathbf{1} \rangle \leq \tau \cdot m\}.$$

We would like to majorize the quantity in (3.20) uniformly over all wedges W by the quantity $\frac{1}{4}\lambda_{\min}(\mathbb{E}_{\mu_{\mathbf{A}}}\mathbf{a}\mathbf{a}^T)$. In order to do this, we define a stochastic process $(Y_{\mathbf{s},\mathbf{v}})$ indexed by $\mathbf{s} \in \mathcal{S}_{2\theta/\pi}$ and $\mathbf{v} \in B_2^n$, setting

$$Y_{\mathbf{s},\mathbf{v}} := n\mathbf{v}^T\mathbf{A}^T\text{diag}(\mathbf{s})\mathbf{A}\mathbf{v} = \sum_{i=1}^m s_i \langle \sqrt{n}\mathbf{a}_i, \mathbf{v} \rangle^2. \tag{3.21}$$

If we condition on the good set in Lemma 3.5.1, it is clear that

$$\sup_{W \in \mathcal{W}_\theta} \frac{1}{m}\lambda_{\max}(\mathbf{A}^T\mathbf{S}_{W,\mathbf{A}}\mathbf{A}) \leq \frac{1}{nm} \sup_{\mathbf{s} \in \mathcal{S}_{2\theta/\pi}, \mathbf{v} \in B_2^n} Y_{\mathbf{s},\mathbf{v}},$$

so it suffices to bound the quantity on the right. We will do this using Theorem 2.4.2, which requires us to show that our process $(Y_{\mathbf{s},\mathbf{v}})$ has mixed tail increments.

Lemma 3.5.2 ($(Y_{\mathbf{s},\mathbf{v}})$ has mixed tail increments)**.** *Let* $(Y_{\boldsymbol{s},\boldsymbol{v}})$ *be the process defined in* (3.21)*. Define the metrics* d_1 *and* d_2 *on* $\mathcal{S}_{2\theta/\pi} \times B_2^n$ *using the norms* $|||(\boldsymbol{w}, \boldsymbol{v})|||_1 = \max\{\|\boldsymbol{w}\|_\infty, \|\boldsymbol{v}\|_2\}$ *and* $|||(\boldsymbol{w}, \boldsymbol{v})|||_2 = \max\{\|\boldsymbol{w}\|_2, \sqrt{2m\theta/\pi}\|\boldsymbol{v}\|_2\}$*. Then the process has mixed tail increments with respect to* (d_1, d_2)*.*

Proof. The main tool that we use is Bernstein's inequality [114] for sums of subexponential random variables. Observe that each $\sqrt{n}\mathbf{a}_i$ is a subgaussian random vector with bounded subgaussian norm $\|\sqrt{n}\mathbf{a}_i\|_{\psi_2} \leq C$, where C by an absolute constant. As such, for any $\mathbf{v} \in B_2^n$, $\langle\sqrt{n}\mathbf{a}_i, \mathbf{v}\rangle^2$ is a subexponential random variable with bounded subexponential norm $\|\langle\sqrt{n}\mathbf{a}_i, \mathbf{v}\rangle^2\|_{\psi_1} \leq C^2$ [114].

Now fix $\mathbf{v}$ and let $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_{2\theta/\pi}$. Then

$$Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}} = \sum_{i=1}^{m}(s_i - s_i')\langle\sqrt{n}\mathbf{a}_i, \mathbf{v}\rangle^2.$$

Using Bernstein, we have

$$\mathbb{P}(|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}}| \geq u) \leq 2\exp(-c\min\{u^2/\|\mathbf{s} - \mathbf{s}'\|_2^2, u/\|\mathbf{s} - \mathbf{s}'\|_\infty\}). \tag{3.22}$$

Similarly, if we fix $\mathbf{s} \in \mathcal{S}_{2\theta/\pi}$ and let $\mathbf{v}, \mathbf{v}' \in B_2^n$, then

$$\begin{aligned} Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s},\mathbf{v}'} &= \sum_{i=1}^{m} s_i(\langle\sqrt{n}\mathbf{a}_i, \mathbf{v}\rangle^2 - \langle\sqrt{n}\mathbf{a}_i, \mathbf{v}'\rangle^2) \\ &= \sum_{i=1}^{m} s_i\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} - \mathbf{v}'\rangle\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} + \mathbf{v}'\rangle. \end{aligned}$$

We can bound the subexponential norm of each summand via

$$\begin{aligned} \|s_i\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} - \mathbf{v}'\rangle\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} + \mathbf{v}'\rangle\|_{\psi_1} &\leq s_i\|\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} - \mathbf{v}'\rangle\|_{\psi_2} \cdot \|\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} + \mathbf{v}'\rangle\|_{\psi_1} \\ &\leq Cs_i\|\mathbf{v} - \mathbf{v}'\|_2. \end{aligned}$$

As such,

$$\sum_{i=1}^{m}\|s_i\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} - \mathbf{v}'\rangle\langle\sqrt{n}\mathbf{a}_i, \mathbf{v} + \mathbf{v}'\rangle\|_{\psi_1}^2 \leq C\|\mathbf{v} - \mathbf{v}'\|_2^2\sum_{i=1}^{m} s_i^2 \leq C(2\theta/\pi)m\|\mathbf{v} - \mathbf{v}'\|_2^2.$$

Applying Bernstein as before, we get

$$\mathbb{P}(|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s},\mathbf{v}'}| \geq u) \leq 2\exp(-c\min\{u^2/(2\theta/\pi)m\|\mathbf{v}-\mathbf{v}'\|_2^2, u/\|\mathbf{v}-\mathbf{v}'\|_2\}). \qquad (3.23)$$

Now, recall the simple observation that for any numbers $a, b \in \mathbb{R}$, we have

$$\max\{|a|,|b|\} \leq |a| + |b| \leq 2\max\{|a|,|b|\}.$$

As such, for any $u > 0$, given $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_{2\theta/\pi}$, $\mathbf{v}, \mathbf{v}' \in B_2^n$, we have

$$\begin{aligned}\sqrt{u}|||(\mathbf{s},\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_2 + u|||(\mathbf{s},\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_1 &\geq \frac{1}{2}(\sqrt{u}\|\mathbf{s}-\mathbf{s}'\|_2 + \sqrt{u}\sqrt{2m\theta/\pi}\|\mathbf{v}-\mathbf{v}'\|_2 \\ &\quad + u\|\mathbf{s}-\mathbf{s}'\|_\infty + u\|\mathbf{v}-\mathbf{v}'\|_2) \\ &\geq \frac{1}{2}\max\{\sqrt{u}\|\mathbf{s}-\mathbf{s}'\|_2 + u\|\mathbf{s}-\mathbf{s}'\|_\infty, \\ &\quad \sqrt{u}\sqrt{2m\theta/\pi}\|\mathbf{s}-\mathbf{s}'\|_2 + u\|\mathbf{v}-\mathbf{v}'\|_2\}.\end{aligned}$$

Since

$$|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}'}| \leq |Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}}| + |Y_{\mathbf{s}',\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}'}|,$$

we have that if

$$|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}'}| \geq c\big(\sqrt{u}|||(\mathbf{s},\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_2 + u|||(,\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_1\big),$$

then either

$$|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}}| \geq \frac{c}{4}\big(\sqrt{u}\|\mathbf{s}-\mathbf{s}'\|_2 + u\|\mathbf{s}-\mathbf{s}'\|_\infty\big)$$

or

$$|Y_{\mathbf{s}',\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}'}| \geq \frac{c}{4}\Big(\sqrt{u}\sqrt{2m\theta/\pi}\|\mathbf{v}-\mathbf{v}'\|_2 + u\|\mathbf{v}-\mathbf{v}'\|_2\Big).$$

We can then combine the bounds (3.23) and (3.22) to get

$$\mathbb{P}\big(|Y_{\mathbf{s},\mathbf{v}} - Y_{\mathbf{s}',\mathbf{v}'}| \geq c\big(\sqrt{u}|||(\mathbf{s},\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_2 + u|||(\mathbf{s},\mathbf{v}) - (\mathbf{s}',\mathbf{v}')|||_1\big)\big) \leq 4e^{-u}.$$

Hence, the process $(Y_{\mathbf{s},\mathbf{v}})$ satisfies the definition (2.11) for having mixed tail increments.

□

We next bound the γ_1 and γ_2 functions for $\mathcal{S}_{2\theta/\pi} \times B_2^n$.

Lemma 3.5.3. *We may bound the γ_1 functional of $\mathcal{S}_{2\theta/\pi} \times B_2^n$ by*

$$\gamma_1(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_1) \leq C((2\theta/\pi)\log(\pi/2\theta)m + n).$$

Proof. The proof of the bound uses metric entropy Dudley's inequality. Recall that $\mathcal{S}_{2\theta/\pi}$ is the set of all $\{0,1\}$ vectors with fewer than $2\theta/\pi$ ones. For convenience, let us assume that $2m\theta/\pi$ is an integer. We then have the inclusion

$$\mathcal{S}_{2\theta/\pi} \subset \bigcup_{I\in\mathcal{I}} [0,1]^I,$$

where $\mathcal{I}$ is the collection of all subsets of $[m]$ of size $2m\theta/\pi$, and $[0,1]^I$ denotes the unit cube in the coordinate set I. We may then also write

$$\mathcal{S}_{2\theta/\pi} \times B_2^n \subset \bigcup_{I\in\mathcal{I}} ([0,1]^I \times B_2^n).$$

Note that a union of covers for each $[0,1]^I \times B_2^n$ gives a cover for $\mathcal{S}_{2\theta/\pi} \times B_2^n$. This, together with the symmetry of $\|\cdot\|_\infty$ with respect to permutation of the coordinates gives

$$N(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_1, u) \le |\mathcal{I}| \cdot N([0,1]^I \times B_2^n, |||\cdot|||_1, u)$$

for some fixed index set I.

We next generalize the notion of covering numbers slightly. Given two sets T and K, we let $N(T,K)$ denote the number of translates of K needed to cover the set T. It is easy to see that we have $N(T,d,u) = N(T, uB_d)$, where B_d is the unit ball with respect to the metric d. Since the unit ball for $|||\cdot|||_1$ is $B_\infty^m \times B_2^n$, we therefore have

$$\begin{aligned} N([0,1]^I \times B_2^n, |||\cdot|||_1, u) &= N([0,1]^I \times B_2^n, u(B_\infty^m \times B_2^n)) \\ &\le N(B_\infty^{(2\theta/\pi)m} \times B_2^n, u(B_\infty^{(2\theta/\pi)m} \times B_2^n)). \end{aligned}$$

Such a quantity can be bounded using a volumetric argument. Generally, for any centrally symmetric convex body K in $\mathbb{R}^n$, we have (see Corollary 4.1.15 in [5])

$$N(K, uK) \le (3/u)^n. \tag{3.24}$$

This implies that

$$\log N([0,1]^I \times S^{n-1}, |||\cdot|||_1, u) \le \log(3/u)((2\theta/\pi)m + n).$$

Finally, observe that

$$\log|\mathcal{I}| = \log \binom{m}{(2\theta/\pi)m} \le (2\theta/\pi)m \log(e\pi/2\theta).$$

We can thus plug these last two bounds into Dudley's inequality with $\alpha = 1$ (Lemma 2.4.3), noting that the integrand is zero for $u \geq 1$ to get

$$\begin{aligned}\gamma_1(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_1) &\leq C \int_0^1 (2\theta/\pi) m \log(e\pi/2\theta) + \log(3/u)((2\theta/\pi)m + n) du \\ &\leq C((2\theta/\pi)\log(\pi/2\theta)m + n)\end{aligned}$$

as was to be shown. □

Lemma 3.5.4. *We may bound the* γ_2 *functional of* $\mathcal{S}_{2\theta/\pi} \times B_2^n$ *by*

$$\gamma_2(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_2) \leq C\sqrt{2\theta/\pi}(m + \sqrt{mn}).$$

Proof. Since $\alpha = 2$, we may appeal directly to the theory of Gaussian complexity [113]. However, since we have already introduced some of the theory of metric entropy in the previous lemma, we might as well continue down this path. In this case, the Dudley bound states that

$$\gamma_2(T, d) \leq C \int_0^\infty \sqrt{\log N(T, d, u)} du \tag{3.25}$$

for any metric space (T, d).

Observe that the unit ball for $|||\cdot|||_2$ is $B_2^m \times (2m\theta/\pi)^{-1/2} B_2^n$. On the other hand, we conveniently have

$$\mathcal{S}_{2\theta/\pi} \times B_2^n \subset \sqrt{2m\theta/\pi} B_2^m \times B_2^n.$$

As such, we have

$$\begin{aligned}N(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_2, u) &\leq N(\sqrt{2m\theta/\pi} B_2^m \times B_2^n, |||\cdot|||_2, u) \\ &= N(\sqrt{2m\theta/\pi} B_2^m \times B_2^n, u(B_2^m \times (2m\theta/\pi)^{-1/2} B_2^n)) \\ &= N(T, (2m\theta/\pi)^{-1/2} uT),\end{aligned}$$

where $T = \sqrt{2m\theta/\pi} B_2^m \times B_2^n$.

Plugging this into (3.25) and subsequently using the volumetric bound (3.24), we get

$$\begin{aligned}\gamma_2(\mathcal{S}_{2\theta/\pi} \times B_2^n, |||\cdot|||_2) &\leq C \int_0^\infty \sqrt{\log N(T, (2m\theta/\pi)^{-1/2} uT)} du \\ &= C\sqrt{2m\theta/\pi} \int_0^\infty \sqrt{\log N(T, uT)} du \\ &\leq C\sqrt{2m\theta/\pi}\sqrt{m + n},\end{aligned}$$

which is clearly equivalent to the bound that we want. □

At this stage, we can put everything together to bound the supremum of our stochastic process.

Theorem 3.5.5 (Bound on supremum of $(Y_{\mathbf{s},\mathbf{v}})$)**.** *Let* $(Y_{\boldsymbol{s},\boldsymbol{v}})$ *be the process defined in* (3.21)*. Let* $0 < \delta < 1/e$*, let* θ *be an acute angle, and suppose* $m \geq \max\{n, \log(1/\delta)\pi/2\theta\}$*. Then with probability at least* $1-\delta$*, the supremum of the process satisfies*

$$\sup_{\boldsymbol{s}\in\mathcal{S}_{2\theta/\pi},\boldsymbol{v}\in B_2^n} Y_{\boldsymbol{s},\boldsymbol{v}} \leq C\sqrt{2\theta/\pi}\cdot m \tag{3.26}$$

Proof. It is easy to see that we have

$$\operatorname{diam}(\mathcal{S}_{2\theta/\pi}\times B_2^n, |||\cdot|||_1) = 2,$$

and

$$\operatorname{diam}(\mathcal{S}_{2\theta/\pi}\times B_2^n, |||\cdot|||_2) = 2\sqrt{2m\theta/\pi}.$$

Also observe that we have $Y_{\mathbf{s},0} = 0$ for any $s \in \mathcal{S}_{2\theta/\pi}$.

Using these, together with the previous two lemmas bounding the γ_1 and γ_2 functionals, we may apply Lemma 2.4.2 to see that

$$\sup_{\mathbf{s}\in\mathcal{S}_{2\theta/\pi},\mathbf{v}\in B_2^n} Y_{\mathbf{s},\mathbf{v}} \leq C(((2\theta/\pi)\log(\pi/2\theta)m+n)+\sqrt{2\theta/\pi}(m+\sqrt{mn})+u+\sqrt{u}\sqrt{2m\theta/\pi}).$$

with probability at least $1-e^{-u}$.

Using our assumptions on m, we may simplify this bound to obtain (3.26). □

Finally, we show that $\frac{1}{m}\sum_{i=1}^m \mathbf{a}_i\mathbf{a}_i^T$ is well-behaved.

Lemma 3.5.6. *Let* $\delta > 0$*. Then if* $m \geq C(n+\sqrt{\log(1/\delta)})$*, with probability at least* $1-\delta$*, we have*

$$\left\| \frac{n}{m}\sum_{i=1}^m \boldsymbol{a}_i\boldsymbol{a}_i^T - \boldsymbol{I}_n \right\| \leq 0.1$$

Proof. Note, as before, that the $\sqrt{n}\mathbf{a}_i$'s are isotropic subgaussian random variables with subgaussian norm bounded by an absolute constant. The claim then follows immediately from covariance estimation (Lemma 2.3.2). □

Theorem 3.5.7 (Finite measurement sets satisfy ACW condition)**.** *There is some* $\theta_0 > 0$ *and an absolute constant* C *such that for all angles* $0 < \theta \leq \theta_0$*, for all dimensions* n*, and any* $\delta > 0$*, if* m *satisfies*

$$m \geq C(\pi/2\theta)^2(n\log(m/n)+\log(1/\delta)), \tag{3.27}$$

then with probability at least $1 - \delta$, the measurement set A comprising m independent random vectors drawn uniformly from S^{n-1} satisfies the ACW(θ, α) *condition with $\alpha = 1/2$.*

Proof. Fix $n, \delta > 0$. Choose θ_0 such that the constant C in the statement in Theorem 3.5.5 satisfies $C\sqrt{2\theta_0/\pi} \leq 0.1$. Fix $0 < \theta \leq \theta_0$, and let Ω_1, Ω_2, and Ω_3 denote the good events in Lemma 3.5.1, Theorem 3.5.5, and Lemma 3.5.6 with this choice of θ. Whenever m satisfies our assumption (3.27), the intersection of these events occurs with probability at least $1 - 3\delta$ by the union bound.

Let us condition on being in the intersection of these events. For any wedge $W \in \mathcal{W}_\theta$ (i.e of angle less than θ), Lemma 3.5.1 tells us that its associated selector vector satisfies $s_{W,A} \in \mathcal{S}_{2\theta/\pi}$ (i.e. that it has at most $2m\theta/\pi$ ones,). By Theorem 3.5.5 and our assumption on θ_0, we then have

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^m \mathbf{a}_i\mathbf{a}_i^T 1_W(\mathbf{a}_i)\right) \leq \frac{1}{nm} \sup_{\mathbf{s}\in\mathcal{S}_{2\theta/\pi}, v\in B_2^n} Y_{\mathbf{s},\mathbf{v}} \leq \frac{0.1}{n}.$$

On the other hand, Lemma 3.5.6 guarantees that

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \mathbf{a}_i\mathbf{a}_i^T\right) \geq \frac{0.9}{n}.$$

Combining these, we get

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \mathbf{a}_i\mathbf{a}_i^T - \frac{4}{m}\sum_{i=1}^m \mathbf{a}_i\mathbf{a}_i^T 1_W(\mathbf{a}_i)\right) \geq \frac{1}{2n},$$

which was to be shown. □

3.6 Proof and discussion of Theorem 3.1.2

We restate the theorem here for convenience.

Theorem 3.6.1. *Fix $\epsilon > 0$, $0 < \delta_1 \leq 1/2$, and $0 < \delta, \delta_2 \leq 1$. There are absolute constants $C, c > 0$ such that if*

$$m \geq C(n\log(m/n) + \log(1/\delta)),$$

then with probability at least $1 - \delta$, m sampling vectors selected uniformly and independently from the unit sphere S^{n-1} form a set such that the following holds: Let $x \in \mathbb{R}^n$ be a *signal vector and let $\boldsymbol{x}_0$ be an initial estimate satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq c\sqrt{\delta_1}\|\boldsymbol{x}_*\|_2$. Then for any $\epsilon > 0$, if*

$$K \geq 2(\log(1/\epsilon) + \log(2/\delta_2))n,$$

then the K-th step randomized Kaczmarz estimate $\boldsymbol{x}_K$ satisfies $\|\boldsymbol{x}_K - \boldsymbol{x}_*\|_2^2 \leq \epsilon\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2$ with probability at least $1 - \delta_1 - \delta_2$.

Proof. Let A be our m by n measurement matrix. By Theorem 3.5.7, there is an angle θ_0, and a constant C such that for $m \geq C(n\log(m/n) + \log(1/\delta))$, A is ACW$(\theta_0, 1/2)$ with probability at least $1 - \delta$.

We can then use Corollary 3.4.4 to guarantee that with probability at least $1 - \delta_1 - \delta_2$, running the randomized Kaczmarz update K times gives an estimate $\mathbf{x}_K$ satisfying

$$\|\mathbf{x}_K - \mathbf{x}_*\|_2^2 \leq \epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2.$$

This completes the proof of the theorem. □

Inspecting the statement of the theorem, we see that we can make the failure probability δ as small as possible by making m large enough. Likewise, we can do the same with δ_2 by adjusting K. Proposition 3.7.1 shows that we can also make δ_2 smaller by increasing m. However, while the dependence of m and K on δ and δ_2 respectively is logarithmic, the dependence of m on δ_1 is *polynomial* (we need $m \gtrsim 1/\delta_1^2$). This is rather unsatisfactory, but can be overcome by a simple ensemble method. We encapsulate this idea in the following algorithm.

Algorithm 1 ENSEMBLE RANDOMIZED KACZMARZ

```
Input: Measurements $b_1, \ldots, b_m$, sampling vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$, relative error tolerance $\epsilon$,
    iteration count $K$, trial count $L$.
Output: An estimate $\hat{\mathbf{x}}_*$ for $\mathbf{x}_*$.
Obtain an initial estimate $\mathbf{x}_0$ using the truncated spectral initialization method (see
    Section 3.7).
for $l = 1, \ldots, L$, run $K$ randomized Kaczmarz update steps starting from $\mathbf{x}_0$ to obtain
    an estimate $\mathbf{x}_K^{(l)}$.
for $l = 1, \ldots, L$, do
    if $|B(\mathbf{x}_K^{(l)}, 2\sqrt{\epsilon}) \cap \{\mathbf{x}_K^{(1)}, \ldots, \mathbf{x}_K^{(L)}\}| \geq L/2$
        return $\hat{\mathbf{x}}_* := \mathbf{x}_K^{(l)}$.
```

Proposition 3.6.2 (Guarantee for ensemble method)**.** *Given the assumptions of Theorem 3.6.1, further assume that $\delta_1 + \delta_2 \leq 1/3$. For any $\delta' > 0$, there is an absolute constant*

C such that if $L \geq C\log(1/\delta')$, then the estimate $\hat{\boldsymbol{x}_*}$ given by Algorithm 1 satisfies $\|\hat{\boldsymbol{x}_*} - \boldsymbol{x}_*\|_2^2 \leq 9\epsilon\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2$ with probability at least $1-\delta'$.

Proof. For $1 \leq l \leq L$, let χ_l be the indicator variable for $\|\mathbf{x}_K^{(l)} - \mathbf{x}_*\|_2^2 \leq \epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2$. Then $\chi_1, \ldots, \chi_L$ are i.i.d. Bernoulli random variables each with success probability at least $2/3$. Let I be the set of indices l for which $\chi_l = 1$. Using a Chernoff bound [113], we see that with probability at least $1 - e^{-cL}$, $|I| \geq L/2$. Now let I' be the set of indices for which $|B(\mathbf{x}_K^{(l)}, 2\epsilon) \cap \{\mathbf{x}_K^{(1)}, \ldots, \mathbf{x}_K^{(L)}\}| \geq L/2$. Observe that for all $l, l' \in I$, we have

$$\|\mathbf{x}_K^{(l)} - \mathbf{x}_K^{(l')}\|_2 \leq \|\mathbf{x}_K^{(l)} - \mathbf{x}_*\|_2 + \|x - \mathbf{x}_K^{(l')}\|_2 \leq 2\sqrt{\epsilon}.$$

This implies that $I \subset I'$, so $I' \neq \emptyset$. Furthermore, for all $l' \in I'$, there is $l \in I$ for which $\|\mathbf{x}_K^{(l)} - \mathbf{x}_K^{(l')}\|_2 \leq 2\sqrt{\epsilon}$. As such, we have

$$\|\mathbf{x}_K^{(l')} - \mathbf{x}_*\|_2 \leq \|\mathbf{x}_K^{(l')} - \mathbf{x}_K^{(l)}\|_2 + \|\mathbf{x}_K^{(l)} - \mathbf{x}_*\|_2 \leq 3\sqrt{\epsilon}.$$

Now, observe that the estimate $\hat{\mathbf{x}_*}$ returned by Algorithm 1 is precisely some $\mathbf{x}_K^{(l')}$ for which $l' \in I'$. This shows that on the good event, we indeed have $\|\hat{\mathbf{x}_*} - \mathbf{x}_*\|_2^2 \leq 9\epsilon\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2$. By our assumption on L, we see that the failure probability is bounded by δ'. □

In practice however, the ensemble method is not required. Numerical experiments show that the randomized Kaczmarz method always eventually converges from any initial estimate.

3.7 Initialization

Several different schemes have been proposed for obtaining initial estimates for *PhaseMax* and gradient descent methods for phase retrieval. Surprisingly, these are all spectral in nature: the initial estimate $\mathbf{x}_0$ is obtained as the leading eigenvector to a matrix that is constructed out of the sampling vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ and their associated measurements $b_1, \ldots, b_m$ [20, 26, 126, 119].

There seems to be empirical evidence, at least for Gaussian measurements, that the best performing method is the *orthogonality-promoting method* of [119]. Nonetheless, for any given relative error tolerance, all the methods seem to require sample complexity of the same order. Hence, we focus on the *truncated spectral method* of [26] for expositional clarity, and refer the reader to the respective papers on the other methods for more details.

The truncated spectral method initializes $\mathbf{x}_0 := \lambda_0\tilde{\mathbf{x}}_0$, where $\lambda_0 = \sqrt{\frac{1}{m}\sum_{i=1}^m b_i^2}$, and $\tilde{\mathbf{x}}_0$

is the leading eigenvector of

$$Y = \frac{1}{m}\sum_{i=1}^{m} b_i^2 \mathbf{a}_i \mathbf{a}_i^T 1(b_i \leq 3\lambda_0).$$

Note that when constructing Y, we sum up only those sampling vectors whose corresponding measurements satisfy $b_i \leq 3\lambda_0$. The point of this is to remove the influence of unduly large measurements, and allow for good concentration estimates, as we shall soon demonstrate.

Suppose from now on that the $\mathbf{a}_i$'s are independent standard Gaussian vectors. In [26], the authors prove that with probability at least $1 - \exp(-\Omega(m))$, we have $\|\tilde{\mathbf{x}}_0 - \mathbf{x}_*\|_2 \leq \epsilon\|\mathbf{x}_*\|_2$ for any fixed relative error tolerance ϵ (see their Proposition 3). They do not, however, examine the dependence of the probability bound on ϵ. Nonetheless, by examining the proof more carefully, we can make this dependence explicit. In doing so, we obtain the following proposition.

Proposition 3.7.1 (Relative error guarantee for initialization). *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$, $b_1, \ldots, b_m$, Y and $\boldsymbol{x}_0$ be defined as in the preceding discussion. Fix $\epsilon > 0$ and $0 < \delta < 1$. Then with probability at least $1-\delta$, we have $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \epsilon\|\boldsymbol{x}_*\|_2$ so long as $m \geq C(\log(1/\delta)+n)/\epsilon^2$.*

Proof. We simply make the following observations while following the proof of Proposition 3 in [26]. First, since all quantities are 2-homogeneous in $\|\mathbf{x}_*\|_2$, we may assume without loss of generality that $\|\mathbf{x}_*\|_2 = 1$. Next, there is some absolute constant c such that if we define $\mathbf{Y}_1$ and $\mathbf{Y}_2$ by choosing $\gamma_1 = 3 + c\epsilon$, $\gamma_2 = 3 - c\epsilon$, we have the bound $\|\mathbb{E}\mathbf{Y}_1 - \mathbb{E}\mathbf{Y}_2\| \leq C\epsilon$. Note also that the deviation estimates $\|\mathbf{Y}_1 - \mathbb{E}\mathbf{Y}_1\|$, $\|\mathbf{Y}_2 - \mathbb{E}\mathbf{Y}_2\|$ are bounded by $C\epsilon$ given our assumptions on m. This implies that with high probability,

$$\|\mathbf{Y} - \beta_1\mathbf{x}_*\mathbf{x}_*^T - \beta_2\mathbf{I}_n\| \leq C\epsilon.$$

Adjust our constants so that C in the last equation is bounded by $\beta_1 - \beta_2$. We may then apply Davis-Kahan [34] to get

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_*\|_2 \leq \frac{\|\mathbf{Y} - \beta_1\mathbf{x}_*\mathbf{x}_*^T - \beta_2\mathbf{I}_n\|}{\beta_1 - \beta_2} \leq \epsilon$$

as we wanted. □

By examining the proof carefully, the astute reader will observe that the crucial properties that we used were the rotational invariance of the $\mathbf{a}_i$'s (to compute the formulas for $\mathbb{E}\mathbf{Y}_1$ and $\mathbb{E}\mathbf{Y}_2$) and their subgaussian tails (to derive the deviation estimates). These properties also hold for sampling vectors that are uniformly distributed on the sphere. As such, a more lengthly and tedious calculation can be done to show that the guarantee also holds for such sampling vectors. If the reader has any residual doubt, perhaps this can be assuaged by noting that a uniform sampling vector and its associated measurement $(\mathbf{a}_i, b_i)$ can be turned into an honest real Gaussian vector by multiplying both quantities by an independent χ^2 random variable with n degrees of freedom.

3.8 Comments and open questions

3.8.1 Arbitrary initialization

In order to obtain a convergence guarantee, we used a truncated spectral initialization to obtain an initial estimate before running randomized Kaczmarz updates. Since the number of steps that we require is only linear in the dimension, and each step requires only linear time, the iteration phase of the algorithm only requires $O(n^2)$ time, and furthermore does not need to see all the data in order to start running.

The spectral initialization on the other hand requires one to see all the data. Forming the matrix from which we obtain the estimate involves adding m rank 1 matrices, and hence naively requires $O(mn^2)$ time. There is hence an incentive to do away with this step altogether, and ask whether the randomized Kaczmarz algorithm works well even if we start from an arbitrary initialization.

We have some numerical evidence that this is indeed true, at least for real Gaussian measurements. Unfortunately, we do not have any theoretical justification for this phenomenon, and it will be interesting to see if any results can be obtained in this direction.

3.8.2 Complex Gaussian measurements

We have proved our main results for measurement systems comprising random vectors drawn independently and uniformly from the sphere, or equivalently, for real Gaussian measurements. These are not the measurement sets that are used in practical applications, which often deal with imaging and hence make use of complex measurements.

While most theoretical guarantees for phase retrieval algorithms are in terms of real Gaussian measurements, some also hold for complex Gaussian measurements, even with identical proofs. This is the case for *PhaseMax* [22] and for Wirtnger flow [20]. We believe that a similar situation should hold for the randomized Kaczmarz method, but are not yet able to recalibrate our tools to handle the complex setting.

It is easy to adapt the randomized Kaczmarz update formula (3.3) itself: we simply replace the sign of $\langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle$ with its phase (i.e. $\frac{\langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle}{|\langle \mathbf{a}_{r(k)}, \mathbf{x}_{k-1} \rangle|}$). Numerical experiments also show that convergence does occur for complex Gaussian measurements (and even CDP measurements) [121]. Nonetheless, in trying to adapt the proof to this situation, we meet an obstacle at the first step: when computing the error term, we can no longer simply sum up the influence of "bad measurements" as we did in Lemma 3.2.1. Instead, *every* term contributes an error that scales with the phase difference

$$\frac{\langle \mathbf{a}_i, \mathbf{z} \rangle}{|\langle \mathbf{a}_i, \mathbf{z} \rangle|} - \frac{\langle \mathbf{a}_i, \mathbf{x}_* \rangle}{|\langle \mathbf{a}_i, \mathbf{x}_* \rangle|}.$$

Since the argument of Jeong and Güntürk also heavily relies on the decomposition of the measurement set into "good" and "bad" measurements, their method likewise does not easily generalize to cover the complex setting. We leave it to future work to prove convergence in this setting, whether by adapting our methods, or by proposing completely new ones.

3.8.3 Deterministic constructions of measurement sets

The theory that we have developed in this chapter does not apply solely to Gaussian measurements, and generalizes to any measurement sets that satisfy the ACW condition that we introduced in Section 3.5. It will be interesting to investigate what natural classes of measurement sets satisfy this condition.

CHAPTER 4

Sparse, Misspecified Phase Retreival

4.1 Introduction

4.1.1 Sparse phase retrieval

Recall that the phase retrieval problem is that of solving a system of quadratic equations

$$|\langle \mathbf{a}_i, \mathbf{x} \rangle^2| = y_i, \qquad i = 1, 2, \ldots, m \tag{4.1}$$

where $\mathbf{a}_i \in \mathbb{R}^n$ (or $\mathbb{C}^n$) are known sampling vectors, $y_i > 0$ are observed measurements, and $\mathbf{x} \in \mathbb{R}^n$ (or $\mathbb{C}^n$) is the decision variable.

In the previous chapter, we saw that there has been much recent theoretical success in the studying this problem. There has also been some work on this problem in the high-dimensional regime. In this setting, it is assumed that the true signal $\mathbf{x}_*$ is s-sparse, and one would like to estimate $\mathbf{x}_*$ accurately with much fewer measurements than the ambient dimension, in analogy with what is possible for sparse linear regression. Work in this direction has mostly comprised straightforward adaptations of algorithms for unconstrained phase retrieval: Both *PhaseLift* and *PhaseMax* have been be adapted by adding l_1 regularizers to their respective objective functions [82, 52]. Meanwhile, the gradient descent schemes *Truncated Wirtinger Flow* and *Truncated Amplitude Flow* have been adapted to alternate gradient steps with either soft- or hard-thresholding [109, 120, 97]. These methods have been mostly shown to accurately recover $\mathbf{x}_*$ with sample complexity $m = \mathrm{O}^*(s^2)$.

4.1.2 Single index models and model agnostic recovery

Phase retrieval is an example of a single index model. In this more general setting, the measurements and the sampling vectors are related by the formula

$$f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle) = y_i, \qquad i = 1, 2, \ldots, m \tag{4.2}$$

where f is a (possibly random) link function. Such models have been studied for some time in the statistics community (see [69] and the references therein). Classically, it is assumed that the link function f is unknown to the observer, and it is of interest to estimate both $\mathbf{x}_*$ (the index parameter) and f. Standard theoretical results in this body of work include asymptotic minimax rates of various estimators.

In this chapter, we take a slightly different approach to the problem. We continue to assume that f is unknown, but now treat $\mathbf{x}_*$ as the only parameter of interest. On the other hand, we are interested in algorithms that are provably efficient from both a statistical as well as a *computational* point of view. Furthermore, we want our algorithms to be able to exploit a sparsity prior and thus work in the high-dimensional regime. The motivation for such an approach comes from the observation that real data almost never obeys a precise algebraic relationship. In other words, the neat relationships we postulate, such as (4.1), are often *misspecified*.

Recently, Plan and Vershynin [87] made significant progress on this problem in the setting of misspecified linear regression. They showed that if $\text{Cov}(g, f(g)) \neq 0$, then the standard *Lasso* algorithm for sparse linear regression is able to estimate $\mathbf{x}_*$ accurately up to scaling, and with a sample complexity of $\text{O}(s \log n)$, the same order as that in the case of no model misspecification. Here, $g \sim \text{N}(0, 1)$ is a standard Gaussian random variable.

In the misspecified phase retrieval (MPR) setting, the first work was done by Neykov, Wang and Liu [81]. They proposed a two stage algorithm that works as follows. First, they form the reweighted sample covariance matrix

$$\hat{\mathbf{\Sigma}} := \frac{1}{m} \sum_{i=1}^{m} y_i (\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I}_n), \tag{4.3}$$

and apply the standard SDP relaxation of *Sparse PCA* to this matrix. Next, they use the leading eigenvector of the solution to formulate a *Lasso*-type program. The solution to this program is their final estimate. The assumptions they make are that

$$\mu = \mu(g, f) := \text{Cov}(g^2, f(g)) > 0, \quad \|f(g)\|_{\psi_1} \leq C, \tag{$A_{f,g}$}$$

under which, they were able to prove that the algorithm recovers x_* accurately when given $m = \text{O}^*(s^2)$ independent standard Gaussian sampling vectors. Again, this is the same order as the guarantees for sparse phase retrieval in the case of no misspecification.

4.1.3 Chapter summary and notes

This chapter is based on the paper “Sparse Phase Retrieval via Sparse PCA Despite Model Misspecification: A Simplified and Extended Analysis” [106]. The goal is twofold. First, we prove that *Sparse PCA*, the first step of the algorithm proposed in [81], suffices to recover the signal vector $\mathbf{x}_*$ accurately with the same sample complexity as the full two-step algorithm given in their paper. We provide a simplified and more flexible analysis that is adapted from [85]. This analysis has the further benefit of generalizing to the case where the prior assumption on $\mathbf{x}_*$ is not that it is sparse, but that it lies in a geometric set $\mathcal{K}$.

Second, we provide a guarantee for *Sparse PCA* to recover $\mathbf{x}_*$ accurately when the sampling vectors are not Gaussian, but are instead drawn from distributions with independent subgaussian entries. In particular, we show that the method works for Rademacher random variables. Although this is a realistic sampling model, to our knowledge, it has not been analyzed in any prior work on phase retrieval. This guarantee requires two conditions. Unsurprisingly, we require the link function f to satisfy a correlation condition similar to $(A_{f,g})$, but adapted to the given subgaussian distribution. Second, we require $\mathbf{x}_*$ to have entries of equal magnitude over its support. This second condition is relatively stringent, but can probably be relaxed in future work.

4.2 Main results

We shall work with the single index model (4.2). We assume that the sampling vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are independent copies of a random vector $\mathbf{a}$ satisfying the following distributional assumption:

Assumption 4.2.1 (Sampling vector distribution)**.** The coordinates of $\mathbf{a}$ are independent copies of a random variable Z that is centered, symmetric, of unit variance, and with subgaussian norm $\|Z\|_{\psi_2}$ bounded by an absolute constant C.

For convenience, we shall hide the dependence on C in our results and in our analysis. We do not assume that we know the link function f. The algorithm we propose to estimate $\mathbf{x}_*$ is the following.

```
Algorithm 2 SPARSE PCA FOR MPR
Input: Measurements $y_1, \dots, y_m$, sampling vectors $\mathbf{a}_1, \dots, \mathbf{a}_m$, sparsity level $s$.
Output: An estimate $\hat{\mathbf{x}}$ for $\mathbf{x}_*$.
1: Compute $\hat{\boldsymbol{\Sigma}}$ as defined in (4.3).
2: Let $\hat{\mathbf{X}}$ be the solution to
       $$\max_{X \succeq 0} \langle \mathbf{X}, \hat{\boldsymbol{\Sigma}} \rangle \quad \text{subject to} \quad \mathrm{Tr}(\mathbf{X}) = 1,\ \|\mathbf{X}\|_1 \leq s. \tag{4.4}$$
3: Let $\hat{\mathbf{x}}$ be the leading eigenvector to $\hat{\mathbf{X}}$.
```

This program is precisely the SDP relaxation of Sparse PCA proposed by d'Aspremont et al. [31] and later analyzed by Amini and Wainwright [2] and Berthet and Rigollet [11]. These two papers analyzed the performance of the algorithm as applied to sparse principal component detection in the spiked covariance model. Since the matrix $\hat{\boldsymbol{\Sigma}}$ does not follow this model a priori, one requires further analysis to show that the algorithm succeeds.

In [81], the authors propose using the Lagrangian version of this program as the first step of their algorithm. Their analysis (see Lemma C.1. therein) shows that when the sampling vectors follow a Gaussian distribution, one can obtain a constant error approximation to $\mathbf{x}_*$ using $O(s^2 \log n)$ samples. Using our methods, we prove a stronger version of this guarantee.

Theorem 4.2.2 (Sparse recovery for Gaussian measurements)**.** *Suppose $\boldsymbol{a}$ is a standard Gaussian in $\mathbb{R}^n$, and suppose Assumption $(A_{f,g})$ holds. Then there is an absolute constant C such that for any s-sparse, unit norm signal $\boldsymbol{x}_*$ and any $\epsilon, \delta > 0$, the output $\hat{\boldsymbol{x}}$ to Algorithm 2 satisfies $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_*\|_2 \leq \epsilon$ with probability at least $1 - \delta$ so long as the sample size m satisfies*

$$m \geq C \max\left\{ \frac{s^2\big(\log(n/\delta) + \log^4(s/\delta)\big)}{\mu(f,g)^2 \epsilon^4}, \frac{s}{\delta}, \frac{\log(n/\delta)}{\log^2 m} \right\}.$$

Although this result is not entirely novel, we prove it in a different way compared to [81]. This method is simple and amenable to generalization to the situation where the sampling vectors are not Gaussian. In the non-Gaussian case, we first fix the sparsity parameter s. Let $\bar{Z}_s := \frac{1}{s}\sum_{i=1}^s Z_i$ and $\mathbf{r}_{s,Z} := (Z_1 - \bar{Z}_s, \dots, Z_s - \bar{Z}_s)$ denote the mean of s independent copies of Z and the vector of residuals respectively. With this notation, we make the assumption:

$$\begin{aligned} \mu(f,Z,s) &:= \mathrm{Cov}((\sqrt{s}\bar{Z}_s)^2, f(\sqrt{s}\bar{Z}_s)) > 0, \\ \sigma(f,Z,s) &:= \mathrm{Cov}(\|\mathbf{r}_{s,Z}\|_2^2, f(\sqrt{s}\bar{Z}_s)) \leq 0, \quad \|f(\sqrt{s}\bar{Z}_s)\|_{\psi_1} \leq C. \end{aligned} \tag{$A_{f,Z,s}$}$$

Furthermore, we say that a unit norm signal vector $\mathbf{x}_*$ is *admissible* if it has entries of equal magnitude across its support. In other words, there is a index set $I \subset [n]$ of cardinality $|I| \leq s$, such that

$$(\mathbf{x}_*)_j = \begin{cases} \pm\frac{1}{\sqrt{|I|}} & j \in I \\ 0 & \text{otherwise.} \end{cases}$$

Using this definition, we have the following analogue of Theorem 4.2.2.

Theorem 4.2.3 (Sparse recovery for non-Gaussian measurements)**.** *There is an absolute constant C such that the following holds. Fix a sparsity parameter s, suppose $\boldsymbol{x}_*$ is admissible and suppose Assumption $(A_{f,Z,s})$ holds. Then for any $\epsilon, \delta > 0$, the output $\hat{\boldsymbol{x}}$ to Algorithm 2 satisfies $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_*\|_2 \leq \epsilon$ with probability at least $1-\delta$ so long as the sample size m satisfies*

$$m \geq \frac{Cs^2\big(\log(n/\delta) + \log^4(s/\delta)\big)}{\mu(f,Z,s)^2\epsilon^4} + \frac{Cs}{\delta} + \frac{C\log(n/\delta)}{\log^2 m}. \tag{4.5}$$

Note that when Z is standard Gaussian, Assumption $(A_{f,Z,s})$ reduces to Assumption $(A_{f,g})$. To see this, observe that for any s, $\sqrt{s}\bar{Z}$ is a standard Gaussian random variable, while $\sigma(f,g,s) = 0$ by the independence property of orthogonal Gaussian marginals. This fact points to the assumption being the right generalization of $(A_{f,g})$.

Furthermore, it is intuitive that the second condition should hold whenever Z has a reasonable distribution and when $\mu(f,Z,s) > 0$: if f is *positively* correlated with the magnitude of $\bar{Z}$, then it should be *negatively* correlated with the norm of the residual vector. Indeed, this can be shown to be true whenever Z is a Rademacher random variable. We thus have a simpler result for Rademacher random variables:

Corollary 4.2.4 (Sparse recovery for Rademacher measurements)**.** *There is an absolute constant C such that the following holds. Fix a sparsity parameter s, suppose $\boldsymbol{x}_*$ is admissible, let Z denote a Rademacher random variable. Suppose $\mu(f,Z,s) > 0$ and $\|f(\sqrt{s}\bar{Z}_s)\|_{\psi_1} \leq C$. Then for any $\epsilon, \delta > 0$, the output $\hat{\boldsymbol{x}}$ to Algorithm 2 satisfies $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_*\|_2 \leq \epsilon$ with probability at least $1-\delta$ so long as the sample size m satisfies* (4.5).

In the Gaussian setting, the recovery guarantee continues to hold even if we relax our constraint on $\mathbf{x}_*$ slightly and instead assume that $\|\mathbf{x}_*\|_1 \leq \sqrt{s}$. This condition is *geometric*: it can equivalently expressed as $\mathbf{x}_* \in \mathcal{K}$, where $\mathcal{K} = \sqrt{s}B_1^n$ is the l_1 norm ball. It is thus an interesting theoretical question to ask whether one can construct efficient algorithms for estimating $\mathbf{x}_*$ that exploit prior knowledge that $\mathbf{x}_* \in \mathcal{K}$ for a *general* convex set $\mathcal{K}$.

There has been some work on proving *statistical* efficiency guarantees for various algorithms. In the misspecified linear regression setting, Plan and Vershynin showed that the $\mathcal{K}$*-Lasso* succeeds whenever the number of measurements m is of the order $w(\mathcal{K})^2$, where $w(\mathcal{K})$ denotes the Gaussian width of $\mathcal{K}$ [87]. In the phase retrieval setting, Soltanolkotabi showed that Projected Amplitude Flow also succeeds whenever $m \gtrsim w(\mathcal{K})^2$. On the other hand, it is hard to remark on the *computational* efficiency of these methods, because this depends on the properties of the set $\mathcal{K}$.

The final main result of this chapter is a guarantee of a similar spirit.

Theorem 4.2.5 (Recovery using general geometric constraints)**.** *Suppose $\boldsymbol{x}_*\boldsymbol{x}_*^T \in \mathcal{K}$, where $\mathcal{K}$ is a convex subset of the space of unit trace PSD matrices in $\mathbb{R}^{n\times n}$. Suppose $\boldsymbol{a}$ is a standard Gaussian in $\mathbb{R}^n$, and suppose Assumption $A_{f,g}$ holds. Then for any $\epsilon, \delta > 0$, the output $\hat{\boldsymbol{x}}$ to Algorithm 3 satisfies $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_*\|_2 \leq \epsilon$ with probability at least $1-\delta$ so long as the sample size m satisfies*

$$m \geq \frac{C\big(w(\mathcal{K})^2 + \log^4(1/\delta) + \log m(\gamma_1(\mathcal{K}, \|\cdot\|) + \log(1/\delta))\big)}{\mu(f,g)^2\epsilon^4} + \frac{C}{\delta}. \tag{4.6}$$

Here, $w(\mathcal{K})$ and $\gamma_1(\mathcal{K}, \|\cdot\|)$ respectively denote the Gaussian width of $\mathcal{K}$ and its γ_1-functional with respect to the operator norm, while C is a universal constant.

Organization of chapter and outline of proof strategy

We prove Theorem 4.2.2 and Theorem 4.2.3 in Section 4.3. The strategy we take comprises two steps. First, we compute the expected objective function used in Algorithm 2, and show that it has sufficient curvature on the feasible set around the ground truth matrix, $\mathbf{x}_*\mathbf{x}_*^T$. This shows that feasible solutions having large expected objective value must also be close to $\mathbf{x}_*\mathbf{x}_*^T$. This computation is done in Section 4.4.

Next, we argue that the empirical objective function is uniformly close to the expected objective function with high probability, so that a solution to the SDP program actually has large expected objective value. This is proved in Section 4.5. Finally, we use the same strategy for Theorem 4.2.5, but replace the objective function concentration analysis with a more sophisticated chaining argument. Due to its more technical nature, we defer the details to Section 4.7.

4.3 Proof of results for sparse recovery

Proof of Theorem 4.2.2. Let $\mathbf{X}$ be the solution to Algorithm 2. Since $\mathbf{x}_*\mathbf{x}_*^T$ is also feasible for the program, we have by optimality that

$$0 \leq \langle \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T, \hat{\mathbf{\Sigma}} \rangle = \langle \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T, \mathbf{\Sigma} \rangle + \langle \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T, \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \rangle. \tag{4.7}$$

Using Lemma 4.4.3, the first term satisfies the bound

$$\langle \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T, \mathbf{\Sigma} \rangle \leq -\frac{\mu(f,g)}{2} \|\mathbf{x}_*\mathbf{x}_*^T - \mathbf{X}\|_F^2.$$

For the second term, we use Hölder's inequality to write

$$\langle \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T, \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \rangle \leq \left\| \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T \right\|_1 \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty.$$

Next, we have by assumption that

$$\left\| \mathbf{x}_*\mathbf{x}_*^T \right\|_1 = \sum_{i,j=1}^m |(\mathbf{x}_*)_i (\mathbf{x}_*)_j| = \|\mathbf{x}_*\|_1^2 \leq s.$$

Meanwhile, by construction, we also know that $\|\mathbf{X}\|_1 \leq s$. Rearranging (4.7), we therefore get

$$\frac{\mu(f,g)}{2} \|\mathbf{x}_*\mathbf{x}_*^T - \mathbf{X}\|_F^2 \leq 2s \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty.$$

Using Proposition 4.5.1 to bound the right hand side, we get

$$\|\mathbf{x}_*\mathbf{x}_*^T - \mathbf{X}\|_F^2 \leq \frac{Cs\Big(\sqrt{\log(n/\delta)} + \log^2(s/\delta)\Big)}{\mu(f,g)\sqrt{m}}.$$

If $\hat{\mathbf{x}}$ denotes the leading eigenvector of $\mathbf{X}$, we use Davis-Kahan's eigenvector perturbation theorem [34] to conclude that $\|\hat{\mathbf{x}} - \mathbf{x}_*\|_2^2$ satisfies the same bound. Finally, we plug in our assumption on m to show that this bound is less than ϵ^2. □

Proof of Theorem 4.2.3. Exactly the same as for the Theorem 4.2.2. □

Proof of Corollary 4.2.4. Observe that $\|\mathbf{r}_{s,Z}\|_2^2 + (\sqrt{s}\bar{Z}_s)^2 = \|\mathbf{a}\|_2^2 = s$. We have

$$\begin{aligned}\sigma(f, Z, s) &= \text{Cov}(f(\sqrt{s}\bar{Z}_s), \|\mathbf{r}_{s,Z}\|_2^2) \\ &= \text{Cov}(f(\sqrt{s}\bar{Z}_s), s - (\sqrt{s}\bar{Z}_s)^2) \\ &= -\text{Cov}(f(\sqrt{s}\bar{Z}_s), (\sqrt{s}\bar{Z}_s)^2) = \mu(f, Z, s).\end{aligned}$$

The corollary now follows from Theorem 4.2.3. □

4.4 Objective function in expectation

In this section, we compute expressions for the expected reweighted covariance matrix $\mathbf{\Sigma} = \mathbb{E}\hat{\mathbf{\Sigma}}$. Note that we may also write

$$\mathbf{\Sigma} = \mathbb{E}y(\mathbf{a}\mathbf{a}^T - \mathbf{I}_n).$$

Lemma 4.4.1 (Expected covariance for Gaussian distribution)**.** *Suppose* $\boldsymbol{a} \sim \mathrm{N}(0, \boldsymbol{I}_n)$. *Then for any* $\boldsymbol{x}_* \in S^{n-1}$*, we have*

$$\mathbf{\Sigma} = \mu(f, g)\boldsymbol{x}_*\boldsymbol{x}_*^T.$$

Proof. Decompose $\mathbf{a} = \langle \mathbf{a}, \mathbf{x}_* \rangle \mathbf{x}_* + \mathbf{a}^\perp$, where $\mathbf{a}^\perp$ is the projection of $\mathbf{a}$ to the orthogonal complement of $\mathbf{x}_*$. Using this, we write

$$\begin{aligned}\mathbb{E}\{y\mathbf{a}\mathbf{a}^T\} &= \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)(\langle \mathbf{a}, \mathbf{x}_* \rangle \mathbf{x}_* + \mathbf{a}^\perp)(\langle \mathbf{a}, \mathbf{x}_* \rangle \mathbf{x}_* + \mathbf{a}^\perp)^T\} \\ &= \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\langle \mathbf{a}, \mathbf{x}_* \rangle^2 \mathbf{x}_*\mathbf{x}_*^T\} + \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\mathbf{a}^\perp(\mathbf{a}^\perp)^T\} \\ &\quad + \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\langle \mathbf{a}, \mathbf{x}_* \rangle \mathbf{x}_*(\mathbf{a}^\perp)^T\} + \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\langle \mathbf{a}, \mathbf{x}_* \rangle \mathbf{a}^\perp \mathbf{x}_*^T\}.\end{aligned}$$

Because $\mathbf{a}$ is a standard Gaussian, $\langle \mathbf{a}, \mathbf{x}_* \rangle$ and $\mathbf{a}^\perp$ are independent. This means that the third and fourth terms in this sum are zero. Furthermore, the second term can be written as the product of two expectations $\mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\}$ and $\mathbb{E}\{\mathbf{a}^\perp(\mathbf{a}^\perp)^T\}$. We now use standard computations for Gaussians to continue writing

$$\begin{aligned}\mathbb{E}\{y(\mathbf{a}\mathbf{a}^T - \mathbf{I}_n)\} &= \mathbb{E}\{f(g)g^2\}x_*x_*^T + \mathbb{E}\{f(g)\}(\mathbf{I}_n - x_*x_*^T) + \mathbb{E}\{f(g)\}\mathbf{I}_n \\ &= \mu(f, g)\mathbf{x}_*\mathbf{x}_*^T.\end{aligned}$$

This completes the proof. □

Lemma 4.4.2 (Expected covariance for non-Gaussian distributions)**.** *Suppose* $\boldsymbol{a}$ *is a random vector in* $\mathbb{R}^n$ *that satisfies Assumption 4.2.1. Let* $2 \leq s \leq n$ *be an integer, and let* $\boldsymbol{x}_*$ *be an admissible signal vector. We have*

$$\boldsymbol{\Sigma} = \mu(f, Z, s)\boldsymbol{x}_*\boldsymbol{x}_*^T + \frac{\sigma(f, Z, s)}{s-1}(\boldsymbol{P}_I - \boldsymbol{x}_*\boldsymbol{x}_*^T).$$

Proof. Let $\mathbf{P}_I$ and $\mathbf{P}_I^\perp$ denote the orthogonal projections to the coordinates in I and I^c respectively. Then $\mathbf{P}_I\mathbf{a}$ and $\mathbf{P}_I^\perp\mathbf{a}$ are independent. Using a similar calculation as in the previous lemma, we see that $\mathbf{P}_I\Sigma\mathbf{P}_I^\perp = \mathbf{P}_I^\perp\Sigma\mathbf{P}_I = \mathbf{P}_I^\perp\Sigma\mathbf{P}_I^\perp = 0$. We may hence assume WLOG that $s = n$ and $I = [n]$. By the symmetry of the distribution of $\mathbf{a}$, we may also assume that $\mathbf{x}_* = \frac{\mathbf{1}}{\sqrt{n}}$, where $\mathbf{1}$ is the all ones vector.

Next, notice that $\langle \mathbf{a}, \mathbf{x}_* \rangle$ is invariant to permutation of the coordinate indices. Meanwhile, the *distributions* of $\mathbf{a}^\perp$ and $\mathbf{a}$ are both symmetric with respect to such transformations. Let $\mathbf{Q}$ be a permutation matrix. Then

$$\begin{aligned}
\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T &= \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\mathbf{Q}\mathbf{a}(\mathbf{Q}\mathbf{a})^T\} \\
&= \mathbb{E}\{f(\langle \mathbf{Q}\mathbf{a}, \mathbf{x}_* \rangle)\mathbf{Q}\mathbf{a}(\mathbf{Q}\mathbf{a})^T\} \\
&= \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\mathbf{a}(\mathbf{a})^T\}.
\end{aligned}$$

In other words, we have

$$\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T = \boldsymbol{\Sigma}. \tag{4.8}$$

One can check that a matrix satisfying (4.8) for all permutation matrices $\mathbf{Q}$ must have the same value for all diagonal entries, and the same value for all off-diagonal entries. In other words, $\boldsymbol{\Sigma}$ must be of the form

$$\boldsymbol{\Sigma} = \frac{\alpha}{n}\mathbf{1}\mathbf{1}^T + \beta(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \tag{4.9}$$

for some values of α and β.

Let us now compute the values of α and β using the fact that $\mathbf{x}_* = \frac{\mathbf{1}}{\sqrt{n}}$. We have

$$\alpha = \mathbf{x}^T\boldsymbol{\Sigma}\mathbf{x}_* = \mu(f, Z, n).$$

Next, we apply traces to (4.9) to get

$$\alpha + (n-1)\beta = \mathrm{Tr}(\boldsymbol{\Sigma}) = \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_* \rangle)\mathrm{Tr}(\mathbf{a}\mathbf{a}^T - \mathbf{I}_n)\}.$$

Observe further that

$$\mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_*\rangle)\mathrm{Tr}(\mathbf{a}\mathbf{a}^T - \mathbf{I}_n)\} = \mathbb{E}\{f(\langle \mathbf{a}, \mathbf{x}_*\rangle)(\|\mathbf{a}\|_2^2 - n)\} = \sigma(f, Z, n) + \mu(f, Z, n).$$

As such, we have $\beta = \frac{\sigma(f,Z,n)}{n-1}$ as we wanted. □

Lemma 4.4.3 (Curvature of objective function)**.** *Suppose the hypotheses of Lemma 4.4.1 (respectively Lemma 4.4.2) hold. For any* $\boldsymbol{X} \succeq 0$ *such that* $\mathrm{Tr}(\boldsymbol{X}) = 1$*, we have*

$$\langle \boldsymbol{\Sigma}, \boldsymbol{x}_*\boldsymbol{x}_*^T - \boldsymbol{X}\rangle \geq \frac{\mu}{2}\|\boldsymbol{x}_*\boldsymbol{x}_*^T - \boldsymbol{X}\|_F^2, \tag{4.10}$$

where $\mu = \mu(f, g)$ *(respectively* $\mu = \mu(f, Z, s)$*).*

Proof. We shall prove the case where the hypotheses of Lemma 4.4.2 hold. The other case is similar and even easier. First, observe that

$$\langle \boldsymbol{\Sigma}, \mathbf{x}_*\mathbf{x}_*^T\rangle = \mu(f, Z, s)\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{x}_*\mathbf{x}_*^T\rangle = \mu(f, Z, s)\|\mathbf{x}_*\|_2^4.$$

We also have

$$\begin{aligned}
\langle \boldsymbol{\Sigma}, \mathbf{X}\rangle &= \mu(f, Z, s)\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle + \frac{\sigma(f, Z, s)}{s-1}\langle \mathbf{P}_I - \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle \\
&= \mu(f, Z, s)\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle + \frac{\sigma(f, Z, s)}{s-1}\mathrm{Tr}(\mathbf{X}\big|_{\mathbb{R}^I \cap \mathbf{x}_*^\perp}) \\
&\leq \mu(f, Z, s)\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle.
\end{aligned}$$

Here, the last inequality follows from the fact that $\mathbf{X}$ is positive semidefinite, which implies that any partial trace has to be non-negative.

Now, the assumptions on $\mathbf{X}$ also imply that $\|\mathbf{X}\|_F \leq 1 = \|\mathbf{x}_*\|_2$. We can thus combine our calculations to get

$$\begin{aligned}
\langle \boldsymbol{\Sigma}, \mathbf{x}_*\mathbf{x}_*^T - \mathbf{X}\rangle &\geq \mu(f, Z, s)\|\mathbf{x}_*\|_2^4 - \mu(f, Z, s)\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle \\
&\geq \frac{\mu(f, Z, s)}{2}\big(\|\mathbf{x}_*\mathbf{x}_*^T\|_F + \|\mathbf{X}\|_F^2 - 2\langle \mathbf{x}_*\mathbf{x}_*^T, \mathbf{X}\rangle\big) \\
&= \frac{\mu(f, Z, s)}{2}\|\mathbf{x}_*\mathbf{x}_*^T - \mathbf{X}\|_F^2.
\end{aligned}$$

This completes the proof. □

4.5 Concentration of objective function

The goal of this section is to prove the following concentration theorem for the reweighted sample covariance matrix $\hat{\mathbf{\Sigma}}$.

Proposition 4.5.1 (Concentration of sample matrix)**.** *There is a universal constant C so that the following holds. Fix a sparsity parameter s, let **a** be defined using Assumption 4.2.1. Suppose Assumption ($A_{f,Z,s}$) holds, and let $\boldsymbol{x}_*$ be a unit norm vector. If **a** is non-Gaussian, further assume that $\boldsymbol{x}_*$ is admissible. Then for any $\delta > 0$, we have*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \frac{C(\sqrt{\log(n/\delta)} + \log^2(s/\delta))}{\sqrt{m}}$$

with probability at least $1 - \delta$, provided $m \geq C \max\{s/\delta, \log(n/\delta)\log^2 m\}$.

Proof. Without loss of generality, assume that the support of $\mathbf{x}_*$ is contained in the first s coordinates. Let $\mathbf{P}_s$ denote the projection to the first s coordinates. We write

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty = \max\Big\{\|\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s\|_\infty, \|\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^\perp\|_\infty, \|\mathbf{P}_s^\perp(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^\perp\|_\infty\Big\}, \quad (4.11)$$

and bound each of the terms on the right separately.

For the first term, we shall use the fact that each entry is the mean of m i.i.d. $\psi_{1/2}$ random variables (see Section 2.2). This tail decay gives us a relatively strong large deviation inequality, which we can use together with a union bound. In more detail, let $1 \leq k, l \leq s$. Then

$$(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})_{kl} = \frac{1}{m}\sum_{i=1}^m \Big[(\mathbf{a}_i)_k(\mathbf{a}_i)_l f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle) - \mathbb{E}\{(\mathbf{a}_i)_k(\mathbf{a}_i)_l f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle)\}\Big].$$

We now use Proposition 2.2.6 followed by Proposition 2.2.5 twice to get

$$\begin{aligned}\|(\mathbf{a})_k(\mathbf{a})_l f(\langle \mathbf{a}, \mathbf{x}_*\rangle) - \mathbb{E}\{(\mathbf{a})_k(\mathbf{a})_l f(\langle \mathbf{a}, \mathbf{x}_*\rangle)\}\|_{\psi_\alpha} &\lesssim \|(\mathbf{a})_k(\mathbf{a})_l f(\langle \mathbf{a}, \mathbf{x}_*\rangle)\|_{\psi_{1/2}} \\ &\lesssim \|(\mathbf{a})_k(\mathbf{a})_l\|_{\psi_1}\|f(\langle \mathbf{a}, \mathbf{x}_*\rangle)\|_{\psi_1} \\ &\lesssim \|(\mathbf{a})_k\|_{\psi_2}\|(\mathbf{a})_l\|_{\psi_2}\|f(\langle \mathbf{a}, \mathbf{x}_*\rangle)\|_{\psi_1}.\end{aligned}$$

Each of the terms in the product on the right hand side is bounded by an absolute constant by assumption. As such, the quantity on the left is also bounded by an absolute constant. We may thus use Proposition 2.2.9 to see that

$$\mathbb{P}\{|(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})_{kl}| > t/\sqrt{m}\} \leq 2\exp(-c\sqrt{t})$$

for $t > 0$ large enough. Pick $t \sim \log^2(s/\delta)$. Then we can take a union bound over all s^2 choices of k and l to get

$$\|\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s\|_\infty \lesssim \frac{\log^2(s/\delta)}{\sqrt{m}}$$

with probability at least $1 - \delta/4$.

We next bound the other two quantities in (4.11) via a conditioning argument similar to that in [81]. The key idea is to condition on the probability $1 - \delta/4$ event over which the three statements in Lemma 4.5.3 hold, and to observe that this event is *independent* of the random variables $(\mathbf{a}_i)_k$ for $1 \leq i \leq m$, $s < k \leq n$. Hence, conditioning on the event does not alter the joint distribution of this set of random variables.

We consider a typical entry in $\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^\perp$, which is of the form

$$\frac{1}{m}\sum_{i=1}^m (\mathbf{a}_i)_k(\mathbf{a}_i)_l f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle), \quad 1 \leq k \leq s,\ s < l \leq n. \tag{4.12}$$

Fixing all randomness apart from $(\mathbf{a}_i)_l$ for all indices $1 \leq i \leq m$, $s < l \leq n$, we can use Hoeffding's inequality (Proposition 2.2.2) to conclude that for each l, (4.12) is a subgaussian random variable with variance $\frac{1}{m^2}\sum_{i=1}^m (\mathbf{a}_i)_k^2 f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle)^2$. By the second statement of Lemma 4.5.3, this is bounded by C/m, so that

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^m (\mathbf{a}_i)_k(\mathbf{a}_i)_l f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle)\right| > \frac{t}{\sqrt{m}}\right\} \leq 2\exp(-ct^2) \tag{4.13}$$

Choosing $t \sim \sqrt{\log(n/\delta)}$ and taking a union bound over $s < l \leq n$ gives

$$\|\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^\perp\|_\infty \lesssim \sqrt{\frac{\log(n/\delta)}{m}}$$

with probability at least $1 - \delta/4$.

Finally, each entry of $\mathbf{P}_s^\perp(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^\perp$ is of the form

$$\frac{1}{m}\sum_{i=1}^m f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle)\big[(\mathbf{a}_i)_k(\mathbf{a}_i)_l - \mathbb{E}\{(\mathbf{a}_i)_k(\mathbf{a}_i)_l\}\big], \quad s < k, l \leq n. \tag{4.14}$$

We again fix all randomness apart from $(\mathbf{a}_i)_l$ for all indices $1 \leq i \leq m$, $s < l \leq n$. Observe that $(\mathbf{a}_i)_k(\mathbf{a}_i)_l - \mathbb{E}\{(\mathbf{a}_i)_k(\mathbf{a}_i)_l\}$, $s < k, l \leq n$, are centered subexponential random variables. We may thus use Bernstein's inequality (Proposition 2.2.3) together with the second and third statements of Lemma 4.5.3 to obtain the tail bound:

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle)\big[(\mathbf{a}_i)_k(\mathbf{a}_i)_l - \mathbb{E}\{(\mathbf{a}_i)_k(\mathbf{a}_i)_l\}\big]\right| > \frac{t}{\sqrt{m}}\right\} \leq 2e^{-c\min\left\{t^2, \frac{t\sqrt{m}}{\log m}\right\}} \tag{4.15}$$

Once again, choosing $t \sim \sqrt{\log(n/\delta)}$ and taking a union bound over $s < k, l \leq n$ gives

$$\|\mathbf{P}_s(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_s^{\perp}\|_\infty \lesssim \sqrt{\frac{\log(n/\delta)}{m}},$$

with probability at least $1 - \delta/4$ provided that $m \gtrsim \log(n/\delta)\log^2 m$. □

Remark 4.5.2. When **a** is a standard Gaussian, [81] gave the bound

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \frac{C\sqrt{\log(n/\delta)}}{\sqrt{m}}$$

with roughly the same tail probability. Hence, the only price to having more distributional generality is the additional $\log^2(s/\delta)$ term in the numerator.

Lemma 4.5.3. *Let the hypotheses of Proposition 4.5.1 hold. There is an absolute constant C such that the following holds. Let I denote the support of $\boldsymbol{x}_*$. Then for any $\delta > 0$, so long as $m \geq Cs/\delta$, the following three statements hold simultaneously with probability at least $1 - \delta/4$.*

1. $\sum_{i=1}^{m} f(\langle \boldsymbol{a}_i, \boldsymbol{x}_*\rangle)^2 \leq Cm.$

2. $\max_{k\in I}\sum_{i=1}^{m} (\boldsymbol{a}_i)_k^2 f(\langle \boldsymbol{a}_i, \boldsymbol{x}_*\rangle)^2 \leq Cm.$

3. $\max_{1\leq i\leq m} f(\langle \boldsymbol{a}_i, \boldsymbol{x}_*\rangle) \leq C\log m.$

Proof. By Assumption ($A_{f,Z,s}$), we know that $\|f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle)\|_{\psi_1}$ is bounded by an absolute constant. As such, Proposition 2.2.4 implies that both its second and fourth moments are also bounded. Furthermore, we have

$$\mathrm{Var}(f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle)^2) \leq \mathbb{E}\{f(\langle \mathbf{a}_i, \mathbf{x}_*\rangle)^4\} \leq C,$$

where C is an absolute constant. Using Chebyshev's inequality together with the second moment bound, we thus get

$$\mathbb{P}\left\{\sum_{i=1}^{m} f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle)^2 \geq m(C+t)\right\} \leq \frac{C}{mt^2}. \tag{4.16}$$

We can use the same argument together with a union bound over $k \in I$ to get

$$\mathbb{P}\left\{\max_{k \in I} \sum_{i=1}^{m} (\mathbf{a}_i)_k^2 f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle)^2 \geq m(C+t)\right\} \leq \frac{Cs}{mt^2}. \tag{4.17}$$

Finally, we again use the union bound and the subexponential tail bound to get

$$\mathbb{P}\left\{\max_{1 \leq i \leq m} f(\langle \mathbf{a}_i, \mathbf{x}_* \rangle) \geq t \log m\right\} \leq 2m \exp(-ct \log m) = 2m^{1-ct}. \tag{4.18}$$

Choose t to be any fixed constant in (4.16) and (4.17), and choose t to be a constant larger than $2/c$ in (4.5). Then each of these probability bounds is of the order $O(1/m)$, so that $m \gtrsim s/\delta$ suffices for all three statements to hold with probability at least $1 - \delta/4$. □

4.6 Comments and open questions

In this chapter, we have analyzed the problem of misspecified phase retrieval, and improved upon the work of Neykov et al. in [81]. In particular, we have shown that the first stage of their algorithm suffices for signal recovery with the same sample complexity, and extended the analysis to non-Gaussian measurements. Furthermore, we showed how the algorithm can be generalized to recover a signal vector $\mathbf{x}_*$ efficiently given geometric prior information other than sparsity.

Experts in compressed sensing may have observed that while the sample complexity for algorithms for misspecified linear regression scales linearly with the sparsity parameter, our sample complexity bounds for misspecified phase retrieval scale instead with the square of the parameter. In [81], the authors showed numerical evidence that this discrepancy is due to the statistical inefficiency of the algorithm, and not merely a slackness in the mathematical analysis.

This s^2 scaling is also observed in all other efficient algorithms for sparse phase retrieval, and it is an open question whether there exist computationally efficient algorithms that can do better. The authors of [81] conjecture that the answer is in the negative. This is supported by results by Berthet and Rigollet, who show that computationally efficient algorithms for the related problem of detecting sparse principal components, using $O(s^{2-\epsilon})$

samples for any $\epsilon > 0$, will lead to computationally efficient algorithms for solving hard instances of the planted clique problem [11, 10]. This is widely conjectured to be impossible.

It will also be interesting to investigate whether there is slackness in the sample complexity bound for signal recovery using general geometric constraints (Theorem 4.2.5). In particular, I do not know how to bound $\gamma_1(\mathcal{K}, \|\cdot\|)$ where $\mathcal{K}$ is the set of unit trace PSD matrices $\mathbf{X}$ with $\|\mathbf{X}\|_1 \leq s$. Hence, it is not yet clear whether Theorem 4.2.2 can be derived from Theorem 4.2.5.

Finally, the literature on high-dimensional signal recovery from non-Gaussian measurements is still fairly limited. In this work, we have proved a recovery guarantee for *admissible* signal vectors in the case of misspecified phase retrieval. Hopefully, this guarantee can be extended to larger classes of signal vectors in the near future.

4.7 Recovery using general geometric signal constraints

The goal of this section is to prove Theorem 4.2.5, and to collate the necessary theoretical apparatus for doing so. First, we state the algorithm we propose for estimating $\mathbf{x}_*$ given general geometric constraints. We call this algorithm $\mathcal{K}$-PCA.

```
Algorithm 3 $\mathcal{K}$-PCA FOR MPR
Input: Measurements $y_1, \dots, y_m$, sampling vectors $\mathbf{a}_1, \dots, \mathbf{a}_m$.
Output: An estimate $\hat{\mathbf{x}}$ for $\mathbf{x}_*$.
1: Compute $\hat{\mathbf{\Sigma}}$ as defined in (4.3).
2: Let $\hat{\mathbf{X}}$ be the solution to
       $$\max_{X \succeq 0} \langle \mathbf{X}, \hat{\mathbf{\Sigma}} \rangle \quad \text{subject to} \quad \mathbf{X} \in \mathcal{K}. \qquad (4.19)$$
3: Let $\hat{\mathbf{x}}$ be the leading eigenvector to $\hat{\mathbf{X}}$.
```

This can be seen as a tensorized version of the 1-bit sensing algorithm proposed in [85]. Our analysis will be also be similar to that in [85], but we will require a more general concentration result (Lemma 4.7.2) that is derived via chaining. In particular, we will make use of the bound on the suprema of mixed tail processes that we used in the previous chapter.

We have a subset $\mathcal{K}$ of unit trace PSD matrices in $\mathbb{R}^{n \times n}$. Fix $\mathbf{x}_*\mathbf{x}_*^T$, and real numbers $z_1, \dots, z_m$. We define a process on the set $\mathcal{K}$ as follows. Let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be standard Gaussians. For each $\mathbf{X} \in \mathcal{K}$, we set

$$Y_{\mathbf{X}} = \langle \textstyle\sum_{i=1}^{m} z_i(\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I}_n), \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T \rangle.$$

We claim the following.

Lemma 4.7.1 (Process increments)**.** *The process $Y_{\boldsymbol{X}}$ has mixed tail increments with respect to (d_1, d_2), where $d_2(\boldsymbol{X}, \boldsymbol{X}') = (\sum_{i=1}^{m} z_i^2)^{1/2}\|\boldsymbol{X} - \boldsymbol{X}'\|_F$, and $d_1(\boldsymbol{X}, \boldsymbol{X}') = \max_i|z_i| \cdot \|\boldsymbol{X} - \boldsymbol{X}'\|$.*

Proof. Fix $\mathbf{X}, \mathbf{X}' \in \mathcal{K}$, and for convenience, denote $\mathbf{H} = \mathbf{X} - \mathbf{X}'$. Then

$$\begin{aligned} Y_{\mathbf{X}} - Y_{\mathbf{X}'} &= \sum_{i=1}^{m} z_i\big(\mathbf{a}_i^T\mathbf{H}\mathbf{a}_i - \mathbb{E}\{\mathbf{a}_i^T\mathbf{H}\mathbf{a}_i\}\big) \\ &= \sum_{i=1}^{m} z_i \sum_{j=1}^{n} \lambda_j\big(\langle \mathbf{a}_i, \mathbf{v}_j\rangle^2 - \mathbb{E}\{\langle \mathbf{a}_i, \mathbf{v}_j\rangle^2\}\big) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} z_i\lambda_j\big(\langle \mathbf{a}_i, \mathbf{v}_j\rangle^2 - 1\big), \end{aligned}$$

where $\mathbf{H} = \sum_{i=1}^{m} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ is the eigendecomposition of $\mathbf{H}$. Next, observe that by the independence of orthogonal Gaussian marginals, $\{(\langle \mathbf{a}_i, \mathbf{v}_j\rangle^2 - 1) : 1 \leq i \leq m,\ 1 \leq j \leq n\}$ are independent, centered subexponential random variables with bounded subexponential norm. We may thus apply Bernstein's inequality to get

$$\mathbb{P}\{|Y_{\mathbf{X}} - Y_{\mathbf{X}'}| \geq t\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} z_i^2\lambda_j^2}, \frac{t}{\max_{i,j}|z_i\lambda_j|}\right\}\right).$$

Finally, observe that $\sum_{j=1}^{n} z_j^2 = \|\mathbf{H}\|_F^2$ and $\max_{1\leq j\leq n}|z_j| = \|\mathbf{H}\|$. One can now check that (2.11) is satisfied with respect to our chosen d_1 and d_2. □

Lemma 4.7.2 (Uniform deviation bound)**.** *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ be independent standard Gaussians, and suppose that Assumption $(A_{f,g})$ holds. Let $\mathcal{K}$ be a convex subset of the space of unit trace PSD matrices in $\mathbb{R}^{n\times n}$. For any $\epsilon, \delta > 0$, if m satisfies the lower bound* (4.6)*, then with probability at least $1 - \delta$, we have*

$$\sup_{\boldsymbol{X}\in\mathcal{K}}\left|\langle \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}, \boldsymbol{X} - \boldsymbol{x}_*\boldsymbol{x}_*^T\rangle\right| \leq \epsilon^2,$$

Proof. The proof of this concentration bound follows the same strategy as that in [87]. A priori, the process we are trying to control has heavy tails. To overcome this, we will use a decoupling argument together with conditioning.

For each $\mathbf{X} \in \mathcal{K}$, denote $\mathbf{H} = \mathbf{X} - \mathbf{x}_*\mathbf{x}_*^T$ for convenience. Let $\mathbf{P}_{\mathbf{x}_*}$ and $\mathbf{P}_{\mathbf{x}_*^\perp}$ denote projection onto $\mathbf{x}_*$ and its orthogonal complement respectively. We can then write

$$\begin{aligned}\langle \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}, \mathbf{H} \rangle &= \langle \mathbf{P}_{\mathbf{x}_*}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle + 2\langle \mathbf{P}_{\mathbf{x}_*^\perp}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle + \langle \mathbf{P}_{\mathbf{x}_*^\perp}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*^\perp}, \mathbf{H} \rangle \\ &= \langle \mathbf{P}_{\mathbf{x}_*}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle + 2\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle + \langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*^\perp}, \mathbf{H} \rangle. \end{aligned} \tag{4.20}$$

We shall bound the three terms on the right separately.

Recalling that $\mathbf{P}_{\mathbf{x}_*} = \mathbf{x}_*\mathbf{x}_*^T$, we see that the first term can be written as

$$\begin{aligned}\langle \mathbf{P}_{\mathbf{x}_*}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle &= \mathbf{x}_*^T(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}_*\mathbf{x}_*^T\mathbf{H}\mathbf{x}_* \\ &= \left[\frac{1}{m}\sum_{i=1}^m y_i(\langle \mathbf{a}_i, \mathbf{x}_* \rangle^2 - 1) - \mathbb{E}\{y(\langle \mathbf{a}_i, \mathbf{x}_* \rangle^2 - 1)\}\right]\mathbf{x}_*^T\mathbf{H}\mathbf{x}_*.\end{aligned}$$

Notice that $\mathbf{x}_*^T\mathbf{H}\mathbf{x}_* \leq 1$. Meanwhile, the term in the square brackets is the average of independent, centered, $\psi_{1/2}$ random variables. Using Proposition 2.2.9, we have a probability at least $1 - \delta/4$ event over which the following bound holds:

$$\sup_{\mathbf{X} \in \mathcal{K}} \left| \langle \mathbf{P}_{\mathbf{x}_*}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{P}_{\mathbf{x}_*}, \mathbf{H} \rangle \right| \leq \frac{C\log(1/\delta)^2}{\sqrt{m}}. \tag{4.21}$$

For the third term in (4.20), we write

$$\mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*^\perp} = \frac{1}{m}\sum_{i=1}^m y_i\big((\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{a}_i)(\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{a}_i)^T - \mathbf{P}_{\mathbf{x}_*^\perp}\big).$$

Since $\mathbf{y}_i$ and $\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{a}_i$ are independent, we may *decouple* them. In other words, we replace each $\mathbf{a}_i$ with a fully independent copy $\tilde{\mathbf{a}}_i$. We can therefore write

$$\begin{aligned}\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*^\perp}, \mathbf{H} \rangle &= \big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m y_i\big((\mathbf{P}_{\mathbf{x}_*^\perp}\tilde{\mathbf{a}}_i)(\mathbf{P}_{\mathbf{x}_*^\perp}\tilde{\mathbf{a}}_i)^T - \mathbf{P}_{\mathbf{x}_*^\perp}\big), \mathbf{H} \big\rangle \\ &= \big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m y_i(\tilde{\mathbf{a}}_i\tilde{\mathbf{a}}_i^T - \mathbf{I}_n), \mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{H}\mathbf{P}_{\mathbf{x}_*^\perp} \big\rangle\end{aligned}$$

Fix the randomness with respect to $\mathbf{a}_1, \ldots, \mathbf{a}_m$, $y_1, \ldots, y_m$, conditioning on the probability $1 - \delta/4$ event that the three statements in Lemma 4.7.3 hold. With respect to the $\tilde{\mathbf{a}}_i$'s, Lemma 4.7.1 shows us that this process indexed over $\mathbf{X} \in \mathcal{K}$ has mixed tails. Furthermore, since $\mathcal{K}$ is a subset of the nuclear norm ball, its diameter with respect to both the Frobenius and operator norms is bounded by 2. Lemma 2.4.2 then gives us another probability $1-\delta/4$

event over which

$$\begin{aligned}\sup_{\mathbf{X}\in\mathcal{K}}\Big|\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*^\perp},\mathbf{H}\rangle\Big| &\leq \frac{C}{m}\Bigg((\sum_{i=1}^m y_i^2)^{1/2}\Big(\gamma_2(\mathcal{K},\|\cdot\|_2)+\sqrt{\log(1/\delta)}\Big)\\ &\qquad+\max_{1\leq i\leq m}|y_i|(\gamma_1(\mathcal{K},\|\cdot\|)+\log(1/\delta))\Bigg)\\ &\leq C\Bigg(\frac{\gamma_2(\mathcal{K},\|\cdot\|_2)+\sqrt{\log(1/\delta)}}{\sqrt{m}}+\frac{\log m(\gamma_1(\mathcal{K},\|\cdot\|)+\log(1/\delta))}{m}\Bigg).\end{aligned}\tag{4.22}$$

Finally, for the second term in (4.20), we again decouple, writing

$$\begin{aligned}\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*},\mathbf{H}\rangle &= \big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m y_i\langle \mathbf{a}_i,\mathbf{x}_*\rangle\mathbf{x}_*(\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{a}_i)^T,\mathbf{H}\big\rangle\\ &= \big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m y_i\langle \mathbf{a}_i,\mathbf{x}_*\rangle\mathbf{x}_*(\mathbf{P}_{\mathbf{x}_*^\perp}\tilde{\mathbf{a}}_i)^T,\mathbf{H}\big\rangle\\ &= \big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m y_i\langle \mathbf{a}_i,\mathbf{x}_*\rangle\tilde{\mathbf{a}}_i,\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{H}\mathbf{x}_*\big\rangle.\end{aligned}\tag{4.23}$$

Again, fix the randomness with respect to $\mathbf{a}_1,\ldots,\mathbf{a}_m$, $y_1,\ldots,y_m$, and remember that we have conditioned on the event that the three statements in Lemma 4.7.3 hold. Then with respect to the $\tilde{\mathbf{a}}_i$'s, the quantity on the right in (4.23) is a centered Gaussian random variable with variance

$$\sigma^2=\frac{1}{m^2}\sum_{i=1}^m y_i^2\langle \mathbf{a}_i,\mathbf{x}_*\rangle^2\|\mathbf{P}_{\mathbf{x}_*^\perp}\mathbf{H}\mathbf{x}_*\|_2^2\leq\frac{C\|\mathbf{H}\|^2}{m}.$$

Therefore $(\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*},\mathbf{H}\rangle)_{\mathbf{X}}$ is a process indexed by $\mathbf{X}$ which has subgaussian increments with respect to the operator norm. We use an analogue of Theorem 2.4.2 (see Theorem 3.2 in [38]) to obtain a probability $1-\delta/4$ event over which

$$\sup_{\mathbf{X}\in\mathcal{K}}\Big|\langle \mathbf{P}_{\mathbf{x}_*^\perp}\hat{\mathbf{\Sigma}}\mathbf{P}_{\mathbf{x}_*},\mathbf{H}\rangle\Big|\leq C\Bigg(\frac{\gamma_2(\mathcal{K},\|\cdot\|_2)+\sqrt{\log(1/\delta)}}{\sqrt{m}}\Bigg).\tag{4.24}$$

Combining the bounds (4.21), (4.24), and (4.22) gives us the statement we want. □

Lemma 4.7.3. *Let $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_m$ be independent standard Gaussians, and suppose that Assumption $(A_{f,g})$ holds. Then for any $\delta>0$, so long as $m\geq C/\delta$, the following three statements hold simultaneously with probability at least $1-\delta/4$.*

1. $\sum_{i=1}^m f(\langle\boldsymbol{a}_i,\boldsymbol{x}_*\rangle)^2\leq Cm.$

2. $\sum_{i=1}^{m} y_i^2 f(\langle \boldsymbol{a}_i, \boldsymbol{x}_* \rangle)^2 \leq Cm.$

3. $\max_{1 \leq i \leq m} f(\langle \boldsymbol{a}_i, \boldsymbol{x}_* \rangle) \leq C \log m.$

Proof. Exactly the same as in Lemma 4.5.3. □

Proof of Theorem 4.2.5. We repeat the argument of Theorem 4.2.2, but replace the Hölder's inequality bound of $\langle \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}, \mathbf{X} - \mathbf{x}_* \mathbf{x}_*^T \rangle$ therein with the uniform deviation bound supplied by Lemma 4.7.2. □

CHAPTER 5

Moment Methods and Energy Minimization

5.1 Introduction

Amongst all Borel probability measures μ on $\mathbb{R}^n$ having the same radial distribution, we seek a minimizer for the energy integral

$$I_k(\mu) := \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \langle \mathbf{x}, \mathbf{y} \rangle^k d\mu(\mathbf{x}) d\mu(\mathbf{y}). \tag{5.1}$$

In this chapter, we will introduce a tensorization trick, thereby proving that the integral is minimized by the rotationally invariant measure, μ_{rot}. More precisely, for any integer k, we define the *k-th eccentricity tensor* of a measure μ. The gap between $I_k(\mu)$ and $I_k(\mu_{rot})$ is then given by the squared Euclidean norm of this tensor. Specializing to Borel probability measures on the sphere, we see that (5.1) is minimized by the uniform measure. Moreover, we may also adapt the proof to obtain an analogous result for the uniform measure on the sphere in $\mathbb{C}^n$.

These facts have several interesting applications, the first of which concerns the well-known Welch bounds in the signal processing literature. Using the complex case of our result, we recover the original Welch bounds, while using the real case, we are able to improve upon them for collections of real vectors. In our opinion, this proof is more illuminating than the existing ones. It shows one view the Welch bounds as saying that the average cross-correlation of signal sets cannot beat that of the uniform distribution.

Next, we are able to obtain new proofs of Björck's theorem from the 1950s and the recent theorem by Bilyk-Dai-Matzke. These theorems characterize optimizers of two one-parameter families of energy integrals, and were proved using methods from potential theory and spherical harmonics. Our methods have the benefit of being more elementary. Furthermore, our proof scheme for both theorems is very similar, and sheds light on the phase transition phenomenon discussed in [14]. Indeed, we are able to show why the phase transition occurs, and why it happens for different parameter values for the two families.

Finally, we use the theory to establish a new method of testing multi-dimensional Gaussian distributions: Given a random vector in $\mathbb{R}_n$, we are able to test whether it follows a normal distribution by considering the moments of its norm and those of the dot product of the random vector with an independent copy. Since Gaussian vectors are usually characterized in terms of their marginal moments, this reduces the problem of testing an uncountable family of one-dimensional distributions to that of testing just two one-dimensional distributions. This will be the basis for a new algorithm for Non-Gaussian Component Analysis (NGCA), which we will analyze and discuss in the next chapter.

The plan of the rest of this chapter is as follows. In Section 4.2, we define the eccentricity tensors and use the tensorization trick to prove the energy minimization property of rotationally invariant measures. In Section 4.3, we discuss the Welch bounds, show how they may be improved, and present some consequences of this improvement. In Section 4.4, we show how our results imply the two theorems on energy optimization on the sphere, and discuss their relevance to the phase transition phenomenon. We will consider Gaussian testing in Section 5.5.

5.2 Eccentricity tensors and the tensorization trick

In this section, we shall introduce the tensorization trick, define eccentricity tensors, and prove that rotationally invariant measures minimize (5.1). For notational as well as intuition purposes, however, it is more convenient to work with random vectors than with measures. We hence do so for the rest of this chapter, being careful to assert the independence of collections of random vectors where necessary.

The tensorization trick is to write the integral (5.1) as the squared Euclidean norm of the k-th moment tensor of μ.

Notation 5.2.1. Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$. For any positive integer k, let

$$\mathbf{M}_{\mathbf{X}}^k := \mathbb{E}\mathbf{X}^{\otimes k}$$

denote its k-th moment tensor if all entries are finite.

Recall the following fact from linear algebra. For any positive integer k, we may identify the k-th tensor product $T^k(\mathbb{R}^n) = \mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n$ with $\mathbb{R}^{n^k}$ by picking as a basis the vectors $\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \cdots \otimes \mathbf{e}_{i_k}\}_{1 \leq i_1, \ldots i_k \leq n}$. With this choice, the Euclidean inner product between any two pure tensors $\mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_k$ and $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_k$ can be written as

$$\langle \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_k, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_k \rangle = \prod_{i=1}^{k} \langle \mathbf{u}_i, \mathbf{v}_i \rangle.$$

In particular, for power tensors $\mathbf{u}^{\otimes k}$ and $\mathbf{v}^{\otimes k}$, we have the formula

$$\langle \mathbf{u}^{\otimes k}, \mathbf{v}^{\otimes k} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle^k. \tag{5.2}$$

Now if $\mathbf{X}$ and $\mathbf{Y}$ are two independent random vectors, we may rewrite the k-th moment of their inner product as an inner product between their k-th moment tensors. Namely, we have

$$\mathbb{E}(\langle \mathbf{X}, \mathbf{Y} \rangle^k) = \mathbb{E}\langle \mathbf{X}^{\otimes k}, \mathbf{Y}^{\otimes k} \rangle = \left\langle \mathbb{E}\mathbf{X}^{\otimes k}, \mathbb{E}\mathbf{Y}^{\otimes k} \right\rangle = \left\langle \mathbf{M}_{\mathbf{X}}^k, \mathbf{M}_Y^k \right\rangle, \tag{5.3}$$

where the first equality follows from equation (5.2). For independent copies $\mathbf{X}$ and $\mathbf{X}'$ of the same random vector having distribution μ, $\mathbf{M}_{\mathbf{X}}^k = \mathbf{M}_{X'}^k$, so

$$I_k(\mu) = \mathbb{E}(\langle \mathbf{X}, \mathbf{X}' \rangle^k) = \|\mathbf{M}_{\mathbf{X}}^k\|^2. \tag{5.4}$$

Here and in the rest of this chapter, we will use $\|\cdot\|$ to denote the vector Euclidean norm. No other norms are used, so there should be no risk of confusion.

We next introduce the notion of the rotation symmetrization of a random vector.

Definition 5.2.2. For any random vector $\mathbf{X}$ in $\mathbb{R}^n$, let $\mathbf{X}_{rot}$ denote a random vector that is independent of $\mathbf{X}$, has the same radial distribution as $\mathbf{X}$, and whose distribution is rotationally invariant (i.e. $\mathbf{Q}\mathbf{X}_{rot} \overset{d}{=} X_{rot}$ for all $\mathbf{Q} \in O(n)$). We call $\mathbf{X}_{rot}$ the *rotational symmetrization* of $\mathbf{X}$.

Comparing the moment tensors of a random vector and those of its rotational symmetrization give rise to what we shall call eccentricity tensors.

Definition 5.2.3. Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$ with finite moments of all orders. For any positive integer k, define its *k-th eccentricity tensor* to be

$$\mathbf{E}_{\mathbf{X}}^k := \mathbf{M}_{\mathbf{X}}^k - \mathbf{M}_{X_{rot}}^k. \tag{5.5}$$

Since $\mathbf{X} \overset{d}{=} \mathbf{X}_{rot}$ if and only if $\mathbf{X}$ is rotationally invariant, we see that the eccentricity tensors of $\mathbf{X}$ are quantitative measures of how far its distribution is from being rotationally invariant. This interpretation is further supported by the following observation.

Lemma 5.2.4 (Orthogonality)**.** *Let X be a random vector in* $\mathbb{R}^n$ *with finite moments of all orders. Its eccentricity tensors are orthogonal to the moment tensors of its rotational symmetrization. In other words, for any positive integer* k*,*

$$\langle \boldsymbol{E}^k_{\boldsymbol{X}}, \boldsymbol{M}^k_{\boldsymbol{X}_{rot}} \rangle = 0 \tag{5.6}$$

and

$$\left\|\boldsymbol{M}^k_{\boldsymbol{X}}\right\|^2 = \left\|\boldsymbol{M}^k_{\boldsymbol{X}_{rot}}\right\|^2 + \left\|\boldsymbol{E}^k_{\boldsymbol{X}}\right\|^2. \tag{5.7}$$

Proof. Let $\mathbf{Q}$ be a random orthogonal matrix chosen according to the Haar measure on $O(n)$. For any fixed vector $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{Qv}$ is uniformly distributed on the sphere of radius $\|\mathbf{v}\|$, so if $\mathbf{Y}$ is any random vector independent of $\mathbf{Q}$, applying $\mathbf{Q}$ to $\mathbf{Y}$ preserves its radial distribution but makes $\mathbf{QY}$ rotationally invariant.

Now choose $\mathbf{Q}$ to be independent of $\mathbf{X}$ and $\mathbf{X}_{rot}$. Our previous discussion implies that

$$\mathbf{Q}^T\mathbf{X} \overset{d}{=} \mathbf{X}_{rot} \overset{d}{=} \mathbf{Q}\mathbf{X}_{rot}.$$

We use this to compute

$$\mathbb{E}(\langle \mathbf{X}, \mathbf{X}_{rot}\rangle^k) = \mathbb{E}(\langle \mathbf{X}, \mathbf{Q}\mathbf{X}_{rot}\rangle^k) = \mathbb{E}(\langle \mathbf{Q}^T\mathbf{X}, \mathbf{X}_{rot}\rangle^k) = \mathbb{E}(\langle \mathbf{X}'_{rot}, \mathbf{X}_{rot}\rangle^k), \tag{5.8}$$

where $\mathbf{X}'_{rot}$ is an independent copy of $\mathbf{X}_{rot}$. We may then apply identities (5.3) and (5.4) to rewrite the above equation as

$$\left\langle \mathbf{M}^k_{\mathbf{X}}, \mathbf{M}^k_{\mathbf{X}_{rot}} \right\rangle = \left\langle \mathbf{M}^k_{\mathbf{X}_{rot}}, \mathbf{M}^k_{\mathbf{X}_{rot}} \right\rangle. \tag{5.9}$$

Subtracting the right hand side from the left hand side gives (5.6), from which (5.7) is an immediate corollary. □

The fact that the integral (5.1) is minimized by rotationally invariant measures is then an easy consequence of the previous lemma. To show that these are the *unique* minimizers, we need further assumptions on our random vectors to ensure that they are determined by their moment tensors. A sufficient condition is that of being subexponential (see Section 2.2).

Lemma 5.2.5. *Let X be a subexponential random vector in* $\mathbb{R}^n$*. Then the distribution of X is determined by its moment tensors.*

Proof. By the definition of being subexponential, we have the following moment growth condition [114]:

$$\sup_{\mathbf{v}\in S^{n-1}} \limsup_{r\to\infty} \frac{\left(\mathbb{E}|\langle \mathbf{X}, \mathbf{v}\rangle|^r\right)^{1/r}}{r} < \infty. \tag{5.10}$$

Let $\phi_{\mathbf{X}}(\mathbf{v}) = \mathbb{E}e^{i\langle \mathbf{X},\mathbf{v}\rangle}$ denote the characteristic function of $\mathbf{X}$. The above condition implies that for each $v \in S^{n-1}$, the function $t \mapsto \mathbb{E}e^{it\langle \mathbf{X},\mathbf{v}\rangle}$ can be written as a power series with coefficients $\frac{\mathbb{E}\langle \mathbf{X},\mathbf{v}\rangle^r}{r!}$ [12], so $\phi_{\mathbf{X}}(v)$ is determined by the moments $\mathbb{E}\langle \mathbf{X}, \mathbf{v}\rangle^r$. By (5.3), $\mathbb{E}\langle \mathbf{X}, \mathbf{v}\rangle^r = \langle \mathbf{M}^r_{\mathbf{X}}, \mathbf{v}^{\otimes r}\rangle$, so these are functions of the moment tensors. Finally, it is a fact from elementary probability that a random vector in $\mathbb{R}^n$ determined by its characteristic function (see exercise 2.36 in [28]). □

We can thus summarize our results so far in the following theorem.

Theorem 5.2.6. *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. Then*

a) (Minimization) If $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$, and $\boldsymbol{X}_{rot}, \boldsymbol{X}'_{rot}$ are independent copies of its rotational symmetrization, we have

$$\mathbb{E}(\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^k) \geq \mathbb{E}(\langle \boldsymbol{X}_{rot}, \boldsymbol{X}'_{rot}\rangle^k) \tag{5.11}$$

for any positive integer k so long as $\boldsymbol{X}$ has finite k-th moment.

b) (Uniqueness) Furthermore, if equality holds in (5.11) *for all k and we assume that $\boldsymbol{X}$ has a subexponential distribution, then $\boldsymbol{X}$ is rotationally invariant.*

Proof. Using identity (5.4), we rewrite the first claim as

$$\left\|\mathbf{M}^k_{\mathbf{X}}\right\|^2 \geq \left\|\mathbf{M}^k_{X_{rot}}\right\|^2,$$

and this follows immediately from equation (5.7).

If equality holds for all positive integers k, then by (5.7), $\mathbf{E}^k_{\mathbf{X}} = 0$ for all k, implying that $\mathbf{X}$ and $\mathbf{X}_{rot}$ have the same moment tensors of all orders. If we assume that $\mathbf{X}$ is subexponential, Lemma 5.2.5 implies that $\mathbf{X}$ and $\mathbf{X}_{rot}$ have the same distribution. □

For the remainder of this chapter, we specialize to the case of distributions on the sphere. Using Lemma 2.3.3, we immediately get the following bound.

Corollary 5.2.7. *Let $\boldsymbol{\theta}$ have the uniform distribution on the sphere S^{n-1}, and let $\boldsymbol{X}$ be any random vector taking values on the sphere. Let $\boldsymbol{\theta}'$ and $\boldsymbol{X}'$ be independent copies of $\boldsymbol{\theta}$ and*

***X** respectively. Then*

$$\mathbb{E}(\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^{2k}) \geq \mathbb{E}(\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle^{2k}) = \frac{1 \cdot 3 \cdots (2k-1)}{n \cdot (n+2) \cdots (n+2k-2)} \tag{5.12}$$

*for any positive integer k. Furthermore, if equality holds for all k, **X** has the uniform distribution.*

Proof. Clearly $\boldsymbol{\theta} \overset{d}{=} \mathbf{X}_{rot}$, and is subexponential. The inequality and the characterization statement then follows immediately from Theorem 5.2.6. By uniformity, we have $\mathbb{E}(\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle^{2k}) = \mathbb{E}(\langle \boldsymbol{\theta}, \mathbf{v} \rangle^{2k})$ for any unit vector $\mathbf{v} \in S^{n-1}$, and the explicit computation for $\mathbb{E}(\langle \boldsymbol{\theta}, \mathbf{v} \rangle^{2k})$ is the content of the next lemma. □

5.3 Applications to dictionary incoherence and the Welch bounds

Given a collection of m unit vectors $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ in $\mathbb{C}^n$, we are often interested in the quantity

$$c_{max} = \max_{i \neq j} |\langle \mathbf{z}_i, \mathbf{z}_j \rangle|.$$

If we think of the vectors as dictionary elements, then c_{max} measures the coherence or maximum cross-correlation of the dictionary. It is well known in the sparse approximation literature that the larger the value of c_{max}, the worse the collection Z performs when we try to recover a sparse representation of a vector as a linear combination of the $\mathbf{z}_j$'s [39]. As such, it is an important question in the design of communication systems to know how well we can do theoretically, and how we may find collections that achieve the theoretical minimum value of c_{max}.

In 1974, Welch gave a family of lower bounds on c_{max} in terms of m and n.

Theorem 5.3.1 (Welch, 1974 [122])**.** *Let Z and c_{max} be defined as above. Then for each positive integer k, we have*

$$(c_{max})^{2k} \geq \frac{1}{m-1}\left(\frac{m}{\binom{n+k-1}{k}} - 1\right). \tag{5.13}$$

Welch proved this theorem by bounding the average cross-correlation (also sometimes called the p-frame potential, with $p = 2k$ [41]).

Lemma 5.3.2 (Welch). *Let $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$ be unit vectors in $\mathbb{C}^n$, then*

$$\frac{1}{m^2}\sum_{i,j=1}^{m}|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle|^{2k} \geq \binom{n+k-1}{k}^{-1}. \tag{5.14}$$

By separating the diagonal terms from the sum and rearranging the summands, it is easy to see how (5.14) implies (5.13). Welch's original proof of (5.14) was combinatorial in nature. In 2003, Alon [1] provided a geometric proof based on examining the Gram matrix associated to Z and dimension counting. The proof was reproduced by Datta et al. [32] in 2012, who were apparently unaware of the earlier paper.

Both arguments are agnostic to whether the vectors are real or complex, and it is a natural question whether one may improve the bound when we restrict to the case of real vectors. Using the energy minimization property of rotationally invariant distributions, we are able to show that this is indeed the case.

Lemma 5.3.3. *Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ be unit vectors in $\mathbb{R}^n$. Then*

$$\frac{1}{m^2}\sum_{i,j=1}^{m}|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle|^{2k} \geq \frac{1\cdot 3\cdots(2k-1)}{n\cdot(n+2)\cdots(n+2k-2)}. \tag{5.15}$$

Remark 5.3.4. Since

$$\binom{n+k-1}{k}^{-1} = \frac{1\cdot 2\cdots k}{n\cdot(n+1)\cdots(n+k-1)},$$

we see that the new bound (5.15) is equal to the old one (5.14) for $k = 1$, and is strictly larger for $k > 1$.

Proof. Let $\mathbf{X}$ be uniformly distributed on the set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. Corollary 5.2.7 applies and we have

$$\mathbb{E}(\langle \mathbf{X}, \mathbf{X}'\rangle^{2k}) \geq \frac{1\cdot 3\cdots(2k-1)}{n\cdot(n+2)\cdots(n+2k-2)}$$

for any positive integer k. On the other hand, we also have

$$\mathbb{E}(\langle \mathbf{X}, \mathbf{X}'\rangle^{2k}) = \frac{1}{m^2}\sum_{i,j=1}^{m}|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|^{2k}.$$

□

Remark 5.3.5. This result stated by Ehler and Okoudjou in [41], and they attribute it to Venkov. The proof in [41], however, proceeds via spherical harmonics and not the tensorization machinery we have used here.

Let us illustrate the improved bound by revisiting an example from [32].

Example 5.3.6. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_7$ be the columns of

$$\begin{bmatrix} 0.99 & 0.14 & 0.56 & -0.68 & 0.93 & -0.86 & 0.30 \\ 0.08 & 0.99 & 0.83 & 0.73 & -0.36 & -0.50 & 0.95 \end{bmatrix}.$$

This collection achieves the $k = 1$ Welch bound (5.14), and its energy[1]

$$\sum_{i,j=1}^{7} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|^6 = 15.3128$$

was experimentally observed to be minimal over all collections of 7 unit vectors in $\mathbb{R}^2$. However, the $k = 3$ Welch bound gives a lower bound of 12.25 for the energy, so there was a gap between theory and experiment. Using our improved bound (5.15), we get 15.3125, thereby bridging this gap completely.

Although the improved bounds do not hold for complex collections of vectors, we are nonetheless able to recover the original Welch bounds using the same circle of ideas and making a few adjustments.

Definition 5.3.7. For any random vector $\mathbf{X}$ in $\mathbb{C}^n$, let $\mathbf{X}_{uni}$ denote a random vector that is independent of $\mathbf{X}$, has the same radial distribution as $\mathbf{X}$, and whose distribution is invariant under unitary transformations. We call $\mathbf{X}_{uni}$ the *unitary symmetrization* of $\mathbf{X}$.

With this definition, we can state the following complex version of Theorem 5.2.6.

Theorem 5.3.8. *Let $\boldsymbol{X}$ be a random vector in $\mathbb{C}^n$ with finite moments of all orders. Then if $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$, and $\boldsymbol{X}_{uni}, \boldsymbol{X}'_{uni}$ are independent copies of its unitary symmetrization, we have*

$$\mathbb{E}|\langle \boldsymbol{X}, \boldsymbol{X}' \rangle|^{2k} \geq \mathbb{E}|\langle \boldsymbol{X}_{uni}, \boldsymbol{X}'_{uni} \rangle|^{2k} \tag{5.16}$$

for any positive integer k.

Proof. By considering the moment tensors

$$\mathbf{M}_{\mathbf{X}}^{2k} := \mathbb{E}\mathbf{X}^{\otimes k} \otimes (\mathbf{X}^*)^{\otimes k},$$

[1] To compute this value, we renormalized the vectors $\mathbf{x}_1, \dots, \mathbf{x}_7$ in order to reduce roundoff error.

we may define a complex version of eccentricity tensors. Next, we replace $\mathbf{Q} \sim \text{Haar}(O(n))$ with $\mathbf{U} \sim \text{Haar}(U(n))$ in Lemma 5.2.4 to prove an orthogonality result analogous to (5.7). With this result, (5.16) follows immediately. □

We are now able to complete the proof of (5.14) with the help of the following version of Lemma 2.3.3.

Lemma 5.3.9 (Moments of complex spherical marginals)**.** *Let $\boldsymbol{\theta}$ be uniformly distributed on the complex sphere $S^{2n-1} \subset \mathbb{C}^n$. Then for any unit vector $\boldsymbol{v} \in S^{2n-1}$ and any positive integer k, we have*

$$\mathbb{E}|\langle \boldsymbol{\theta}, \boldsymbol{v} \rangle|^{2k} = \binom{n+k-1}{k}^{-1}. \tag{5.17}$$

Proof. Let γ and $\mathbf{g}$ denote standard complex Gaussians in 1 dimension and n dimensions respectively. Then

$$\mathbb{E}|\langle \boldsymbol{\theta}, \mathbf{v} \rangle|^{2k} = \frac{\mathbb{E}|\gamma|^{2k}}{\mathbb{E}\|\mathbf{g}\|^{2k}}.$$

Since $|\gamma|$ is the norm of a two-dimensional standard real Gaussian, while $\|\mathbf{g}\|$ is the norm of a $2n$-dimensional standard real Gaussian, (5.17) follows from the calculations of Gaussian integrals done in Lemma 2.3.3. □

Remark 5.3.10. Given $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ a set of unit vectors in a Hilbert space $\mathbb{H}$, k a positive integer, define the set

$$Z^{(k)} = \{\mathbf{z}_1^{\otimes k}, \mathbf{z}_2^{\otimes k}, \ldots, \mathbf{z}_m^{\otimes k}\} \subset \text{Sym}^k(\mathbb{H}).$$

Datta et al.'s paper [32] characterized sets Z achieving equality in the k-th Welch average cross-correlation bound (5.14) as those for which $Z^{(k)}$ forms a tight frame for $\text{Sym}^k(\mathbb{H})$. Since our results show that this bound is not tight when $\mathbb{H}$ is a real Hilbert space and $k > 1$, we have proved that tight frames of the form $Z^{(k)}$ do not exist for symmetric spaces of real tensors with $k > 1$. Indeed, this also holds true for generalized frames as defined by the same authors.

Remark 5.3.11. Datta et al. [32] showed that the analogous statement for complex vector spaces is false. In fact, if $\boldsymbol{\theta}$ is distributed uniformly on the complex sphere $S^{2n-1} \subset \mathbb{C}^n$, then

$$\mathbb{E}\boldsymbol{\theta}^{\otimes k} \otimes (\boldsymbol{\theta}^*)^{\otimes k} = \binom{n+k-1}{k}^{-1} I_{\text{Sym}^k(\mathbb{C}^n)}.$$

5.4 Applications to energy optimization on the sphere

In two recent papers [13, 14], Bilyk et al. presented a theorem characterizing probability measures minimizing geodesic distance energy integrals. This is an analogue of Björck's theorem from 1956 which characterized probability measures minimizing energy integrals based on Euclidean distance [15]. Björck proved his theorem by considering Riesz potentials, while Bilyk et al. proved their result using spherical harmonic expansions and the hermisphere Stolarsky principle. In this section, we show how to derive both results using the tensorization trick and the energy minimization property of the uniform distribution on the sphere.

Theorem 5.4.1 (Bilyk-Dai-Matzke, 2016)**.** *For $\delta > 0$, define the geodesic energy integral*

$$G_\delta(\mu) := \int_{S^{n-1}} \int_{S^{n-1}} d(\boldsymbol{x}, \boldsymbol{y})^\delta d\mu(\boldsymbol{x}) d\mu(\boldsymbol{y}), \tag{5.18}$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ denotes the geodesic distance between x and y. The maximizers of this energy integral over Borel probability measures on S^{n-1} can be characterized as follows:

a) $0 < \delta < 1$*: the unique maximizer of $G_\delta(\mu)$ is $\mu = \sigma$, the uniform measure.*

b) $\delta = 1$*: $G_\delta(\mu)$ is maximized if and only if μ is centrally symmetric.*

c) $\delta > 1$*: $G_\delta(\mu)$ is maximized if and only if $\mu = \frac{1}{2}(\delta_{\boldsymbol{p}} + \delta_{-\boldsymbol{p}})$, i.e. the mass is supported equally by two antipodal points.*

Proof. Observe that the geodesic distance $d(\mathbf{x}, \mathbf{y})$ is simply the angle between x and y. As such, we have $d(\mathbf{x}, \mathbf{y}) = \arccos(\langle \mathbf{x}, \mathbf{y} \rangle)$. We may thus rewrite (5.18) as

$$G_\delta(\mu) = \mathbb{E}\arccos(\langle \mathbf{X}, \mathbf{X}' \rangle)^\delta,$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors with distribution μ.

Let us start by proving part b). It is an exercise to show that the even derivatives of $\arccos$ vanish at 0, while the odd derivatives are strictly negative at 0. For $-1 < t < 1$ may hence write $\arccos$ as its Taylor series

$$\arccos(t) = \frac{\pi}{2} - \sum_{k=0}^{\infty} a_{2k+1} t^{2k+1} \tag{5.19}$$

where $a_{2k+1} > 0$ for all k. We claim that in fact, the above formula holds for all t in the *closed* interval $[-1, 1]$, and furthermore that the series is absolutely convergent. This is the content of Lemma 5.4.2 to come. As a result, we may use Fubini to interchange sums and expectations, thereby writing

$$\mathbb{E}\arccos(\langle \mathbf{X}, \mathbf{X}' \rangle) = \frac{\pi}{2} - \sum_{k=0}^{\infty} a_{2k+1} \mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle^{2k+1}.$$

Since $\mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle^{2k+1} \geq 0$ for each k by identity (5.4), this last expression is maximized if and only if $\mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle^{2k+1} = 0$ for every non-negative integer k. By the same identity, this happens if and only if all odd moments of $\mathbf{X}$ are zero, i.e. if and only if $\mathbf{X}$ is centrally symmetric. This proves the case $\delta = 1$.

Now let $0 < \delta < 1$. We claim that for $-1 \leq t \leq 1$, we may write

$$\arccos(t)^{\delta} = \left(\frac{\pi}{2}\right)^{\delta} - \sum_{k=1}^{\infty} a_k t^k \tag{5.20}$$

where $a_k > 0$ for all $k > 0$, and that the series is absolutely convergent. Lemma 5.4.3 (to come) tells us that the Taylor series of $\arccos(t)^{\delta}$ has this form, which combined with Lemma 5.4.2 proves this claim. As such, we may again use Fubini to write

$$\mathbb{E}\arccos(\langle \mathbf{X}, \mathbf{X}' \rangle)^{\delta} = \left(\frac{\pi}{2}\right)^{\delta} - \sum_{k=1}^{\infty} a_k \mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle^{k}. \tag{5.21}$$

By identity (5.4), $\mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle^{k} \geq 0$ for any distribution, while by Corollary 5.2.7, the uniform measure uniquely minimizes all of these moments simultaneously. As such, we see that it is the unique maximizer of $G_{\delta}(\mu)$.

The remaining case where $\delta > 1$ is easy and does not require a proof using our methods. For completeness, we repeat the proof given by the original authors [14]. Since $d(\mathbf{x}, \mathbf{y}) \leq \frac{\pi}{2}$, we have

$$G_{\delta}(\mu) \leq \left(\frac{\pi}{2}\right)^{\delta - 1} \int_{S^{n-1}} \int_{S^{n-1}} d(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y}) \leq \left(\frac{\pi}{2}\right)^{\delta}.$$

The first inequality is tight whenever $d(\mathbf{x}, \mathbf{y})$ only takes the values $\frac{\pi}{2}$ and 0, while by part b), the second inequality becomes equality when μ is centrally symmetric. Together, these imply that $\mu = \frac{1}{2}(\delta_{\mathbf{p}} + \delta_{-\mathbf{p}})$ for some $\mathbf{p} \in S^{n-1}$. □

Lemma 5.4.2. *Let f be a function that is continuous on $[-1, 1]$ and that agrees with its Taylor series at 0 on the open interval $(-1, 1)$. Suppose further that all but finitely many of its derivatives at 0 have the same sign. Then the series is absolutely convergent over the closed interval $[-1, 1]$, and agrees with f over the interval.*

Proof. By subtracting off polynomials and negating the function if necessary, we may assume without loss of generality that the Taylor series for $f(t)$ is given by $\sum_{k=0}^{\infty} c_k t^k$ where $c_k \geq 0$ for all k. By the monotone convergence theorem, together with our assumptions on f, we have

$$\sum_{k=0}^{\infty} c_k = \lim_{t \to 1^-} \sum_{k=0}^{\infty} c_k t^k = \lim_{t \to 1^-} f(t) = f(1).$$

As such, the series $\sum_{k=0}^{\infty} c_k$ is absolutely convergent, and the Taylor series is also absolutely convergent on the closed interval $[-1, 1]$. Finally, we can apply the dominated convergence theorem to see that $f(-1) = \sum_{k=0}^{\infty} c_k (-1)^k$. □

Lemma 5.4.3. *Let f be a function that has derivatives of all orders at 0 and let $0 < \alpha < 1$. Suppose $f(0) > 0$ and $f'(0) < 0$, while all higher derivatives f at 0 are non-positive, then all derivatives of f^α at 0 are strictly negative.*

Proof. Let $F(t) = f(t)^\alpha$. By induction, one may observe that for any positive integer k, $F^{(k)}(t)$ is a sum of 2^{k-1} terms of the form

$$g_{\vec{n}}(t) := f(t)^{\alpha - j} \left(\prod_{i=0}^{j-1} (\alpha - i) \right) \left(\prod_{i=0}^{j-1} f^{(n_i)}(t) \right),$$

where $1 \leq j \leq k$, and $\vec{n} = (n_0, n_1, \ldots, n_{j-1})$ is a vector of positive integers. If there is some index i such that $f^{(n_i)}(0) = 0$, then $g_{\vec{n}}(0) = 0$. Otherwise, $\prod_{i=0}^{j-1} f^{(n_i)}(0)$ is a product of j negative numbers and so has sign $(-1)^j$. On the other hand, our assumption on α imply that $\left(\prod_{i=0}^{j-1} (\alpha - i) \right)$ is a product of one positive number and $j-1$ negative numbers, and so has sign $(-1)^{j-1}$. As such, $g_{\vec{n}}(0) \leq 0$.

Finally, notice that $F^{(k)}(0)$ always contains the term

$$g_{(1,1,\ldots,1)}(0) = f(t)^{\alpha - k} \left(\prod_{i=0}^{k-1} (\alpha - i) \right) f'(0)^k.$$

Since we have assumed that $f'(0) < 0$, this term is strictly negative. As such, $F^{(k)}(0)$ is also negative, as was to be shown. □

In the course of proving the previous theorem, we have in fact proved the following more general result.

Theorem 5.4.4. *Let F be a function on on $[-1, 1]$ that is given by the power series*

$$F(t) = a_0 - \sum_{k=1}^{\infty} a_k t^k, \tag{5.22}$$

where $a_k \geq 0$ for all $k > 0$. Then the energy integral

$$I_F(\mu) := \int_{S^{n-1}} \int_{S^{n-1}} F(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) d\mu(\boldsymbol{x}) d\mu(\boldsymbol{y}) \tag{5.23}$$

is maximized over all Borel probability measures on S^{n-1} by the uniform measure. Furthermore, if $a_k > 0$ for all $k > 0$, then the maximizer is unique.

Let us see how we may apply this more general theorem to recover Björck's original result.

Theorem 5.4.5 (Björck, 1956)**.** *For $\delta > 0$, define the energy integral*

$$E_\delta(\mu) = \int_{S^{n-1}} \int_{S^{n-1}} \|\boldsymbol{x} - \boldsymbol{y}\|^\delta d\mu(\boldsymbol{x}) d\mu(\boldsymbol{y}). \tag{5.24}$$

The maximizers of this energy integral over Borel probability measures on S^{n-1} can be characterized as follows:

1. *$0 < \delta < 2$: the unique maximizer of $E_\delta(\mu)$ is $\mu = \sigma$, the uniform measure.*

2. *$\delta = 2$: $E_\delta(\mu)$ is maximized if and only if the center of mass of μ is at the origin.*

3. *$\delta > 2$: $E_\delta(\mu)$ is maximized if and only if $\mu = \frac{1}{2}(\delta_{\boldsymbol{p}} + \delta_{-\boldsymbol{p}})$, i.e. the mass is supported equally by two antipodal points.*

Proof. We rewrite (5.24) as

$$E_\delta(\mu) = \mathbb{E}\|\mathbf{X} - \mathbf{X}'\|^\delta$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors with distribution μ. The easy case $\delta > 2$ is proved exactly as in Theorem 5.4.1. The case $\delta = 2$ is also clear, for we may write $\|\mathbf{X} - \mathbf{X}'\|^2 = 2 - 2\langle \mathbf{X}, \mathbf{X}' \rangle$, and by identity (5.4),

$$E_2(\mu) = 2 - \mathbb{E}\langle \mathbf{X}, \mathbf{X}' \rangle = 2 - \|\mathbb{E}\mathbf{X}\|^2.$$

This is maximized if and only if $\mathbb{E}X = 0$.

For $0 < \delta < 2$, we set $f(t) = 2 - 2t$ and $F(t) = f(t)^{\delta/2}$. Then f and (5.7) $= \delta/2$ satisfy the hypotheses of Lemma 5.4.3, so $F^{(k)}(0) < 0$ for all positive integers k. This, together with Lemma 5.4.2, implies that F satisfies the hypothesis of Theorem 5.4.4. Since

$$E_\delta(\mu) = \int_{S^{n-1}} \int_{S^{n-1}} (2 - 2\langle \mathbf{x}, \mathbf{y} \rangle)^{\delta/2} d\mu(\mathbf{x}) d\mu(\mathbf{y}) = \int_{S^{n-1}} \int_{S^{n-1}} F(\langle \mathbf{x}, \mathbf{y} \rangle) d\mu(\mathbf{x}) d\mu(\mathbf{y}),$$

we can conclude that $E_\delta(\mu)$ is uniquely maximized by the uniform measure. □

Remark 5.4.6. In their paper [14], Bilyk et al. remarked that while the Euclidean and geodesic distances are both metrics on the sphere, the phase transition for the behavior of their energy integrals is different. In the Euclidean case, Björck's theorem shows that it occurs at $\delta = 2$, while in the geodesic case, Bilyk et al.'s theorem shows that it occurs at $\delta = 1$. This peculiar phenomenon is explained by our unified proof of both results.

In both cases, the existence of a phase transition as we let δ decrease to 0 is asserted by Lemma 5.4.3 and Theorem 5.4.4. If the integrand satisfies the hypotheses of Lemma 5.4.3 for some δ_0, then for all $0 < \delta < \delta_0$, the integrand will satisfy the hypothesis of Theorem 5.4.4, from which we can conclude that the unique maximizer is the uniform measure. For the Euclidean integral, we have $\delta_0 = 2$, while for the geodesic integral, we have $\delta_0 = 1$.

Remark 5.4.7. Bilyk et al. were also interested in understanding continuous functions F for which the uniform measure σ is the unique minimizer of I_F as defined in (5.23). They managed to characterize these functions as those for which all non-constant Gegenbauer coefficients are strictly positive, i.e.

$$\hat{F}(k, \lambda) > 0$$

for all positive integers k, and where $\lambda = \frac{n}{2} - 1$. On the other hand, by flipping signs, Theorem 5.4.4 implies that a sufficient condition for this to happen is to require all non-constant Taylor series coefficents to be strictly positive.

5.5 Testing multi-dimensional Gaussian distributions

Theorem 5.5.1 (First Gaussian test)**.** *Suppose* $\boldsymbol{X}$ *has the same radial distribution as* $\mathbf{g}$*, i.e.* $\|\boldsymbol{X}\|_2$ *and* $\|\boldsymbol{g}\|_2$ *are identically distributed. If* $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ *has the same distribution as* $\langle \mathbf{g}, \mathbf{g}' \rangle$*, then* $\boldsymbol{X}$ *has the same distribution as* $\mathbf{g}$*, i.e. the standard Gaussian distribution.*

Proof. If $\mathbf{X}$ has the same radial distribution as $\mathbf{g}$, then $\mathbf{g}$ is the rotational symmetrization of $\mathbf{X}$. The claim is then a direct application of the uniqueness portion of Theorem 5.2.6. □

By considering more carefully the orthogonal decomposition of moment tensors, we can make this characterization theorem quantitative. This will be useful if we seek to use the characterization for learning problems in which we only have access to empirical estimates of the moment tensors. For instance, this will be the case for our analysis in Chapter 6.

Lemma 5.5.2. *Let* $\boldsymbol{X}$ *be a random vector in* $\mathbb{R}^n$*. Let* $\boldsymbol{\theta}$ *be uniformly distributed on the sphere* S^{n-1}*. Then the following hold for any positive integer* r*:*

a) $\boldsymbol{M}^r_{\boldsymbol{X}_{rot}} = \mathbb{E}\{\|\boldsymbol{X}\|_2^r\}\boldsymbol{M}^r_{\boldsymbol{\theta}}$.
b) $\|\boldsymbol{E}^r_{\boldsymbol{X}}\|_2^2 = (\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle\}^r) - (\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^2(\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle\}^r)$.
c) For any unit vector $v \in \mathbb{R}^n$*,*

$$|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \le |\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}|\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \\ + \Big(\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - (\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\Big)^{1/2}. \tag{5.25}$$

d) In particular, when r *is odd,*

$$|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \le (\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\})^{1/2} = |\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - (\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\})|^{1/2}. \tag{5.26}$$

Proof. For the first statement, observe that $\mathbf{X}_{rot} = \|\mathbf{X}\|_2\boldsymbol{\theta}$, with $\|\mathbf{X}\|_2$ and $\boldsymbol{\theta}$ independent. We thus have

$$\mathbf{M}^r_{\mathbf{X}_{rot}} = \mathbb{E}\{(\|\mathbf{X}\|_2\boldsymbol{\theta})^{\otimes r}\} = \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbb{E}\{\boldsymbol{\theta}^{\otimes r}\} = \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}}.$$

Next, rewrite (5.7) as $\|\mathbf{E}^r_{\mathbf{X}}\|_2^2 = \|\mathbf{M}^r_{\mathbf{X}}\|_2^2 - \|\mathbf{M}^r_{\mathbf{X}_{rot}}\|_2^2$. By definition, we have $\|\mathbf{M}^r_{\mathbf{X}}\|_2^2 = \mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\}$ and using a), we get $\|\mathbf{M}^r_{\mathbf{X}_{rot}}\|_2^2 = (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$.

To prove part c), fix v and write

$$\mathbb{E}\{\langle \mathbf{X}, \mathbf{v}\rangle^r\} - \mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} = \langle \mathbf{M}^r_{\mathbf{X}} - \mathbf{M}^r_g, \mathbf{v}^{\otimes r}\rangle = \langle \mathbf{M}^r_{\mathbf{X}_{rot}} - \mathbf{M}^r_g, \mathbf{v}^{\otimes r}\rangle + \langle \mathbf{E}^r_{\mathbf{X}}, \mathbf{v}^{\otimes r}\rangle.$$

We use a) to write

$$\begin{aligned}\langle \mathbf{M}^r_{\mathbf{X}_{rot}} - \mathbf{M}^r_g, v^{\otimes r}\rangle &= \langle \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}}, v^{\otimes r}\rangle \\ &= (\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\})\mathbb{E}\{\langle \boldsymbol{\theta}, \mathbf{v}\rangle^r\}.\end{aligned}$$

Notice that $\mathbb{E}\{\langle \boldsymbol{\theta}, \mathbf{v}\rangle^r\} = \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$. We then combine the last two equations with b) and Cauchy-Schwarz to get (5.25). Finally, to get the last claim, we use the fact that $\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} = \mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\} = 0$ whenever r is odd. □

By balancing the two terms on the right hand side in part c), we obtain the following lemma, which will again be useful in Chapter 6.

Lemma 5.5.3. *Let* $\boldsymbol{X}$ *be a random vector in* $\mathbb{R}^n$ *for* $n \ge 2$*. Suppose there is a unit vector* $v \in S^{n-1}$*, an even integer* $r \ge 2$*, and a positive number* $0 < \delta \le 1$ *such that*

$|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \geq \delta \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}$. *Then either*

$$|\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}| \geq \frac{\delta^2}{4}\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}, \text{ or}$$

$$|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}| \geq \frac{15\delta^2}{64}(\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\})^2.$$

Proof. Observe that (5.25) gives the bound

$$\begin{aligned} \delta \mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} \leq |\mathbb{E}\{\|\mathbf{X}\|_2^r\} &- \mathbb{E}\{\|\mathbf{g}\|_2^r\}|\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \\ &+ \Big(\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2(\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\})\Big)^{1/2}. \end{aligned} \tag{5.27}$$

Suppose $|\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}| \leq \frac{\delta^2}{4}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}$. Then the second term on the right in equation (5.27) has to be large. Indeed, since $\delta \leq 1$ and $\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \leq 1/2$ for $r, n \geq 2$, we have

$$\begin{aligned} \Big(\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\Big)^{1/2} &\geq \delta \mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} - \frac{\delta^2}{4}\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} \\ &\geq \frac{7\delta}{8}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}. \end{aligned}$$

Now, applying the fact that $\mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\} = (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$, we use the reverse triangle inequality and the above bound to write

$$\begin{aligned} |\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - \mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\}| &\geq \Big|\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\Big| \\ &\qquad - \Big|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\Big|\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \\ &\geq \left(\frac{7\delta}{8}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}\right)^2 \\ &\qquad - \Big|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\Big|\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}. \end{aligned} \tag{5.28}$$

Next, notice that

$$\begin{aligned} \Big|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\Big| &= |\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}|(\mathbb{E}\{\|\mathbf{X}\|_2^r\} + \mathbb{E}\{\|\mathbf{g}\|_2^r\}) \\ &= |\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}| \cdot 2\mathbb{E}\{\|\mathbf{g}\|_2^r\} \\ &\qquad + (\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\})^2, \end{aligned}$$

so by the assumption on $|\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}|$, we have

$$\begin{aligned}\left|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\right|\mathbb{E}\{\langle\boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} &\le \frac{\delta^2}{4}\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\}\cdot 2\mathbb{E}\{\|\mathbf{g}\|_2^r\}\cdot\mathbb{E}\{\langle\boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} \\ &\quad + \left(\frac{\delta^2}{4}\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\}\right)^2\mathbb{E}\{\langle\boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} \\ &= \frac{\delta^2}{2}(\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\})^2 \\ &\quad + \left(\frac{\delta^2}{4}\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\}\right)^2\mathbb{E}\{\langle\boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} \\ &\le \frac{17\delta^2}{32}(\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\})^2. \end{aligned} \tag{5.29}$$

We can now substitute (5.29) into (5.28) to get

$$|\mathbb{E}\{\langle\mathbf{X},\mathbf{X}'\rangle^r\} - \mathbb{E}\{\langle\mathbf{g},\mathbf{g}'\rangle^r\}| \ge \frac{15\delta^2}{64}(\mathbb{E}\{\langle\mathbf{g},\mathbf{v}\rangle^r\})^2.$$

□

5.6 Comments and open questions

This chapter is based on the paper “Energy optimization for distributions on the sphere and improvement to the Welch bounds” [105]. After submitting the first version of that paper, I became aware that a partial version of Corollary 5.2.7 was proved by Ehler and Okoudjou in [41] (see Theorem 4.10 therein). Their result gives the inequality portion of the corollary but not the uniqueness part of it. They also do not prove any other part of Theorem 5.2.6, which applies to more general random vectors, and for all positive integer moments (as opposed to just even integer moments).

Like Bilyk et al., Ehler and Okoudjou obtained their result using spherical harmonics, and in particular, by considering the Gegenbauer coefficients of monomial functions. This is more evidence that there should be a close relationship between the theory of eccentricity tensors and that of spherical harmonics, and it will be interesting to investigate this connection further.

CHAPTER 6

Non-Gaussian Component Analysis

6.1 Introduction

6.1.1 Non-Gaussian Component Analysis

Dimension reduction is a necessary step for much of modern data analysis, the principle being that the structure or "interestingness" of a collection of data points is contained in a geometric structure which has much lower dimension than the ambient vector space. We consider the case where the geometric structure in question is a linear subspace. In other words, we are in the situation where the variation of the data points within this subspace contains some information which we would like to extract, while their variation in the complementary directions constitute mere noise.

In many cases, it is reasonable to think of the noise as being Gaussian. Formally, we then have the following generative model. Let E be an unknown d-dimensional subspace of $\mathbb{R}^n$, and let $E^\perp$ be the orthogonal complement of E. Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$, which we can decompose into two independent components: a non-Gaussian component $\tilde{\mathbf{X}}$ that takes values in E, and a Gaussian component $\mathbf{g}$ that takes values in $E^\perp$. In other words, we let $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{g}) \in E \oplus E^\perp$.[1]

Our goal is to recover the subspace E from a sample of independent realizations of $\mathbf{X}$. This is precisely the framework of the problem of Non-Gaussian Component Analysis (NGCA). We make no assumption on the relative magnitudes of $\tilde{\mathbf{X}}$ and $\mathbf{g}$. When the noise component is much smaller, which is a reasonable assumption in some real world applications, E can be recovered using the standard Principal Component Analysis (PCA). However, PCA manifestly fails when the signal to noise ratio is small, i.e. when $\tilde{\mathbf{X}}$ has lower magnitude than $\mathbf{g}$.

[1] It is not necessary to assume that the Gaussian and non-Gaussian subspaces are perpendicular. They automatically become perpendicular if we apply a whitening transformation.

With mild distributional assumptions, applying a whitening transformation to the data points can be done efficiently with sample size linear in the dimension (see [114]). As such, we might as well assume that the distribution is already whitened (i.e. isotropic). In other words, for the rest of this chapter, we work with the model:

Definition 6.1.1 (Isotropic NGCA model)**.**

$$\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{g}) \in E \oplus E^{\perp}, \quad \mathbb{E}\mathbf{X} = 0, \quad \mathbb{E}\mathbf{X}\mathbf{X}^T = \mathbf{I}_n. \tag{6.1}$$

The NGCA problem is closely related to the problem of Independent Component Analysis (ICA), but generalizes it in a crucial way. ICA assumes the existence of a latent variable s with independent coordinates, whereas in our case, the distribution of $\tilde{\mathbf{X}}$ is allowed to have any manner of dependencies amongst its entries.

6.1.2 Quantifying "non-Gaussianness"

In order to provide a guarantee for an algorithm for NGCA, one needs to quantify the deviation of $\tilde{\mathbf{X}}$ from being Gaussian. We will do so in terms of its moments.

Definition 6.1.2. We say that $\tilde{\boldsymbol{X}}$ *is* (m, η)*-moment-identifiable* along a unit vector $\mathbf{v} \in E$ if there is some $1 \leq r \leq m$ for which

$$\left|\mathbb{E}\{\langle \tilde{\mathbf{X}}, \mathbf{v} \rangle^r\} - \gamma_r\right| \geq \eta. \tag{6.2}$$

Here $\gamma_r := (2r-1)!!$ is the r-th moment of a $\mathcal{N}(0,1)$ random variable. The r-th moment distance of $\tilde{\mathbf{X}}$ from a standard Gaussian is defined as the quantity

$$D_{\tilde{\mathbf{X}},r} := \sup_{\mathbf{v} \in S^{n-1} \cap E} \left|\mathbb{E}\{\langle \tilde{\mathbf{X}}, \mathbf{v} \rangle^r\} - \gamma_r\right|. \tag{6.3}$$

There are three reasons why we take such an approach. First, it allows us to analyze our proposed algorithm more easily, since the algorithm is a moment method, and second, it allows us to quantify the "non-Gaussianness" of distributions that possibly do not have densities. This would not be possible had we chosen a notion like the total variation distance for instance. Finally, by the classical moment problem, if $D_{\tilde{\mathbf{X}},r} = 0$ for all positive integers r, then $\tilde{\mathbf{X}}$ has the standard Gaussian distribution.

Nonetheless, readers may be concerned about how the moment-identifiability condition squares with other notions of distribution distance. This was investigated somewhat by [112], who proved the following result for log-concave distributions on $\mathbb{R}$.

Fact 6.1.3 (Lemma 1 in [112]). *Let G be the density of a standard Gaussian random variable, F the density of an isotropic log-concave distribution. Suppose G is not (m, η)-moment-identifiable, i.e. for $r = 1, \ldots, m$, $|\mathbb{E}_F\{X^r\} - \gamma_r| < \eta$. Then there is a universal constant C such that*

$$\|F - G\|_1 \leq C \frac{\log m}{m^{1/16}} + \eta m e^m.$$

We note that the log-concave assumption is simply to obtain a tail bound for the characteristic function for F. Hence, the result also holds for any distribution with a C^1 density, albeit with possibly a different constant in the bound. Furthermore, the method for proving the result can easily be generalized to multivariate distributions.

6.1.3 Notes

This chapter is based on the paper [108]. As far as we know, the NGCA problem was first formulated and studied by [16]. They observed that whenever $\mathbf{X}$ satisfies the NGCA model (6.1), then for any smooth function h, we have

$$\boldsymbol{\beta}(h) := \mathbb{E}\{\mathbf{X}h(\mathbf{X})\} - \mathbb{E}\{\nabla h(\mathbf{X})\} \in E. \tag{6.4}$$

This suggests that if we can find a rich enough collection of functions $\mathcal{H}$, then one should be able to recover E as the span of $\{\boldsymbol{\beta}(h) \colon h \in \mathcal{H}\}$. Hence, the authors proposed first forming empirical estimates $\hat{\boldsymbol{\beta}}(h)$ using the given i.i.d. samples of $\mathbf{X}$, and then running PCA on this collection of vectors. Inspired by the FastICA algorithm of [57], they suggested picking test functions of the form $h_{a,\boldsymbol{\omega}}(\mathbf{x}) = \tilde{h}_a(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$ where $\boldsymbol{\omega} \in S^{n-1}$ and $\{\tilde{h}_a \colon a \in \mathbb{R}\}$ is a one-parameter family of smooth functions. They called this approach *Multi-index Projection Pursuit*.

Subsequent papers have built upon this in several ways. [64] investigated the situation when the contrast functions h_i's are chosen to be radial kernel functions, and when these are adapted to the data in an iterative fashion. [37, 36] replaced the PCA step with a semidefinite program, thereby yielding an approach they call *Sparse NGCA*.

All the papers in this line of research suffer from the defect that the performance of the algorithms all depend experimentally and theoretically on some "good" behavior of the $\boldsymbol{\beta}(h)$'s. Clearly, how "good" the $\boldsymbol{\beta}(h)$'s are depends intimately on how the chosen contrast functions interact with the particular way in which $\tilde{\mathbf{X}}$ deviates from being Gaussian. None of these papers are able to quantify this dependence theoretically, and instead simply assume the "good" behavior (see for instance Assumption 1 in [36]), so their proposed algorithms cannot be said to have polynomial time and sample complexity guarantees.

Indeed, prior to our work, the only algorithm with such guarantees was proposed and studied by [112]. Their strategy was to adapt [44]'s work on ICA to higher moments. For each positive integer r, they defined the marginal moment function $f_r(\mathbf{v}) := \mathbb{E}\{\langle \mathbf{X}, \mathbf{v}\rangle^r\}$, and noted that the strict local optima of f_r would have to lie in E. Furthermore, for each r, the r-th moment tensors of $\mathbf{X}$ defining f_r can be approximated up to ϵ accuracy in each of its entries with enough samples. These therefore yield empirical estimates $\hat{f}_r$ that have local optima that are close to those of f_r. Finally, they showed how to identify a local optima of $\hat{f}_r$ using a 2nd order local search. The samples are then projected onto the orthogonal complement of this direction, and the algorithm is applied recursively on the projection. They were able to prove that whenever $\tilde{\mathbf{X}}$ is (m, η)-moment-identifiable along all unit vectors $\mathbf{v} \in E$, then their algorithm recovers a subspace $\hat{E}$ close enough to E with time and sample complexity polynomial in n, η, $1/\epsilon$, and $\log(1/\delta)$, where δ is the failure probability. The degree of the polynomial however grows linearly in m and d. [2]

Other work on NGCA include [63, 64, 65, 94]. These papers have limited theoretical analysis, and we omit a discussion of these because of space constraints.

6.2 Main results

The principle that underlies our approach to NGCA is the new characterization of multivariate Gaussian distributions developed in Section 5.5 of the previous chapter. Throughout this section, $\mathbf{X}$ denotes a random vector in $\mathbb{R}^n$ and $\mathbf{g}$ is a standard Gaussian random vector in $\mathbb{R}^n$. By $\mathbf{X}'$ we will always denote an independent copy of $\mathbf{X}$.

Theorem 5.5.1 tells us how to identify non-Gaussian distributions. This result by itself does not address the NGCA problem, in which we are looking to identify non-Gaussian *directions* in the distribution of $\mathbf{X}$. To this end, we propose a matrix version of the first Gaussian test. Pick a parameter $\alpha > 0$ and consider the *test matrices*

$$\mathbf{\Phi}_{\mathbf{X},\alpha} := \frac{1}{Z_{\mathbf{\Phi}}}\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\mathbf{X}\mathbf{X}^T\} \quad \text{and} \quad \mathbf{\Psi}_{\mathbf{X},\alpha} := \frac{1}{Z_{\mathbf{\Psi}}}\mathbb{E}\{e^{-\alpha\langle \mathbf{X},\mathbf{X}'\rangle}\mathbf{X}(\mathbf{X}')^T\}, \tag{6.5}$$

where the normalizing quantities $Z_{\mathbf{\Phi}} = Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) := \mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\}$ and $Z_{\mathbf{\Psi}} = Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) := \mathbb{E}\{e^{-\alpha\langle \mathbf{X},\mathbf{X}'\rangle}\}$ resemble partition functions in statistical mechanics.

For a standard Gaussian random vector $\mathbf{g}$, a straightforward computation (see Lemma

[2] We are of course omitting numerous details of their work. In addition, their statement of their guarantee (see Theorem 1 in their paper) is also somewhat different from how we have stated it here: they have both a slightly weaker assumption on $\tilde{\mathbf{X}}$ ((m, η)-moment-distinguishability) and a slightly weaker conclusion on $\hat{E}$ (in terms of moment distance). We believe there is a mistake in their proof of the theorem, but nonetheless, their intermediate results are sufficient to prove the version that we have stated in the main text.

6.8.3) shows that both test matrices are multiples of the identity, namely

$$\mathbf{\Phi}_{\mathbf{g},\alpha} = (2\alpha+1)^{-1}\mathbf{I}_n \quad \text{and} \quad \mathbf{\Psi}_{\mathbf{g},\alpha} = \alpha(\alpha^2-1)^{-1}\mathbf{I}_n. \tag{6.6}$$

Our second test guarantees that the non-Gaussianness of $\mathbf{X}$ is captured by one of the test matrices, and moreover that their eigenvectors reveal the non-Gaussian directions of $\mathbf{X}$.

Theorem 6.2.1 (Second Gaussian test)**.** *Consider a random vector $\boldsymbol{X}$ which follows the isotropic NGCA model* (6.1)*. Then, for any $|\alpha|$ small enough, either $\mathbf{\Phi}_{\boldsymbol{X},\alpha}$ has an eigenvalue not equal to $(2\alpha+1)^{-1}$ or $\mathbf{\Psi}_{\boldsymbol{X},\alpha}$ has an eigenvalue not equal to $\alpha(\alpha^2-1)^{-1}$. Furthermore, all eigenvectors corresponding to such eigenvalues lie in E.*[3]

In Section 6.3, we will show how to derive the second Gaussian test from the first using a block diagonalization formula for each of the matrices $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$. Again, it is easy to see that $\mathbf{\Phi}_{\mathbf{X},\alpha}$ is not sufficient by itself to identify non-Gaussian directions: Take $\mathbf{X} = \|\mathbf{g}\|_2\boldsymbol{\theta}$ as before, and this time that assume that $\boldsymbol{\theta}$ is uniform on $\{\pm\mathbf{e}_i\}_{1=1}^{N}$. The symmetry implies that $\mathbf{\Phi}_{\mathbf{X},\alpha}$ is a scalar matrix, and computing its trace shows that it is equal to $(2\alpha+1)^{-1}\mathbf{I}_n$.

Both the first and second Gaussian tests for population rather than for finite samples; they involve taking expectations over the entire distribution of $\mathbf{X}$ which is typically unknown in practice. However, both tests are quite robust and work provably well on finite (polynomially large) samples. Robust versions of Gaussian tests can be formulated in terms of our definition of moment distance (see (6.3)).

Theorem 6.2.2 (First Gaussian test, robust)**.** *There is a universal constant $c > 0$ such that for each positive integer r, we have either*

$$|\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}| \geq c\eta_r^2/\tilde{\gamma}_r \quad \textit{or} \quad |\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}| \geq c\eta_r^2.$$

Here $\tilde{\gamma}_r = \mathbb{E}\{|g|^r\}$ is the r-th absolute moment of a standard Gaussian random variable, and $\eta_r = \min\{D_{\boldsymbol{X},r}, \tilde{\gamma}_r\}$.

Proof. If r is odd, then the statement follows from (5.26). If r is even, set $\delta = \frac{D_{\mathbf{X},r}}{\mathbb{E}\langle \mathbf{g},\mathbf{v}\rangle^r}$ in Lemma 5.5.3. □

There is a similar robust version of the second Gaussian test, which we will skip here but state and prove in Section 6.3.

[3] The matrix $\mathbf{\Phi}_{\mathbf{X},\alpha}$ always exists, but when $\tilde{\mathbf{X}}$ is not sub-Gaussian (i.e. can be rescaled so that marginals have tails lighter than a standard Gaussian), $\mathbf{\Psi}_{\mathbf{X},\alpha}$ may not be well-defined even for small α. In that case, $\|\tilde{\mathbf{X}}\|_2$ has a different distribution from $\|\mathbf{g}\|_2$, so that $\mathbf{\Phi}_{\mathbf{X},\alpha}$ has non-Gaussian eigenvalues. We can hence think of $\mathbf{\Phi}_{\mathbf{X},\alpha}$ as the primary test matrix, and $\mathbf{\Psi}_{\mathbf{X},\alpha}$ being an auxiliary that is only required in hard (effectively adversarial) cases.

Robustness allows us to use finite sample averages instead of expectations in the Gaussian tests, which is critical for practical applications. Indeed, consider a sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$, $\mathbf{X}'_1, \ldots, \mathbf{X}'_N$ of $2N$ i.i.d. realizations of a random variable $\mathbf{X}$. We can then define the sample versions of the test matrices in (6.5) in an obvious way:

$$\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} = \frac{1}{\hat{Z}_{\mathbf{\Phi}}} \sum_{i=1}^{N} e^{-\alpha\|\mathbf{X}_i\|_2^2} \mathbf{X}_i \mathbf{X}_i^T \quad \text{and} \quad \hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} = \frac{1}{\hat{Z}_{\mathbf{\Psi}}} \sum_{i=1}^{N} e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}'_i\rangle} (\mathbf{X}_i(\mathbf{X}'_i)^T + \mathbf{X}'_i\mathbf{X}_i^T), \tag{6.7}$$

with the normalizing quantities $\hat{Z}_{\mathbf{\Phi}} := \sum_{i=1}^{N} e^{-\alpha\|\mathbf{X}_i\|_2^2}$ and $\hat{Z}_{\mathbf{\Psi}} := 2\sum_{i=1}^{N} e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}'_i\rangle}$.

The second Gaussian test leads to the following straightforward algorithm for solving NGCA problem based on a finite sample: Use the sample to compute the test matrices $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$; select the eigenspaces corresponding to the eigenvalues that significantly deviate from the Gaussian eigenvalues. Then all vectors in both eigenspaces will be close to the non-Gaussian subspace E which we are trying to find. Let us state this algorithm and its guarantee precisely.

Algorithm 4 REWEIGHTED PCA($\mathbf{X}$,α_1,α_2,β_1,β_2)

```
Input: Data points $\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}'_1, \ldots, \mathbf{X}'_N$, scaling parameters $\alpha_1, \alpha_2 \in \mathbb{R}$, tolerance
    parameters $\beta_1, \beta_2 > 0$.
Output: Two estimates $\hat{E}_{\mathbf{\Phi}}$ and $\hat{E}_{\mathbf{\Psi}}$ for $E$.
  Compute test matrices $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha_2}$.
  Compute the eigenspace $\hat{E}_{\mathbf{\Phi}}$ of $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ corresponding to the nonzero eigenvalues that
    are farther than $\beta_1$ from the value $(2\alpha_1 + 1)^{-1}$.
  Compute the eigenspace $\hat{E}_{\mathbf{\Psi}}$ of $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha_2}$ corresponding to the nonzero eigenvalues that
    are farther than $\beta_2$ from the value $\alpha_2(\alpha_2^2 - 1)^{-1}$.
```

Theorem 6.2.3 (Finding one non-Gaussian direction)**.** *Let X be a sub-Gaussian*[4] *random vector which follows the isotropic NGCA model* (6.1)*, and with sub-Gaussian norm bounded above by* $K \geq 1$*. Let* r *be the minimum integer for which the* r*-th moment distance* $D_{\tilde{X},r} =: D > 0$*. Then for any* $\delta, \epsilon \in (0, 1)$*, with probability at least* $1 - \delta$*, if we run* Reweighted *PCA with a choice of parameters* $\alpha_1, \alpha_2, \beta_1, \beta_2$ *that is optimal up to constant multiples, at least one of* $\hat{E}_{\mathbf{\Phi}}$ *and* $\hat{E}_{\mathbf{\Psi}}$ *is non-trivial, and any unit vector in their union is* ϵ*-close to one in* E*, so long as the sample size* N *is greater than poly*$_r(n, 1/\epsilon, \log(1/\delta), 1/D, K)$*. Here, poly*$_r$ *is a polynomial whose total degree depends linearly on* r*.*

[4] For a formal definition of sub-Gaussian random vectors and an introduction to their properties, please see [114].

The idea of the proof is to use eigenvector perturbation theory from [34]. The robust version of the second Gaussian test exerts the existence of a gap between Gaussian and non-Gaussian eigenvalues. By bounding the deviation of the test matrix estimators $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$ from their expectation, we can thus show that their eigenstructures are similar. We will prove this theorem formally in Section 6.4.

The next step is to obtain a good estimate for the entire non-Gaussian subspace. To do so, we follow [112]'s strategy of projecting the sample points onto the orthogonal complement of the found directions, and recursing our algorithm on the new sample. After a set number of iterations, we collate all the found directions into a basis spanning candidate subspace $\hat{E}$. To state our guarantee for this procedure, we use the following notion of distance between subspaces. Note that this is equal to the $\sin(\mathbf{\Theta})$ distance of [34].

Definition 6.2.4 (Subspace distance)**.** Let F and F' be subspaces of $\mathbb{R}^n$ of dimensions m. Let $\mathbf{U}$ and $\mathbf{U}'$ be matrices whose columns form an orthonormal basis for F and F' respectively. The distance between F and F' is defined to be $d(F, F') := \|\mathbf{U}\mathbf{U}^T - \mathbf{U}'(\mathbf{U}')^T\|_F$.

Theorem 6.2.5 (Finding all non-Gaussian directions)**.** *Let $\boldsymbol{X}$ be a sub-Gaussian random vector which follows the isotropic NGCA model* (6.1)*, and with sub-Gaussian norm bounded above by $K \geq 1$. Suppose that $\tilde{\boldsymbol{X}}$ is (m, η)-moment-identifiable along all unit vectors $\boldsymbol{v} \in E$. Then running* Reweighted PCA *recursively (i.e. Algorithm 5) produces an estimate $\hat{E}$ such that $d(\hat{E}, E) < \epsilon$ so long as the sample size N is greater than poly$_{m,d}(n, 1/\epsilon, \log(1/\delta), 1/D, K)$. Here, poly$_{m,d}$ is a polynomial whose total degree depends linearly on m and d.*

We shall prove this theorem in Section 6.11. The theorem gives a polynomial time and sample complexity guarantee that REWEIGHTED PCA solves the NGCA problem, so long as m and d are assumed to be constants, while making exactly the same assumptions as [112]. This means that theoretically, both algorithms do just as well. On the other hand, REWEIGHTED PCA is a simple spectral algorithm, which is easier and faster to implement than local search.

Furthermore, while local search discovers one non-Gaussian direction at a time, the algorithm REWEIGHTED PCA possibly discovers multiple directions in each iteration. Most importantly, there is hope that all non-Gaussian directions can be discovered in the very first iteration. This is probably what will happen in practice with real data, and we may moreover prove that this is the case for special distributions. For instance, we can prove the following guarantee for finding a planted sphere.

Corollary 6.2.6 (Finding a sphere)**.** *Let $\tilde{\boldsymbol{X}}$ be uniformly distributed on the scaled unit sphere $\sqrt{d}S^{d-1}$ in E. Suppose we are given a sample of size $N \gtrsim dn^2(n + \log(1/\delta))/\epsilon^2$, then* *running the first two steps of* REWEIGHTED PCA *with a choice of* $\alpha \in [c_1/n, c_2/n]$*, and* $\beta = \alpha/3$ *yields a subspace* $\hat{E}_{\mathbf{\Phi}}$ *so that* $d(\hat{E}_{\mathbf{\Phi}}, E) < \epsilon$*. Here,* c_1 *and* c_2 *are absolute constants.*

6.2.1 Reweighted PCA in other contexts

The name of the algorithm stems from the first test matrix, which can be seen as a PCA matrix for the reweighted sample obtained when each point $\mathbf{X}_i$ is given the weight $e^{-\alpha\|\mathbf{X}_i\|_2^2}$. As mentioned in the previous section, $\mathbf{\Phi}_{\mathbf{X},\alpha}$ reveals at least one non-Gaussian direction in all but adversarial situations, and so can be considered the primary test matrix.

The idea of doing PCA with weight functions that are non-linear in the sample points can be traced back at least as far as [18]. In that paper, the authors similarly use Gaussian weights, but do so in order to handle clustering for Gaussian mixture models that are highly non-spherical. In a later paper, [48] used Fourier weights to handle ICA. While our analysis is radically different, the idea for the algorithm was directly inspired by these two papers.

6.2.2 Organization of chapter and notation

In Section 6.3, we will prove the second Gaussian test and state a robust version needed for proving our guarantee for Reweighted PCA. The guarantee for finding one direction is proved in Section 6.4. The guarantees for finding all directions, and the special case of finding a sphere are proved in Sections 6.11 and 6.12 respectively. To enhance the flow of the chapter, many technical details are also deferred to the later sections. Throughout the chapter, scalars are denoted in standard font, while vectors and matrices are denoted with bold font. C and c denote absolute constants whose value may change from line to line. We let $\mathbf{g}_n$ denote the standard Gaussian vector in $\mathbb{R}^n$. The subscript is omitted whenever the dimension is obvious. In addition, for each r, we let γ_r and $\tilde{\gamma}_r$ denote the r-th moment and r-th absolute moment of a standard Gaussian random variable.

6.3 Proof of the second Gaussian test

In this section, we return to the setting where $\mathbf{X}$ follows the NGCA model (6.1). We further assume that the non-Gaussian component $\tilde{\mathbf{X}}$ is a sub-Gaussian random vector with sub-Gaussian norm bounded by K. In order not to break the flow of the chapter, most of the proofs are deferred to Section 6.7.

The first step in proving the test is to notice that the independence of the Gaussian and non-Gaussian components allows us to block diagonalize the test matrices.

Lemma 6.3.1 (Block diagonalization for $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$)**.** *Assume E is spanned by the first d basis vectors. Then the test matrices $\mathbf{\Phi}_{\boldsymbol{X},\alpha}$ and $\mathbf{\Psi}_{\boldsymbol{X},\alpha}$ decompose into blocks in the following manner:*

$$\mathbf{\Phi}_{\boldsymbol{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha} & 0 \\ \hline 0 & \mathbf{\Phi}_{\boldsymbol{g},\alpha} \end{array}\right), \quad \mathbf{\Psi}_{\boldsymbol{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha} & 0 \\ \hline 0 & \mathbf{\Psi}_{\boldsymbol{g},\alpha} \end{array}\right). \tag{6.8}$$

We then observe that the trace of the test matrices are conveniently equal to the negated log derivatives of their respective partition functions.

Lemma 6.3.2 (Trace of $\mathbf{\Phi}_{\mathbf{Y},\alpha}$ and $\mathbf{\Psi}_{\mathbf{Y},\alpha}$)**.** *Let **Y** be any random vector in $\mathbb{R}^n$. Then* $\mathrm{Tr}(\mathbf{\Phi}_{\boldsymbol{Y},\alpha}) = -(\log Z_{\mathbf{\Phi},\boldsymbol{Y}})'(\alpha)$ *and* $\mathrm{Tr}(\mathbf{\Psi}_{\boldsymbol{Y},\alpha}) = -(\log Z_{\mathbf{\Psi},\boldsymbol{Y}})'(\alpha)$.

Our next lemma shows that for α small enough, the partition functions themselves differentiate between Gaussian and non-Gaussian random vectors. This is obvious once we realize that they are just the moment generating functions of $\|\mathbf{X}\|_2^2$ and $\langle \mathbf{X}, \mathbf{X}' \rangle$, and that these are analytic in a small neighborhood around 0.

Lemma 6.3.3 (Partition functions characterize Gaussian distributions)**.** *The following hold for any sub-Gaussian random vector **Y**:*

*a) If $Z_{\mathbf{\Phi},\boldsymbol{Y}}(\alpha_k) = Z_{\mathbf{\Phi},\boldsymbol{g}}(\alpha_k)$ for a sequence of values α_k converging to 0, then **Y** has the same radial distribution as **g**.*

*b) If in addition, $Z_{\mathbf{\Psi},\boldsymbol{Y}}(\beta_k) = Z_{\mathbf{\Psi},\boldsymbol{g}}(\beta_k)$ for a sequence of values β_k converging to 0, then **X** has the standard Gaussian distribution.*

We are now in a position to prove the second Gaussian test.

Proof of Theorem 6.2.1. Let $\mathbf{g}_d$ denote the standard Gaussian in $\mathbb{R}^d$. By Lemma 6.3.3, either $Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha) \neq Z_{\mathbf{\Phi},\mathbf{g}_d}(\alpha)$ for $|\alpha|$ small enough, or $Z_{\mathbf{\Psi},\tilde{\mathbf{X}}}(\alpha) \neq Z_{\mathbf{\Psi},\mathbf{g}_d}(\alpha)$ for $|\alpha|$ small enough. As such, either $(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Phi},\mathbf{g}_d})'(\alpha)$ or $(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Psi},\mathbf{g}_d})'(\alpha)$. Assume the former holds, and let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha}$. Since we may write $\mathbf{\Phi}_{\mathbf{X},\alpha}$ in a block form, these eigenvalues are either those of $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}$ or $\mathbf{\Phi}_{\mathbf{g},\alpha}$. Without loss of generality, we may assume that $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}$, and $\lambda_{d+1}, \ldots, \lambda_n$ are those of $\mathbf{\Phi}_{\mathbf{g},\alpha}$.

Lemma 6.8.3 tells us that $\lambda_{d+1} = \cdots = \lambda_n = (2\alpha + 1)^{-1}$. On the other hand, by Lemma 6.3.2,

$$\sum_{i=1}^{d} \lambda_i = \mathrm{Tr}(\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}) = -(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'(\alpha).$$

By Lemma 6.8.2, $-(\log Z_{\mathbf{\Phi},\mathbf{g}_d})'(\alpha) = d(2\alpha+1)^{-1}$, so we have $\sum_{i=1}^{d} \lambda_i \neq d(2\alpha+1)^{-1}$. Dividing through by d, we get $\frac{1}{d}\sum_{i=1}^{d} \lambda_i \neq (2\alpha+1)^{-1}$, which implies that at least one λ_i differs from this value for $1 \leq i \leq d$.

If it were the case that $(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Psi},\mathbf{g}_d})'(\alpha)$, a similar argument involving $\mathbf{\Psi}_{\mathbf{X},\alpha}$ gives the alternate conclusion. □

It is tedious but not too difficult to make the second Gaussian test quantitative. We do this by tracking how the non-Gaussian moments for $\|\tilde{\mathbf{X}}\|_2$ and $\langle \tilde{\mathbf{X}}, \tilde{\mathbf{X}}' \rangle$ contribute to the power series expansions for $-(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'$ and $-(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'$ around 0. This yields the following theorem.

Theorem 6.3.4 (Second Gaussian test, robust)**.** *Let r be the integer such that $D_{\tilde{\mathbf{X}},r} > 0$ and $D_{\tilde{\mathbf{X}},r'} = 0$ for all $r' < r$. Then either*

a) for $|\alpha| \leq \eta_r^2 r/(CK^2)^r(d^{r+1} + (r+1)!)$, we have

$$\left| \frac{1}{d}\sum_{i=1}^{d} \lambda_i(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{\alpha}{\alpha^2 - 1} \right| \geq \frac{c\eta_r^2}{d(r-1)!}|\alpha|^{r-1}, \tag{6.9}$$

b) or for $|\alpha| \leq \eta_r^2 r/(CK^2)^{r/2}\tilde{\gamma}_r(d^{r/2+1} + (r/2+1)!)$, we have

$$\left| \frac{1}{d}\sum_{i=1}^{d} \lambda_i(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{1}{2\alpha+1} \right| \geq \frac{c\eta_r^2}{d(r/2-1)!\tilde{\gamma}_r}|\alpha|^{r/2-1}. \tag{6.10}$$

Here $\tilde{\gamma}_r = \mathbb{E}|\langle \boldsymbol{g}, \boldsymbol{v}\rangle|^r$ for an arbitrary vector $v \in S^{n-1}$ and $\eta_r = \min\{D_{\boldsymbol{X},r}, \tilde{\gamma}_r\}$.

6.4 Proof of guarantee for Reweighted PCA

The second Gaussian test tells us how we can recover non-Gaussian directions from $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$. Our guarantee for Reweighted PCA algorithm shows that we can do the same with the plug-in estimators $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$. To this end, we first provide concentration bounds for these estimators, whose proofs can be found in Section 6.9.

Theorem 6.4.1 (Concentration for $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$)**.** *There is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, and any $0 \leq \alpha < 1/CK^2 n$, we have $\mathbb{P}\Big\{\|\hat{\mathbf{\Phi}}_{\boldsymbol{X},\alpha} - \mathbf{\Phi}_{\boldsymbol{X},\alpha}\| > \epsilon\Big\} \leq \delta$ so long as $N \geq CK^2(n + \log(1/\delta))\epsilon^{-2}$.*

Theorem 6.4.2 (Concentration for $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$)**.** *There is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, if $N \geq CK^2(n + \log(1/\delta))\epsilon^{-2}$ and $|\alpha| \leq 1/CK^2\tau(n+\tau)$, we have $\mathbb{P}\Big\{\|\hat{\mathbf{\Psi}}_{\boldsymbol{X},\alpha} - \mathbf{\Psi}_{\boldsymbol{X},\alpha}\| > \epsilon\Big\} \leq \delta$. Here, $\tau = \log^{1/2}(N/\min\{\delta, K\epsilon\})$.*

Lemma 6.4.3 (Guarantee for $\hat{E}_{\mathbf{\Phi}}$). *Suppose the moments of $\|\tilde{\boldsymbol{X}}\|_2^2$ and $\|\boldsymbol{g}_d\|_2^2$ agree up to order $r-1$, but there is a number $\Delta > 0$ such that $\left|\mathbb{E}\{\|\tilde{\boldsymbol{X}}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}_d\|_2^{2r}\}\right| \geq \Delta$. For any $\delta, \epsilon \in (0,1)$, pick α_1 such that $0 < \alpha_1 < \min\{\Delta r/(CK^2)^r(d^{r+1} + (r+1)!), 1/CK^2 n\}$, and $\beta_1 = \Delta\alpha_1^{r-1}/4d(r-1)!$. Then with probability at least $1-\delta$,* R*eweighted PCA with $2N \geq CK^2 d^{3/2}(n + \log(1/\delta))/\beta_1^2\epsilon^2$ samples together with this choice of α_1 and β_1 produces a nontrivial estimate $\hat{E}_{\mathbf{\Phi}}$ of dimension $1 \leq \hat{d}_{\mathbf{\Phi}} \leq d$, such that there is a $\hat{d}_{\mathbf{\Phi}}$-dimensional subspace $E_{\mathbf{\Phi}} \subset E$ satisfying $d(\hat{E}_{\mathbf{\Phi}}, E_{\mathbf{\Phi}}) \leq \epsilon$.*

Proof. Combining Lemmas 6.3.1, 6.3.2, and 6.8.3 tells us that in the right coordinates, $\mathbf{\Phi}_{\mathbf{X},\alpha}$ block diagonalizes as

$$\mathbf{\Phi}_{\mathbf{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha} & 0 \\ \hline 0 & (2\alpha+1)^{-1}\mathbf{I}_{n-d}, \end{array}\right). \tag{6.11}$$

Next, label the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We can find $0 \leq p \leq q \leq n$ such that the eigenvalues corresponding to the $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha_1}$ block are $\lambda_1, \lambda_2, \ldots, \lambda_p, \lambda_{q+1}, \ldots, \lambda_n$. Using Theorem 6.7.4, we then have

$$\left|\frac{1}{d}\left(\sum_{i=1}^{p} \lambda_i + \sum_{i=q+1}^{n} \lambda_i\right) - \frac{1}{2\alpha_1+1}\right| \geq \frac{\Delta}{2d(r-1)!}\alpha_1^{r-1} = 2\beta_1. \tag{6.12}$$

In particular, we have $\frac{1}{p}\sum_{i=1}^{p}\lambda_i - 1/(2\alpha_1+1) \geq 2\beta_1$, and $1/(2\alpha_1+1) - \frac{1}{n-q}\sum_{i=q+1}^{n}\lambda_i \geq 2\beta_1$. Since at least one of these sums of eigenvalues is non-empty, truncating the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ at the β_1 level gives us a non-trivial subspace of E.

In order to show that our empirical estimate $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ also has an approximation to this property, we will need to use the eigenvector perturbation theory explained in Section 6.10. First, we need to bound from below the "eigengap" in $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$. Suppose first that $p \geq 1$, i.e. that there are eigenvalues larger than $(2\alpha_1+1)^{-1}$. Then by the pigeonhole principle, one can find i such that $(2\alpha_1+1)^{-1} + \beta_1/2 \geq \lambda_{i+1} \geq (2\alpha_1+1)^{-1}$ and $\lambda_i - \lambda_{i+1} \geq \beta_1/2d$. Similarly, if $q \leq n-1$, then we can find j such that $(2\alpha_1+1)^{-1} \geq \lambda_{j-1} \geq (2\alpha_1+1)^{-1} - \beta_1/2$ and $\lambda_{j-1} - \lambda_j \geq \beta_1/2d$.

Now let F be the span of the eigenvectors of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ corresponding to $\lambda_1, \ldots, \lambda_i, \lambda_j, \ldots, \lambda_n$, and let $\hat{F}$ be the eigenvectors of $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ corresponding to $\hat{\lambda}_1, \ldots, \hat{\lambda}_i, \hat{\lambda}_j, \ldots, \hat{\lambda}_n$. By Theorem 6.4.1, with probability at least $1-\delta$, we have

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq \frac{\beta_1\epsilon}{4\sqrt{2}d^{3/2}}. \tag{6.13}$$

We may then use Theorem 6.10.4 to see that $d(\hat{F}, F) \leq \epsilon$.

We are not yet done, because we do not have access to $\hat{F}$. Nonetheless, we can show that $\hat{F}$ contains $\hat{E}_{\mathbf{\Phi}}$. Using eigenvalue perturbation inequalities together with equation (6.13) tells us that we have

$$\hat{\lambda}_{i+1} \leq \lambda_{i+1} + \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} + \frac{\beta_1}{2} + \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} + \frac{2\beta_1}{3}, \tag{6.14}$$

and similarly that

$$\hat{\lambda}_{j-1} \leq \lambda_{j-1} - \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} - \frac{\beta_1}{2} - \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} - \frac{2\beta_1}{3}. \tag{6.15}$$

Let $\hat{I}_{\mathbf{\Phi}} = \{i : |\hat{\lambda}_i - (1 - 2\alpha_1)^{-1}| > \beta_1\}$. We see that this set does not contain any index between $i + 1$ and $j - 1$, so $\hat{E}_{\mathbf{\Phi}}$, which comprises the span of the eigenvectors to these eigenvalues, does not contain any eigenvector that $\hat{F}$ does not contain, as was to be shown. The inclusion then implies that we may find a subspace $E_{\mathbf{\Phi}} \subset F$ such that $d(\hat{E}_{\mathbf{\Phi}}, E_{\mathbf{\Phi}}) \leq \epsilon$.

Finally, we observe that $\dim \hat{E}_{\mathbf{\Phi}} \geq 1$, since

$$\frac{1}{p}\sum_{i=1}^{p} \hat{\lambda}_i - \frac{1}{2\alpha_1 + 1} \geq \frac{1}{p}\sum_{i=1}^{p} \lambda_i - \frac{\beta_1 \epsilon}{2d} - \frac{1}{2\alpha_1 + 1} > \beta_1, \tag{6.16}$$

and

$$\frac{1}{2\alpha_1 + 1} - \frac{1}{n - q}\sum_{i=q+1}^{n} \hat{\lambda}_i \geq \frac{1}{2\alpha_1 + 1} - \frac{1}{n - q}\sum_{i=q+1}^{n} \lambda_i - \frac{\beta_1 \epsilon}{2d} > \beta_1. \tag{6.17}$$

□

Lemma 6.4.4 (Guarantee for $\hat{E}_{\mathbf{\Psi}}$)**.** *Suppose the moments of $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ and $\langle \boldsymbol{g}, \boldsymbol{g}' \rangle$ agree up to order $r - 1$ but $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}' \rangle^r\}| \geq \Delta$. For any $\delta, \epsilon, \tau \in (0, 1)$, pick $0 < \alpha_2 < \min\{\Delta r/(CK^2)^r(d^{r+1} + (r + 1)!), 1/CK^2 n^{1+\tau}\}$, and $\beta_2 = \Delta \alpha_2^{r-1}/4d(r - 1)!$. Then with probability at least $1 - \delta$,* R*eweighted PCA with sample size $2N$ satisfying $\exp(n^{2\tau}) \min\{\delta, K\epsilon\} \geq 2N \geq CK^2 d^{3/2}(n + \log(1/\delta))/\beta_2^2 \epsilon^2$, together with this choice of α_2 and β_2 produces a nontrivial estimate $\hat{E}_{\mathbf{\Psi}}$ of dimension $1 \leq \hat{d}_{\mathbf{\Psi}} \leq d$, such that there is a $\hat{d}_{\mathbf{\Psi}}$-dimensional subspace $E_{\mathbf{\Psi}} \subset E$ satisfying $d(\hat{E}_{\mathbf{\Psi}}, E_{\mathbf{\Psi}}) \leq \epsilon$.*

Proof. The proof is completely analogous to that for the previous theorem, except that we replace our estimates and identities for $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ with those for $\mathbf{\Psi}_{\mathbf{X},\alpha_2}$ wherever necessary. □

Proof of Theorem 6.2.3. Combine the last two lemmas with Theorem 6.3.4 from the last section. □

Remark 6.4.5 (Selecting optimal parameters). If the problem parameters d, n, r, K and $D_{\tilde{\mathbf{X}},r}$ were known before hand, then in principle, one could compute the optimal tuning parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$. In practice, however, one rarely is in this situation, so one would have to estimate the problem parameters as a first step to solving the NGCA problem. Nonetheless, one can do this by the doubling/halving trick. In other words, we start with some fixed initial choice of α_1 and α_2. Using Theorems 6.4.1 and 6.4.2, we can detect whether there are any outlier eigenvalues with high probability. If there are none, we halve α_1 and α_2 and try again, repeating this process until outliers show up. The number of iterations is then the base 2 logarithm of the final α_1 and α_2, plus an additive constant. This is at most polynomial in all the problem parameters, so the algorithm remains efficient.

6.5 Comments and open questions

We have presented and analyzed an algorithm that is guaranteed to return at least one non-Gaussian direction efficiently, with sample and time complexity a polynomial in the problem parameters for a fixed r, where r is the smallest order at which $\tilde{\mathbf{X}}$ has positive r-th moment distance from a standard Gaussian. Furthermore, if $\tilde{\mathbf{X}}$ is (m, η)-moment-identifiable, then the algorithm estimates the d-dimensional non-Gaussian subspace efficiently with polynomial time and sample complexity for fixed m and d.

Since the degree of the polynomial increases linearly in r, it would seem that the algorithm is practically useless if r is larger than a small constant. However, note that having all third and fourth moments equal those of a Gaussian is a condition that is already stringent in one dimension, and which becomes even more so in higher dimensions. As such, unless $\tilde{\mathbf{X}}$ has some kind of adversarial distribution, r will be either 4 or 3, depending on whether $\tilde{\mathbf{X}}$ is centrally symmetric or not.

The algorithm also often delivers much more than is guaranteed for several reasons. First, in order to bound the subspace perturbation by ϵ, we used a very crude estimate of the eigengap, bounding it from below using the pigeonhole principle, which in the worst case assumes that the eigenvalues are spread out at regular intervals. This should not happen in practice, and we expect the non-Gaussian eigenvalues to instead cluster relatively tightly around their average. If this happens, the sample complexity requirement can be relaxed by a factor of d.

Second, just as it is extremely unlikely for r to be higher than 4, for a general non-Gaussian $\tilde{\mathbf{X}}$ and a small, random α, it is extremely unlikely for any of the non-Gaussian values of $\mathbf{\Phi}_{\mathbf{X},\alpha}$ to be equal to the Gaussian one on the dot. This means that even though the guarantee for a single run of the base algorithm is for one direction, in practice we most probably can recover the entire subspace E simultaneously with just $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ alone (as in the case in Corollary 6.2.6), albeit with a more sophisticated truncation technique.

6.5.1 Conjectures

We conjecture that REWEIGHTED PCA actually recovers the entire non-Gaussian subspace E with in polynomial time and sample complexity if we fix m, but now allow d to vary. This would improve upon both our result and that of [112]. The first Gaussian test for a random vector $\mathbf{X}$ using the distribution of its norm and dot product pairing also leads to further questions. For a fixed nonzero real number t, both of these appear in the formula for $\|\mathbf{Y}_t\|_2^2$, where we set $\mathbf{Y}_t := \mathbf{X} + t\mathbf{X}'$, so it is natural to ask whether Reweighted PCA works with $\mathbf{\Phi}_{\mathbf{Y}_t,\alpha}$ alone for some t. In particular, does it work for $t = -1$? It is also an open question whether $\langle \mathbf{X}, \mathbf{X}' \rangle$ alone is sufficient to test whether $\mathbf{X}$ is standard Gaussian.

6.6 Equivalence of NGCA models

In this section, we note the equivalence of several formulations of the NGCA model used in the literature. First, the isotropic NGCA model (6.1.1) can be written equivalently as

$$F(\mathbf{x}) = H(\mathbf{P}_E(\mathbf{x}))G(\mathbf{P}_{E^\perp}(\mathbf{x})),$$

where F is the distribution of $\mathbf{X}$, H is the distribution of $\tilde{\mathbf{X}}$, and G is the standard normal distribution. This is the way in which [112] stated the NGCA model.

Next, consider the model

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{g},$$

where now $\tilde{\mathbf{X}} \in E$ as before, but $\mathbf{g}$ is a centered Gaussian in $\mathbb{R}^n$ with arbitrary covariance. As a special case of this, we have $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{g}) \in E \oplus E'$, where E and E' are complementary but not necessarily orthogonal. Let $\mathbf{\Sigma} = \text{Cov}(\mathbf{X})$, and consider the whitened distribution $\mathbf{\Sigma}^{-1/2}\mathbf{X} = \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}} + \mathbf{\Sigma}^{-1/2}\mathbf{g}$. Now the non-Gaussian subspace is $\mathbf{\Sigma}^{-1/2}E$, which we assume without loss of generality to be the span of the first d coordinate vectors. This means that $\text{Cov}(\mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}})$ only has nonzero entries in its top left d by d block. Since we can decompose

$$\mathbf{I}_n = \text{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{X}) = \text{Cov}(\mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}}) + \text{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{g}),$$

this in turn implies that we can write

$$\mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{g}) = \left(\begin{array}{c|c} \mathbf{A} & 0 \\ \hline 0 & \mathbf{I}_{n-d} \end{array}\right),$$

where $\mathbf{A}$ is a PSD matrix such that $\mathbf{A} = \mathbf{I}_d - \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}}$. Because of this structure, we have $\mathbf{\Sigma}^{-1/2}\mathbf{g} = (\tilde{\mathbf{h}}, \mathbf{h}) \in E \oplus E^{\perp}$, with $\tilde{\mathbf{h}} \sim \mathcal{N}(\mathbf{0}, \mathbf{A})$ and $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-d})$. Since these two Gaussian components have zero correlation, they are independent. Since a non-Gaussian distribution remains non-Gaussian after convolution with a Gaussian, if we set $\tilde{\mathbf{Y}} := \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}} + \tilde{\mathbf{h}}$ to be our new non-Gaussian component, we see that we have again produced an instance of (6.1.1).

This additive model seems to be the most common formulation of NGCA in the literature (see [16, 64], etc.). It can also be equivalently written as

$$F(\mathbf{x}) = H(\mathbf{P}_E(\mathbf{x}))G(\mathbf{x}), \tag{6.18}$$

where G is now a centered Gaussian density with arbitrary covariance, and H is now just some function. See Lemma 1 in [16] for more details.

6.7 Details for Section 6.3

Proof of Lemma 6.3.1. The decompositions follow easily from the independence of the two components of the mixed vector, $\tilde{\mathbf{X}}$ and $\mathbf{g}$, as well as the unconditional symmetry of the Gaussian component. Let us illustrate this by proving the decomposition for $\mathbf{\Phi}_{\mathbf{X},\alpha}$. First, note that $e^{-\alpha\|\mathbf{X}\|_2^2} = e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2} e^{-\alpha\|\mathbf{g}\|_2^2}$, so that $Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) = Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha) Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)$. The top left d by d block is hence given by

$$\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha)} = \mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}.$$

The bottom right d' by d' block is also computed similarly. Finally, any entry outside these two blocks is of the form

$$\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i\mathbf{g}_j\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i(-\mathbf{g}_j)\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = -\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i\mathbf{g}_j\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = 0.$$

□

Proof of Lemma 6.3.2. We have

$$\mathrm{Tr}(\boldsymbol{\Phi}_{\mathbf{X},\alpha}) = \frac{\mathbb{E}\|\mathbf{X}\|_2^2 e^{-\alpha\|\mathbf{X}\|_2^2}}{\mathbb{E}e^{-\alpha\|\mathbf{X}\|_2^2}} = \frac{-Z'_{\boldsymbol{\Phi},\mathbf{X}}(\alpha)}{Z_{\boldsymbol{\Phi},\mathbf{X}}(\alpha)} = -(\log Z_{\boldsymbol{\Phi},\mathbf{X}})'(\alpha).$$

The calculation for $\boldsymbol{\Psi}_{\mathbf{X},\alpha}$ is similar. □

In order to prove Lemma 6.3.3, we first need to establish the analyticity for the two partition functions.

Lemma 6.7.1 (Analyticity for $Z_{\boldsymbol{\Phi},\mathbf{X}}$ and $Z_{\boldsymbol{\Psi},\mathbf{X}}$)**.** *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded by $K \geq 1$. The functions $Z_{\boldsymbol{\Phi},\boldsymbol{X}}$ and $Z_{\boldsymbol{\Psi},\boldsymbol{X}}$ are both analytic on $(-1/CK^2, 1/CK^2)$. They are given by the formulae $Z_{\boldsymbol{\Phi},\boldsymbol{X}}(\alpha) = \sum_{r=0}^{\infty}\mathbb{E}\{\|\boldsymbol{X}\|_2^{2r}\}(-\alpha)^r/r!$ and $Z_{\boldsymbol{\Psi},\boldsymbol{X}}(\alpha) = \sum_{r=0}^{\infty}\mathbb{E}\{\langle\boldsymbol{X},\boldsymbol{X}'\rangle^r\}(-\alpha)^r/r!$. Furthermore, by choosing C sufficiently large, on this interval they satisfy the bounds*

$$|Z_{\boldsymbol{\Phi},\boldsymbol{X}}(\alpha)|, |Z_{\boldsymbol{\Psi},\boldsymbol{X}}(\alpha)| \leq e^{CK^2n|\alpha|} + \frac{CK^2|\alpha|}{1-CK^2|\alpha|}. \tag{6.19}$$

Proof. Let us first prove the bounds in (6.19). Observe that

$$\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\} \leq \mathbb{E}\{e^{|\alpha|\|\mathbf{X}\|_2^2}\} = \sum_{n=0}^{\infty}\frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2n}\}}{n!}|\alpha|^n. \tag{6.20}$$

Here, Tonelli allows us to interchange the sum and expectation. We next use Lemma 2.3.1 to bound the terms of this series. Indeed, using the equivalent estimate (2.5), we have

$$\mathbb{E}\{\|\mathbf{X}\|_2^{2r}\} \leq C^r K^{2r}(n^r + r!)$$

for some universal constant C. Substituting this into (6.20) and using $|\alpha| \leq 1/CK^2$, we have

$$\begin{aligned}\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\} &\leq \sum_{r=0}^{\infty}\frac{(CK^2)^r(n^r+r!)}{r!}|\alpha|^r \\ &= \sum_{r=0}^{\infty}\frac{(CK^2n|\alpha|)^r}{r!} + \sum_{r=1}^{\infty}(CK^2|\alpha|)^r \\ &= e^{CK^2n|\alpha|} + \frac{CK^2|\alpha|}{1-CK^2|\alpha|}.\end{aligned}$$

One may prove the bound for $Z_{\boldsymbol{\Psi},\mathbf{X}}$ by doing the same computation but using (2.2) instead of (2.1).

We next handle analyticity of $Z_{\mathbf{\Phi},\mathbf{X}}$. We shall prove by induction on r that we may differentiate under the integral sign to get the formula

$$Z^{(r)}_{\mathbf{\Phi},\mathbf{X}}(\alpha) = (-1)^r \mathbb{E}\{\|\mathbf{X}\|_2^{2r} e^{-\alpha\|\mathbf{X}\|_2^2}\}. \tag{6.21}$$

Assume the formula is true for all $r' < r$. Then

$$\begin{align} Z^{(r)}_{\mathbf{\Phi},\mathbf{X}}(\alpha) &= (-1)^{r-1} \lim_{h\to 0} \mathbb{E}\{\frac{\|\mathbf{X}\|_2^{2r-2} e^{-(\alpha+h)\|\mathbf{X}\|_2^2} - \|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2}}{h}\} \tag{6.22}\\ &= (-1)^r \lim_{h\to 0} \mathbb{E}\{\|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2} \frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h}\} \tag{6.23}\end{align}$$

Next, note that the integrand is positive and by the mean value theorem, for a fixed value of $\|\mathbf{X}\|_2^2$, we have

$$\frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h} = \|\mathbf{X}\|_2^2 e^{-h'\|\mathbf{X}\|_2^2}$$

for some $h' \in [0, h]$ if $h > 0$ and $h' \in [h, 0]$ otherwise. As such, we have

$$\|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2} \frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h} \leq \|\mathbf{X}\|_2^{2r} e^{(|h|-\alpha)\|\mathbf{X}\|_2^2}$$

For $|h| - \alpha \leq 1/CK^2$, one can easily show that this is integrable by expanding this as a power series in $\|\mathbf{X}\|_2^2$ and bounding the growth of the coefficients as above. As such, we may apply the Dominated Convergence Theorem to push the limit inside the expectation in (6.22), thereby yielding (6.21).

In particular, differentiating $Z_{\mathbf{\Phi},\mathbf{X}}$ at 0, we see that its Taylor series at 0 is given by

$$Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) \sim \sum_{r=0}^{\infty} \frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2r}\}}{r!}(-\alpha)^r. \tag{6.24}$$

The formula above shows that the Taylor series is absolutely convergent on our chosen interval. We next need to show that $Z_{\mathbf{\Phi},\mathbf{X}}$ agrees with its Taylor series on this interval, meaning we have to show that the remainder term for the r-th Taylor polynomial goes to zero pointwise. The Lagrange form of the remainder term is written as

$$R_{Z_{\mathbf{\Phi},\mathbf{X}},r}(\alpha) = \frac{Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}(\alpha')}{(r+1)!}\alpha^{r+1}$$

where $0 < |\alpha'| < |\alpha|$. Applying Cauchy-Schwarz to the formula (6.21), we get

$$|Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}(\alpha)| \le \big(\mathbb{E}\{\|\mathbf{X}\|_2^{4r+2}\}\big)^{1/2}\Big(\mathbb{E}\{e^{-2\alpha\|\mathbf{X}\|_2^2}\}\Big)^{1/2}. \tag{6.25}$$

Lemma 2.3.1 again allows us to compute

$$\big(\mathbb{E}\{\|\mathbf{X}\|_2^{4r+2}\}\big)^{1/2} \le (CK^2)^{r+1}(n^{r+1} + (r+1)!).$$

This implies that for any $C' > 2C$,

$$\begin{aligned}
\|R_{Z_{\mathbf{\Phi},\mathbf{X}},r}\|_{L_\infty([-1/C'K^2,1/C'K^2])} &\le \frac{\|Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}\|_{L_\infty([-1/C'K^2,1/C'K^2])}}{(C'K^2)^{r+1}(r+1)!} \\
&\le \frac{(CK^2)^{r+1}(n^{r+1}+(r+1)!)}{(C'K^2)^{r+1}(r+1)!}\Big(\mathbb{E}\{e^{2\|\mathbf{X}\|_2^2/C'K^2}\}\Big)^{1/2} \\
&\le \left(\frac{C}{C'}\right)^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\left(e^{2nC/C'} + \frac{2C/C'}{1-2C/C'}\right).
\end{aligned}$$

Using the fact that $r! \sim \left(\frac{r}{e}\right)^r$, this last expression decays to zero as r tends to ∞. Finally, to prove the claim for $Z_{\mathbf{\Psi},\mathbf{X}}$, we repeat the same arguments. □

Note that in the course of proving the last lemma, we have also proved the following result to be used elsewhere in the chapter.

Lemma 6.7.2 (Taylor remainder terms for $Z_{\mathbf{\Phi},\mathbf{X}}$ and $Z_{\mathbf{\Psi},\mathbf{X}}$)**.** *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded above by $K \ge 1$. There is an absolute constant C such that for all $0 < \alpha < 1/CK^2$, on the interval $[-\alpha, \alpha]$, the remainder terms for the r-th degree Taylor polynomials for $Z_{\mathbf{\Phi},\boldsymbol{X}}$ and $Z_{\mathbf{\Psi},\boldsymbol{X}}$ at 0 satisfy the uniform bound*

$$\|R_{Z_{\mathbf{\Phi},X},r}\|_\infty, \|R_{Z_{\mathbf{\Psi},X},r}\|_\infty \le (CK^2)^{r+1}\alpha^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\left(e^{CK^2\alpha n} + \frac{CK^2\alpha}{1-CK^2\alpha}\right) \tag{6.26}$$

Proof of Lemma 6.3.3. By Lemma 6.7.1, all four functions are analytic in a neighborhood of 0. Now recall that two different analytic functions cannot agree on a sequence with an accumulation point. □

We now move on to proving Theorem 6.3.4. This requires the following technical lemma.

Lemma 6.7.3. *Let $\boldsymbol{X}$ be sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded above by $K \ge 1$. Suppose the moments of $\|\boldsymbol{X}\|_2^2$ and $\|\boldsymbol{g}\|_2^2$ agree up to order*

$r-1$, *but there is a number* $\Delta > 0$ *such that* $\left|\mathbb{E}\{\|\boldsymbol{X}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^{2r}\}\right| \geq \Delta$*, then there is an absolute constant* C *such that for* $|\alpha| \leq \Delta r/(CK^2)^r(n^{r+1} + (r+1)!)$*, we have*

$$|(\log Z_{\boldsymbol{\Phi},\boldsymbol{X}})'(\alpha) - (\log Z_{\boldsymbol{\Phi},\boldsymbol{g}})'(\alpha)| \geq \frac{\Delta}{2(r-1)!}|\alpha|^{r-1}. \tag{6.27}$$

Similarly, suppose the moments of $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ *and* $\langle \boldsymbol{g}, \boldsymbol{g}' \rangle$ *agree up to order* $r-1$ *but* $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}' \rangle^r\}| \geq \Delta$*, then for* $|\alpha| \leq \Delta r/(CK^2)^r(n^{r+1} + (r+1)!)$*, we have*

$$|(\log Z_{\boldsymbol{\Psi},\boldsymbol{X}})'(\alpha) - (\log Z_{\boldsymbol{\Psi},\boldsymbol{g}})'(\alpha)| \geq \frac{\Delta}{2(r-1)!}|\alpha|^{r-1}. \tag{6.28}$$

Proof. Let us first prove (6.27). For every positive integer k, let

$$p_{\mathbf{X},k}(\alpha) = \sum_{j=0}^{k} \mathbb{E}\{\|\mathbf{X}\|_2^{2j}\}\alpha^j/j!$$

denote the k-th Taylor polynomial of $Z_{\mathbf{\Phi},\mathbf{X}}$, and define $p_{\mathbf{g},k}$ analogously. For convenience, also denote the k-th Taylor remainder term as $R_{\mathbf{X},k} := R_{Z_{\mathbf{\Phi},\mathbf{X}},k}$. For any α, we then have

$$(\log Z_{\mathbf{\Phi},\mathbf{X}})'(\alpha) - (\log Z_{\mathbf{\Phi},\mathbf{g}})'(\alpha) = \frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}, \tag{6.29}$$

which we can then bound using

$$\begin{aligned} \left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right| \geq \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)} - \frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}\right| - \left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| \\ - \left|\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)} - \frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right|. \end{aligned} \tag{6.30}$$

We now bound each of these three terms individually. First, we need upper and lower bounds for $p_{\mathbf{X},k}(\alpha)$. Using the $\|\mathbf{X}\|_2^2$ moment bound (2.5), we have

$$\begin{aligned}
|p_{\mathbf{X},k}(\alpha) - 1| &\leq \sum_{j=1}^{k} \frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2j}\}|\alpha|^j}{j!} \\
&\leq \sum_{j=1}^{k} \frac{(CK^2)^j(n^j + j!)|\alpha|^j}{j!} \\
&= \sum_{j=1}^{k} \frac{(CK^2 n|\alpha|)^j}{j!} + \sum_{j=1}^{k} (CK^2|\alpha|)^j \\
&\leq e^{CK^2 n|\alpha|} - 1 + \frac{CK^2|\alpha|}{1 - CK^2|\alpha|}.
\end{aligned}$$

By sharpening the constant C in our assumption on $|\alpha|$ if necessary, we may thus ensure that

$$|p_{\mathbf{X},k}(\alpha) - 1| \leq \frac{1}{2} \tag{6.31}$$

By the same argument, we can also ensure that

$$\left|p'_{\mathbf{X},k}(\alpha) - \mathbb{E}\|\mathbf{X}\|_2^2\right| \leq \frac{1}{2}. \tag{6.32}$$

By our assumptions on the moments of $\|\mathbf{X}\|_2^2$ and $\|\mathbf{g}\|_2^2$, we have $p_{\mathbf{X},r-1} \equiv p_{\mathbf{g},r-1}$. Furthermore, only the leading terms of $p'_{\mathbf{X},r}$ and $p'_{\mathbf{g},r}$ differ. This, together with (6.31) implies that

$$\begin{aligned}
\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)} - \frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}\right| &\geq \frac{2}{3}\left|p'_{\mathbf{X},r}(\alpha) - p'_{\mathbf{g},r}(\alpha)\right| \\
&\geq \frac{2\Delta|\alpha|^{r-1}}{3(r-1)!}.
\end{aligned} \tag{6.33}$$

Next, we have

$$\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| \leq \left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| + \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right|. \tag{6.34}$$

Again we bound these two terms individually. Using the identity $p_{\mathbf{X},r}(\alpha) = p_{\mathbf{X},r-1}(\alpha) +$

$\mathbb{E}\|\mathbf{X}\|_2^{2r}(-\alpha)^r/r!$, we get

$$\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| = \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right|\left|1 - \frac{p_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right|$$
$$= \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)p_{\mathbf{X},r-1}(\alpha)}\right|\frac{\mathbb{E}\|\mathbf{X}\|_2^{2r}|\alpha|^r}{r!} \tag{6.35}$$

Using the bounds on $p_{\mathbf{X},r}$ and $p'_{\mathbf{X},r}$ (6.31) and (6.32), together with the $\|\mathbf{X}\|_2^2$ moment bound (2.5), we get

$$\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)p_{\mathbf{X},r-1}(\alpha)}\right|\frac{\mathbb{E}\|\mathbf{X}\|_2^{2r}|\alpha|^r}{r!} \leq \frac{3}{8}\frac{(CK^2)^r(n^r + r!)|\alpha|^r}{r!}. \tag{6.36}$$

For the first term in (6.34), we write

$$\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| = \left|(\log Z_{\mathbf{\Phi},\mathbf{X}}(\alpha))' - (\log p_{\mathbf{X},r}(\alpha))'\right|$$
$$= \left|\frac{d}{d\alpha}\log\left(\frac{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right|$$
$$= \left|\frac{d}{d\alpha}\log\left(1 + \frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right|. \tag{6.37}$$

Using Lemma 6.7.2 together with our assumptions on $|\alpha|$, we observe that

$$|R_{\mathbf{X},r}(\alpha)| \leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!} + 1\right)\left(e^{CK^2|\alpha|n} + \frac{CK^2|\alpha|}{1 - CK^2|\alpha|}\right)$$
$$\leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!} + 1\right). \tag{6.38}$$

In particular, by sharpening the constant C in our assumption on $|\alpha|$ if necessary, we can ensure that this quantity is less than $\frac{1}{4}$. In this case, we have

$$\left|\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| \leq \frac{1}{2},$$

so that

$$\begin{aligned}\left|\frac{d}{d\alpha}\log\left(1+\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right| &= \left|\log'\left(1+\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right|\left|\left(\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)'\right| \\ &\leq 2\left|\left(\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)'\right| \\ &\leq 2\left(\left|\frac{R'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| + \left|\frac{R_{\mathbf{X},r}(\alpha)p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)^2}\right|\right). \qquad (6.39)\end{aligned}$$

By our bounds on these functions (6.31), (6.32), and (6.38), we have

$$\left|\frac{R_{\mathbf{X},r}(\alpha)p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)^2}\right| \leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right). \qquad (6.40)$$

Furthermore, by using the moment bounds (2.5) as before, one can show that

$$|R'_{\mathbf{X},r}(\alpha)| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right).$$

so that the first term is also bounded according to

$$\left|\frac{R'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \qquad (6.41)$$

As such, combining (6.37) and (6.39) tells us that

$$\begin{aligned}\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right) + (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right) \\ &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \qquad (6.42)\end{aligned}$$

We can now use this estimate together with (6.36) to continue (6.34), writing

$$\begin{aligned}\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right) + \frac{(CK^2)^r(n^r+r!)|\alpha|^r}{r!} \\ &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \qquad (6.43)\end{aligned}$$

Notice that same methods also give us

$$\left|\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)} - \frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \qquad (6.44)$$

We may therefore finally substitute these last two bounds, together with (6.33), into (6.30).

This yields

$$|(\log Z_{\mathbf{\Phi},\mathbf{X}})'(\alpha) - (\log Z_{\mathbf{\Phi},\mathbf{g}})'(\alpha)| \geq \frac{2\Delta|\alpha|^{r-1}}{3(r-1)!} - C(CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!} + r + 1\right). \quad (6.45)$$

We now claim that with our assumptions on $|\alpha|$, the first term dominates the second. This is a simple calculation, thereby competing the proof of (6.27). To prove (6.28), we repeat the entire argument, but using the relevant estimates for $Z_{\mathbf{\Psi},\mathbf{X}}$ instead of those for $Z_{\mathbf{\Phi},\mathbf{X}}$. □

Applying the previous lemma in the setting of our NGCA model, we get the following result.

Theorem 6.7.4 (Robustness for non-Gaussian eigenvalues)**.** *Let $\boldsymbol{X}$ be a sub-Gaussian random vector satisfying the NGCA model* (6.1)*, and with sub-Gaussian norm bounded above by $K \geq 1$. Let $\lambda_1(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}), \ldots, \lambda_d(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha})$ denote the eigenvalues of $\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}$. Suppose the moments of $\|\tilde{\boldsymbol{X}}\|_2^2$ and $\|\boldsymbol{g}_d\|_2^2$ agree up to order $r-1$, but there is a number $\Delta > 0$ such that $\left|\mathbb{E}\{\|\tilde{\boldsymbol{X}}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}_d\|_2^{2r}\}\right| \geq \Delta$, then there is an absolute constant C such that for $|\alpha| \leq \Delta r/(CK^2)^r(d^{r+1} + (r+1)!)$, we have*

$$\left|\frac{1}{d}\sum_{i=1}^{d}\lambda_i(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{1}{2\alpha+1}\right| \geq \frac{\Delta}{2d(r-1)!}|\alpha|^{r-1}. \quad (6.46)$$

Similarly, let $\lambda_1(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}), \ldots, \lambda_d(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha})$ denote the eigenvalues of $\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}$, and suppose the moments of $\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$ and $\langle \boldsymbol{g}, \boldsymbol{g}'\rangle$ agree up to order $r-1$ but $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}| \geq \Delta$. Then for $|\alpha| \leq \Delta r/(CK^2)^r(d^{r+1} + (r+1)!)$, we have

$$\left|\frac{1}{d}\sum_{i=1}^{d}\lambda_i(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{\alpha}{\alpha^2-1}\right| \geq \frac{\Delta}{2d(r-1)!}|\alpha|^{r-1}. \quad (6.47)$$

Proof. This is simply a translation of the previous theorem with the help of Lemma 6.3.2, which tells us that the log derivatives of the partition functions are equal to the traces of $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$, and that of Lemma 6.8.3, which tells us what the Gaussian eigenvalue is. □

Proof of Theorem 6.3.4. Combine the previous Corollary with Theorem 6.2.2. □

6.8 Identities for Gaussian test matrices

In this section, we let g denote a standard Gaussian random variable, and $\mathbf{g}_n$, a standard Gaussian random vector in $\mathbb{R}^n$. First, notice that independence gives $Z_{\mathbf{\Phi},\mathbf{g}_n}(\alpha) = Z_{\mathbf{\Phi},g}(\alpha)^n$

and $Z_{\mathbf{\Psi},\mathbf{g}_n}(\alpha) = Z_{\mathbf{\Psi},g}(\alpha)^n$.

Lemma 6.8.1. *We have the identities* $Z_{\mathbf{\Phi},\boldsymbol{g}_n}(\alpha) = (2\alpha+1)^{-n/2}$ *when* $\alpha > -1/2$ *and* $Z_{\mathbf{\Psi},\boldsymbol{g}_n}(\alpha) = (1-\alpha^2)^{-n/2}$ *when* $|\alpha| < 1$.

Proof. By the remarks above, it suffices to prove the formula when $n = 1$. These are then simple exercises in calculus. Notice that

$$Z_{\mathbf{\Phi},g}(\alpha) = \mathbb{E}\{e^{-\alpha g^2}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\alpha t^2} e^{-\frac{t^2}{2}}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(2\alpha+1)t^2}{2}}dt.$$

Now substitute $u = \sqrt{2\alpha+1}\cdot t$ to arrive at the formula for $Z_{\mathbf{\Phi},g}$. For the next formula, we use conditional expectations to write

$$Z_{\mathbf{\Psi},g}(\alpha) = \mathbb{E}\{e^{-\alpha g g'}\} = \mathbb{E}\{\mathbb{E}\{e^{-\alpha g g'}|g\}\}. \tag{6.48}$$

The inner expectation can be computed as

$$\mathbb{E}\{e^{-\alpha g g'}|g\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\alpha g t} e^{-\frac{t^2}{2}}dt = e^{\frac{(\alpha g)^2}{2}}.$$

Substituting this back into (6.48) and using the same technique as above gives us what we want. □

Lemma 6.8.2. *We have the identities* $-(\log Z_{\mathbf{\Phi},\boldsymbol{g}_n})'(\alpha) = n(2\alpha+1)^{-1}$ *when* $\alpha > -1/2$ *and*
$-(\log Z_{\mathbf{\Psi},\boldsymbol{g}_n})'(\alpha) = n\alpha(\alpha^2-1)^{-1}$ *when* $|\alpha| < 1$.

Lemma 6.8.3. *We have the identities* $\mathbf{\Phi}_{\boldsymbol{g}_n,\alpha} = (2\alpha+1)^{-1}\boldsymbol{I}_n$ *when* $\alpha > -1/2$ *and* $\mathbf{\Psi}_{\boldsymbol{g}_n,\alpha} = \alpha(\alpha^2-1)^{-1}\boldsymbol{I}_n$ *when* $|\alpha| < 1$. *Here,* $\boldsymbol{I}_n$ *is the* n*-dimensional identity matrix.*

Proof. By rotational symmetry, we know that both matrices are multiples of the identity. To compute these scalars, it hence suffices to find the trace of both matrices. But

$$\mathrm{Tr}(\mathbf{\Phi}_{\mathbf{g}_n,\alpha}) = \frac{\mathbb{E}\{e^{-\alpha\|\mathbf{g}_n\|_2^2}\|\mathbf{g}_n\|_2^2\}}{\mathbb{E}\{e^{-\alpha\|\mathbf{g}_n\|_2^2}\}} = -(\log Z_{\mathbf{\Phi},\mathbf{g}_n})'(\alpha).$$

Dividing by n and using the previous lemma gives us what we want. □

6.9 Concentration of sample test matrices

Proof of Theorem 6.4.1. Let $\mathbf{Y} = e^{-\alpha\|\mathbf{X}\|_2^2}\mathbf{X}$. Then $\mathbf{Y}$ is a sub-Gaussian random vector with $\|\mathbf{Y}\|_{\psi_2} \leq K$. Let $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ denote its covariance and empirical covariance matrices respectively. Then $\|\mathbf{\Sigma}\| \leq 1$ and by Lemma 2.3.2, we have $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq \epsilon/2$ with probability at least $1 - \delta/2$. Next, observe that $\mathbf{\Phi}_{\mathbf{X},\alpha} = Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}\mathbf{\Sigma}$ and $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} = \hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}\hat{\mathbf{\Sigma}}$, where $\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha) = \sum_{j=1}^{N} e^{-\alpha\|\mathbf{X}_j\|_2^2}/N$. As such, we have

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq |\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\|. \tag{6.49}$$

Combining our lower bound on α with the power series formula for $Z_{\mathbf{\Phi}}$ from Lemma 6.7.1, we have $Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) \geq 1/2$. Furthermore, we may apply Hoeffding's inequality to see that $|\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha) - Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)| \leq \epsilon/2$ with probability at least $1 - \delta/2$. We can now combine all of this together to get the probability bound. □

Proof of Theorem 6.4.2. First, define

$$\mathbf{\Sigma} = \mathbb{E}\{e^{-\alpha\langle \mathbf{X},\mathbf{X}'\rangle}\mathbf{X}(\mathbf{X}')^T\}, \text{ and}$$

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{N} e^{-\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}(\mathbf{X}_i(\mathbf{X}_i')^T + \mathbf{X}_i'\mathbf{X}_i^T)/2N,$$

so that $\mathbf{\Psi}_{\mathbf{X},\alpha} = Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}\mathbf{\Sigma}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} = \hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}\hat{\mathbf{\Sigma}}$. As in the previous theorem, we can write

$$\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| \leq |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\|. \tag{6.50}$$

This time however, we cannot immediately invoke Lemma 2.3.2 because we can no longer view $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ as the covariance and empirical covariance matrices of a random vector. Nonetheless, we can follow the same proof scheme with a few adjustments.

The basic idea is to use a net argument to transform the operator deviation bound into a scalar bound for random variables. Let $\mathcal{N}$ be a $\frac{1}{4}$-net on S^{n-1}. By a volumetric argument, we may pick $\mathcal{N}$ to have size no more than 9^n (see [114]). For any n by n real symmetric matrix $\mathbf{M}$, we then have

$$\|\mathbf{M}\| = \sup_{\mathbf{v}\in S^{n-1}} |\langle \mathbf{v}, \mathbf{M}\mathbf{v}\rangle| \leq 2\sup_{\mathbf{v}\in\mathcal{N}}|\langle \mathbf{v}, \mathbf{M}\mathbf{v}\rangle|. \tag{6.51}$$

As such, by taking a union bound, we can hope to bound $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$ by bounding $|\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\rangle|$ for a fixed unit vector $v \in S^{n-1}$. Let us do just this. We have

$$\langle \mathbf{v}, \hat{\mathbf{\Sigma}}\mathbf{v}\rangle = \frac{1}{N}\sum_{i=1}^{N} e^{-\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle,$$

so that

$$\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\rangle = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}Y_i), \tag{6.52}$$

where

$$Y_i = e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle. \tag{6.53}$$

Observe that the Y_i's are i.i.d. random variables. At this point in the proof of covariance estimation, one observes that the resulting random variables are subexponential, so one may apply Bernstein's inequality. Unfortunately, our Y_i's are not subexponetial because of the $e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}$ factor. The way we overcome this is to condition on the size of these factors being uniformly small. Indeed, by Lemma 6.9.1 to come, we have $e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle} \leq e$ for all samples i with probability at least $1 - \delta$. We call this event A.

Next, define $\tilde{Y}_i := Y_i 1_A$. The $\tilde{Y}_i$'s are i.i.d random variables with subexponential norm bounded by eK^2. We can then apply Bernstein and our assumption on the sample size N to get

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i - \mathbb{E}\tilde{Y}_i)\right| > \epsilon\right\} \leq e^{-N\epsilon^2/CK^4} \leq \frac{\delta}{9^n}. \tag{6.54}$$

Conditioning on the set A, we have $Y_i = \tilde{Y}_i$ for each i. We can also rewrite the bound on the right hand side using our assumption on N. Doing this gives us

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}\tilde{Y}_i)\right| > \epsilon \;\middle|\; A\right\} \leq \frac{\delta}{9^n}. \tag{6.55}$$

We would like to replace $\mathbb{E}\tilde{Y}_i$ with $\mathbb{E}Y_i$, but the two quantities are not necessarily equal. Nonetheless, we can bound their difference as follows. We have

$$\mathbb{E}Y_i - \mathbb{E}\tilde{Y}_i = \mathbb{E}\{Y 1_{A^c}\} = \mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle 1_{A^c}\}. \tag{6.56}$$

We apply generalized Hölder to write

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle 1_{A^c}\}| \leq \Big(\mathbb{E}\{e^{-4\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\}\Big)^{1/4}\big(\mathbb{E}\{\langle \mathbf{X}_i, \mathbf{v}\rangle^4\langle \mathbf{X}_i', \mathbf{v}\rangle^4\}\big)^{1/4}\mathbb{P}\{A^c\}^{1/2}. \tag{6.57}$$

We now use the moment bounds for sub-Gaussian random variables and Lemma 6.9.2 to bound the first two multiplicands on the right. This gives us

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\langle \mathbf{X}_i,\mathbf{v}\rangle\langle \mathbf{X}_i',\mathbf{v}\rangle 1_{A^c}\}| \leq CK^2\mathbb{P}\{A^c\}^{1/2}. \tag{6.58}$$

Next, we use Lemma 6.9.1 together with our assumption on $|\alpha|$, tightening the constant if necessary, to see that $\mathbb{P}\{A^c\} \leq \epsilon^2/C^2K^4$. We combine this together with the last few equations to obtain $|\mathbb{E}Y_i - \mathbb{E}\tilde{Y}_i| \leq \epsilon$, and combining this with (6.55), we obtain

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}Y_i)\right| > 2\epsilon \;\middle|\; A\right\} \leq \frac{\delta}{9^n}. \tag{6.59}$$

Recall that Y_i's were defined for a fixed $\mathbf{v} \in \mathcal{N}$. We can take a union bound over all vectors in $\mathcal{N}$ to get

$$\mathbb{P}\left\{\sup_{\mathbf{v}\in\mathcal{N}}|\langle \mathbf{v}, (\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{v}\rangle| > 2\epsilon \;\middle|\; A\right\} \leq \delta. \tag{6.60}$$

Combining this with (6.51) then gives

$$\mathbb{P}\left\{\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| > 4\epsilon \;\middle|\; A\right\} \leq \delta. \tag{6.61}$$

Let us continue to bound the other terms in (6.50) conditioned on the set A. Notice that on this set, $\hat{Z}_{\boldsymbol{\Psi},\mathbf{X}}(\alpha)$ is an average of terms that are each bounded in absolute value by e. Using Hoeffding's inequality together with a similar argument as above to bound $|\mathbb{E}\hat{Z}_{\boldsymbol{\Psi},\mathbf{X}}(\alpha)1_A - Z_{\boldsymbol{\Psi},\mathbf{X}}(\alpha)|$, one may show that

$$\mathbb{P}\left\{|\hat{Z}_{\boldsymbol{\Psi},\mathbf{X}}(\alpha) - Z_{\boldsymbol{\Psi},\mathbf{X}}(\alpha)| > \epsilon/2 \;\middle|\; A\right\} \leq \delta. \tag{6.62}$$

We may also use the power series formula for $Z_{\boldsymbol{\Psi},\mathbf{X}}$ from Lemma 6.7.1 together with our bound on $|\alpha|$ to show that $Z_{\boldsymbol{\Psi},\mathbf{X}}(\alpha) \geq \frac{1}{2}$.

It remains to bound $\|\boldsymbol{\Sigma}\|$. To do this, we let v again be an arbitrary unit vector, and use the Cauchy-Schwarz inequality to compute

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\langle \mathbf{X}_i,\mathbf{v}\rangle\langle \mathbf{X}_i',\mathbf{v}\rangle\}| \leq \left(\mathbb{E}e^{-2\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\right)^{1/2}\left(\mathbb{E}\{\langle \mathbf{X}_i,\mathbf{v}\rangle^2\langle \mathbf{X}_i',\mathbf{v}\rangle^2\}\right)^{1/2}. \tag{6.63}$$

We have already seen that moment bounds and Lemma 6.9.2 imply that this is bounded by an absolute constant C. In fact, we can take $C = 3$.

Putting everything together, we see that on the set A, we can continue writing (6.50) as

$$\begin{aligned}\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| &\leq |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\| \\ &\leq C\epsilon.\end{aligned}$$

Using our bound for $\mathbb{P}\{A\}$, we can therefore uncondition to get

$$\mathbb{P}\Big\{\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| > C\epsilon\Big\} \leq \delta + \mathbb{P}\{A\} \leq 2\delta. \tag{6.64}$$

Finally, note that we can massage the constants so that the multiplying constants in front of ϵ and δ disappear. □

Lemma 6.9.1. *For any* $0 < \delta < 1$ *and* $N \in \mathbb{N}$*, if* α *satisfies*

$$|\alpha| \leq \Big(CK^2\sqrt{\log(N/\delta)}(\sqrt{n} + \sqrt{\log(N/\delta)})\Big)^{-1},$$

then we have the probability bound

$$\mathbb{P}\bigg\{\sup_{1\leq i\leq N} e^{-\alpha\langle X_i, X_i'\rangle} > e\bigg\} \leq \delta. \tag{6.65}$$

Proof. Without loss of generality, assume that $\alpha > 0$. Using the union bound, it suffices to prove that

$$\mathbb{P}\{\langle \mathbf{X}, \mathbf{X}'\rangle < -1/\alpha\} = \mathbb{P}\Big\{e^{-\alpha\langle \mathbf{X},\mathbf{X}'\rangle} > e\Big\} \leq \frac{\delta}{N}. \tag{6.66}$$

To compute this, we first condition on $\mathbf{X}'$ and use the sub-Gaussian tail of $\mathbf{X}$ to get

$$\mathbb{P}\{\langle \mathbf{X}, \mathbf{X}'\rangle < -1/\alpha \mid \mathbf{X}'\} \leq \exp\left(-\frac{1}{CK^2\alpha^2\|\mathbf{X}'\|_2^2}\right),$$

and integrating out $\mathbf{X}'$, then gives

$$\mathbb{P}\{\langle \mathbf{X}, \mathbf{X}'\rangle < -1/\alpha\} \leq \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}}\}. \tag{6.67}$$

To compute this expectation, let A be the event that $\|\mathbf{X}'\|_2 \leq CK(\sqrt{n} + \sqrt{\log(N/\delta)})$. Then by equation (2.4) in Theorem 2.3.1, we have $\mathbb{P}\{A^c\} \leq \delta/N$. As such, we can break up the expectation into the portion over A and the the portion over A^c to obtain

$$\begin{aligned}\mathbb{E}e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} &= \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A\}\mathbb{P}\{A\} + \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A^c\}\mathbb{P}\{A^c\} \\ &\leq \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A\} + \mathbb{P}\{A^c\} \\ &\leq \exp\bigg(-\frac{1}{CK^4\alpha^2(n+\log(N/\delta))}\bigg) + \frac{\delta}{N}. \qquad (6.68)\end{aligned}$$

As such, we just need the first term to be less than δ/N, which corresponds to the requirement that

$$\frac{1}{CK^4\alpha^2(n+\log(N/\delta))} \geq \log(N/\delta).$$

This is simply a rearrangement of our assumption on $|\alpha|$. □

Lemma 6.9.2 (Better bound for $Z_{\mathbf{\Psi}}$). *There is an absolute constant C such that if $|\alpha| \leq 1/CK^2\sqrt{n}$, then $Z_{\mathbf{\Psi},\boldsymbol{X}}(\alpha) \leq 3$.*

Proof. The idea of the proof is similar to that of the previous lemma. We first condition on $\mathbf{X}'$ and use the sub-Gaussian nature of $\mathbf{X}$ to bound its Laplace transform, thereby obtaining

$$\mathbb{E}\{e^{-\alpha\langle\mathbf{X},\mathbf{X}'\rangle} \mid \mathbf{X}'\} \leq e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2}.$$

Integrating out $\mathbf{X}'$ gives

$$\begin{aligned}Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) &\leq \mathbb{E}\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2}\} \\ &= \int_0^\infty \mathbb{P}\Big\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\Big\}dt \\ &\leq e + \int_e^\infty \mathbb{P}\Big\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\Big\}dt \qquad (6.69)\end{aligned}$$

Next, we use our assumption on $|\alpha|$ to write

$$\begin{aligned}\mathbb{P}\Big\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\Big\} &= \mathbb{P}\bigg\{\|\mathbf{X}'\|_2 > \frac{\sqrt{\log t}}{CK|\alpha|}\bigg\} \\ &\leq \mathbb{P}\Big\{\|\mathbf{X}'\|_2 > \sqrt{\log t}CK\sqrt{n}\Big\}. \qquad (6.70)\end{aligned}$$

For $t > e$, we have $\sqrt{\log t} > 1$, so we may apply (2.4) to get

$$\mathbb{P}\Big\{\|\mathbf{X}'\|_2 > \sqrt{\log t}CK\sqrt{n}\Big\} \leq e^{-\log tCn} = t^{-Cn}. \qquad (6.71)$$

Plugging this into (6.69) gives

$$Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) \leq e + \frac{e^{-Cn}}{Cn} \leq 3 \tag{6.72}$$

if we choose C to be large enough. □

6.10 Eigenvector perturbation theory

If two n by n matrices are close in spectral norm, one can use minimax identities to show that their eigenvalues are also close. It is less trivial to show that their eigenvectors are also close, which is the case in the presence of an "eigengap". This was addressed by [34].

Definition 6.10.1. Let E and $\hat{E}$ be two subspaces of $\mathbb{R}^n$ of dimension d. Let $\mathbf{V}$ and $\hat{\mathbf{V}}$ be n by d matrices with orthonormal columns forming a basis for E and $\hat{E}$ respectively. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ be the singular values of $\mathbf{V}^T\hat{\mathbf{V}}$. We define the *principal angles* of E and $\hat{E}$ to be $\boldsymbol{\theta}_i(E, \hat{E}) = \arccos \sigma_i$ for $1 \leq i \leq d$.

Lemma 6.10.2. *Let E, $\hat{E}$, $\mathbf{V}$ and $\hat{\mathbf{V}}$ be as in the previous definition. We have*

$$\|\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F^2 = 2\sum_{i=1}^{d} \sin^2 \boldsymbol{\theta}_i(E, \hat{E}). \tag{6.73}$$

In particular, the quantity depends only on E and $\hat{E}$ and not the choice of bases.

Proof. We expand

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 = \|\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|_F^2 + \|\mathbf{V}\mathbf{V}^T\|_F^2 - 2\langle \mathbf{V}\mathbf{V}^T, \hat{\mathbf{V}}\hat{\mathbf{V}}^T\rangle. \tag{6.74}$$

Observe that

$$\|\mathbf{V}\mathbf{V}^T\|_F^2 = \mathrm{Tr}(\mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T) = \mathrm{Tr}(\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}) = \mathrm{Tr}(\mathbf{I}_d) = d. \tag{6.75}$$

Similarly, we have

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|_F^2 = d. \tag{6.76}$$

Next, we compute

$$\langle \mathbf{V}\mathbf{V}^T, \hat{\mathbf{V}}\hat{\mathbf{V}}^T\rangle = \mathrm{Tr}(\mathbf{V}\mathbf{V}^T\hat{\mathbf{V}}\hat{\mathbf{V}}^T) = \mathrm{Tr}(\hat{\mathbf{V}}^T\mathbf{V}\mathbf{V}^T\hat{\mathbf{V}}) = \|\hat{\mathbf{V}}^T\mathbf{V}\|_F^2. \tag{6.77}$$

Next, we use the fact that the squared Frobenius norm of a matrix is the sum of squares of its singular values to write

$$\|\hat{\mathbf{V}}^T\mathbf{V}\|_F^2 = \sum_{i=1}^{d}\sigma_i^2 = \sum_{i=1}^{d}\cos^2\boldsymbol{\theta}_i(E,\hat{E}). \tag{6.78}$$

We may then combine these identities to write

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 = 2\sum_{i=1}^{d}(1-\cos^2\boldsymbol{\theta}_i(E,\hat{E})) = 2\sum_{i=1}^{d}\sin^2\boldsymbol{\theta}_i(E,\hat{E}). \tag{6.79}$$

as was to be shown. □

Using the previous lemma, it is easy to see that the distance between subspaces is preserved under taking orthogonal complements.

Lemma 6.10.3. *Let F and F' be subspaces of $\mathbb{R}^n$ of dimensions m, and let F' and $F'^{\perp}$ denote their orthogonal complements. We have $d(F,F') = d(F^{\perp},F'^{\perp})$.*

We can now use these observations to state Theorem 2 from [125] in a convenient form.

Theorem 6.10.4. *Let $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ be two n by n symmetric real matrices, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$. Fix $1 \leq r \leq s \leq n$, and assume that $\min\{\lambda_r - \lambda_{r+1}, \lambda_s - \lambda_{s+1}\} > 0$, where we define $\lambda_0 = \infty$ and $\lambda_{n+1} = -\infty$. Let $d = r+n-s$, and let $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r, \boldsymbol{v}_{s+1}, \ldots, \boldsymbol{v}_n)$ and $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \ldots, \hat{\boldsymbol{v}}_r, \hat{\boldsymbol{v}}_{s+1}, \ldots, \hat{\boldsymbol{v}}_n)$ be n by d matrices whose columns are orthonormal eigenvectors to $\lambda_1, \lambda_2, \ldots, \lambda_r, \lambda_{s+1}, \ldots, \lambda_n$ and $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_r, \hat{\lambda}_{s+1}, \ldots, \hat{\lambda}_n$ respectively. Then*

$$\|\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F \leq \frac{2\sqrt{2d}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|}{\min\{\lambda_r - \lambda_{r+1}, \lambda_s - \lambda_{s+1}\}}. \tag{6.80}$$

6.11 Proof of Theorem 6.2.5

Before we prove the guarantee, we state our proposed algorithm more formally.

Algorithm 5 ITERATED REWEIGHTED PCA($\mathbf{X}$,d,$\boldsymbol{\alpha}_1$,$\boldsymbol{\alpha}_2$,$\boldsymbol{\beta}_1$,$\boldsymbol{\beta}_2$)

```
Input: Data points $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_N, \mathbf{X}'_1, \dots, \mathbf{X}'_N]$, scaling parameters $\alpha_1, \alpha_2 \in \mathbb{R}$, tolerance parameters $\beta_1, \beta_2 > 0$.
Output: Output $\hat{E}$ for $E$.
1: Initialize $\check{E} := 0$.
2: for $k = 1, \dots, d$ do:
3:     $F_1, F_2 :=$ REWEIGHTED PCA($\mathbf{P}_{\check{E}^\perp}\mathbf{X}$,$\alpha_1^{(k)}$,$\alpha_2^{(k)}$,$\beta_1^{(k)}$,$\beta_2^{(k)}$).
4:     if $F_1 \neq 0$, then $\check{E} := \check{E} \oplus F_1$.
5:     else $\check{E} := \check{E} \oplus F_2$.
6:     if $\dim(\check{E}) = d$ return $\hat{E} := \check{E}$.
```

Proof. We provide an outline of the proof, omitting details that are similar to those in the proof of Theorem 6.2.3. Suppose we are at Step 3, having just completed k iterations, and have found $\check{E}$ so that $\dim(\check{E}) = d_k$ and $d(\check{E}, E_k) < \epsilon_k$ for some subspace $E_k \subset E$. Call $\mathbf{Y} := \mathbf{P}_{E^\perp}\mathbf{X}$, and $\check{\mathbf{Y}} := \mathbf{P}_{\check{E}^\perp}\mathbf{X}$.

By Lemma 6.11.2, the remaining non-Gaussian part of $\mathbf{Y}$ is either $(m, c\eta^2/\tilde{\gamma}_m)$-norm-moment-identifiable or it is $(m, c\eta^2)$-product-moment-identifiable (see Definition 6.11.1 below). Let us assume that the former holds since the other case is similar. For convenience, we denote $\alpha = \alpha_1^{(k+1)}$, $\beta = \beta_1^{(k+1)}$ to be the scaling and tolerance parameters for the $k+1$-th iteration.

By Theorem 6.7.4, we observe the existence of non-Gaussian eigenvalues in the $\mathbf{\Phi}$ matrix for $\mathbf{Y}$ for α small enough (specifically, $\alpha < \min\{c\eta^2 r/(CK^2)^r\tilde{\gamma}_m(d^{r+1} + (r+1)!), 1/CK^2 n\}$):

$$\left| \frac{1}{d-d_0}\sum_{i=1}^{d} \lambda_i(\mathbf{\Phi}_{\mathbf{P}_E\mathbf{Y},\alpha}) - \frac{1}{2\alpha+1} \right| \geq \frac{c\eta^2}{d(m-1)!\tilde{\gamma}_m}\alpha^{m-1}. \tag{6.81}$$

It remains to see that this signal is not destroyed by the noise stemming from our estimation of $\mathbf{\Phi}_{\mathbf{Y},\alpha}$ by $\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}$. Note that we have

$$\begin{aligned} |\mathbb{E}\{e^{-\alpha\|\mathbf{Y}\|_2^2}\} - \mathbb{E}\{e^{-\alpha\|\check{\mathbf{Y}}\|_2^2}\}| &\leq \mathbb{E}\{(e^{-\alpha\|\mathbf{Y}\|_2^2} + e^{-\alpha\|\check{\mathbf{Y}}\|_2^2})|\alpha\|\mathbf{Y}\|_2^2 - \alpha\|\check{\mathbf{Y}}\|_2^2|\} \\ &\leq \alpha\mathbb{E}\{|\|\mathbf{Y}\|_2^2 - \|\check{\mathbf{Y}}\|_2^2|\} \\ &= \alpha\mathbb{E}\{|\mathbf{X}^T(\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp})\mathbf{X}|\} \\ &\leq \alpha\|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\|\mathbb{E}\{\|\mathbf{X}\|_2^2\} \\ &\leq n\alpha\epsilon_k. \end{aligned}$$

Here, the first inequality follows from Lemma 6.11.3, while the last one follows from the fact that

$$\|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\| \leq \|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\|_F = d(\check{E}, E_k).$$

By doing several computations similar to the above, we obtain

$$\|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \leq \text{poly}_m(n)\epsilon_k. \tag{6.82}$$

Meanwhile, Theorems 6.4.1 and 6.4.2 imply that with high probability,

$$\|\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \leq \epsilon_0 \tag{6.83}$$

We may combine (6.82) and (6.83) to get

$$\begin{aligned}\|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}\| &\leq \|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| + \|\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \\ &\leq \text{poly}_m(n)\epsilon_k + \epsilon_0.\end{aligned}$$

Suppose ϵ_0 and ϵ_k are small enough so that

$$\text{poly}_m(n)\epsilon_k + \epsilon_0 \lesssim \frac{\eta^2}{d(m-1)!\tilde{\gamma}_m}\alpha^{m-1}. \tag{6.84}$$

Then the non-Gaussian eigenvalues continue to be outlier eigenvalues of $\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}$, and can be discovered via truncation. One can formalize this using same argument as in the proof of Lemma 6.4.3. Finally, we again imitate the proof of Lemma 6.4.3 and appeal to Theorem 6.10.4. This tells us that the eigenspace F_1 corresponding to the found eigenvalues is ϵ' close to that of the "true" eigenspace F' in E if

$$\text{poly}_m(n)\epsilon_k + \epsilon_0 \lesssim \frac{\beta\epsilon'}{d^{3/2}}, \tag{6.85}$$

where we pick $\beta \asymp \eta^2\alpha^{m-1}/d(r-1)!\tilde{\gamma}_r$. If this is the case, we have have

$$d(\check{E} \oplus F_1, E_k \oplus F') = \|\mathbf{P}_{\check{E}} + \mathbf{P}_{F_1} - \mathbf{P}_{E_k} + \mathbf{P}_{F'}\|_F \leq \epsilon_k + \epsilon' =: \epsilon_{k+1}.$$

Suppose the algorithm terminates in l steps. Then $l \leq d$, and if we fix a desired $\epsilon_l < 1$, then iterating the inequalities (6.84) and (6.85) shows us that we just require

$$\epsilon_0 \leq \epsilon_l/\text{poly}_m(n)^d = \epsilon_l/\text{poly}_{m,d}(n).$$

By Theorem 6.4.1, this condition can be met with a sample size that grows according to $\text{poly}_{m,d}(n)$. □

Definition 6.11.1. Let $\tilde{\mathbf{X}}$ be an isotropic random vector in $\mathbb{R}^d$. For any positive integer m, $\gamma_r > \eta > 0$, we say that $\tilde{\mathbf{X}}$ is (m, μ)-norm-moment-identifiable if

$$\left|\mathbb{E}\{\|\tilde{\mathbf{X}}\|_2^{2r}\} - \mathbb{E}\{\|\mathbf{g}\|_2^{2r}\}\right| \geq \mu$$

for some integer $r \leq m/2$. Similarly, we say that $\tilde{\mathbf{X}}$ is (m, μ)-product-moment-identifiable if

$$\left|\mathbb{E}\{\langle\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'\rangle^r\} - \mathbb{E}\{\langle\mathbf{g}, \mathbf{g}'\rangle^r\}\right| \geq \mu$$

for some integer $r \leq m$.

Lemma 6.11.2. *In the NGCA model* (6.1.1)*, suppose $\tilde{\mathbf{X}}$ is (m, η)-moment-identifiable along every direction $\mathbf{v} \in E$ for some $\eta \in (0, 1)$. Then for any proper subspace E_k of E, $\mathbf{P}_{E_k^\perp}\tilde{\mathbf{X}}$ is either $(m, c\eta^2/\tilde{\gamma}_r)$-norm-moment-identifiable or it is $(m, c\eta^2)$-product-moment-identifiable.*

Proof. Note that $\mathbf{P}_{E_k^\perp}\tilde{\mathbf{X}}$ is still (m, η)-moment-identifiable along every direction in $E \cap E_k^\perp$. As such, we may apply Theorem 6.2.2 to conclude. □

Lemma 6.11.3. *For any real numbers a and b, we have $|e^b - e^a| \leq (e^a + e^b)|b - a|$.*

Proof. Use the fact that $e^x(x - 1) + 1 \geq 0$ for all real x. □

6.12 Proof of Corollary 6.2.6

Proof. By symmetry, we know that $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha} = c_0\mathbf{I}_d$ is a scalar matrix. To compute c_0, we write

$$c_0 = \frac{1}{d}\text{Tr}(\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}) = \frac{1}{d}\frac{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\|\tilde{\mathbf{X}}\|_2^2\}}{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\}} = \frac{1}{d}\frac{e^{-\alpha d}d}{e^{-\alpha d}} = 1.$$

Combining this with Lemma 6.8.3 and 6.3.1 allows us to write

$$\mathbf{\Phi}_{\mathbf{X},\alpha} = \left(\begin{array}{c|c} \mathbf{I}_d & 0 \\ \hline 0 & (2\alpha+1)^{-1}\mathbf{I}_{n-d}, \end{array}\right).$$

By our choice of α, this gives an eigengap of

$$1 - \frac{1}{1+2\alpha} \geq \alpha = \frac{c}{n}.$$

Our assumption that $N \gtrsim dn^2(n + \log(1/\delta))/\epsilon^2$ together with Theorem 6.4.1 then guarantees that

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq \frac{c\epsilon}{\sqrt{d}n} \tag{6.86}$$

with high probability. We may now apply Theorem 6.10.4 to see that $d(F, E) \leq \epsilon$ where F is the subspace spanned by the top d eigenvectors of $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$.

It remains to see that F is discovered by the algorithm. But then (6.86) implies that

$$\lambda_i(\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} \geq \lambda_i(\mathbf{\Phi}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} - \frac{c\epsilon}{\sqrt{d}n} \geq \frac{\alpha}{2}$$

for $1 \leq i \leq d$, and similarly,

$$\lambda_i(\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} \leq \lambda_i(\mathbf{\Phi}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} + \frac{c\epsilon}{\sqrt{d}n} \leq \frac{\alpha}{4}$$

for $d + 1 \leq i \leq n$. The final inequality in both lines holds after choosing c to be small enough. We therefore see that the top d eigenvalues are indeed those that are identified by truncating at level $\beta = \alpha/3$. □

BIBLIOGRAPHY

[1] N. Alon. Problems and results in extremal combinatorics – I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[2] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5 B):2877–2921, 2009.

[3] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, Oct 2014.

[4] S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ICA with Unknown Gaussian Noise, and Implications for Gaussian Mixtures and Autoencoders. *Algorithmica*, 72(1):215–236, 2015.

[5] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic Geometric Analysis, Part I.* AMS, Providence, RI, 2015.

[6] S. Bahmani and J. Romberg. Phase Retrieval Meets Statistical Learning Theory: A Flexible Convex Relaxation. pages 1–17, 2016. arXiv:1610.04210.

[7] S. Bahmani and J. Romberg. Solving Equations of Random Convex Functions via Anchored Regression. pages 1–15, 2017. arXiv:1702.05327.

[8] D. M. Bean. *Non-Gaussian Component Analysis*. PhD thesis, University of California, Berkeley, 2014.

[9] M. Belkin and K. Sinha. Polynomial learning of distribution families. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 103–112, 2010.

[10] Q. Berthet and P. Rigollet. Complexity Theoretic Lower Bounds for Sparse Principal Component Detection. *JMLR: Workshop and Conference Proceedings*, 30(2012):1–21, 2013.

[11] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, 2013.

[12] P. Billingsley. *Probability and Measure - Third Edition*. Wiley, 1995.

[13] D. Bilyk and F. Dai. Geodesic distance Riesz energy on the sphere. pages 1–24, 2016. arXiv:1612.08442.

[14] D. Bilyk, F. Dai, and R. Matzke. Stolarsky principle and energy optimization on the sphere. 04702:1–30, 2016. arXiv:1611.04420.

[15] G. Björck. Distributions of positive mass, which maximize a certain generalized energy integral. *Arkiv för matematik*, 3(3):255–269, 1956.

[16] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In Search of Non-Gaussian Components of a High-Dimensional Distribution. *Journal of Machine Learning Research*, 7(2):247–282, 2006.

[17] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[18] S. C. Brubaker and S. S. Vempala. Isotropic PCA and affine-invariant clustering. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 551–560, 2008.

[19] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

[20] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[21] E. J. Candes and B. Recht. Exact low-rank matrix completion via convex optimization. *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 806–812, 2008.

[22] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[23] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[24] J.-F. Cardoso. High-Order Contrasts for Independent Component Analysis. *Neural Computation*, 11(1):157–192, 1999.

[25] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing Any Low-rank Matrix, Provably. *Journal of Machine Learning Reasearch*, 16(c):2999–3034, 2013.

[26] Y. Chen and E. J. Candès. Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[27] Y. Chi and Y. M. Lu. Kaczmarz Method for Solving Quadratic Equations. *IEEE Signal Processing Letters*, 23(9):1183–1187, 2016.

[28] E. Çınlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer New York, 2011.

[29] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994.

[30] K. Cukier and V. Mayer-Schoenberger. Rise of Big Data: How it's Changing the Way We Think about the World, The. *Foreign Affairs*, 92(3):28–40, 2013.

[31] A. D'Aspremont, F. R. Bach, and L. E. Ghaoui. Full regularization path for sparse principal component analysis. *Proceedings of the 24th International Conference on Machine Learning (2007)*, 9(1):177–184, 2007.

[32] S. Datta, S. Howard, and D. Cochran. Geometry of the Welch bounds. *Linear Algebra and Its Applications*, 437(10):2455–2470, 2012.

[33] Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. pages 1–14, 2014. arXiv:1406.2572.

[34] C. Davis and W. M. Kahan. The Rotation of Eigenvectors by a Pertubation III. *SIAM Journal of Numerical Analysis*, 7(1):1–46, 1970.

[35] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. 2017. arXiv:1711.03247.

[36] E. Diederichs, A. Juditsky, A. Nemirovski, and V. Spokoiny. Sparse non Gaussian component analysis by semidefinite programming. *Machine Learning*, 91(2):211–238, 2013.

[37] E. Diederichs, A. Juditsky, V. Spokoiny, and C. Schütte. Sparse non-Gaussian component analysis. *IEEE Transactions on Information Theory*, 56(6):3033–3047, 2010.

[38] S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29, 2015.

[39] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[40] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. 1(2):1–49, 2017. arXiv:1705.02356.

[41] M. Ehler and K. A. Okoudjou. Minimization of the probabilistic p-frame potential. *Journal of Statistical Planning and Inference*, 142(3):645–659, 2012.

[42] Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[43] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758, 1982.

[44] A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. *Proceedings of 37th Conference on Foundations of Computer Science*, (0):359–368, 1996.

[45] R. Ganti, N. Rao, R. M. Willett, and R. Nowak. Learning Single Index Models in High Dimensions. pages 1–16, jun 2015. arXiv:1506.08910.

[46] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition. 2015. arXiv:1503.02101.

[47] T. Goldstein and C. Studer. PhaseMax: Convex Phase Retrieval via Basis Pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, apr 2018.

[48] N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA and Robust Tensor Decomposition. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing - STOC '14*, pages 584–593, New York, New York, USA, 2013. ACM Press.

[49] O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[50] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, Jan 2011.

[51] P. Hand and T. Huynh. Robust phaselift for phase retrieval under corruptions. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, (3):1039–1042, 2017.

[52] P. Hand and V. Voroninski. An Elementary Proof of Convex Phase Retrieval in the Natural Parameter Space via the Linear Program PhaseMax. pages 1–8, nov 2016. arXiv:1611.03935.

[53] R. Heckel and H. Bölcskei. Robust Subspace Clustering via Thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

[54] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science - ITCS '13*, page 11, New York, New York, USA, 2013. ACM Press.

[55] P. J. Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435–475, 2007.

[56] A. Hyvärinen. New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit. *Advances in Neural Information Processing Systems*, 10(1):273–279, 1998.

[57] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999.

[58] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[59] M. A. Iwen, B. Preskitt, R. Saab, and A. Viswanathan. Phase Retrieval from Local Measurements: Improved Robustness via Eigenvector-Based Angular Synchronization. 2016. arXiv:1612.01182.

[60] K. Jaganathan, S. Oymak, and B. Hassibi. Sparse phase retrieval: Convex algorithms and limitations. *IEEE International Symposium on Information Theory - Proceedings*, pages 1022–1026, 2013.

[61] H. Jeong and C. S. Güntürk. Convergence of the randomized Kaczmarz method for phase retrieval. pages 1–13, 2017. arXiv:1706.10291.

[62] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to Escape Saddle Points Efficiently. *Journal of Geometric Analysis*, 26(1):231–251, Mar 2017.

[63] M. Kawanabe. Linear dimension reduction based on the fourth-order cumulant tensor. *Proc. of Artifical Neural Networks – ICANN 2005*, pages 151–156, 2005.

[64] M. Kawanabe, M. Sugiyama, G. Blanchard, and K.-R. Müller. A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75, 2006.

[65] M. Kawanabe and F. J. Theis. Estimating non-Gaussian subspaces by characteristic functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3889 LNCS:157–164, 2006.

[66] M. Kawanabe and F. J. Theis. Joint low-rank approximation for extracting non-Gaussian subspaces. *Signal Processing*, 87(8):1890–1903, 2007.

[67] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[68] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, Feb 2017.

[69] A. K. Kuchibhotla, R. K. Patra, and B. Sen. Efficient Estimation in Convex Single Index Models. (1989), 2017. arXiv:1708.00145.

[70] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 1991.

[71] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient Descent Converges to Minimizers. (Equation 1):1–11, 2016. arXiv:1602.04915.

[72] J. Lipor, D. Hong, D. Zhang, and L. Balzano. Subspace Clustering using Ensembles of K-Subspaces. pages 1–12, 2017. arXiv:1709.04744.

[73] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion and Blind Deconvolution. 2017. arXiv:1711.10467.

[74] T. Ma and A. Wigderson. Sum-of-Squares Lower Bounds for Sparse PCA. *NIPS Proceedings*, pages 1612–1620, 2015.

[75] J. A. Mingo and R. Speicher. *Free Probability and Random Matrices*, volume 35 of *Fields Institute Monographs*. Springer New York, New York, NY, 2017.

[76] A. Moitra. Algorithmic Aspects of Machine Learning. 2014. http://people.csail.mit.edu/moitra/docs/bookex.pdf.

[77] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 93–102, 2010.

[78] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.

[79] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016.

[80] D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and Its Applications*, 441(August):199–221, 2014.

[81] M. Neykov, Z. Wang, and H. Liu. Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems 29*, pages 4089–4097. 2016.

[82] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *IFAC Proceedings Volumes*, 16(Part 1):89–94, 2012.

[83] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano. Algebraic variety models for high-rank matrix completion. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2691–2700, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[84] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[85] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[86] Y. Plan and R. Vershynin. Dimension Reduction by Random Hyperplane Tessellations. *Discrete and Computational Geometry*, 51(2):438–461, 2014.

[87] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

[88] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints: Table 1. *Information and Inference*, (1103909):iaw015, 2016.

[89] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. 2017. arXiv:1702.03849.

[90] B. Recht. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*, pages 1–13, 2009.

[91] A. Rostamizadeh, A. Talwalkar, and M. Mohri. Boston, MA.

[92] M. Rudelson and R. Vershynin. Small Ball Probabilities for Linear Images of High-Dimensional Distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2015.

[93] H. Sasaki, A. Hyvärinen, and M. Sugiyama. Clustering via mode seeking by direct estimation of the gradient of a log-density. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8726 LNAI(Part 3):19–34, 2014.

[94] H. Sasaki, G. Niu, and M. Sugiyama. Non-Gaussian Component Analysis with Log-Density Gradient Estimation. *International Conference on Artificial Intelligence and Statistics*, 51:1–20, 2016.

[95] Y. Shechtman, A. Beck, and Y. C. Eldar. GESPAR: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4):928–938, 2014.

[96] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase Retrieval with Application to Optical Imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015.

[97] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. pages 1–47, feb 2017. arXiv:1702.06175.

[98] M. Soltanolkotabi and E. J. Candés. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[99] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

[100] L. Song, S. Vempala, J. Wilmes, and B. Xie. On the Complexity of Learning Neural Networks. pages 1–21, Jul 2017. arXiv:1707.04615.

[101] Statista. Number of smartphone users worldwide from 2014 to 2020 (in millions), 2016. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

[102] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[103] J. Sun, Q. Qu, and J. Wright. A Geometric Analysis of Phase Retrieval. *Foundations of Computational Mathematics*, pages 1–68, Feb 2016.

[104] M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics. Springer Berlin Heidelberg, 2005.

[105] Y. S. Tan. Energy optimization for distributions on the sphere and improvement to the welch bounds. *Electronic Communications in Probability*, 22(43):1–12, 2017.

[106] Y. S. Tan. Sparse Phase Retrieval via Sparse PCA despite Model Misspecification: A Simplified and Extended Analysis. pages 1–20, 2017. arXiv:1712.04106v2.

[107] Y. S. Tan and R. Vershynin. Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees. *Information and Inference*, (to appear).

[108] Y. S. Tan and R. Vershynin. Polynomial Time and Sample Complexity for Non-Gaussian Component Analysis: Spectral Methods. 2017. arXiv:1704.01041.

[109] T. Tony Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *Annals of Statistics*, 44(5):2221–2251, 2016.

[110] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996.

[111] V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencie of events to their probabilities. *Theory Probab. Appl.*, 16(2):264–280, Jan 1971.

[112] S. S. Vempala and Y. Xiao. Structure from Local Optima: Learning Subspace Juntas via Higher Order PCA. *arXiv:1108.3329*, 2011.

[113] R. Vershynin. *High-Dimensional Probability*. Book draft, available at https://www.math.uci.edu/{~}rvershyn/papers/HDP-book/HDP-book.html.

[114] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, Cambridge, 2009.

[115] R. Vershynin. Estimation in High Dimensions: A Geometric Perspective. In *Sampling Theory, a Renaissance*, pages 3–66. Birkhauser, Basel, 2015.

[116] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[117] J. Virta, K. Nordhausen, and H. Oja. Projection Pursuit for non-Gaussian Independent Components. (grant 268703):1–47, 2016. arXiv:1612.05445.

[118] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2006.

[119] G. Wang, G. B. Giannakis, and J. Chen. Scalable Solvers of Random Quadratic Equations via Stochastic Truncated Amplitude Flow. *IEEE Transactions on Signal Processing*, 65(8):1961–1974, may 2017.

[120] G. Wang, G. B. Giannakis, J. Chen, and M. Akcakaya. SPARTA: Sparse phase retrieval via Truncated Amplitude flow. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3974–3978, 2017.

[121] K. Wei. Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, Dec 2015.

[122] L. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

[123] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, nov 2016.

[124] H. Wu and R. R. Nadakuditi. Free component analysis. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, pages 85–89, 2017.

[125] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[126] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. Reshaped Wirtinger Flow and Incremental Algorithm for Solving Quadratic System of Equations. *NIPS Proceedings*, pages 2622–2630, 2016.