A geometric perspective on some topics in statistical learning

by

Yuting Wei

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin Wainwright, Co-chair
Professor Adityanand Guntuboyina, Co-chair
Professor Peter Bickel
Professor Venkat Anantharam

Spring 2018

A geometric perspective on some topics in statistical learning

Abstract

A geometric perspective on some topics in statistical learning

by

Yuting Wei

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Martin Wainwright, Co-chair

Professor Adityanand Guntuboyina, Co-chair

Modern science and engineering often generate data sets with a large sample size and a comparably large dimension which puts classic asymptotic theory into question in many ways. Therefore, the main focus of this thesis is to develop a fundamental understanding of statistical procedures for estimation and hypothesis testing from a non-asymptotic point of view, where both the sample size and problem dimension grow hand in hand. A range of different problems are explored in this thesis, including work on the geometry of hypothesis testing, adaptivity to local structure in estimation, effective methods for shape-constrained problems, and early stopping with boosting algorithms.

Our treatment of these different problems shares the common theme of emphasizing the underlying geometric structure. To be more specific, in our hypothesis testing problem, the null and alternative are specified by a pair of convex cones. This cone structure makes it possible for a sharp characterization of the behavior of Generalized Likelihood Ratio Test (GLRT) and its optimality property. The problem of planar set estimation based on noisy measurements of its support function, is a non-parametric problem in nature. It is interesting to see that estimators can be constructed such that they are more efficient in the case when the underlying set has a simpler structure, even without knowing the set beforehand. Moreover, when we consider applying boosting algorithms to estimate a function in reproducing kernel Hibert space (RKHS), the optimal stopping rule and the resulting estimator turn out to be determined by the localized complexity of the space.

These results demonstrate that, on one hand, one can benefit from respecting and making use of the underlying structure (optimal early stopping rule for different RKHS); on the other hand, some procedures (such as GLRT or local smoothing estimators) can achieve better performance when the underlying structure is simpler, without prior knowledge of the structure itself.

To evaluate the behavior of any statistical procedure, we follow the classic minimax framework and also discuss about more refined notion of local minimaxity.

To my parents and grandmother.

Contents

List of Figures

Acknowledgments

Before entering college, I never dreamt that I would fly to the other side of the world, complete a Ph.D. in statistics and be so accepted, understood, supported, and loved in the way from people within Berkeley and through a greater academic community, have shown me. I cannot begin to thank adequately those who helped me in the preparation of this thesis and made my past five years probably the most wonderful journey of my life.

First and foremost, I am grateful to have two most amazing advisors that a graduate student can ever hope for, Martin Wainwright and Adityanand Guntuboyina. I first met Aditya through taking a graduate class with him on theoretical statistics. His class greatly intrigued my interest and equipped me with tools to work on statistics theory, primarily due to the extraordinary clarity of his teaching, as well as his passion for the material (who would know I came to Berkeley with the intention to work on applied statistics). After that we started to work together and I wrote my first real paper with him. As an advisor, Aditya is incredibly generous with his ideas and time, and has influenced me greatly with his genuine feature of humility, despite of his great talent and expertise. I also started to talk to Martin more frequently during my second year and was fortunate enough to visit him for three months in my third year when he was on sabbatical to ETH Zürich. During my interaction with Martin, I was (and I still am now) constantly amazed by his mathematical sharpness; his ability of distilling the essence of a problem so rapidly; his broad knowledge and deep understanding of so many subjects—statistics, optimization, information theory and computing; and by his care, his humor and aesthetical appreciation of coffee. It was one of the best things that could ever happen to me, to have worked with both of them over an intensive period of time. Over these years, they guided me about how to approach research, give talks, write, taught me what is good research, and helped me to believe in my potential and make most of it. It changed me completely.

I also benefited a lot from interactions with other faculty members in both statistics and EECS departments. Prof. Peter Bickel's knowledge and kindness are unparalleled; Prof. Bin Yu is a source of life wisdom; Prof. Noureddine El Karoui's research and appreciation of music has been an inspiration. I also thank Prof. Micheal Jordan for introducing me to non-parametric statistics through the weekly reading group on a book by Tsybakov. I thank Prof. Peng Ding for teaching me everything I know about causal inference and being so supportive of me when I was reluctant about being on job market. I am also thankful to Prof. Venkat Anantharam to be on my committee and to provide me with very helpful feedback during my qualifying exam and in our subsequent interactions. Besides, I was also lucky enough to have some wonderful teachers with whom I learned a lot from in Berkeley—Steve Evans, Allan Sly, Bin Yu, Noureddine El Karoui, Peng Ding, Peter Bartlett, Ben Recht, Ravi Kannan, Fraydoun Rezakhanlou, Alessandro Chiesa, Aditya Guntuboyina, Martin Wainwright—who gifted me with oars for sailing in the ocean of research.

In my earlier graduate years, I was very fortunate to collaborate with Prof. Tony Cai through my advisor Aditya. The problem that we worked on together got me into the field of shape-constraints methods where a lot of beautiful mathematical theories lie in. Besides being an extraordinary researcher and statistics encyclopedia, Tony has been very inspiring and supportive of young researchers. I would also like to thank members of Seminar für Statistik at ETH Zürich for their warm welcome during my visit in my third year of grad school, in particular, many thanks to Peter Bühlmann for making my stay so enjoyable. I am also indebted to Martin for getting me an invitation to the workshop in Oberwolfach on "Statistical Recovery of Discrete, Geometric and Invariant Structures", where I met many great scholars of our field for the first time. During this workshop, I received many encouraging feedbacks of my work and was overwhelmed in a good way by enlightening talks, so for me that was a huge eye-opening and inspiring experience, of which I will always be grateful. I also want to thank Liza Levina for continuous support, Bodhisattva Sen for helpful discussions, Richard Nickl for sharing classical music and Sivaraman Balakrishnan for many confused and aha moments we have had together. Before coming to Berkeley, I spent a summer at City University of Hong Kong, working on bioinformatics with Prof. Steven Smale from whom I learnt my first step of research, which was an exceedingly helpful foundation for me.

To all my friends that I made throughout my journey at Berkeley: It has been a luxury to have you in my life, and without you my 20s would lose half of its color. I must thank Po-Ling Loh for being an extremely supportive and caring academic sister and friend, who I always felt safe to turn to. And thanks to my older academic brothers, in Martin's group or from China: Nihar Shah, Yuchen Zhang, Mert Pilanci, Xiaodong Li, Zhiyu Wang, Ruixiang Zhang, Jiantao Jiao —your encouragement and advice at each critical step in my grad career are invaluable to me. I thank Yumeng Zhang who kept me accompanied through those ups and downs and fed me using her perfect cooking skills; Hye Soo Choi, who had the magic of turning my sad moments into smiley days. I thank all my friends in statistics department, in particular, Siqi Wu, Lihua Lei, Yuansi Chen, Xiao Li, Billy Fang for the academic and non-academic conversations. I am also very thankful for all my friends in Wi-Fo/Bliss lab as I moved to Cory Hall during my last two years of grad school. I thank Fanny Yang, for being a great roommate and an inspiring figure who is always on her way to perfection; Orhan Ocal, for those wonderful lunch/coffee/boba time with me; Raaz Dwivedi, with one of whom I visited beautiful Prague and shared a lot laughters. Special thanks to Varun Jog, Rashmi Korlakai Vinayak, Vasuki Narasimha Swamy, Reinhard Heckel, Vidya Muthukumar, Ashwin Pananjady, Sang Min Han, Soham Phade among others. I am also grateful for all the care and support from you during these years—thanks to Jiequn Han, Jiajun Tong, Song Mei, Ze Xu, Jingxue Fu, Shiman Ding, Ben Zhang, Kyle Yang, Haoran Tang, Ruoxi Jia, Qian Zhong, Chang Liu and Zhe Ji—with whom I have shared some of my fondest memories, whether it was moments of sadness, of tears, of sickness, or happiness or silly.

Above all, I owe the most to my family, in particular to my parents and grandmother, to whom this thesis is dedicated, for your unconditional love and heoric researves of patience. Every achievement of mine past and future, if there is any, is all because of you.

Part I

Introduction and background

Chapter 1

Introduction

With thousands of hundreds of data being collected everyday from modern science and engineering, statistics has entered a new era. While the cost or time for data collection has constrained the previous scientific studies, advanced technology allows for obtaining extremely large and high-dimensional data. These data sets often have dimension of the same order or even larger than the sample size, which often puts the class asymptotic theory into question and a non-asymptotic point of view is called for in modern statistics.

The main focus of this thesis is to develop a fundamental understanding of statistical procedures for high-dimensional testing and estimation, and brings together a combination of techniques from statistics, optimization and information theory. In this thesis, a range of different problems are explored, including work on the geometry of hypothesis testing, adaptivity to local structure in estimation, effective methods for shape-constrained problems, and early stopping with boosting algorithms. A common theme underlying much of this work is the underlying geometric structure of the problem. In the following sections, we outline some of the core problems and key ideas that will be developed in the remainder of this thesis.

1.1 Geometry of high-dimensional hypothesis testing

Hypothesis testing, along with the closely associated notion of a confidence region, has long played a central role in statistical inference. While research on hypothesis testing dates back to the seminal work of Neyman and Pearson, high-dimensional and structured testing problems have drawn attention in recent years, motivated by the large amounts of data generated by experimental sciences and technological applications.

The generalized likelihood ratio test (GLRT) is a standard approach to composite testing problems. Despite the wide-spread use of the GLRT, its properties have yet to be fully understood. When is it optimal, and when can it be improved upon? How does its performance depend on the null and alternative hypotheses? In this thesis, we provide answers to these and other questions for the case where the null and alternative are specified by a pair of closed, convex cones. Such cone testing problems arise in various applications, including detection of treatment effects, trend detection in econometrics, signal detection in radar processing, and shape-constrained inference in non-parametric statistics.

The main contribution of this study is to provide a sharp characterization of the GLRT testing radius purely in terms of the geometric structure of the underlying convex cones. When applied to concrete examples, our result reveals some fundamental phenomena that do *not* arise in the analogous problem of estimation under convex constraints. In particular, in contrast to estimation error, the testing error no longer depends only on the problem instance via a volume-based measure such as metric entropy or Gaussian complexity; instead, other geometric properties of the cones also play an important role. In order to address the issue of optimality, we proved information-theoretic lower bounds for the minimax testing radius again in terms of geometric quantities. These lower bounds applies to any test function thus providing a sufficient condition for the GLRT to be an optimal test.

These general theorems are illustrated by examples including the cases of monotone and orthant cones, and involve some results of independent interest. It is worthwhile to note that these newfound connections between the hardness of hypothesis testing and the local geometry of the underlying structures have many implications. In particular, as we pointed out, they reveal the intrinsic similarities and differences between estimation and hypothesis testing.

1.2 Shape-constrained problems

Research on estimation and testing under shape constraints started in the 1950s. A non-parametric problem is said to be shape-constrained if the underlying density or function is required to satisfy constraints such as monotonicity, unimodality, or convexity (e.g., [70]). Shape-constrained methods have their own merits in many ways, first of all, being non-parametric, these methods are more robust than standard parametric approaches; on the other hand, although these methods deal with infinite-dimensional models, shape constraints may be implemented without tuning parameters (such as bandwidth, or penalization parameter).

Recent years have witnessed renewed interest in shape-constrained problems, motivated by applications in areas such as medical research and econometrics. Here, in the second part, we consider the problem of estimating an unknown planar convex set from noisy measurements of its support function. For a given direction, the support function of a convex set measures the distance between the origin and the supporting hyperplane that is perpendicular to that direction. Set recovery from support functions is used in areas such as computational tomography, tactical sensing in robotics, and projection magnetic resonance imaging [115].

For this problem, we construct a local smoothing estimator with an explicit data-driven choice of bandwidth parameter. The main contribution is to establish the interesting fact that, in every direction, this estimator adapts to the local geometry of the underlying set, and it does so without any pre-knowledge of the set itself. Using a decision-theoretic framework tailored to specific functions first introduced in Cai and Low [29], we establish the optimality of our estimator in a strong pointwise sense. From these point estimators, we also construct a set estimator that is both adaptive to polytopes with a bounded number of extreme points, and achieves the globally optimal minimax rate.

Similarly to other shape-constrained problems, results developed for this problem also exhibit a form of adaptivity to local problem structure, with methods performing better for certain instances than suggested by a global minimax analysis. We will make these points more concrete in our later chapter. In this general area, there are many problems that still remain open. For example, there is only very limited theory on estimating multivariate functions under shape constraints. The absence of a natural order structure in $\mathbb{R}^d$ for $d > 1$ presents a significant obstacle to such a generalization. Moreover, relative to estimation, it is less clear how one can construct optimal and adaptive confidence intervals or regions (in the multi-dimensional case) in these scenarios.

1.3 Optimization and early-stopping

Many methods for statistical estimation and testing, including maximum likelihood and the generalized likelihood ratio test, are based on optimizing a suitable data-dependent objective function. It is well-understood that procedures for fitting non-parametric models must involve some form of regularization to prevent overfitting to the noisy data. The classical approach is to add a penalty term to the objective function, leading to the notion of a penalized estimator.

An alternative approach is to apply an iterative optimization algorithm to the original objective, and then stop it after a pre-specified number of steps, thereby terminating it prior to convergence. To be more specific, suppose based on the observations, we construct empirical loss function $\mathcal{L}_n(f)$. A optimization algorithm is based on taking gradient steps

$$f^{t+1} = f^t - \alpha^t g^t,$$

to minimize this loss function. We want to specify the number of steps T, such that f^T is as close to the minimizer of the population loss as possible.

Relative to our rich and detailed understanding of regularization via penalization (e.g., [138, 63]), our understanding of early stopping regularization is not as well-developed. In particular, for penalized estimators, it is now well-understood that complexity measures such as the localized Gaussian width, or its Rademacher analogue, can be used to characterize their achievable rates.

In this part, we show that such sharp characterizations can also be obtained for a broad class of boosting algorithms with early stopping, including L^2-boost, LogitBoost, and AdaBoost, among others. This result, to our best knowledge, is the first one to establish a precise connection between early stopping and regularized estimation in a general setting. Since boosting algorithms are used broadly in data analysis, understanding this connection provides direct guidance in many applications for obtaining more generalizable and stable statistical estimates.

1.4 Thesis overview

We want to note that although the emphasis to date has been primarily methodological and theoretical, all of this work is motivated by applications arising from areas such as computational imaging, statistical signal processing, and treatment effects which will be further pursued in the future.

The remainder of this thesis is organized as follows. We begin with the basic statistical notation and terminology in Chapter 2. It introduces important criteria to evaluate both hypothesis testing and estimation procedures that will be used through out the thesis. Chapter 3 is devoted to discuss a hypothesis testing problem where the null and alternative are both specified both convex cones. It is based on my joint work with A. Guntuboyina and M. Wainwright [149]. In Chapter 4, we consider the problem of estimating a planar set based on noisy measurements of it support function. The estimators are constructed based on locally smoothing and we focus on their adaptive behaviors when the underlying geometry varies. This part is based on joint work with T. Cai and A. Guntuboyina [28]. In Chapter 5, we explore a type of algorithmic regularization, where an optimal early stopping rule is purposed for boosting algorithms applied to reproducing kernel Hibert space. The result of this chapter is based on the joint work with F. Yang and M. Wainwright [150]. Finally we close in Chapter 6, with discussions on possible future directions and open problems, as a supplementary to the discussions in each Chapter. Proofs of more technical lemmas are deferred to the appendices.

Chapter 2

Background

Understanding the fundamental limits of estimation and testing problems is worthwhile for multiple reasons. Firstly, it provides insights of the hardness of these tasks, regardless of what procedures we are using. From a mathematical point of view, it often reveals some intrinsic properties of the problems themselves. On the other hand, exhibiting fundamental limits of performance also makes it possible to guarantee that an estimator/testing procedure is optimal, so that there are limited pay-offs in searching for another procedure with lower statistical error, although it might still be interesting to study other procedures with better performance in other metrics.

In this chapter, our first goal is to set up the basic minimax frameworks for both estimation and hypothesis testing, which are regarded as standards for discussing about the optimality of estimation and testing procedures in later chapters. Our second goal is to introduce the standard setting of non-parametric estimation, of which we will discuss about an important class of functions called reproducing kernel Hilbert space. It worth noting that this chapter only includes some basic statistical notion and terminology, and for more detailed descriptions, we refer the readers to examine the introductory material of individual chapters.

2.1 Evaluating statistical procedures

Our first step here is to establish the minimax framework we use throughout the thesis. Depending on the problem we work on, we use either the minimax risk or minimax testing radius to evaluate optimality of our statistical procedures. Our treatment here is essentially standard and more references can be found (e.g. [153, 156, 135, 81, 82, 49, 132, 96]).

Throughout, let $\mathcal{P}$ denote a class of distributions, and θ denote a functional on the space $\mathcal{P}$—a mapping from every distribution $\mathbb{P}$ to a parameter $\theta(\mathbb{P})$ taking value in some space Θ. In some scenarios, the underlying distribution $\mathbb{P}$ is uniquely determined by the quantity $\theta(\mathbb{P})$, namely, $\theta(\mathbb{P}_0) = \theta(\mathbb{P}_1)$ if and only if $\mathbb{P}_0 = \mathbb{P}_1$. In these cases, θ provides a parameterization of the family of distributions, and we write $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ for such classes.

2.1.1 Minimax estimation framework

Suppose now, we are given i.i.d observations X_i drawn from a distribution $\mathbb{P} \in \mathcal{P}$ for which $\theta(\mathbb{P}) = \theta^*$. From these observation $X^n \equiv \{X_i\}_n$, our goal is to estimate the unknown parameter θ^* and an estimator $\widehat{\theta}$ to do so is a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$. In order to evaluate the quality of any estimator, let $\rho : \Theta \times \Theta \to [0, \infty)$ be a semi-metric and we consider the quantity $\rho(\widehat{\theta}, \theta^*)$. Note that here θ^* is a fixed but unknown quantity, whereas $\widehat{\theta} \equiv \widehat{\theta}(X^n)$ is a random quantity. So we then assess the quality of the estimator by taking expectations over the randomness in X_i, which gives us

$$\mathbb{E}_{\mathbb{P}}\, \rho(\widehat{\theta}(X_1, \ldots, X_n), \theta^*). \tag{2.1}$$

As the parameter θ^* varies, this quantity also changes accordingly, which referred to as the risk function associated with the parameter. Of course, for any θ^*, we can always estimate it by ignoring the data completely and simply returning θ^*. This estimator will have zero loss when evaluated at θ^* but is likely to behave badly for other choices of the parameter.

In order to deal with the risk in a more uniform sense, let us look at the minimax principle, first suggested by Wald [145]. For any estimator $\widehat{\theta}$, its behavior is evaluated in an adversarial manner, meaning we compute its worst-case behavior $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\theta, \theta(\mathbb{P}))]$ and compare estimators according to this criterion. The optimal estimator in this sense defines the *minimax* risk—

$$\mathfrak{M}(\theta(\mathcal{P}), \rho) = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\rho(\widehat{\theta}(X_1^n), \theta(\mathbb{P}))\right], \tag{2.2}$$

where the infimum is taken over all possible estimators. Often the case, we are interested in evaluating the risk through some function of a norm—by letting $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$), then a generalization of the ρ-minimax risk can be defined as

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(\mathbb{P})))\right]. \tag{2.3}$$

For instance, if $\rho(\theta, \theta') = \|\theta - \theta'\|_2$ and $\Phi(t) = t^2$, it corresponds to the minimax risks for the mean squared error.

2.1.2 Minimax testing framework

Suppose again we are given observation X from $\mathbb{P}$, a goodness-of-fit testing problem is to decide whether the null-hypothesis $\theta(\mathbb{P}) \in \Theta_0$ holds or instead the alternative $\theta(\mathbb{P}) \in \Theta_1$ holds. Here both sets Θ_0 and Θ_1 are subsets of Θ. Usually the set Θ_0 corresponds to some desirable properties of the object of study. When both Θ_0 and Θ_1 consist of only one point, we called the hypothesis *simple*, otherwise it is called *composite*.

We want to construct a decision rule with the values 1 when the null-hypothesis is rejected, or 0 when the null-hypothesis is accepted. The decision rule $\psi : \mathcal{X} \to \{0, 1\}$ is a measurable function of an observation and it is called a *test*. Two types of errors are considered in hypothesis testing literature. The type I error is made if the null is rejected whenever it is true and the type II error is made if the null is accepted whenever it does not hold. We refer the readers to Lehmann and Romano [94] for more details.

For any test function ψ, two types of error are clearly defined when the testing problem is simple, however for a composite testing problem, we measure its performance in terms of its *uniform error*

$$\mathcal{E}(\psi; \Theta_0, \Theta_1, \epsilon) := \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi(y)] + \sup_{\theta \in \Theta_1 \backslash B_2(\epsilon; \Theta_0)} \mathbb{E}_\theta[1 - \psi(y)], \tag{2.4}$$

which controls the worst-case error over both null and alternative. Here, for a given $\epsilon > 0$, we define the *ϵ-fattening* of the set Θ_0 as

$$B_2(\Theta_0; \epsilon) := \big\{ \theta \in \mathbb{R}^d \mid \min_{u \in \Theta_0} \|\theta - u\|_2 \leq \epsilon \big\}, \tag{2.5}$$

corresponding to the set of vectors in Θ that are at most Euclidean distance ϵ from some element of Θ_0.

The reason to do is because our formulation of the testing problem allows for the possibility that θ lies in the set $\Theta_1 \backslash \Theta_0$, but is arbitrarily close to some element of Θ_0. Thus, under this formulation, it is not possible to make any non-trivial assertions about the power of any other test in a uniform sense. Accordingly, so as to be able to make quantitative statements about the performance of different statements, we exclude a certain ϵ-ball from the alternative. This procedure leads to the notion of the *minimax testing radius* associated this composite decision problem. This minimax formulation was introduced in the seminal work of Ingster and co-authors [81, 82]; since then, it has been studied by many authors (e.g., [49, 132, 96, 97, 7]).

For a given error level $\rho \in (0, 1)$, we are interested in the smallest setting of ϵ for which some test ψ has uniform error at most ρ. More precisely, we define

$$\epsilon_{\text{OPT}}(\Theta_0, \Theta_1; \rho) := \inf \Big\{ \epsilon \mid \inf_{\psi} \mathcal{E}(\psi; \Theta_0, \Theta_1, \epsilon) \leq \rho \Big\}. \tag{2.6}$$

When the sets (Θ_0, Θ_1) are clear from the context, we occasionally omit this dependence, and write $\epsilon_{\text{OPT}}(\rho)$ instead. We refer to these two quantities as the *minimax testing radius*.

By definition, the minimax testing radius ϵ_{OPT} corresponds to the smallest separation ϵ at which there exists *some test* that distinguishes between the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ with uniform error at most ρ. Thus, it provides a fundamental characterization of the statistical difficulty of the hypothesis testing. Similar to the definition of minimax estimation risk, defined in (2.6), the minimax testing radius also characterize the best possible worst-case guarantee.

2.2 Non-parametric estimation

In this section, we move beyond the parametric setting, where $\mathbb{P}$ is uniquely determined by a lower dimensional functional $\theta(\mathbb{P})$. We instead consider the problem of nonparametric regression, in which the goal is to estimate a (possibly non-linear) function on the basis of noisy observations.

Suppose we are given covariates $x \in \mathcal{X}$, along with a response variable $y \in \mathcal{Y}$. Through out this thesis, unless it is particularly mentioned, we focus our attention on the case of real-valued response variables, where the space $\mathcal{Y}$ is the real-line or or some subset of the real line. Given a class of functions $\mathcal{F}$, our goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ in $\mathcal{F}$, such that the error between y and $f(x)$ is as small as possible.

Consider a cost function $\phi : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, where the non-negative scalar $\phi(y, \theta)$ denotes the cost associated with predicting θ when the true response is y. Some common examples of loss functions ϕ that we consider in later sections include:

- the *least-squares loss* $\phi(y, \theta) := \frac{1}{2}(y - \theta)^2$
- the *logistic regression loss* $\phi(y, \theta) = \ln(1 + e^{-y\theta})$, and
- the *exponential loss* $\phi(y, \theta) = \exp(-y\theta)$.

In the *fixed design* version of regression, only the response is a random quantity, in which case it is reasonable to measure the quality of any f in terms of its error

$$\mathcal{L}(f) := \mathbb{E}_{Y^n} \left[\frac{1}{n} \sum_{i=1}^{n} \phi\big(Y_i, f(x_i)\big)\right]. \tag{2.7}$$

Accordingly, we can define $\bar{\mathcal{L}}(f)$ for the *random design* case, where the expectation is taken over both the responses and the covariates. Note that with the covariates $\{x_i\}_{i=1}^n$ fixed, the functional $\mathcal{L}$ is a non-random object. In function space $\mathcal{F}$, the optimal function minimizes the population cost functional—that is

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathcal{L}(f). \tag{2.8}$$

As a standard example, when we adopt the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, the population minimizer f^* corresponds to the conditional expectation $x \mapsto \mathbb{E}[Y \mid x]$.

Since we do not have access to the population distribution of the responses however, the computation of f^* is impossible. Given our samples $\{Y_i\}_{i=1}^n$, we consider instead some procedure applied to the *empirical loss*

$$\mathcal{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i, f(x_i)), \tag{2.9}$$

where the population expectation has been replaced by an empirical expectation. For example, when $\mathcal{L}_n$ corresponds to the log likelihood of the samples with $\phi(Y_i, f(x_i)) = \log[\mathbb{P}(Y_i; f(x_i))]$, direct unconstrained minimization of $\mathcal{L}_n$ would yield the maximum likelihood estimator.

2.2.1 Adaptive minimax risk

In this section, let us consider the case when the response $\{y_i\}_{i=1}^n$ is generated through

$$y_i = f^*(x_i) + w_i \qquad \text{for } i = 1, 2, \ldots, n, \tag{2.10}$$

where w_i is a random variable characterizing the noise in the measurements, with mean zero. Now, based on these noisy responses, our goal is to find a function f (in the function class $\mathcal{F}$) such that $f : \mathcal{X} \to \mathbb{R}$ is as close as f^* as possible.

For each estimator $\widehat{f}$, recall that its performance is measured by the loss function (2.7), where

$$\mathcal{L}(\widehat{f}, f^*) = \mathbb{E}_{Y^n} \left[\frac{1}{n}\sum_{i=1}^n \phi\big(Y_i, f(x_i)\big)\right].$$

Note that here, the response is generated from model (2.10) so the loss is also a function of f^*. Of course, for each $f^* \in \mathcal{F}$, we can always estimate it by omitting the data and simply returning f^*. This will give us a zero loss at f^* but possibly huge loss for other choices of functions. So analogous to our Section 2.1.1, we compare estimators of f^* by their worst-case behavior, namely

$$R(\mathcal{F}, \mathcal{F}_0, \phi) = \inf_{\widehat{f}\in\mathcal{F}} \sup_{f^*\in\mathcal{F}_0} \mathcal{L}(\widehat{f}, f^*). \tag{2.11}$$

Here the infimum is taken over all possible estimators in function class $\mathcal{F}$ and the supremum is taken over the space $\mathcal{F}_0$ that f^* lies in. If there is no side knowledge of f^*, we may take $\mathcal{F}_0$ to be all possible functions.

Note that in this classic minimax risk framework, estimator are compared via their worst-case behavior as measured by performance over the entire problem class. When the risk function is near to constant over the set, then the global minimax risk is reflective of the typical behavior. If not, then one is motivated to seek more refined ways of characterizing the hardness of different problems, and the performance of different estimators.

One way of doing so is by studying the notion of an adaptive estimator, meaning one whose performance automatically adapts to some (unknown) property of the underlying function being estimated. For instance, estimators using wavelet bases are known to be adaptive to unknown degree of smoothness [44, 45]. Similarly, in the context of shape-constrained problems, there is a line of work showing that for functions with simpler structure, it is possible to achieve faster rates than the global minimax ones (e.g. [109, 158, 39]).

To discuss the optimality in this adaptive or local sense, we review the notion of local minimax framework here where the focus is on the performance at every function, instead of the maximum risk over a large parameter space as in the conventional minimax theory. This framework, first introduced in Cai and Low ([29, 30]) for shape constrained regression, provides a much more precise characterization of the performance of an estimator than the conventional minimax theory does.

For a given function $f \in \mathcal{F}_0$, we choose the other function, say g, to be the one which is most difficult to distinguish from f in the ϕ-loss. This benchmark is defined as

$$R_n(f) = \sup_{g \in \mathcal{F}_0} \inf_{\widehat{f}} \max \left\{ \mathcal{L}(\widehat{f}, f),\ \mathcal{L}(\widehat{f}, g) \right\}. \tag{2.12}$$

Cai and Low [29] demonstrates that this is an useful benchmark in the context of estimating convex functions, namely $\mathcal{F}_0$ denotes the class of convex functions. They established some interesting properties, such as $R_n(f)$ varies considerably over the collection of convex functions and outperforming the benchmark $R_n(f)$ at some convex function f leads to worse performance at other functions. We want to point out that without saying this is a very useful benchmark to evaluate the optimality of adaptive estimators, but there can be other reasonable definitions of local minimax framework that are suitable in other contexts.

2.2.2 Reproducing kernel Hilbert spaces

In this section, we provide some background on a particular class of functions that will be used in our later chapters—a class of function-based Hilbert spaces that are defined by reproducing kernels. These function spaces have many attractive properties from both the computational and statistical points of view.

A reproducing kernel Hilbert space $\mathscr{H}$ (short as RKHS, see standard sources [143, 73, 128, 17]), consisting of functions mapping a domain $\mathcal{X}$ to the real line $\mathbb{R}$. Any RKHS is defined by a bivariate symmetric *kernel function* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is required to be positive semidefinite, i.e. for any integer $N \geq 1$ and a collection of points $\{x_j\}_{j=1}^N$ in $\mathcal{X}$, the matrix $[\mathbb{K}(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$ is positive semidefinite.

The associated RKHS is the closure of the linear span of functions in the form $f(\cdot) = \sum_{j \geq 1} \omega_j \mathbb{K}(\cdot, x_j)$, where $\{x_j\}_{j=1}^\infty$ is some collection of points in $\mathcal{X}$, and $\{\omega_j\}_{j=1}^\infty$ is a real-valued sequence. We can also define the inner product of two functions in the space. For two functions $f_1, f_2 \in \mathscr{H}$ which can be expressed as a finite sum $f_1(\cdot) = \sum_{i=1}^{\ell_1} \alpha_i \mathbb{K}(\cdot, x_i)$ and $f_2(\cdot) = \sum_{j=1}^{\ell_2} \beta_j \mathbb{K}(\cdot, x_j)$, the inner product is defined as

$$\langle f_1, f_2 \rangle_{\mathscr{H}} = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \alpha_i \beta_j \mathbb{K}(x_i, x_j)$$

with induced norm $\|f_1\|_{\mathscr{H}}^2 = \sum_{i=1}^{\ell_1} \alpha_i^2 \mathbb{K}(x_i, x_i)$. For each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to $\mathscr{H}$, and satisfies the reproducing relation

$$\langle f,\, \mathbb{K}(\cdot, x) \rangle_{\mathscr{H}} = f(x) \quad \text{for all } f \in \mathscr{H}. \tag{2.13}$$

This property is known as the kernel reproducing property for the Hilbert space, and it gives the power of RKHS methods in practice.

Moreover, when the covariates X_i are drawn i.i.d. from a distribution $\mathbb{P}_X$ with compact domain $\mathcal{X}$, we can invoke Mercer's theorem which states that any function in $\mathscr{H}$ can be represented as

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_k(x'), \tag{2.14}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the *eigenvalues* of the kernel function $\mathbb{K}$ and $\{\phi_k\}_{k=1}^{\infty}$ are eigenfunctions of $\mathbb{K}$ which form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P}_X)$ with the inner product $\langle f, g \rangle := \int_{\mathcal{X}} f(x) g(x) \mathrm{d}\mathbb{P}_X(x)$. We refer the reader to the standard sources [143, 73, 128, 17] for more details on RKHSs and their properties.

Part II

Statistical inference and estimation

Chapter 3

Hypothesis testing over convex cones

3.1 Introduction

Composite testing problem arise in a wide variety of applications and the generalized likelihood ratio test (GLRT) is a general purpose approach to such problem. The basic idea of the likelihood ratiotest dates back to the early works of Fisher, Neyman and Pearson; it attracted further attention following the work of Edwards [48], who emphasized likelihood as a general principle of inference. Recent years have witnessed a great amount of work on the GLRT in various contexts, including the papers [94, 112, 93, 51, 50]. However, despite the wide-spread use of the GLRT, its optimality properties have yet to be fully understood. For suitably regular problem, there is a great deal of asymptotic theory on the GLRT, and in particular when its distribution under the null is independent of nuisance parameters (e.g., [9, 120, 117]). On the other hand, there are some isolated cases in which the GLRT can be shown to dominated by other tests (e.g., [147, 107, 106, 93]).

In this chapter, we undertake an in-depth study of the GLRT in application to a particular class of composite testing problem of a geometric flavor. In this class of testing problem, the null and alternative hypotheses are specified by a pair of closed convex cones C_1 and C_2, taken to be nested as $C_1 \subset C_2$. Suppose that we are given an observation of the form $y = \theta + w$, where w is a zero-mean Gaussian noise vector. Based on observing y, our goal is to test whether a given parameter θ belongs to the smaller cone C_1—corresponding to the null hypothesis—or belongs to the larger cone C_2. Cone testing problem of this type arise in many different settings, and there is a fairly substantial literature on the behavior of the GLRT in application to such problem (e.g., see the papers and books [18, 89, 118, 117, 119, 122, 110, 107, 108, 47, 130, 147], as well as references therein).

3.1.1 Some motivating examples

Before proceeding, let us consider some concrete examples so as to motivate our study.

Example 1 (Testing non-negativity and monotonicity in treatment effects)**.** Suppose that we have a collection of d treatments, say different drugs for a particular medical condition. Letting $\theta_j \in \mathbb{R}$ denote the mean of treatment j, one null hypothesis could be that none of treatments has any effect—that is, $\theta_j = 0$ for all $j = 1, \ldots, d$. Assuming that none of the treatments are directly harmful, a reasonable alternative would be that θ belongs to the *non-negative orthant cone*

$$K_+ := \big\{\theta \in \mathbb{R}^d \mid \theta_j \geq 0 \quad \text{for all } j = 1, \ldots, d\big\}. \tag{3.1}$$

This set-up leads to a particular instance of our general set-up with $C_1 = \{0\}$ and $C_2 = K_+$. Such orthant testing problem have been studied by Kudo [89] and Raubertas et al. [117], among other people.

In other applications, our treatments might consist of an ordered set of dosages of the same drug. In this case, we might have reason to believe that if the drug has any effect, then the treatment means would obey a monotonicity constraint—that is, with higher dosages leading to greater treatment effects. One would then want to detect the presence or absence of such a dose response effect. Monotonicity constraints also arise in various types of econometric models, in which the effects of strategic interventions should be monotone with respect to parameters such as market size (e.g.,[42]). For applications of this flavor, a reasonable alternative would be specified by the *monotone cone*

$$M := \big\{\theta \in \mathbb{R}^d \ \mid \ \theta_1 \leq \theta_2 \leq \cdots \leq \theta_d\big\}. \tag{3.2}$$

This set-up leads to another instance of our general problem with $C_1 = \{0\}$ and $C_2 = M$. The behavior of the GLRT for this particular testing problem has also been studied in past works, including papers by Barlow et al. [9], and Raubertas et al. [117].

As a third instance of the treatment effects problem, we might like to include in our null hypothesis the possibility that the treatments have some (potentially) non-zero effect but one that remains constant across levels—i.e., $\theta_1 = \theta_2 = \cdots = \theta_d$. In this case, our null hypothesis is specified by the *ray cone*

$$R := \big\{\theta \in \mathbb{R}^d \ \mid \ \theta = c\mathbf{1} \quad \text{for some } c \in \mathbb{R}\big\}. \tag{3.3}$$

Supposing that we are interested in testing the alternative that the treatments lead to a monotone effect, we arrive at another instance of our general set-up with $C_1 = R$ and $C_2 = M$. This testing problem has also been studied by Bartholomew [10, 11] and Robertson et al. [121] among other researchers.

In the preceding three examples, the cone C_1 was linear subspace. Let us now consider two more examples, adapted from Menendnez et al. [108], in which C_1 is not a subspace. As before, suppose that component θ_i of the vector $\theta \in \mathbb{R}^d$ denotes the expected response of treatment i. In many applications, it is of interest to test equality of the expected responses of a subset S of the full treatment set $[d] = \{1, \ldots, d\}$. More precisely, for a given subset S containing the index 1, let us consider the problem of testing the the null hypothesis

$$C_1 \equiv E(S) := \Big\{\theta \in \mathbb{R}^d \mid \theta_i = \theta_1 \ \forall\, i \in S, \text{ and } \theta_j \geq \theta_1 \ \forall\, j \notin S \ \Big\} \tag{3.4}$$

versus the alternative $C_2 \equiv G(S) = \{\theta \in \mathbb{R}^d \mid \theta_j \geq \theta_1 \; \forall \; j \in [d]\}$. Note that C_1 here is not a linear subspace.

As a final example, suppose that we have a factorial design consisting of two treatments, each of which can be applied at two different dosages (high and level). Let (θ_1, θ_2) denote the expected responses of the first treat at the low and high dosages, respectively, with the pair (θ_3, θ_4) defined similarly for the second treatment. Suppose that we are interesting in testing whether the first treatment at the lowest level is more effective than the second treatment at the highest level. This problem can be formulated as testing the null cone

$$\begin{aligned} &C_1 := \{\theta \in \mathbb{R}^4 \mid \theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4\} \quad \text{versus the alternative} \\ &C_2 := \{\theta \in \mathbb{R}^4 \mid \theta_1 \leq \theta_2, \text{ and } \; \theta_3 \leq \theta_4\}. \end{aligned} \tag{3.5}$$

As before, the null cone C_1 is not a linear subspace.

Example 2 (Robust matched filtering in signal processing)**.** In radar detection problem [126], a standard goal is to detect the presence of a known signal of unknown amplitude in the presence of noise. After a matched filtering step, this problem can be reduced to a vector testing problem, where the known signal direction is defined by a vector $\gamma \in \mathbb{R}^d$, whereas the unknown amplitude corresponds to a scalar pre-factor $c \geq 0$. We thus arrive at a ray cone testing problem: the null hypothesis (corresponding to the absence of signal) is given $C_1 = \{0\}$, whereas the alternative is given by the positive ray cone $R_+ = \big\{\theta \in \mathbb{R}^d \mid \theta = c\gamma \text{ for some } c \geq 0\big\}$.

In many cases, there may be uncertainty about the target signal, or jamming by adversaries, who introduce additional signals that can be potentially confused with the target signal γ. Signal uncertainties of this type are often modeled by various forms of cones, with the most classical choice being a subspace cone [126]. In more recent work (e.g., [18, 66]), signal uncertainty has been modeled using the *circular cone* defined by the target signal direction, namely

$$C(\gamma; \alpha) := \big\{\theta \in \mathbb{R}^d \;\mid\; \langle \gamma, \, \theta \rangle \geq \cos(\alpha) \, \|\gamma\|_2 \|\theta\|_2 \big\}, \tag{3.6}$$

corresponding to the set of all vectors θ that have angle at least α with the target signal. Thus, we are led to another instance of a cone testing problem involving a circular cone.

Example 3 (Cone-constrained testing in linear regression)**.** Consider the standard linear regression model

$$y = X\beta + \sigma Z, \qquad \text{where } Z \sim N(0, I_n), \tag{3.7}$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed and known design matrix. In many applications, we are interested in testing certain properties of the unknown regression vector β, and these can often be encoded in terms of cone-constraints on the vector $\theta := X\beta$. As a very simple example, the problem of testing whether or not $\beta = 0$ corresponds to testing whether $\theta \in C_1 := \{0\}$ versus the alternative that $\theta \in C_2 := \text{range}(X)$. Thus, we arrive at a *subspace testing* *problem*. We note this problem is known as testing the global null in the linear regression literature (e.g., [24]). If instead we consider the case when the p-dimensional vector β lies in the non-negative orthant cone (3.1), then our alternative for the n-dimensional vector θ becomes the *polyhedral cone*

$$P := \big\{\theta \in \mathbb{R}^n \mid \theta = X\beta \quad \text{for some } \beta \geq 0\big\}. \tag{3.8}$$

The corresponding estimation problem with non-negative constraints on the coefficient vector β has been studied by Slawski et al. [131] and Meinshausen [104]; see also Chen et al. [40] for a survey of this line of work. In addition to these preceding two cases, we can also test various other types of cone alternatives for β, and these are transformed via the design matrix X into other types of cones for the parameter $\theta \in \mathbb{R}^n$.

Example 4 (Testing shape-constrained departures from parametric models)**.** Our third example is non-parametric in flavor. Consider the class of functions f that can be decomposed as

$$f = \sum_{j=1}^{k} a_j \phi_j + \psi. \tag{3.9}$$

Here the known functions $\{\phi_j\}_{j=1}^k$ define a linear space, parameterized by the coefficient vector $a \in \mathbb{R}^k$, whereas the unknown function ψ models a structured departure from this linear parametric class. For instance, we might assume that ψ belongs to the class of monotone functions, or the class of convex functions. Given a fixed collection of design points $\{t_i\}_{i=1}^n$, suppose that we make observations of the form $y_i = f(t_i) + \sigma g_i$ for $i = 1, \ldots, n$, where each g_i is a standard normal variable. Defining the shorthand notation $\theta := \big(f(t_1), \ldots, f(t_n)\big)$ and $g = (g_1, \ldots, g_n)$, our observations can be expressed in the standard form $y = \theta + \sigma g$. If, under the null hypothesis, the function f satisfies the decomposition (3.9) with $\psi = 0$, then the vector θ must belong to the subspace $\{\Phi a \mid a \in \mathbb{R}^k\}$, where the matrix $\Phi \in \mathbb{R}^{n \times k}$ has entries $\Phi_{ij} = \phi_j(x_i)$.

Now suppose that the alternative is that f satisfies the decomposition (3.9) with some ψ that is convex. A convexity constraint on ψ implies that we can write $\theta = \Phi a + \gamma$, for some coefficients $a \in \mathbb{R}^k$ and a vector $\gamma \in \mathbb{R}^n$ belonging to the *convex cone*

$$V(\{t_i\}_{i=1}^n) := \Big\{\gamma \in \mathbb{R}^n \mid \frac{\gamma_2 - \gamma_1}{t_2 - t_1} \leq \frac{\gamma_3 - \gamma_2}{t_3 - t_2} \leq \cdots \leq \frac{\gamma_n - \gamma_{n-1}}{t_n - t_{n-1}}\Big\}. \tag{3.10}$$

This particular cone testing problem and other forms of shape constraints have been studied by Meyer [110], as well as by Sen and Meyer [129].

3.1.2 Problem formulation

Having understood the range of motivations for our problem, let us now set up the problem more precisely. Suppose that we are given observations of the form $y = \theta + \sigma g$, where

$\theta \in \mathbb{R}^d$ is a fixed but unknown vector, whereas $g \sim N(0, I_d)$ is a d-dimensional vector of i.i.d. Gaussian entries and σ^2 is a known noise level. Our goal is to distinguish the null hypothesis that $\theta \in C_1$ versus the alternative that $\theta \in C_2 \backslash C_1$, where $C_1 \subset C_2$ are a nested pair of closed, convex cones in $\mathbb{R}^d$.

In this chapter, we study both the fundamental limits of solving this composite testing problem, as well as the performance of a specific procedure, namely the *generalized likelihood ratio test*, or GLRT for short. By definition, the GLRT for the problem of distinguishing between cones C_1 and C_2 is based on the statistic

$$T(y) := -2 \log \left(\frac{\sup_{\theta \in C_1} \mathbb{P}_\theta(y)}{\sup_{\theta \in C_2} \mathbb{P}_\theta(y)} \right). \tag{3.11a}$$

It defines a family of tests, parameterized by a threshold parameter $\beta \in [0, \infty)$, of the form

$$\phi_\beta(y) := \mathbb{I}(T(y) \geq \beta) \;=\; \begin{cases} 1 & \text{if } T(y) \geq \beta \\ 0 & \text{otherwise.} \end{cases} \tag{3.11b}$$

Recall that in our Section 2.1.2, we have set up the minimax testing framework. In order to be able to make quantitative statements about the performance of different statements, we exclude a certain ϵ-ball from the alternative. We consider the testing problem of distinguishing between the two hypotheses

$$\mathcal{H}_0 : \theta \in C_1 \quad \text{and} \quad \mathcal{H}_1 : \theta \in C_2 \backslash B_2(C_1; \epsilon), \tag{3.12}$$

where

$$B_2(C_1; \epsilon) := \big\{ \theta \in \mathbb{R}^d \;\mid\; \min_{u \in C_1} \|\theta - u\|_2 \leq \epsilon \big\}, \tag{3.13}$$

is the ϵ-*fattening* of the cone C_1. To be clear, the parameter $\epsilon > 0$ is a quantity that is used during the course of our analysis in order to titrate the difficulty of the testing problem. All of the tests that we consider, including the GLRT, are not given knowledge of ϵ. Let us introduce shorthand $\mathcal{T}(C_1, C_2; \epsilon)$ to denote this testing problem (3.12).

Obviously, the testing problem (3.12) becomes more difficult as ϵ approaches zero, and so it is natural to study this increase in quantitative terms. Recall that for any (measurable) test function $\psi : \mathbb{R}^d \to \{0, 1\}$, we measure its performance in terms of its *uniform error*

$$\mathcal{E}(\psi; C_1, C_2, \epsilon) := \sup_{\theta \in C_1} \mathbb{E}_\theta[\psi(y)] + \sup_{\theta \in C_2 \backslash B_2(\epsilon; C_1)} \mathbb{E}_\theta[1 - \psi(y)], \tag{3.14}$$

which controls the worst-case error over both null and alternative.

For a given error level $\rho \in (0, 1)$, we are interested in the smallest setting of ϵ for which either the GLRT, or some other test ψ has uniform error at most ρ. More precisely, we define

$$\epsilon_{\text{OPT}}(C_1, C_2; \rho) := \inf \Big\{ \epsilon \;\mid\; \inf_{\psi} \mathcal{E}(\psi; C_1, C_2, \epsilon) \leq \rho \Big\}, \qquad \text{and} \tag{3.15a}$$

$$\epsilon_{\text{GLR}}(C_1, C_2; \rho) := \inf \Big\{ \epsilon \;\mid\; \inf_{\beta \in \mathbb{R}} \mathcal{E}(\phi_\beta; C_1, C_2, \epsilon) \leq \rho \Big\}. \tag{3.15b}$$

When the subspace-cone pair (C_1, C_2) are clear from the context, we occasionally omit this dependence, and write $\epsilon_{\mathrm{OPT}}(\rho)$ and $\epsilon_{\mathrm{GLR}}(\rho)$ instead. We refer to these two quantities as the *minimax testing radius* and the *GLRT testing radius* respectively.

By definition, the minimax testing radius ϵ_{OPT} corresponds to the smallest separation ϵ at which there exists *some test* that distinguishes between the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ in equation (3.12) with uniform error at most ρ. Thus, it provides a fundamental characterization of the statistical difficulty of the hypothesis testing. On the other hand, the GLRT testing radius $\epsilon_{\mathrm{GLR}}(\rho)$ provides us with the smallest radius ϵ for which there exists *some threshold*—say β^*— for which the associated generalized likelihood ratio test ϕ_{β^*} distinguishes between the hypotheses with error at most ρ. Thus, it characterizes the performance limits of the GLRT when an optimal threshold β^* is chosen. Of course, by definition, we always have $\epsilon_{\mathrm{OPT}}(\rho) \leq \epsilon_{\mathrm{GLR}}(\rho)$. We write $\epsilon_{\mathrm{OPT}}(\rho) \asymp \epsilon_{\mathrm{GLR}}(\rho)$ to mean that—in addition to the previous upper bound—there is also a lower bound $\epsilon_{\mathrm{OPT}}(\rho) \geq c_\rho \epsilon_{\mathrm{GLR}}(\rho)$ that matches up to a constant $c_\rho > 0$ depending only on ρ.

3.1.3 Overview of our results

Having set up the problem, let us now provide a high-level overview of the main results of this chapter.

1. Our first main result, stated as Theorem 3.3.1 in Section 3.3.1, gives a sharp characterization—meaning upper and lower bounds that match up to universal constants—of the GLRT testing radius ϵ_{GLR} for cone pairs (C_1, C_2) that are non-oblique (we discuss the non-obliqueness property and its significance at length in Section 3.2.2). We illustrate the consequences of this theorem for a number of concrete cones, include the subspace cone, orthant cone, monotone cone, circular cone and a Cartesian product cone.

2. In our second main result, stated as Theorem 3.3.2 in Section 3.3.2, we derive a lower bound that applies to any testing function. It leads to a corollary that provides sufficient conditions for the GLRT to be an optimal test, and we use it to establish optimality for the subspace cone and circular cone, among other examples. We then revisit the Cartesian product cone, first analyzed in the context of Theorem 3.3.1, and use Theorem 3.3.2 to show that the GLRT is sub-optimal for this particular cone, even though it is in no sense a pathological example.

3. For the monotone and orthant cones, we find that the lower bound established in Theorem 3.3.2 is not sharp, but that the GLRT turns out to be an optimal test. Thus, Section 3.3.3 is devoted to a detailed analysis of these two cases, in particular using a more refined argument to obtain sharp lower bounds.

The remainder of this chapter is organized as follows: Section 3.2 provides background on conic geometry, including conic projections, the Moreau decomposition, and the notion of Gaussian width. It also introduces the notion of a non-oblique pair of cones, which have been studied in the context of the GLRT. In Section 3.3, we state our main results and illustrate their consequences via a series of examples. Sections 3.3.1 and 3.3.2 are devoted, respectively, to our sharp characterization of the GLRT and a general lower bound on the minimax testing radius. Section 3.3.3 explores the monotone and orthant cones in more detail. In Section 3.5, we provide the proofs of our main results, with certain more technical aspects deferred to the appendix sections.

Notation Here we summarize some notation used throughout the remainder of this chapter. For functions $f(\sigma, d)$ and $g(\sigma, d)$, we write $f(\sigma, d) \lesssim g(\sigma, d)$ to indicate that $f(\sigma, d) \leq cg(\sigma, d)$ for some constant $c \in (0, \infty)$ that may only depend on ρ but independent of (σ, d), and similarly for $f(\sigma, d) \gtrsim g(\sigma, d)$. We write $f(\sigma, d) \asymp g(\sigma, d)$ if both $f(\sigma, d) \lesssim g(\sigma, d)$ and $f(\sigma, d) \gtrsim g(\sigma, d)$ are satisfied.

3.2 Background on conic geometry and the GLRT

In this section, we provide some necessary background on cones and their geometry, including the notion of a polar cone and the Moreau decomposition. We also define the notion of a non-oblique pair of cones, and summarize some known results about properties of the GLRT for such cone testing problem.

3.2.1 Convex cones and Gaussian widths

For a given closed convex cone $C \subset \mathbb{R}^d$, we define the Euclidean projection operator $\Pi_C : \mathbb{R}^d \to C$ via

$$\Pi_C(v) := \arg\min_{u \in C} \|v - u\|_2. \tag{3.16}$$

By standard properties of projection onto closed convex sets, we are guaranteed that this mapping is well-defined. We also define the polar cone

$$C^* := \big\{ v \in \mathbb{R}^d \mid \langle v,\, u \rangle \leq 0 \quad \text{for all } u \in C \big\}. \tag{3.17}$$

Figure 3.1(b) provides an illustration of a cone in comparison to its polar cone. Using Π_{C^*} to denote the projection operator onto this cone, Moreau's theorem [111] ensures that every vector $v \in \mathbb{R}^d$ can be decomposed as

$$v = \Pi_C(v) + \Pi_{C^*}(v), \quad \text{and such that } \langle \Pi_C(v),\, \Pi_{C^*}(v) \rangle = 0. \tag{3.18}$$

We make frequent use of this decomposition in our analysis.

Let $S^{-1} := \{u \in \mathbb{R}^d \mid \|u\|_2 = 1\}$ denotes the Euclidean sphere of unit radius. For every set $A \subseteq S^{-1}$, we define its *Gaussian width* as

$$\mathbb{W}(A) := \mathbb{E}\big[\sup_{u \in A} \langle u,\, g \rangle\big] \qquad \text{where } g \sim N(0, I_d). \tag{3.19}$$

This quantity provides a measure of the size of the set A; indeed, it can be related to the volume of A viewed as a subset of the Euclidean sphere. The notion of Gaussian width arises in many different areas, notably in early work on probabilistic methods in Banach spaces [113]; the Gaussian complexity, along with its close relative the Rademacher complexity, plays a central role in empirical process theory [137, 87, 14].

Of interest in this work are the Gaussian widths of sets of the form $A = C \cap S^{-1}$, where C is a closed convex cone. For a set of this form, using the Moreau decomposition (3.18), we have the useful equivalence

$$\mathbb{W}(C \cap S^{-1}) \;=\; \mathbb{E}\big[\sup_{u \in C \cap S^{-1}} \langle u,\, \Pi_C(g) + \Pi_{C^*}(g)\rangle\big] \;=\; \mathbb{E}\|\Pi_C(g)\|_2, \tag{3.20}$$

where the final equality uses the fact that $\langle u,\, \Pi_{C^*}(g)\rangle \leq 0$ for all vectors $u \in C$, with equality holding when u is a non-negative scalar multiple of $\Pi_C(g)$.

For future reference, let us derive a lower bound on $\mathbb{E}\|\Pi_C g\|_2$ that holds for every cone C strictly larger than $\{0\}$. Take some non-zero vector $u \in C$ and let $R_+ = \{cu \mid c \geq 0\}$ be the ray that it defines. Since $R_+ \subseteq C$, we have $\|\Pi_C g\|_2 \geq \|\Pi_{R_+} g\|_2$. But since R_+ is just a ray, the projection $\Pi_{R_+}(g)$ is a standard normal variable truncated to be positive, and hence

$$\mathbb{E}\|\Pi_C g\|_2 \geq \mathbb{E}\|\Pi_{R_+} g\|_2 = \sqrt{\frac{1}{2\pi}}. \tag{3.21}$$

This lower bound is useful in parts of our development.

3.2.2 Cone-based GLRTs and non-oblique pairs

In this section, we provide some background on the notion of non-oblique pairs of cones, and their significance for the GLRT. First, let us exploit some properties of closed convex cones in order to derive a simpler expression for the GLRT test statistic (3.11a). Using the form of the multivariate Gaussian density, we have

$$\begin{aligned} T(y) \;&=\; \min_{\theta \in C_1} \|y - \theta\|_2^2 - \min_{\theta \in C_2} \|y - \theta\|_2^2 = \|y - \Pi_{C_1}(y)\|_2^2 - \|y - \Pi_{C_2}(y)\|_2^2 && (3.22)\\ &= \|\Pi_{C_2}(y)\|_2^2 - \|\Pi_{C_1}(y)\|_2^2, && (3.23) \end{aligned}$$

where we have made use of the Moreau decomposition to assert that

$$\|y - \Pi_{C_1}(y)\|_2^2 = \|y\|_2^2 - \|\Pi_{C_1}(y)\|_2^2, \quad \text{and} \quad \|y - \Pi_{C_2}(y)\|_2^2 = \|y\|_2^2 - \|\Pi_{C_2}(y)\|_2^2.$$

Thus, we see that a cone-based GLRT has a natural interpretation: it compares the squared amplitude of the projection of y onto the two different cones.

When $C_1 = \{0\}$, then it can be shown that under the null hypothesis (i.e., $y \sim N(0, \sigma^2 I_d)$), the statistic $T(y)$ (after rescaling by σ^2) is a mixture of χ^2-distributions (see e.g., [117]). On the other hand, for a general cone pair (C_1, C_2), it is not straightforward to characterize the distribution of $T(y)$ under the null hypothesis. Thus, past work has studied conditions on the cone pair under which the null distribution has a simple characterization. One such condition is a certain non-obliqueness property that is common to much past work on the GLRT (e.g., [147, 107, 108, 80]). The non-obliqueness condition, first introduced by Warrack et al. [147], is also motivated by the fact that are many instances of oblique cone pairs for which the GLRT is known to dominated by other tests. Menendez et al. [106] provide an explanation for this dominance in a very general context; see also the papers [108, 80] for further studies of non-oblique cone pairs.

A nested pair of closed convex cones $C_1 \subset C_2$ is said to be *non-oblique* if we have the successive projection property

$$\Pi_{C_1}(x) = \Pi_{C_1}(\Pi_{C_2}(x)) \qquad \text{for all } x \in \mathbb{R}^d. \tag{3.24}$$

For instance, this condition holds whenever one of the two cones is a subspace, or more generally, whenever there is a subspace L such that $C_1 \subseteq L \subseteq C_2$; see Hu and Wright [80] for details of this latter property. To be clear, these conditions are sufficient—but not necessary—for non-obliqueness to hold. There are many non-oblique cone pairs in which neither cone is a subspace; the cone pairs (3.4) and (3.5), as discussed in Example 1 on treatment testing, are two such examples. (We refer the reader to Section 5 of the paper [108] for verification of these properties.) More generally, there are various non-oblique cone pairs that do not sandwich a subspace L.

The significance of the non-obliqueness condition lies in the following decomposition result. For any nested pair of closed convex cones $C_1 \subset C_2$ that are non-oblique, for all $x \in \mathbb{R}^d$ we have

$$\Pi_{C_2}(x) = \Pi_{C_1}(x) + \Pi_{C_2 \cap C_1^*}(x) \quad \text{and} \quad \langle \Pi_{C_1}(x), \, \Pi_{C_2 \cap C_1^*}(x) \rangle = 0. \tag{3.25}$$

This decomposition follows from general theory due to Zarantonello [157], who proves that for non-oblique cones, we have $\Pi_{C_2 \cap C_1^*} = \Pi_{C_1^*}\Pi_{C_2}$—in particular, see Theorem 5.2 in this paper.

An immediate consequence of the decomposition (3.25) is that the GLRT for any non-oblique cone pair (C_1, C_2) can be written as

$$\begin{aligned} T(y) = \|\Pi_{C_2}(y)\|_2^2 - \|\Pi_{C_1}(y)\|_2^2 \; &= \; \|\Pi_{C_2 \cap C_1^*}(y)\|_2^2 \\ &= \; \|y\|_2^2 - \min_{\theta \in C_2 \cap C_1^*} \|y - \theta\|_2^2. \end{aligned}$$

Consequently, we see that the GLRT for the pair (C_1, C_2) is equivalent to—that is, determined by the same statistic as—the GLRT for testing the *reduced hypothesis*

$$\widetilde{\mathcal{H}}_0 : \theta = 0 \quad \text{versus} \quad \widetilde{\mathcal{H}}_1 : \theta \in \big(C_2 \cap C_1^*\big) \backslash B_2(\epsilon). \tag{3.26}$$

Following the previous notation, write it as $\mathcal{T}(\{0\}, C_2 \cap C_1^*; \epsilon)$ and we make frequent use of this convenient reduction in the sequel.

3.3 Main results and their consequences

We now turn to the statement of our main results, along with a discussion of some of their consequences. Section 3.3.1 provides a sharp characterization of the minimax radius for the generalized likelihood ratio test up to a universal constant, along with a number of concrete examples. In Section 3.3.2, we state and prove a general lower bound on the performance of any test, and use it to establish the optimality of the GLRT in certain settings, as well as its sub-optimality in other settings. In Section 3.3.3, we revisit and study in details two cones of particular interest, namely the orthant and monotone cones.

3.3.1 Analysis of the generalized likelihood ratio test

Let (C_1, C_2) be a nested pair of closed cones $C_1 \subseteq C_2$ that are non-oblique (3.24). Consider the polar cone C_1^* as well as the intersection cone $K = C_2 \cap C_1^*$. Letting $g \in \mathbb{R}^d$ denote a standard Gaussian random vector, we then define the quantity

$$\delta^2_{\mathrm{LR}}(C_1, C_2) := \min\left\{ \mathbb{E}\|\Pi_K g\|_2, \quad \Big(\frac{\mathbb{E}\|\Pi_K g\|_2}{\max\{0, \inf\limits_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g\rangle\}} \Big)^2 \right\}. \tag{3.27}$$

Note that $\delta^2_{\mathrm{LR}}(C_1, C_2)$ is a purely geometric object, depending on the pair (C_1, C_2) via the new cone $K = C_2 \cap C_1^*$, which arises due to the GLRT equivalence (3.26) discussed previously.

Recall that the GLRT is based on applying a threshold, at some level $\beta \in [0, \infty)$, to the likelihood ratio statistic $T(y)$; in particular, see equations (3.11a) and (3.11b). In the following theorem, we study the performance of the GLRT in terms of the the uniform testing error $\mathcal{E}(\phi_\beta; C_1, C_2, \epsilon)$ from equation (3.14). In particular, we show that the critical testing radius for the GLRT is governed by the geometric parameter $\delta^2_{\mathrm{LR}}(C_1, C_2)$.

Theorem 3.3.1. *There are numbers $\{(b_\rho, B_\rho), \rho \in (0, 1/2)\}$ such that for every pair of non-oblique closed convex cones (C_1, C_2) with C_1 strictly contained within C_2:*

(a) For every error probability $\rho \in (0, 0.5)$, we have

$$\inf_{\beta \in [0,\infty)} \mathcal{E}(\phi_\beta; C_1, C_2, \epsilon) \leq \rho \qquad \textit{for all } \epsilon^2 \geq B_\rho \, \sigma^2 \, \delta^2_{LR}(C_1, C_2). \tag{3.28a}$$

(b) Conversely, for every error probability $\rho \in (0, 0.11]$, we have

$$\inf_{\beta \in [0,\infty)} \mathcal{E}(\phi_\beta; C_1, C_2, \epsilon) \geq \rho \qquad \textit{for all } \epsilon^2 \leq b_\rho \, \sigma^2 \, \delta^2_{LR}(C_1, C_2). \tag{3.28b}$$

See Section 3.5.1 for the proof of this result.

Remarks While our proof leads to universal values for the constants B_ρ and b_ρ, we have made no efforts to obtain the sharpest possible ones, so do not state them here. In any case, our main interest is to understand the scaling of the testing radius with respect to σ and the geometric parameters of the problem. In terms of the GLRT testing radius ϵ_{GLR} previously defined (3.15b), Theorem 3.3.1 establishes that

$$\epsilon_{\mathrm{GLR}}(C_1, C_2; \rho) \asymp \sigma \; \delta_{\mathrm{LR}}(C_1, C_2), \tag{3.29}$$

where $\asymp$ denotes equality up to constants depending on ρ, but independent of all other problem parameters. Since ϵ_{GLR} always upper bounds ϵ_{OPT} for every fixed level ρ, we can also conclude from Theorem 3.3.1 that

$$\epsilon_{\mathrm{OPT}}(C_1, C_2; \rho) \lesssim \sigma \; \delta_{\mathrm{LR}}(C_1, C_2).$$

It is worthwhile noting that the quantity $\delta^2_{\mathrm{LR}}(C_1, C_2)$ depends on the pair (C_1, C_2) only via the new cone $K = C_2 \cap C_1^*$. Indeed, as discussed in Section 3.2.2, for any pair of non-oblique closed convex cones, the GLRT for the original testing problem (3.12) is equivalent to the GLRT for the modified testing problem $\mathcal{T}(\{0\}, K; \epsilon)$.

Observe that the quantity $\delta^2_{\mathrm{LR}}(C_1, C_2)$ from equation (3.27) is defined via the minima of two terms. The first term $\mathbb{E}\|\Pi_K g\|_2$ is the (square root of the) Gaussian width of the cone K, and is a familiar quantity from past work on least-squares estimation involving convex sets [139, 37]. The Gaussian width measure the size of the cone K, and it is to be expected that the minimax testing radius should grow with this size, since K characterizes the set of possible alternatives. The second term involving the inner product $\langle \eta, \, \mathbb{E}\Pi_K g\rangle$ is less immediately intuitive, partly because no such term arises in estimation over convex sets. The second term becomes dominant in cones for which the expectation $v^* := \mathbb{E}[\Pi_K g]$ is relatively large; for such cones, we can test between the null and alternative by performing a univariate test after projecting the data onto the direction v^*. This possibility only arises for cones that are more complicated than subspaces, since $\mathbb{E}[\Pi_K g] = 0$ for any subspace K.

Finally, we note that Theorem 1 gives a sharp characterization of the behavior of the GLRT up to a constant. It is different from the usual minimax guarantee. To the best of our knowledge, it is the first result to provide tight upper and lower control on the uniform performance of a specific test.

3.3.1.1 Consequences for convex set alternatives

Although Theorem 3.3.1 applies to cone-based testing problem, it also has some implications for a more general class of problem based on convex set alternatives. In particular, suppose that we are interested in the testing problem of distinguishing between

$$\mathcal{H}_0 : \theta = \theta_0, \quad \text{versus} \quad \mathcal{H}_1 : \theta \in \mathcal{S}, \tag{3.30}$$

where $\mathcal{S}$ is a not necessarily a cone, but rather an arbitrary closed convex set, and θ_0 is some vector such that $\theta_0 \in \mathcal{S}$. Consider the tangent cone of $\mathcal{S}$ at θ_0, which is given by

$$\mathcal{T}_{\mathcal{S}}(\theta) := \big\{ u \in \mathbb{R}^d \;\big|\; \text{there exists some } t > 0 \text{ such that } \theta + tu \in \mathcal{S} \big\}. \tag{3.31}$$

Note that $\mathcal{T}_{\mathcal{S}}(\theta_0)$ contains the shifted set $\mathcal{S} - \theta_0$. Consequently, we have

$$\mathcal{E}(\psi; \{0\}, \mathcal{S} - \theta_0, \epsilon) \leq \mathbb{E}_{\theta=0}[\psi(y)] + \sup_{\theta \in \mathcal{T}_{\mathcal{S}}(\theta_0) \backslash B_2(0;\epsilon)} \mathbb{E}_{\theta}[1 - \psi(y)] \;=\; \mathcal{E}(\psi; \{0\}, \mathcal{T}_{\mathcal{S}}(\theta_0), \epsilon),$$

which shows that the tangent cone testing problem

$$\mathcal{H}_0: \;\theta = 0 \quad \text{versus} \quad \mathcal{H}_1: \;\theta \in \mathcal{T}_{\mathcal{S}}(\theta_0), \tag{3.32}$$

is more challenging than the original problem (3.30). Thus, applying Theorem 3.3.1 to this cone-testing problem (3.32), we obtain the following:

Corollary 1. *For the convex set testing problem* (3.30)*, we have*

$$\epsilon^2_{OPT}(\theta_0, \mathcal{S}; \rho) \lesssim \sigma^2 \min\left\{ \mathbb{E}\|\Pi_{\mathcal{T}_{\mathcal{S}}(\theta_0)} g\|_2, \quad \Big(\frac{\mathbb{E}\|\Pi_{\mathcal{T}_{\mathcal{S}}(\theta_0)} g\|_2}{\max\{0, \inf_{\eta \in \mathcal{T}_{\mathcal{S}}(\theta_0) \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_{\mathcal{T}_{\mathcal{S}}(\theta_0)} g \rangle\}} \Big)^2 \right\}. \tag{3.33}$$

This upper bound can be achieved by applying the GLRT to the tangent cone testing problem (3.32)*.*

This corollary offers a general recipe of upper bounding the optimal testing radius. In Subsection 3.3.1.6, we provide an application of Corollary 1 to the problem of testing

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \mathcal{H}_1 : \theta \in M,$$

where M is the monotone cone (defined in expression (3.2)). When $\theta_0 \neq 0$, this is not a cone testing problem, since the set $\{\theta_0\}$ is not a cone. Using Corollary 1, we prove an upper bound on the optimal testing radius for this problem in terms of the number of constant pieces of θ_0.

In the remainder of this section, we consider some special cases of testing a cone K versus $\{0\}$ in order to illustrate the consequences of Theorem 3.3.1. In all cases, we compute the GLRT testing radius for a constant error probability, and so ignore the dependencies on ρ. For this reason, we adopt the more streamlined notation $\epsilon_{\text{GLR}}(K)$ for the radius $\epsilon_{\text{GLR}}(\{0\}, K; \rho)$.

3.3.1.2 Subspace of dimension k

Let us begin with an especially simple case—namely, when K is equal to a subspace S_k of dimension $k \leq d$. In this case, the projection Π_K is a linear operator, which can be represented by matrix multiplication using a rank k projection matrix. By symmetry of the Gaussian distribution, we have $\mathbb{E}[\Pi_K g] = 0$. Moreover, by rotation invariance of the Gaussian distribution, the random vector $\|\Pi_K g\|_2^2$ follows a χ^2-distribution with k degrees of freedom, whence

$$\frac{\sqrt{k}}{2} \;\leq\; \mathbb{E}\|\Pi_K g\|_2 \;\leq\; \sqrt{\mathbb{E}\|\Pi_K g\|_2^2} \;=\; \sqrt{k}.$$

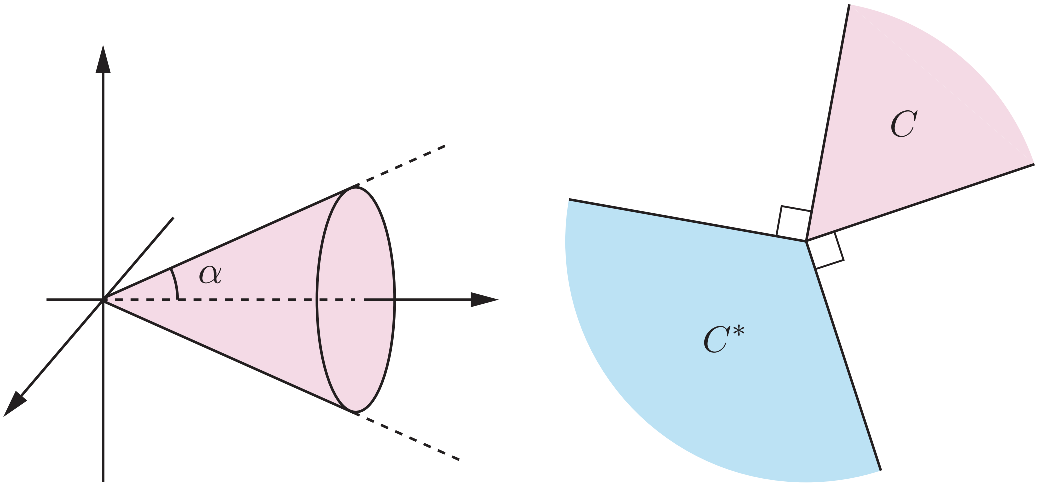
Applying Theorem 3.3.1 then yields that the testing radius of the GLRT scales as

$$\epsilon^2_{\mathrm{GLR}}(S_k) \asymp \sigma^2\sqrt{k}. \tag{3.34}$$

Here our notation $\asymp$ denotes equality up to constants independent of (σ, k); we have omitted dependence on the testing error ρ so as to simplify notation, and will do so throughout our discussion.

3.3.1.3 Circular cone

A circular cone in $\mathbb{R}^d$ with constant angle $\alpha \in (0, \pi/2)$ is given by $\mathrm{Circ}_d(\alpha) := \{\theta \in \mathbb{R}^d \mid \theta_1 \geq \|\theta\|_2 \cos(\alpha)\}$. In geometric terms, it corresponds to the set of all vectors whose angle with the standard basis vector $e_1 = (1, 0, \ldots, 0)$ is at most α radians. Figure 3.1(a) gives an illustration of a circular cone.



(a) (b)

Figure 3.1. (a) A 3-dimensional circular cone with angle α. (b) Illustration of a cone versus its polar cone.

Suppose that we want to test the null hypothesis $\theta = 0$ versus the cone alternative $K = \mathrm{Circ}_d(\alpha)$. We claim that, in application to this particular cone, Theorem 3.3.1 implies that

$$\epsilon^2_{\mathrm{GLR}}(K) \asymp \sigma^2 \min\left\{\sqrt{d}, 1\right\} \;=\; \sigma^2, \tag{3.35}$$

where $\asymp$ denotes equality up to constants depending on (ρ, α), but independent of all other problem parameters.

In order to apply Theorem 3.3.1, we need to evaluate both terms that define the geometric quantity $\delta^2_{\mathrm{LR}}(C_1, C_2)$. On one hand, by symmetry of the cone $K = \mathrm{Circ}_d(\alpha)$ in its last $(d-1)$-coordinates, we have $\mathbb{E}\Pi_K g = \beta e_1$ for some scalar $\beta > 0$ and e_1 denotes the standard

Euclidean basis vector with a 1 in the first coordinate. Moreover, for any $\eta \in K \cap S^{-1}$, we have $\eta_1 \geq \cos(\alpha)$, and hence

$$\inf_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle = \eta_1 \beta \geq \cos(\alpha)\beta \;=\; \cos(\alpha)\|\mathbb{E}\Pi_K g\|_2.$$

Next, we claim that $\|\mathbb{E}\Pi_K g\|_2 \asymp \mathbb{E}\|\Pi_K g\|_2$. In order to prove this claim, note that Jensen's inequality yields

$$\mathbb{E}\|\Pi_K g\|_2 \geq \|\mathbb{E}\Pi_K g\|_2 \;\overset{(a)}{\geq}\; (\mathbb{E}\Pi_{\mathrm{Circ}_d(\alpha)} g)_1 = \mathbb{E}(\Pi_{\mathrm{Circ}_d(\alpha)} g)_1 \;\overset{(b)}{\geq}\; \mathbb{E}\|\Pi_{\mathrm{Circ}_d(\alpha)} g\|_2 \cos(\alpha), \quad (3.36)$$

where in this argument, inequality (a) follows from simply fact that $\|x\|_2 \geq |x_1|$ whereas inequality (b) follows from the definition of circular cone. Plugging into definition $\delta^2_{\mathrm{LR}}(C_1, C_2)$, the corresponding second term equals to a constant. Therefore, the second term in the definition (3.27) of $\delta^2_{\mathrm{LR}}(C_1, C_2)$ is upper bounded by a constant, independent of the dimension d.

On the other hand, from known results on circular cones (see §6.3, [103]), there are constants $\kappa_j = \kappa_j(\alpha)$ for $j = 1, 2$ such that $\kappa_1 d \;\leq\; \mathbb{E}\|\Pi_K g\|_2^2 \;\leq\; \kappa_2 d$. Moreover, we have

$$\mathbb{E}\|\Pi_K g\|_2^2 - 4 \;\overset{(a)}{\leq}\; (\mathbb{E}\|\Pi_K g\|_2)^2 \;\overset{(b)}{\leq}\; \mathbb{E}\|\Pi_K g\|_2^2.$$

Here inequality (b) is an immediate consequence of Jensen's inequality, whereas inequality (a) follows from the fact that $\mathrm{var}(\|\Pi_K g\|_2) \leq 4$—see Lemma A.4.1 in Section 3.5.1 and the surrounding discussion for details. Putting together the pieces, we see that $\mathbb{E}\|\Pi_K g\|_2 \asymp \sqrt{d}$ for the circular cone. Combining different elements of our argument leads to the stated claim (3.35).

3.3.1.4 A Cartesian product cone

We now consider a simple extension of the previous two examples—namely, a convex cone formed by taking the Cartesian product of the real line $\mathbb{R}$ with the circular cone $\mathrm{Circ}_{d-1}(\alpha)$—that is

$$K_\times := \mathrm{Circ}_{d-1}(\alpha) \times \mathbb{R}. \quad (3.37)$$

Please refer to Figure 3.2 as an illustration of this cone in three dimensions.

This example turns out to be rather interesting because—as will be demonstrated in Section 3.3.2.3—the GLRT is sub-optimal by a factor $\sqrt{d}$ for this cone. In order to set up this later analysis, here we use Theorem 3.3.1 to prove that

$$\epsilon^2_{\mathrm{GLR}}(K_\times) \asymp \sigma^2\sqrt{d}. \quad (3.38)$$

Note that this result is strongly suggestive of sub-optimality on the part of the GLRT. More concretely, the two cones that form $K_\times$ are both "easy", in that the GLRT radius scales as

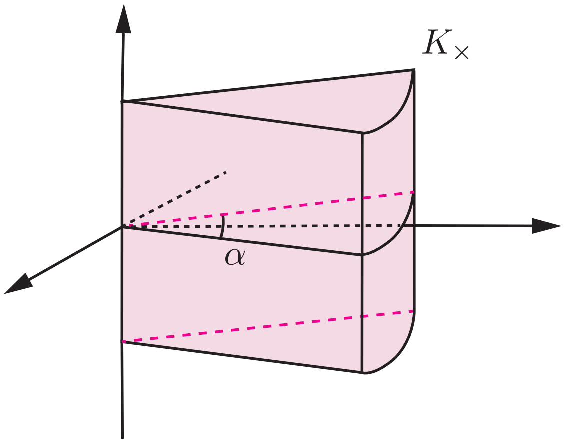


Figure 3.2: Illustration of the product cone defined in equation (3.37).

σ^2 for each. For this reason, one would expect that the squared radius of an optimal test would scale as σ^2—as opposed to the $\sigma^2\sqrt{d}$ of the GLRT—and our later calculations will show that this is indeed the case.

We now prove claim (3.38) as a consequence of Theorem 3.3.1. First notice that projecting to the product cone $K_\times$ can be viewed as projecting the first $d-1$ dimension to circular cone $\text{Circ}_{d-1}(\alpha)$ and the last coordinate to $\mathbb{R}$. Consequently, we have the following inequality

$$\begin{aligned}\mathbb{E}\|\Pi_{\text{Circ}_{d-1}(\alpha)}g\|_2 \;\leq\; \mathbb{E}\|\Pi_{K_\times}g\|_2 &\overset{(a)}{\leq} \sqrt{\mathbb{E}\|\Pi_{K_\times}g\|_2^2} \\ &= \sqrt{\mathbb{E}\|\Pi_{\text{Circ}_{d-1}(\alpha)}g\|_2^2 + \mathbb{E}[g_d^2]}.\end{aligned}$$

where inequality (a) follows by Jensen's inequality. Making use of our previous calculations for the circular cone, we have $\mathbb{E}\|\Pi_{K_\times}g\|_2 \asymp \sqrt{d}$. Moreover, note that the last coordinate of $\mathbb{E}[\Pi_{K_\times}g]$ is equal to 0 by symmetry and the standard basis vector $e_d \in \mathbb{R}^d$, with a single one in its last coordinate, belongs to $K_\times \cap S^{-1}$, we have

$$\inf_{\eta\in K_\times\cap S^{-1}} \langle \eta,\, \mathbb{E}\Pi_{K_\times}(g)\rangle \leq \langle e_d,\, \mathbb{E}\Pi_{K_\times}(g)\rangle = 0.$$

Plugging into definition $\delta^2_{\text{LR}}(C_1, C_2)$, the corresponding second term equals infinity. Therefore, the minimum that defines $\delta^2_{\text{LR}}(C_1, C_2)$ is achieved in the first term, and so is proportional to $\sqrt{d}$. Putting together the pieces yields the claim (3.38).

3.3.1.5 Non-negative orthant cone

Next let us consider the (non-negative) orthant cone given by $K_+ := \big\{\theta \in \mathbb{R}^d \mid \theta_j \geq 0 \quad \text{for } j = 1, \ldots, d\big\}$. Here we use Theorem 3.3.1 to show that

$$\epsilon^2_{\text{GLR}}(K_+) \asymp \sigma^2\sqrt{d}. \tag{3.39}$$

Turning to the evaluation of the quantity $\delta^2_{\text{LR}}(C_1, C_2)$, it is straightforward to see that $[\Pi_{K_+}(\theta)]_j = \max\{0, \theta_j\}$, and hence $\mathbb{E}\Pi_{K_+}(g) = \frac{1}{2}\mathbb{E}|g_1| \; \mathbf{1} = \frac{1}{\sqrt{2\pi}} \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^d$ is a vector of all ones. Thus, we have

$$\|\mathbb{E}\Pi_{K_+}(g)\|_2 = \sqrt{\frac{d}{2\pi}}$$

$$\text{and} \quad \|\mathbb{E}\Pi_{K_+}(g)\|_2 \;\leq\; \mathbb{E}\|\Pi_{K_+}(g)\|_2 \;\leq\; \sqrt{\mathbb{E}\|\Pi_{K_+}(g)\|_2^2} = \sqrt{\frac{d}{2}},$$

where the second inequality follows from Jensen's inequality. So the first term in the definition of quantity $\delta^2_{\text{LR}}(C_1, C_2)$ is proportional to $\sqrt{d}$. As for the second term, since the standard basis vector $e_1 \in K_+ \cap S^{-1}$, we have

$$\inf_{\eta \in K_+ \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g\rangle \leq \langle e_1, \, \frac{1}{\sqrt{2\pi}} \, \mathbf{1}\rangle = \frac{1}{\sqrt{2\pi}}.$$

Consequently, the second term in the definition of quantity $\delta^2_{\text{LR}}(C_1, C_2)$ lower bounded by a universal constant times d. Combining these derivations yields the stated claim (3.39).

3.3.1.6 Monotone cone

As our final example, consider testing in the monotone cone given by $M := \{\theta \in \mathbb{R}^d \mid \theta_1 \leq \theta_2 \leq \cdots \leq \theta_d\}$. Testing with monotone cone constraint has also been studied in different settings before, where it is known in some cases that restricting to monotone cone helps reduce the hardness of the problem to be logarithmically dependent on the dimension (e.g., [16, 148]).

Here we use Theorem 3.3.1 to show that

$$\epsilon^2_{\text{GLR}}(M) \asymp \sigma^2 \sqrt{\log d}. \tag{3.40}$$

From known results on monotone cone (see §3.5, [3]), we know that $\mathbb{E}\|\Pi_M g\|_2 \asymp \sqrt{\log d}$. So the only remaining detail is to control the second term defining $\delta^2_{\text{LR}}(C_1, C_2)$. We claim that the second term is actually infinity since

$$\max\{0, \inf_{\eta \in M \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_M g\rangle\} = 0, \tag{3.41}$$

which can be seen by simply noticing vectors $\frac{1}{\sqrt{d}}\mathbf{1}, -\frac{1}{\sqrt{d}}\mathbf{1} \in M \cap S^{-1}$ and

$$\min\Big\{\langle \frac{\mathbf{1}}{\sqrt{d}}, \, \mathbb{E}\Pi_M g\rangle, \;\; \langle -\frac{\mathbf{1}}{\sqrt{d}}, \, \mathbb{E}\Pi_M g\rangle\Big\} \leq 0.$$

Here $\mathbf{1} \in \mathbb{R}^d$ denotes the vector of all ones. Combining the pieces yields the claim (3.40).

Testing constant versus monotone It is worth noting that the same GLRT bound also holds for the more general problem of testing the monotone cone M versus the linear subspace $L = \mathrm{span}(\mathbf{1})$ of constant vectors, namely:

$$\epsilon^2_{\mathrm{GLR}}(L, M) \asymp \sigma^2 \sqrt{\log d}. \tag{3.42}$$

In particular, the following lemma provides the control that we need:

Lemma 3.3.1. *For the monotone cone M and the linear space $L = span(\mathbf{1})$, there is a universal constant c such that*

$$\inf_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle \leq c, \qquad K := M \cap L^{\perp}.$$

See Appendix A.7.1 for the proof of this lemma.

Testing an arbitrary vector θ_0 versus the monotone cone Finally, let us consider an important implication of Corollary 1 in the context of testing departures in monotone cone. More precisely, for a fixed vector $\theta_0 \in M$, consider the testing problem

$$\mathcal{H}_0 : \theta = \theta_0, \quad \text{versus} \quad \mathcal{H}_1 : \theta \in M, \tag{3.43}$$

Let us define $k(\theta_0)$ as the number of constant *pieces* of θ_0, by which we mean there exist integers $d_1, \ldots, d_{k(\theta_0)}$ with $d_i \geq 1$ and $d_1 + \cdots + d_{k(\theta_0)} = d$ such that θ_0 is a constant on each set $S_i := \{j : \sum_{t=1}^{i-1} d_t + 1 \leq j \leq \sum_{t=1}^{i} d_i\}$, for $1 \leq i \leq k(\theta_0)$.

We claim that Corollary 1 guarantees that the optimal testing radius satisfies

$$\epsilon^2_{\mathrm{OPT}}(\theta_0, M; \rho) \lesssim \sigma^2 \sqrt{k(\theta_0) \log\left(\frac{d}{k(\theta_0)}\right)}. \tag{3.44}$$

Note that this upper bound depends on the structure of θ_0 through how many pieces θ_0 possesses, which reveals the adaptive nature of Corollary 1.

In order to prove inequality (3.44), let us use shorthand k to denote $k(\theta_0)$. First notice that both $\mathbf{1}/\sqrt{d}, -\mathbf{1}/\sqrt{d} \in \mathcal{T}_M(\theta_0)$, then

$$\max\{0, \inf_{\eta \in \mathcal{T}_M(\theta_0) \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_{\mathcal{T}_M(\theta_0)} g \rangle\} \leq 0,$$

which implies the second term for $\delta^2_{\mathrm{LR}}(C_1, C_2)$ equals to infinity. It only remains to calculate $\mathbb{E}\|\Pi_{\mathcal{T}_M(\theta_0)} g\|_2$. Since the tangent cone $\mathcal{T}_M(\theta_0)$ equals to the Cartesian product of k monotone cones, namely $\mathcal{T}_M(\theta_0) = M_{d_1} \times \cdots \times M_{d_k}$, we have

$$\begin{aligned} \mathbb{E}\|\Pi_{\mathcal{T}_M(\theta_0)} g\|_2^2 = \mathbb{E}\|\Pi_{M_{d_1}} g\|_2^2 + \cdots + \mathbb{E}\|\Pi_{M_{d_k}} g\|_2^2 &= \log(d_1) + \cdots + \log(d_k) \\ &\leq k \log\left(\frac{d}{k}\right), \end{aligned}$$

where the last step follows from convexity of the logarithm function. Therefore Jensen's inequality guarantees that

$$\mathbb{E}\|\Pi_{\mathcal{T}_M(\theta_0)}g\|_2 \leq \sqrt{\mathbb{E}\|\Pi_{\mathcal{T}_M(\theta_0)}g\|_2^2} \leq \sqrt{k\log\left(\frac{d}{k}\right)}.$$

Putting the pieces together, Corollary 1 guarantees that the claimed inequality (3.44) holds for the testing problem (3.43).

3.3.2 Lower bounds on the testing radius

Thus far, we have derived sharp bounds for a particular procedure—namely, the GLRT. Of course, it is of interest to understand when the GLRT is actually an optimal test, meaning that there is no other test that can discriminate between the null and alternative for smaller separations. In this section, we use information-theoretic methods to derive a lower bound on the optimal testing radius ϵ_{OPT} for every pair of non-oblique and nested closed convex cones (C_1, C_2). Similar to Theorem 3.3.1, this bound depends on the geometric structure of intersection cone $K := C_2 \cap C_1^*$, where C_1^* is the polar cone to C_1.

In particular, let us define the quantity

$$\delta^2_{\text{OPT}}(C_1, C_2) := \min\left\{\mathbb{E}\|\Pi_K g\|_2, \;\; \left(\frac{\mathbb{E}\|\Pi_K g\|_2}{\sup\limits_{\eta \in K\cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g\rangle}\right)^2\right\}. \tag{3.45}$$

Note that the only difference from $\delta^2_{\text{LR}}(C_1, C_2)$ is the replacement of the infimum over $K\cap S^{-1}$ with a supremum, in the denominator of the second term. Moreover, since the supremum is achieved at $\frac{\mathbb{E}\Pi_K g}{\|\mathbb{E}\Pi_K g\|_2}$, we have $\sup_{\eta\in K\cap S^{-1}}\langle \eta, \, \mathbb{E}\Pi_K g\rangle = \|\mathbb{E}\Pi_K g\|_2$. Consequently, the second term on the right-hand side of equation (3.45) can be also written in the equivalent form $\left(\frac{\mathbb{E}\|\Pi_K g\|_2}{\|\mathbb{E}\Pi_K g\|_2}\right)^2$.
With this notation in hand, are now ready to state a general lower bound for minimax optimal testing radius:

Theorem 3.3.2. *There are numbers $\{\kappa_\rho, \rho \in (0, 1/2]\}$ such that for every nested pair of non-oblique closed convex cones $C_1 \subset C_2$, we have*

$$\inf_{\psi} \mathcal{E}(\psi; C_1, C_2, \epsilon) \geq \rho \qquad \textit{whenever } \epsilon^2 \leq \kappa_\rho \, \sigma^2 \, \delta^2_{OPT}(C_1, C_2), \tag{3.46}$$

In particular, we can take $\kappa_\rho = 1/14$ for all $\rho \in (0, 1/2]$.

See Section 3.5.2 for the proof of this result.

Remarks In more compact terms, Theorem 3.3.2 can be understood as guaranteeing

$$\epsilon_{\text{OPT}}(C_1, C_2; \rho) \gtrsim \sigma \delta_{\text{OPT}}(C_1, C_2),$$

where $\gtrsim$ denotes an inequality up to constants (with ρ viewed as fixed).

Theorem 3.3.2 is proved by constructing a distribution over the alternative $\mathcal{H}_1$ supported only on those points in $\mathcal{H}_1$ that are hard to distinguish from $\mathcal{H}_0$. Based on this construction, the testing error can be lower bound by controlling the total variation distance between two marginal likelihood functions. We refer our readers to our Section 3.5.2 for more details on this proof.

One useful consequence of Theorem 3.3.2 is in providing a sufficient condition for optimality of the GLRT, which we summarize here:

Corollary 2 (Sufficient condition for optimality of GLRT)**.** *Given the cone $K = C_2 \cap C_1^*$, suppose that there is a numerical constant $b > 1$, independent of K and all other problem parameters, such that*

$$\sup_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle \; = \; \|\mathbb{E}\Pi_K g\|_2 \; \leq b \inf_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle. \tag{3.47}$$

Then the GLRT is a minimax optimal test—that is, $\epsilon_{GLR}(C_1, C_2; \rho) \asymp \epsilon_{OPT}(C_1, C_2; \rho)$.

It is natural to wonder whether the condition (3.47) is also necessary for optimality of the GLRT. This turns out not to be the case. The monotone cone, to be revisited in Section 3.3.3.2, provides an instance of a cone testing problem for which the GLRT is optimal while condition (3.47) is violated. Let us now return to these concrete examples.

3.3.2.1 Revisiting the k-dimensional subspace

Let S_k be a subspace of dimension $k \leq d$. In our earlier discussion in Section 3.3.1.2, we established that $\epsilon^2_{\text{GLR}}(S_k) \asymp \sigma^2 \sqrt{k}$. Let us use Corollary 2 to verify that the GLRT is optimal for this problem. For a k-dimensional subspace $K = S_k$, we have $\mathbb{E}\Pi_K g = 0$ by symmetry; consequently, condition (3.47) holds in a trivial manner. Thus, we conclude that $\epsilon^2_{\text{OPT}}(S_k) \asymp \epsilon^2_{\text{GLR}}(S_k)$, showing that the GLRT is optimal over all tests.

3.3.2.2 Revisiting the circular cone

Recall the circular cone $K = \{\theta \in \mathbb{R}^d \mid \theta_1 \geq \|\theta\|_2 \cos(\alpha)\}$ for fixed $0 < \alpha < \pi/2$. In our earlier discussion, we proved that $\epsilon^2_{\text{GLR}}(K) \asymp \sigma^2$. Here let us verify that this scaling is optimal over all tests. By symmetry, we find that $\mathbb{E}\Pi_K g = \beta e_1 \in \mathbb{R}^d$, where e_1 denotes the standard Euclidean basis vector with a 1 in the first coordinate, and $\beta > 0$ is some scalar. For any vector $\eta \in K \cap S^{-1}$, we have $\eta_1 \geq \cos(\alpha)$, and hence

$$\inf_{\eta \in K \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle \geq \cos(\alpha)\beta \; = \; \cos(\alpha)\|\mathbb{E}\Pi_K g\|_2.$$

Consequently, we see that condition (3.47) is satisfied with $b = \frac{1}{\cos(\alpha)} > 0$, so that the GLRT is optimal over all tests for each fixed α. (To be clear, in this example, our theory does not provide a sharp bound uniformly over varying α.)

3.3.2.3 Revisiting the product cone

Recall from Section 3.3.1.4 our discussion of the Cartesian product cone $K_\times = \text{Circ}_{d-1}(\alpha) \times \mathbb{R}$. In this section, we establish that the GLRT, when applied to a testing problem based on this case, is sub-optimal by a factor of $\sqrt{d}$.

Let us first prove that the sufficient condition (3.47) is violated, so that Corollary 2 does *not* imply optimality of the GLRT. From our earlier calculations, we know that $\mathbb{E}\|\Pi_{K_\times} g\|_2 \asymp \sqrt{d}$. On the other hand, we also know that $\mathbb{E}\Pi_{K_\times} g$ is equal to zero in its last coordinate. Since the standard basis vector e_d belongs to the set $K_\times \cap S^{-1}$, we have

$$\inf_{\eta \in K_\times \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_{K_\times} g \rangle \leq \langle e_d, \, \mathbb{E}\Pi_{K_\times} g \rangle \; = \; 0,$$

so that condition (3.47) does not hold.

From this calculation alone, we cannot conclude that the GLRT is sub-optimal. So let us now compute the lower bound guaranteed by Theorem 3.3.2. From our previous discussion, we know that $\mathbb{E}\Pi_{K_\times} g = \beta e_1$ for some scalar $\beta > 0$. Moreover, we also have $\|\mathbb{E}\Pi_{K_\times} g\|_2 = \beta \asymp \sqrt{d}$; this scaling follows because we have $\|\mathbb{E}\Pi_{K_\times} g\|_2 = \|\mathbb{E}\Pi_{\text{Circ}_{d-1}(\alpha)} g\|_2 \asymp \sqrt{d-1}$, where we have used the previous inequality (3.36) for circular cone. Putting together the pieces, we find that Theorem 3.3.2 implies that

$$\epsilon^2_{\text{OPT}}(K_\times) \gtrsim \sigma^2, \tag{3.48}$$

which differs from the GLRT scaling in a factor of $\sqrt{d}$.

Does there exist a test that achieves the lower bound (3.48)? It turns out that a simple truncation test does so, and hence is optimal. To provide intuition for the test, observe that for any vector $\theta \in K_\times \cap S^{-1}$, we have $\theta_1^2 + \theta_d^2 \geq \cos^2(\alpha)$. To verify this claim, note that

$$\frac{1}{\cos^2(\alpha)}\Big(\theta_1^2 + \theta_d^2\Big) \geq \frac{\theta_1^2}{\cos^2(\alpha)} + \theta_d^2 \; \geq \; \sum_{j=1}^{d-1} \theta_j^2 + \theta_d^2 \; = \; 1.$$

Consequently, the two coordinates (y_1, y_d) provide sufficient information for constructing a good test. In particular, consider the truncation test

$$\varphi(y) := \mathbb{I}\big[\|(y_1, y_d)\|_2 \geq \beta\big],$$

for some threshold $\beta > 0$ to be determined. This can be viewed as a GLRT for testing the standard null against the alternative $\mathbb{R}^2$, and hence our general theory guarantees that it will succeed with separation $\epsilon^2 \gtrsim \sigma^2$. This guarantee matches our lower bound (3.48), showing that the truncation test is indeed optimal, and moreover, that the GLRT is sub-optimal by a factor of $\sqrt{d}$ for this particular problem.

We provide more intuition on why the the GLRT sub-optimal and use this intuition to construct a more general class of problem for which a similar sub-optimality is witnessed in Appendix A.1.

3.3.3 Detailed analysis of two cases

This section is devoted to a detailed analysis of the orthant cone, followed by the monotone cone. Here we find that the GLRT is again optimal for both of these cones, but establishing this optimality requires a more delicate analysis.

3.3.3.1 Revisiting the orthant cone

Recall from Section 3.3.1.5 our discussion of the (non-negative) orthant cone

$$K_+ := \{\theta \in \mathbb{R}^d \mid \theta_j \geq 0 \quad \text{for } j = 1, \ldots, d\},$$

where we proved that $\epsilon^2_{\text{GLR}}(K_+) \asymp \sigma^2\sqrt{d}$. Let us first show that the sufficient condition (3.47) does not hold, so that Corollary 2 does *not* imply optimality of the GLRT. As we have computed in our Section 3.3.1.5, quantity $\mathbb{E}\|\Pi_{K_+}(g)\|_2 \asymp \sqrt{d}$ and

$$\inf_{\eta \in K_+ \cap S^{-1}} \langle \eta, \, \mathbb{E}\Pi_K g\rangle \leq \langle e_1, \, \frac{1}{\sqrt{2\pi}}\,\mathbf{1}\rangle = \frac{1}{\sqrt{2\pi}},$$

where use the fact that $\mathbb{E}\Pi_{K_+}(g) = \frac{1}{\sqrt{2\pi}}\,\mathbf{1}$. So that condition (3.47) is violated.

Does this mean the GLRT is sub-optimal? It turns out that the GLRT is actually optimal over all tests, as we can demonstrate by proving a lower bound—tighter than the one given in Theorem 3.3.2—that matches the performance of the GLRT. We summarize it as follows:

Proposition 3.3.1. *There are numbers $\{\kappa_\rho, \rho \in (0, 1/2]\}$ such that for the (non-negative) orthant cone K_+, we have*

$$\inf_{\psi} \mathcal{E}(\psi; \{0\}, K_+, \epsilon) \geq \rho \qquad \textit{whenever } \epsilon^2 \leq \kappa_\rho\, \sigma^2\sqrt{d}. \tag{3.49}$$

See the Section A.3.1 for the proof of this proposition.
From Proposition 3.3.1, we see that the optimal testing radius satisfies $\epsilon^2_{\text{OPT}}(K_+) \gtrsim \sigma^2\sqrt{d}$. Compared to the GLRT radius $\epsilon^2_{\text{GLR}}(K_+)$ established in expression (3.39), it implies the optimality of the GLRT.

3.3.3.2 Revisiting the monotone cone

Recall the monotone cone given by $M := \{\theta \in \mathbb{R}^d \mid \theta_1 \leq \theta_2 \leq \cdots \leq \theta_d\}$. In our previous discussion in Section 3.3.1.6, we established that $\epsilon^2_{\text{GLR}}(M) \asymp \sigma^2\sqrt{\log d}$. We also pointed out that this scaling holds for a more general problem, namely, testing cone M versus linear subspace $L = \text{span}(\mathbf{1})$. In this section, we show that the GLRT is also optimal for both cases.

First, observe that Corollary 2 does not imply optimality of the GLRT. In particular, using symmetry of the inner product, we have shown in expression (3.41) that

$$\max\{0, \inf_{\eta \in M \cap S^{-1}} \langle \eta, \mathbb{E}\Pi_M g \rangle\} = 0,$$

for cone pair $(C_1, C_2) = (\{0\}, M)$. Also note that from Lemma 3.3.1 we know that for cone pair $(C_1, C_2) = (\text{span}(\mathbf{1}), M)$, there is a universal constant c such that

$$\inf_{\eta \in K \cap S^{-1}} \langle \eta, \mathbb{E}\Pi_K g \rangle \leq c, \qquad K := M \cap L^{\perp}.$$

In both cases, since $\mathbb{E}\|\Pi_K g\|_2 \asymp \sqrt{\log d}$, so that the sufficient condition (3.47) for GLRT optimality fails to hold.

It turns out that we can demonstrate a matching lower bound for $\epsilon^2_{\text{OPT}}(M)$ in a more direct way by carefully constructing a prior distribution on the alternatives and control the testing error. Doing so allows us to conclude that the GLRT is optimal, and we summarize our conclusions in the following:

Proposition 3.3.2. *There are numbers $\{\kappa_\rho, \rho \in (0, 1/2]\}$ such that for the monotone cone M and subspace $L = \{0\}$ or span$(\mathbf{1})$, we have*

$$\inf_{\psi} \mathcal{E}(\psi; L, M, \epsilon) \geq \rho \qquad \textit{whenever } \epsilon^2 \leq \kappa_\rho \, \sigma^2 \sqrt{\log(ed)}. \tag{3.50}$$

See Section A.3.2 for the proof of this proposition.

Proposition 3.3.2, equipped with previous achievable results by GLRT (3.40), gives a sharp rate characterization on the testing radius for both problem with regard to monotone cone:

$$\begin{aligned} \mathcal{H}_0 : \theta = 0 \quad &\text{versus} \quad \mathcal{H}_1 : \theta \in M \\ \text{and} \quad \mathcal{H}_0 : \theta \in \text{span}(\mathbf{1}) \quad &\text{versus} \quad \mathcal{H}_1 : \theta \in M. \end{aligned}$$

In both cases, the optimal testing radius satisfies $\epsilon^2_{\text{OPT}}(L, M, \rho) \asymp \sigma^2 \sqrt{\log(ed)}$. As a consequence, the GLRT is optimal up to an universal constant. As far as we know, the problem of testing a zero or constant vector versus the monotone cone as the alternative has not been fully characterized in any past work.

3.4 Discussion

In this chapter, we have studied the the problem of testing between two hypotheses that are specified by a pair of non-oblique closed convex cones. Our first main result provided a characterization, sharp up to universal multiplicative constants, of the testing radius achieved by the generalized likelihood ratio test. This characterization was geometric in nature, depending on a combination of the Gaussian width of an induced cone, and a second geometric parameter. Due to the combination of these parameters, our analysis shows that the GLRT can have very different behavior even for cones that have the same Gaussian width; for instance, compare our results for the circular and orthant cone in Section 3.3.1. It is worth noting that this behavior is in sharp contrast to the situation for estimation problem over convex sets, where it is understood that (localized) Gaussian widths completely determine the estimation error associated with the least-squares estimator [139, 37]. In this way, our analysis reveals a fundamental difference between minimax testing and estimation.

Our analysis also highlights some new settings in which the GLRT is non-optimal. Although past work [147, 107, 112] has exhibited non-optimality of the GLRT in certain settings, in the context of cones, all of these past examples involve oblique cones. In Section 3.3.1.4, we gave an example of sub-optimality which, to the best of our knowledge, is the first for a non-oblique pair of cones—namely, the cone $\{0\}$, and a certain type of Cartesian product cone.

Our work leaves open various questions, and we conclude by highlighting a few here. First, in Section 3.3.2, we proved a general information-theoretic lower bound for the minimax testing radius. This lower bound provides a sufficient condition for the GLRT to be minimax optimal up to constants. Despite being tight in many non-trivial situations, our information-theoretic lower bound is not tight for all cones; proving such a sharp lower bound is an interesting topic for future research. Second, as with a long line of past work on this topic [117, 108, 106, 147], our analysis is based on assuming that the noise variance σ^2 is known. In practice, this may or may not be a realistic assumption, and so it is interesting to consider the extension of our results to this setting.

We note that our minimax lower bounds are proved by constructing prior distributions on $\mathcal{H}_0$ and $\mathcal{H}_1$ and then control the distance between marginal likelihood functions. Following this idea, we can also consider our testing problem in the Bayesian framework. Without any prior preference on which hypothesis to take, we will let $\Pr(\mathcal{H}_0) = \Pr(\mathcal{H}_1) = 1/2$; thus the Bayesian testing procedure makes decision based on quantity

$$B_{01} := \frac{m(y \mid \mathcal{H}_0)}{m(y \mid \mathcal{H}_1)} = \frac{\int_{\theta \in C_1} \mathbb{P}_\theta(y)\pi_1(\theta)d\theta}{\int_{\theta \in C_2} \mathbb{P}_\theta(y)\pi_2(\theta)d\theta}, \tag{3.51}$$

which is often called Bayesian factor in literature. Analyzing the behavior of this statistic is an interesting direction to pursue in the future.

3.5 Proofs of main results

We now turn to the proofs of our main results, with the proof of Theorems 3.3.1 and 3.3.2 given in Sections 3.5.1 and 3.5.2 respectively. In all cases, we defer the proofs of certain more technical lemmas to the appendices.

3.5.1 Proof of Theorem 3.3.1

Since the cones (C_1, C_2) are both invariant under rescaling by positive numbers, we can first prove the result for noise level $\sigma = 1$, and then recapture the general result by rescaling appropriately. Thus, we fix $\sigma = 1$ throughout the remainder of the proof so as to simplify notation. Moreover, let us recall that the GLRT consists of tests of the form $\phi_\beta(y) := \mathbb{I}(T(y) \geq \beta)$, where the likelihood ratio $T(y)$ is given in equation (3.11a). Note here the cut-off $\beta \in [0, \infty)$ is a constant that does not depend on the data vector y.

By the previously discussed equivalence (3.26), we can focus our attention on the simpler problem $\mathcal{T}(\{0\}, K; \epsilon)$, where $K = C_2 \cap C_1^*$. By the monotonicity of the square function for positive numbers, the GLRT is controlled by the behavior of the statistic $\|\Pi_K(y)\|_2$, and in particular how it varies depending on whether y is drawn according to $\mathcal{H}_0$ or $\mathcal{H}_1$.

Letting $g \in \mathbb{R}^d$ denote a standard Gaussian random vector, let us introduce the random variable $Z(\theta) := \|\Pi_K(\theta + g)\|_2$ for each $\theta \in \mathbb{R}^d$. Observe that the statistic $\|\Pi_K(y)\|_2$ is distributed according to $Z(0)$ under the null $\mathcal{H}_0$, and according to $Z(\theta)$ for some $\theta \in K$ under the alternative $\mathcal{H}_1$. The Lemma A.4.1 which is stated and proved in Appendix A.4.1 guarantees random variables of the type $Z(\theta)$ and $\langle \theta, \Pi_K g \rangle$ are sharply concentrated around their expectations.

As shown in the sequel, using the concentration bound (A.15a), the study of the GLRT can be reduced to the problem of bounding the mean difference

$$\Gamma(\theta) := \mathbb{E}\left(\|\Pi_K(\theta + g)\|_2 - \|\Pi_K g\|_2\right) \tag{3.52}$$

for each $\theta \in K$. In particular, in order to prove the achievability result stated in part (a) of Theorem 3.3.1, we need to lower bound $\Gamma(\theta)$ uniformly over $\theta \in K$, whereas a uniform upper bound on $\Gamma(\theta)$ is required in order to prove the negative result in part (b).

3.5.1.1 Proof of GLRT achievability result (Theorem 3.3.1(a))

By assumption, we can restrict our attention to alternative distributions defined by vectors $\theta \in K$ satisfying the lower bound $\|\theta\|_2^2 \geq B_\rho \, \delta_{\mathrm{LR}}^2(\{0\}, K)$, where for every target level $\rho \in (0, 1)$, constant B_ρ is chosen such that

$$B_\rho := \max\left\{32\pi, \;\inf\left(B > 0 \mid \frac{B^{1/2}}{(2^7\pi B)^{1/4} + 16} - \frac{2}{\sqrt{e}} \geq \sqrt{-8\log(\rho/2)}\right)\right\}.$$

Since function $f(x) := \frac{x^{1/2}}{(2^7\pi x)^{1/4}+16} - \frac{2}{\sqrt{e}}$ is strictly increasing and goes to infinity, so that the constant B_ρ defined above is always finite.

We first claim that it suffices to show that for such vector, the difference (3.52) is lower bounded as

$$\Gamma(\theta) \geq \frac{B_\rho^{1/2}}{(2^7\pi B_\rho)^{1/4} + 16} - \frac{2}{\sqrt{e}} = f(B_\rho). \tag{3.53}$$

Taking inequality (3.53) as given for the moment, we claim that the test

$$\phi_\tau(y) = \mathbb{I}[\|\Pi_K(y)\|_2^2 \geq \tau] \qquad \text{with threshold } \tau := (\tfrac{1}{2} f(B_\rho) + \mathbb{E}[\|\Pi_K(g)\|_2])^2$$

has uniform error probability controlled as

$$\mathcal{E}(\phi_\tau; \{0\}, K, \epsilon) := \mathbb{E}_0[\phi_\tau(y)] + \sup_{\theta \in K, \|\theta\|_2^2 \geq \epsilon^2} \mathbb{E}_\theta[1 - \phi_\tau(y)] \;\leq\; 2e^{-f^2(B_\rho)/8} < \rho. \tag{3.54}$$

where the last inequality follows from the definition of B_ρ.

Establishing the error control (3.54) Beginning with errors under the null $\mathcal{H}_0$, we have

$$\begin{aligned}\mathbb{E}_0[\phi_\tau(y)] = \mathbb{P}_0(\|\Pi_K g\|_2 \geq \sqrt{\tau}) &= \mathbb{P}_0\big[\|\Pi_K g\|_2 - \mathbb{E}[\|\Pi_K g\|_2] \geq f(B_\rho)/2\big] \\ &\leq \exp(-f^2(B_\rho)/8),\end{aligned}$$

where the final inequality follows from the concentration bound (A.15a) in Lemma A.4.1, as along as $f(B_\rho) > 0$.

On the other hand, we have

$$\begin{aligned}\sup_{\theta \in K, \|\theta\|_2^2 \geq \epsilon^2} \mathbb{E}_\theta[1 - \phi_\tau(y)] &= \mathbb{P}\Big[\|\Pi_K(\theta + g)\|_2 \leq \frac{1}{2} f(B_\rho) + \mathbb{E}\|\Pi_K g\|_2\Big] \\ &= \mathbb{P}\Big[\|\Pi_K(\theta + g)\|_2 - \mathbb{E}\|\Pi_K(\theta + g)\|_2 \leq \frac{1}{2} f(B_\rho) - \Gamma(\theta)\Big],\end{aligned}$$

where the last equality follows by substituting $\Gamma(\theta) = \mathbb{E}[\|\Pi_K(\theta + g)\|_2] - \mathbb{E}[\|\Pi_K g\|_2]$. Since the lower bound (3.53) guarantees that $\frac{1}{2} f(B_\rho) - \Gamma(\theta) \leq -\frac{1}{2} f(B_\rho)$, we find that

$$\begin{aligned}\sup_{\theta \in K, \|\theta\|_2^2 \geq \epsilon^2} \mathbb{E}_\theta[1 - \phi_\tau(y)] &\leq \mathbb{P}\Big[\|\Pi_K(\theta + g)\|_2 - \mathbb{E}\|\Pi_K(\theta + g)\|_2 \leq -\frac{1}{2} f(B_\rho)\Big] \\ &\leq \exp(-f^2(B_\rho)/8),\end{aligned}$$

where the final inequality again uses the concentration inequality (A.15a). Putting the pieces together yields the claim (3.54).

The only remaining detail is to prove the lower bound (3.53) on the difference (3.52). To prove inequality (3.53), we make use of the following auxiliary Lemma 3.5.1.

Lemma 3.5.1. *For every closed convex cone K and vector $\theta \in K$, we have the lower bounds*

$$\Gamma(\theta) \geq \frac{\|\theta\|_2^2}{2\|\theta\|_2 + 8\mathbb{E}\|\Pi_K g\|_2} - \frac{2}{\sqrt{e}}. \tag{3.55a}$$

Moreover, for any vector θ that also satisfies the inequality $\langle \theta, \mathbb{E}\Pi_K g\rangle \geq \|\theta\|_2^2$, we have

$$\Gamma(\theta) \geq \alpha^2(\theta) \frac{\langle \theta, \mathbb{E}\Pi_K g\rangle - \|\theta\|_2^2}{\alpha(\theta)\|\theta\|_2 + 2\mathbb{E}\|\Pi_K g\|_2} - \frac{2}{\sqrt{e}}, \tag{3.55b}$$

where $\alpha(\theta) := 1 - \exp\left(\frac{-\langle \theta, \mathbb{E}\Pi_K g\rangle^2}{8\|\theta\|_2^2}\right)$.

See Appendix A.4.2 for the proof of this claim.

We now use Lemma 3.5.1 to prove the lower bound (3.53). Note that the inequality $\|\theta\|_2^2 \geq B_\rho \delta_{\mathrm{LR}}^2(\{0\}, K)$ implies that one of the following two lower bounds must hold:

$$\|\theta\|_2^2 \geq B_\rho \mathbb{E}\|\Pi_K g\|_2, \tag{3.56a}$$

$$\text{or} \quad \langle \theta,\, \mathbb{E}\Pi_K g\rangle \geq \sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2. \tag{3.56b}$$

We will analyze these two cases separately.

Case 1 In order to show that the lower bound (3.56a) implies inequality (3.53), we will prove a stronger result—namely, that the inequality $\|\theta\|_2^2 \geq \sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2/2$ implies that inequality (3.53) holds.

From the lower bound (3.55a) and the fact that, for each fixed $a > 0$, the function $x \mapsto x^2/(2x + a)$ is increasing on the interval $[0, \infty)$, we find that

$$\Gamma(\theta) \geq \frac{\sqrt{B_\rho \mathbb{E}\|\Pi_K g\|_2}/2}{\sqrt{2}B_\rho^{1/4} + 8\sqrt{\mathbb{E}\|\Pi_K g\|_2}} - \frac{2}{\sqrt{e}}.$$

Further, because of general bound (3.21) that $\mathbb{E}\|\Pi_K g\|_2 \geq 1/\sqrt{2\pi}$ and the fact that the function $x \mapsto x/(a + x)$ is increasing in x, we obtain

$$\Gamma(\theta) \geq \frac{\sqrt{B_\rho}}{2(8\pi B_\rho)^{1/4} + 16} - \frac{2}{\sqrt{e}},$$

which ensures inequality (3.53).

Case 2 We now turn to the case when inequality (3.56b) is satisfied. We may assume the inequality $\|\theta\|_2^2 \geq \sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2/2$ is violated because otherwise, inequality (3.53) follows immediately. When this inequality is violated, we have

$$\langle \theta,\, \mathbb{E}\Pi_K g\rangle \geq \sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2 \quad \text{and} \quad \|\theta\|_2^2 < \sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2/2. \tag{3.57}$$

Our strategy is to make use of inequality (3.55b), and we begin by bounding the quantity α appearing therein. By combining inequality (3.57) and inequality (3.21)—namely, $\mathbb{E}\|\Pi_K g\|_2 \geq 1/\sqrt{2\pi}$, we find that

$$\alpha \geq 1 - \exp\left(-\frac{\sqrt{B_\rho}\mathbb{E}\|\Pi_K g\|_2}{4}\right) \geq 1 - \exp\left(-\frac{\sqrt{B_\rho}}{4\sqrt{2\pi}}\right) \geq 1/2, \quad \text{whenever } B_\rho \geq 32\pi.$$

Using expression (3.57), we deduce that

$$\Gamma(\theta) \;\geq\; \frac{\alpha^2\sqrt{B_\rho \mathbb{E}\|\Pi_K g\|_2}}{\alpha(4B_\rho)^{1/4} + 4\sqrt{\mathbb{E}\|\Pi_K g\|_2}} - \sqrt{\frac{2}{e}} \;\geq\; \frac{\sqrt{B_\rho \mathbb{E}\|\Pi_K g\|_2}}{(2^6 B_\rho)^{1/4} + 16\sqrt{\mathbb{E}\|\Pi_K g\|_2}} - \sqrt{\frac{2}{e}}.$$

where the second inequality uses the previously obtained lower bound $\alpha > 1/2$, and the fact that the function $x \mapsto x^2/(x + b)$ is increasing in x.

This completes the proof of inequality (3.53) thus completing the GLRT achievability result.

3.5.1.2 Proof of GLRT lower bound (Theorem 3.3.1(b))

We divide our proof into two scenarios, depending on whether or not $\mathbb{E}\|\Pi_K g\|_2$ is less than 128.

Case $\mathbb{E}\|\Pi_K g\|_2 < 128$ We begin by setting $b_\rho = \frac{1}{256}$. The assumed bound $\epsilon^2 \leq \frac{1}{256}\delta^2_{\text{LR}}(\{0\}, K)$ then implies that

$$\epsilon^2 \leq \frac{1}{256}\delta^2_{\text{LR}}(\{0\}, K) \leq \frac{\mathbb{E}\|\Pi_K g\|_2}{256} < \frac{1}{2}.$$

For every $\epsilon^2 \leq \frac{1}{2}$, we claim that $\mathcal{E}(\phi; \{0\}, K, \epsilon) \geq 1/2$. Note that the uniform error $\mathcal{E}(\phi; \{0\}, K, \epsilon)$ is at least as large as the error in the simple binary test

$$\mathcal{H}_0 : y \sim N(0, I_d) \quad \text{versus} \quad \mathcal{H}_1 : y \sim N(\theta, I_d), \tag{3.58a}$$

where $\theta \in K$ is any vector such that $\|\theta\|_2 = \epsilon$. We claim that the error for the simple binary test (3.58a) is lower bounded as

$$\inf_\psi \mathcal{E}(\psi; \{0\}, \{\theta\}, \epsilon) \geq 1/2 \qquad \text{whenever } \epsilon^2 \leq 1/2. \tag{3.58b}$$

The proof of this claim is straightforward: introducing the shorthand $\mathbb{P}_\theta = N(\theta, I_d)$ and $\mathbb{P}_0 = N(0, I_d)$, we have

$$\inf_\psi \mathcal{E}(\psi; \{0\}, \{\theta\}, \epsilon) = 1 - \|\mathbb{P}_\theta - \mathbb{P}_0\|_{\text{TV}}.$$

Using the relation between χ^2 distance and TV-distance in expression (A.1c) and the fact that $\chi^2(\mathbb{P}_\theta, \mathbb{P}_0) = \exp(\epsilon^2) - 1$, we find that the testing error satisfies

$$\inf_\psi \mathcal{E}(\psi; \{0\}, \{\theta\}, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\exp(\epsilon^2) - 1} \geq 1/2, \qquad \text{whenever } \epsilon^2 \leq 1/2.$$

(See Section A.2 for more details on the relation between the TV and χ^2-distances.) This completes the proof under the condition $\mathbb{E}\|\Pi_K g\|_2 < 128$.

Case $\mathbb{E}\|\Pi_K g\|_2 \geq 128$ In this case, our strategy is to exhibit some $\theta \in \mathcal{H}_1$ for which the expected difference $\Gamma(\theta) = \mathbb{E}\left(\|\Pi_K(\theta + g)\|_2 - \|\Pi_K g\|_2\right)$ is small, which then leads to significant error when using the GLRT. In order to do so, we require an auxiliary lemma (Lemma A.5.1) to suitable control $\Gamma(\theta)$ which is stated and proved in Appendix A.5.1.

We now proceed to prove our main claim. Based on Lemma A.5.1, we claim that if $\epsilon^2 \leq b_\rho \delta^2_{\text{LR}}(\{0\}, K)$ for a suitably small constant b_ρ such that

$$b_\rho := \sup\left\{ b_\rho > 0 \mid 12\sqrt{b_\rho} + 3\sqrt{b_\rho}\left(\frac{2}{e}\right)^{1/4} + 24\sqrt{\frac{b_\rho}{2e}} \leq \frac{1}{16} \right\},$$

then

$$\Gamma(\theta) \leq \frac{1}{16}, \qquad \text{for some } \theta \in K, \ \|\theta\|_2 \geq \epsilon. \tag{3.59}$$

We take inequality (3.59) as given for now, returning to prove it in our appendix A.5.2. In summary, then, we have exhibited some $\theta \in \mathcal{H}_1$—namely, a vector $\theta \in K$ with $\|\theta\|_2 \geq \epsilon$—such that $\Gamma(\theta) \leq 1/16$. This special vector θ plays a central role in our proof.

We claim that the GLRT cannot succeed with error smaller than 0.11 no matter how the cut-off β is chosen. In order to see this, firstly the following lemma allows us to relate $\|\Pi_K g\|_2$ to its expectation:

Lemma 3.5.2. *Given every closed convex cone K such that $\mathbb{E}\|\Pi_K g\|_2 \geq 128$, we have*

$$\mathbb{P}(\|\Pi_K g\|_2 > \mathbb{E}\|\Pi_K g\|_2) \ > \ 7/16. \tag{3.60}$$

See Appendix A.5.3 for the proof of this claim.

For future reference, we note that it is relatively straightforward to show that the random variable $\|\Pi_K g\|_2$ is distributed as a mixture of χ-distributions, and indeed, the Lemma 3.5.2 can be proved via this route. Raubertas et al. [117] proved that the squared quantity $\|\Pi_K g\|_2^2$ is a mixture of χ^2 distributions, and a very similar argument yields the analogous statement for $\|\Pi_K g\|_2$.

We are now ready to calculate the testing error for the GLRT given in equation (3.11b). Our goal is to lower bound the error $\mathcal{E}(\phi_\beta; \{0\}, K, \epsilon)$ uniformly over the chosen threshold $\beta \in [0, \infty)$. We divide the choice of β into three cases, depending on the relationship between β and $\mathbb{E}\|\Pi_K g\|_2$, $\mathbb{E}\|\Pi_K(\theta + g)\|_2$. Notice this particular θ is chosen to be the one that satisfies inequality (3.59).

Case 1 First, consider a threshold $\beta \in [0, \ \mathbb{E}\|\Pi_K g\|_2]$. It then follows immediately from inequality (3.60) that the type I error by its own satisfies

$$\text{type I error} = \mathbb{P}_0(\|\Pi_K y\|_2 \geq \beta) \geq \mathbb{P}(\|\Pi_K g\|_2 \geq \mathbb{E}\|\Pi_K g\|_2) \geq \frac{7}{16}.$$

Case 2 Otherwise, consider a threshold $\beta \in \big(\mathbb{E}\|\Pi_K g\|_2, \ \mathbb{E}\|\Pi_K(\theta + g)\|_2\big]$. In this case, we again use inequality (3.60) to bound the type I error, namely

$$\begin{aligned}\text{type I error} &= \mathbb{P}_0(\|\Pi_K y\|_2 \geq \beta) \\ &= \mathbb{P}\Big[\|\Pi_K g\|_2 \geq \mathbb{E}\|\Pi_K g\|_2\Big] - \mathbb{P}\Big[\|\Pi_K g\|_2 \in [\mathbb{E}\|\Pi_K g\|_2, \beta)\Big] \\ &\geq \frac{7}{16} - \max_x\{f_{\|\Pi_K g\|_2}(x)(\beta - \mathbb{E}\|\Pi_K g\|_2)\},\end{aligned}$$

where we use $f_{\|\Pi_K g\|_2}$ to denote the density function of the random variable $\|\Pi_K g\|_2$ As discussed earlier, the random variable $\|\Pi_K g\|_2$ is distributed as a mixture of χ-distributions;

in particular, see Lemma 3.5.2 above and the surrounding discussion for details. As can be verified by direct numerical calculation, any χ_k variable has a density that bounded from above by 4/5. Using this fact, we have

$$\text{type I error} \geq \frac{7}{16} - \frac{4}{5}(\beta - \mathbb{E}\|\Pi_K g\|_2) \overset{(i)}{\geq} \frac{7}{16} - \frac{4}{5}\Gamma(\theta) \overset{(ii)}{>} 3/8,$$

where step (i) follows by the assumption that β belongs to the interval $\big(\mathbb{E}\|\Pi_K g\|_2,\ \mathbb{E}\|\Pi_K(\theta+g)\|_2\big]$, and step (ii) follows since $\Gamma(\theta) \leq 1/16$.

Case 3 Otherwise, given a threshold $\beta \in \big(\mathbb{E}\|\Pi_K(g+\theta)\|_2, \infty\big)$, we define the scalar $x := \beta - \mathbb{E}\|\Pi_K(g+\theta)\|_2$. From the concentration inequality given in Lemma A.4.1, we can deduce that

$$\begin{aligned}\text{type II error} &\geq \mathbb{P}_\theta(\|\Pi_K y\|_2 \leq \beta) \\ &= 1 - \mathbb{P}\Big(\|\Pi_K(\theta+g)\|_2 - \mathbb{E}\|\Pi_K(\theta+g)\|_2 > \beta - \mathbb{E}\|\Pi_K(\theta+g)\|_2\Big) \\ &\geq 1 - \exp(-x^2/2).\end{aligned}$$

At the same time,

$$\begin{aligned}\text{type I error} = \mathbb{P}_0(\|\Pi_K y\|_2 \geq \beta) &= \mathbb{P}(\|\Pi_K g\|_2 \geq \mathbb{E}\|\Pi_K g\|_2) - \mathbb{P}(\|\Pi_K g\|_2 \in [\mathbb{E}\|\Pi_K g\|_2, \beta)) \\ &\geq \frac{7}{16} - \frac{4}{5}(\beta - \mathbb{E}\|\Pi_K g\|_2),\end{aligned}$$

where we again use inequality (3.60) and the boundedness of the density of $\|\Pi_K g\|_2$. Recalling that we have defined $x := \beta - \mathbb{E}\|\Pi_K(g+\theta)\|_2$ as well as $\Gamma(\theta) = \mathbb{E}\big(\|\Pi_K(\theta+g)\|_2 - \|\Pi_K g\|_2\big)$, we have

$$\beta - \mathbb{E}\|\Pi_K g\|_2 = x + \Gamma(\theta) \ \leq \ x + \frac{1}{16},$$

where the last step uses the fact that $\Gamma(\theta) \leq 1/16$. Consequently, the type I error is lower bounded as

$$\text{type I error} \geq \frac{7}{16} - \frac{4}{5}(x + 1/16) \ = \ \frac{31}{80} - \frac{4}{5}x.$$

Combining the two types of error, we find that the testing error is lower bounded as

$$\inf_{x>0}\Big\{(\frac{31}{80} - \frac{4}{5}x)^+ + 1 - \exp(-x^2/2)\Big\} = 1 - \exp(-\frac{31^2}{2\times 64^2}) \geq 0.11.$$

Putting pieces together, the GLRT cannot succeed with error smaller than 0.11 no matter how the cut-off β is chosen.

3.5.2 Proof of Theorem 3.3.2

We now turn to the proof of Theorem 3.3.2. As in the proof of Theorem 3.3.1, we can assume without loss of generality that $\sigma = 1$. Since $0 \in C_1$ and $K := C_2 \cap C_1^* \subseteq C_2$, it suffices to prove a lower bound for the reduced problem of testing

$$\mathcal{H}_0 : \theta = 0, \quad \text{versus} \quad \mathcal{H}_1 : \|\theta\|_2 \geq \epsilon, \ \theta \in K.$$

Let $B(1) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 < 1\}$ denotes the open Euclidean ball of radius 1, and let $B^c(1) := \mathbb{R}^d \backslash B(1)$ denotes its complement.

We divide our analysis into two cases, depending on whether or not $\mathbb{E}\|\Pi_K g\|_2$ is less than 7. In both cases, let us set $\kappa_\rho = 1/14$.

Case 1 Suppose that $\mathbb{E}\|\Pi_K g\|_2 < 7$. In this case,

$$\epsilon^2 \leq \kappa_\rho \delta^2_{\text{OPT}}(\{0\}, K) \leq \kappa_\rho \mathbb{E}\|\Pi_K g\|_2 < 1/2.$$

Similar to our proof of Theorem 3.3.1(b), Case 1, by reducing to the simple verses simple testing problem (3.58a), any test yields testing error no smaller than 1/2 if $\epsilon^2 < 1/2$. So our lower bound directly holds for the case when $\mathbb{E}\|\Pi_K g\|_2 < 7$.

Case 2 Otherwise, suppose we have $\mathbb{E}\|\Pi_K g\|_2 \geq 7$. The following lemma provides a generic way to lower bound the testing error.

Lemma 3.5.3. *For every non-trivial closed convex cone K and probability measure $\mathbb{Q}$ supported on $K \cap B^c(1)$, the testing error is lower bounded as*

$$\inf_\psi \mathcal{E}(\psi; \{0\}, K, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle\eta,\, \eta'\rangle) - 1}, \tag{3.61}$$

where $\mathbb{E}_{\eta,\eta'}$ denotes expectation with respect to an i.i.d pair $\eta, \eta' \sim \mathbb{Q}$.

See Appendix A.6.1 for the proof of this claim.

We apply Lemma 3.5.3 with the probability measure $\mathbb{Q}$ defined as

$$\mathbb{Q}(A) := \mathbb{P}\left(\frac{\Pi_K g}{\mathbb{E}\|\Pi_K g\|_2/2} \in A \;\middle|\; \|\Pi_K g\|_2 \geq \mathbb{E}\|\Pi_K g\|_2/2\right), \tag{3.62}$$

for measurable set $A \subset \mathbb{R}^d$ where g denotes a standard d-dimensional Gaussian random vector i.e., $g \sim N(0, I_d)$. It is easy to check that measure $\mathbb{Q}$ is supported on $K \cap B^c(1)$. We make use of Lemma A.6.1 in Appendix A.6.2 to control $\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle\eta,\, \eta'\rangle)$ and thus upper bounding the testing error.

We now lower bound the testing error when $\epsilon^2 \leq \kappa_\rho\, \delta^2_{\text{OPT}}(\{0\}, K)$. By definition of $\delta^2_{\text{OPT}}(\{0\}, K)$, the inequality $\epsilon^2 \leq \kappa_\rho\, \delta^2_{\text{OPT}}(\{0\}, K)$ implies that

$$\epsilon^2 \leq \kappa_\rho \mathbb{E}\|\Pi_K g\|_2 \qquad \text{and} \quad \epsilon^2 \leq \kappa_\rho \left(\frac{\mathbb{E}\|\Pi_K g\|_2}{\|\mathbb{E}\Pi_K g\|_2}\right)^2.$$

The first inequality above implies, with $\kappa_\rho = 1/14$, that $\epsilon^2 \leq \mathbb{E}\|\Pi_K g\|_2/14 \leq (\mathbb{E}\|\Pi_K g\|_2)^2/32$ (note that $\mathbb{E}\|\Pi_K g\|_2 \geq 7$). Therefore the assumption in Lemma A.6.1 is satisfied so that inequality (A.40) gives

$$\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2 \langle \eta,\, \eta' \rangle) \;\leq\; \frac{1}{a^2} \exp\left(5\kappa_\rho + \frac{40\kappa_\rho^2 \mathbb{E}(\|\Pi_K g\|_2^2)}{(\mathbb{E}\|\Pi_K g\|_2)^2}\right). \tag{3.63}$$

So it suffices to control the right hand side above. From the concentration result in Lemma A.4.1, we obtain

$$a = \mathbb{P}(\|\Pi_K g\|_2 - \mathbb{E}\|\Pi_K g\|_2 \geq -\frac{1}{2}\mathbb{E}\|\Pi_K g\|_2) \;\geq\; 1 - \exp(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}) > 1 - \exp(-6),$$

where the last step uses $\mathbb{E}\|\Pi_K g\|_2 \geq 7$, and

$$\mathbb{E}\|\Pi_K g\|_2^2 = (\mathbb{E}\|\Pi_K g\|_2)^2 + \mathrm{var}(\|\Pi_K g\|_2) \leq (\mathbb{E}\|\Pi_K g\|_2)^2 + 4.$$

Here the last inequality follows from the fact that $\mathrm{var}(\|\Pi_K g\|_2) \leq 4$—see Lemma A.4.1. Plugging these two inequalities into expression (3.63) gives

$$\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2 \langle \eta,\, \eta' \rangle) \;\leq\; \left(\frac{1}{1 - \exp(-6)}\right)^2 \exp\left(5\kappa_\rho + 40\kappa_\rho^2 + \frac{160\kappa_\rho^2}{(\mathbb{E}\|\Pi_K g\|_2)^2}\right),$$

where the right hand side is less than 2 when $\kappa_\rho = 1/14$ and $\mathbb{E}\|\Pi_K g\|_2 \geq 7$. Combining with inequality (3.61) forces the testing error to be lower bounded as

$$\forall \psi, \quad \mathcal{E}(\psi; \{0\}, K, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2 \langle \eta,\, \eta' \rangle) - 1} \geq \frac{1}{2} > \rho,$$

which completes the proof of Theorem 3.3.2.

Chapter 4

Adaptive estimation of planar convex sets

4.1 Introduction

In this chapter, we discuss the problem of nonparametric estimation of an unknown planar compact, convex set from noisy measurements of its support function on a uniform grid. Before describing the details of the problem, let us first introduce the support function. For a compact, convex set K in $\mathbb{R}^2$, its support function is defined by

$$h_K(\theta) := \max_{(x_1,x_2)\in K} (x_1 \cos\theta + x_2 \sin\theta) \qquad \text{for } \theta \in \mathbb{R}.$$

Note that h_K is a periodic function with period 2π. It is useful to think about θ in terms of the direction $(\cos\theta, \sin\theta)$. The line $x_1 \cos\theta + x_2 \sin\theta = h_K(\theta)$ is a support line for K (i.e., it touches K and K lies on one side of it). Conversely, every support line of K is of this form for some θ. The convex set K is completely determined by the its support function h_K because $K = \bigcap_\theta \{(x_1, x_2) : x_1 \cos\theta + x_2 \sin\theta \le h_K(\theta)\}$.

The support function h_K possesses the *circle-convexity* property (see, e.g., [140]): for every $\alpha_1 > \alpha > \alpha_2$ and $0 < \alpha_1 - \alpha_2 < \pi$,

$$\frac{h_K(\alpha_1)}{\sin(\alpha_1 - \alpha)} + \frac{h_K(\alpha_2)}{\sin(\alpha - \alpha_2)} \ge \frac{\sin(\alpha_1 - \alpha_2)}{\sin(\alpha_1 - \alpha)\sin(\alpha - \alpha_2)} h_K(\alpha). \tag{4.1}$$

Moreover the above inequality characterizes h_K, i.e., any periodic function of period 2π satisfying the above inequality equals h_K for a unique compact, convex subset K in $\mathbb{R}^2$. The circle-convexity property (4.1) is clearly related to the usual convexity property. Indeed, replacing $\sin\alpha$ by α in (4.1) leads to the condition for convexity. In spite of this similarity, (4.1) is different from convexity as can be seen from the example of the function $h(\theta) = |\sin\theta|$ which satisfies (4.1) but is clearly not convex.

4.1.1 The problem, motivations, and background

We are now ready to describe the problem studied in this chapter. Let K^* be an unknown compact, convex set in $\mathbb{R}^2$. We study the problem of estimating K^* or h_{K^*} from noisy measurements of h_{K^*}. Specifically, we observe data $(\theta_1, Y_1), \dots, (\theta_n, Y_n)$ drawn according to the model

$$Y_i = h_{K^*}(\theta_i) + \xi_i \qquad \text{for } i = 1, \dots, n \tag{4.2}$$

where $\theta_1, \dots, \theta_n$ are fixed grid points in $(-\pi, \pi]$ and $\xi_1, \dots, \xi_n$ are i.i.d Gaussian random variables with mean zero and known variance σ^2. We focus on the dual problems of estimating the scalar quantity $h_{K^*}(\theta_i)$ for each $1 \le i \le n$ as well as the convex set K^*. In this chapter, we propose data-driven adaptive estimators and establish their optimality for both of these problems.

The problem considered here has a range of applications in engineering. The regression model (4.2) was first proposed and studied by Prince and Willsky [115] who were motivated by an application to Computed Tomography. Lele et al. [95] showed how solutions to this problem can be applied to target reconstruction from resolved laser-radar measurements in the presence of registration errors. Gregor and Rannou [67] considered application to Projection Magnetic Resonance Imaging. It is also a fundamental problem in geometric tomography; see Gardner [57]. Another application domain where this problem might plausibly arise is robotic tactical sensing as has been suggested by Prince and Willsky [115]. Finally this is a natural shape constrained estimation problem and would fit right into the recent literature on shape constrained estimation (e.g. [70]).

Most proposed procedures for estimating K^* in this setting are based on least squares minimization. The least squares estimator $\hat{K}_{\text{ls}}$ is defined as any minimizer of $\sum_{i=1}^n (Y_i - h_K(\theta_i))^2$ as K ranges over all compact convex sets. The minimizer in this optimization problem is not unique and one can always take it to be a polytope. This estimator was first proposed by [115] who also proposed an algorithm for computing it based on quadratic programming. Further algorithms for computing $\hat{K}_{\text{ls}}$ were proposed in Prince et al. [115, 95, 58].

The theoretical performance of the least squares estimator was first considered by Gardner et al. [59] who mainly studied its accuracy for estimating K^* under the natural fixed design loss:

$$L_f(K^*, \hat{K}) := \frac{1}{n} \sum_{i=1}^{n} \left(h_{K^*}(\theta_i) - h_{\hat{K}}(\theta_i)\right)^2. \tag{4.3}$$

The key result of Gardner et al. [59] (specialized to the planar case that we are studying) states that $L_f(K^*, \hat{K}_{\text{ls}}) = O(n^{-4/5})$ as $n \to \infty$ almost surely provided K^* is contained in a ball of bounded radius. This result is complemented by the minimax lower bound in Guntuboyina [74] where it was shown that $n^{-4/5}$ is the minimax rate for this problem. These two results together imply minimax optimality of $\hat{K}_{\text{ls}}$ under the loss function L_f. No other theoretical results for this problem are available outside of those in Gardner et al. [59] and Guntuboyina [74].

As a result, the following basic questions are still unanswered:

1. How to optimally and adaptively estimate $h_{K^*}(\theta_i)$ for a fixed $i \in \{1, \ldots, n\}$? This is the pointwise estimation problem. In the literature on shape constrained estimation, pointwise estimation has been well studied. Prominent examples include [23, 152, 68, 69, 34, 36, 83] for monotonicity constrained estimation and [77, 100, 71, 72, 29] for convexity constrained estimation. For the problem considered here however, nothing is known about pointwise estimation. It may be noted that the result $L_f(K^*, \hat{K}_{\text{ls}}) = O(n^{-4/5})$ of Gardner et al. [59] does not say anything about the accuracy of $h_{\hat{K}_{\text{ls}}}(\theta_i)$ as an estimator for $h_{K^*}(\theta_i)$.

2. How to construct minimax optimal estimators for the set K^* that also adapt to polytopes? Polytopes with a small number of extreme points have a much simpler structure than general convex sets. In the problem of estimating convex sets under more standard observation models different from the one studied here, it is possible to construct estimators that converge at faster rates for polytopes compared to the overall minimax rate (see Brunk [22] for a summary of this theory). Similar kinds of adaptation has been recently studied for monotonicity and convexity constrained estimation problems, see [75, 38, 8]. Based on these results, it is natural to expect minimax estimators that adapt to polytopes in this problem. This has not been addressed previously.

4.1.2 Overview of our results

We will answer both the above questions in the affirmative in this chapter. The main contributions can be summarized as follows:

1. We study the pointwise adaptive estimation problem in detail in the decision theoretic framework where the focus is on the performance at every function, instead of the maximum risk over a large parameter space as in the conventional minimax theory in nonparametric estimation literature. Recall that this framework which has been discussed in our Section 2.2.1, is first introduced in Cai and Low [29] for shape constrained regression and provides a much more precise characterization of the performance of an estimator than the conventional minimax theory does.

   In the context of the present problem, the difficulty of estimating $h_{K^*}(\theta_i)$ at a given K^* and θ_i can be expressed by means of a benchmark $R_n(K^*, \theta)$ which is defined as follows (below $\mathbb{E}_L$ denotes expectation taken with respect to the joint distribution of $Y_1, \ldots, Y_n$ generated according to the model (4.2) with K^* replaced by L):

$$R_n(K^*, \theta) = \sup_{L} \inf_{\tilde{h}} \max \left(\mathbb{E}_{K^*}(\tilde{h} - h_{K^*}(\theta))^2, \; \mathbb{E}_L(\tilde{h} - h_L(\theta))^2 \right), \tag{4.4}$$

   where the supremum above is taken over all compact, convex sets L while the infimum is over all estimators $\tilde{h}$. In our first result for pointwise estimation, we establish, for each $i \in \{1, \ldots, n\}$, a lower bound for the performance of every estimator for estimating $h_{K^*}(\theta_i)$. Specifically, it is shown that

$$R_n(K^*, \theta_i) \geq c \cdot \frac{\sigma^2}{k_*(i) + 1} \tag{4.5}$$

where $k_*(i)$ is an integer for which an explicit formula can be given in terms of K^* and i; and c is a universal positive constant. It will turn out that $k_*(i)$ is related to the smoothness of $h_{K^*}(\theta)$ at $\theta = \theta_i$.

We construct a data-driven estimator, $\hat{h}_i$, of $h_{K^*}(\theta_i)$ based on local smoothing together with an optimization scheme for automatically choosing a bandwidth, and show that the estimator $\hat{h}_i$ satisfies

$$\mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C \cdot \frac{\sigma^2}{k_*(i) + 1} \tag{4.6}$$

for a universal constant $C > 0$. Inequalities (4.5) and (4.6) (see also inequality (4.21)) together imply that $\hat{h}_i$ is, within a constant factor, an optimal estimator of $h_{K^*}(\theta_i)$ for every compact, convex set K^*. This optimality is much stronger than the traditional minimax optimality usually employed in nonparametric function estimation. The quantity $\sigma^2/(k_*(i) + 1)$ depends on the unknown set K^* in a similar way that the Fisher information depends on the unknown parameter in a regular parametric model. In contrast, the optimal rate in the minimax paradigm is given in terms of the worst case performance over a large parameter space and does not depend on individual parameter values.

2. Using the optimal adaptive point estimators $\hat{h}_1, \ldots, \hat{h}_n$, we construct two set estimators $\hat{K}$ and $\hat{K}'$. The details of this construction are given in Section 4.2.2. In Theorems 4.3.3 and 4.3.5, it is shown that $\hat{K}$ is minimax optimal for K^* under the loss function L_f while the estimator $\hat{K}'$ is minimax optimal under the integral squared loss function defined by

$$L(K^*, \hat{K}') := \int_{-\pi}^{\pi} \left(h_{\hat{K}'}(\theta) - h_{K^*}(\theta)\right)^2 d\theta. \tag{4.7}$$

The square root of the above loss function is often referred to as the McClure-Vitale metric on the space of non-empty compact, convex sets (e.g. [102, 43]). In Theorem 4.3.3, we prove that

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) \leq C \left\{ \frac{\sigma^2}{n} + \left(\frac{\sigma^2 \sqrt{R}}{n} \right)^{4/5} \right\} \tag{4.8}$$

provided K^* is contained in a ball of radius R. This, combined with the minimax lower bound in Guntuboyina [74], proves the minimax optimality of $\hat{K}$. An analogous result is shown in Theorem 4.3.5 for $\mathbb{E}_{K^*} L(K^*, \hat{K}')$. For the pointwise estimation problem where the goal is to estimate $h_{K^*}(\theta_i)$, the optimal rate $\sigma^2/(k_*(i)+1)$ can be as large as $n^{-2/3}$. However the bound (4.8) shows that the globally the risk is at most $n^{-4/5}$. The shape constraint given by convexity of K^* ensures that the points where pointwise estimation rate is $n^{-2/3}$ cannot be too many. Note that we make no smoothness assumptions for proving (4.8).

3. We show that our set estimators $\hat{K}$ and $\hat{K}'$ adapt to polytopes with bounded number of extreme points. Already inequality (4.8) implies that $\mathbb{E}_{K^*} L_f(K^*, \hat{K})$ is bounded from above by the parametric risk $C\sigma^2/n$ provided $R = 0$ (note that $R = 0$ means that K^* is a singleton). Because σ^2/n is much smaller than $n^{-4/5}$, the bound (4.8) shows that $\hat{K}$ adapts to singletons. Theorem 4.3.4 extends this adaptation phenomenon to polytopes and we show that $\mathbb{E}_{K^*} L_f(K^*, \hat{K})$ is bounded by the parametric rate (up to a logarithmic multiplicative factor of n) for all polytopes with bounded number of extreme points. An analogous result is also proved for $\mathbb{E}_{K^*} L(K^*, \hat{K}')$ in Theorem 4.3.5. It should be noted that the construction of our estimators $\hat{K}$ and $\hat{K}'$ (described in Section 4.2.2) does not involve any special treatment for polytopes; yet the estimators automatically achieve faster rates for polytopes.

We would like to stress two features of the results in this chapter: (a) we do not make any smoothness assumptions on the boundary of K^* throughout; in particular, note that we obtain the $n^{-4/5}$ rate for the set estimators $\hat{K}$ and $\hat{K}'$ without any smoothness assumptions, and (b) we go beyond the traditional minimax paradigm by considering adaptive estimation in both the pointwise estimation problem and the problem of estimating the entire set K^*. In particular, pointwise estimation is studied in a general non-asymptotic framework, which evaluates the performance of a procedure at eah individual set K^*, not the worst case performance over a large parameter space as in the conventional minimax theory.

The remainder of this chapter is structured as follows. The proposed estimators are described in detail in Section 4.2. The theoretical properties are analyzed in Section 4.3; Section 4.3.1 gives results for pointwise estimation while Section 4.3.2 deals with set estimators. Section 4.4 considers optimal estimation of some special compact convex sets K^* where we explicitly compute the associated rates of convergence. A simulation study is given in Section 4.5 where we compare the performance of our estimators to other existing estimators in the literature. In Section 4.6, we summarize our main results and discuss potential open problems for future work. The proofs of the main results are given in Section 4.7. Proofs of other results together with additional technical results are given in Chapter B.

4.2 Estimation procedures

Recall the regression model (4.2), where we observe noisy measurements $(\theta_1, Y_1), \ldots, (\theta_n, Y_n)$ with $\theta_i = 2\pi i/n - \pi, i = 1, \ldots, n$ being fixed grid points in $(-\pi, \pi]$. In this section, we first describe in detail our estimate $\hat{h}_i$ for $h_{K^*}(\theta_i)$ for each i. Subsequently, we will put together these estimates $\hat{h}_1, \dots, \hat{h}_n$ to yield set estimators for K^*.

4.2.1 Estimators for $h_{K^*}(\theta_i)$ for each fixed i

Fixing $1 \le i \le n$, our construction of the estimator $\hat{h}_i$ for $h_{K^*}(\theta_i)$ is based on the key circle-convexity property (4.1) of the function $h_{K^*}(\cdot)$. Let us define, for $\phi \in (0, \pi/2)$ and $\theta \in (-\pi, \pi]$, the following two quantities:

$$l(\theta, \phi) := \cos\phi \left(h_{K^*}(\theta + \phi) + h_{K^*}(\theta - \phi)\right) - \frac{h_{K^*}(\theta + 2\phi) + h_{K^*}(\theta - 2\phi)}{2}$$

and

$$u(\theta, \phi) := \frac{h_{K^*}(\theta + \phi) + h_{K^*}(\theta - \phi)}{2\cos\phi}.$$

The following lemma states that for every θ, the quantity $h_{K^*}(\theta)$ is sandwiched between $l(\theta, \phi)$ and $u(\theta, \phi)$ for every ϕ. This will be used crucially in defining $\hat{h}$. The proof of this lemma is a straightforward consequence of (4.1) and is given in Section B.1.6.

Lemma 4.2.1. *For every $0 < \phi < \pi/2$ and every $\theta \in (-\pi, \pi]$, we have $l(\theta, \phi) \le h_{K^*}(\theta) \le u(\theta, \phi)$.*

For a fixed $1 \le i \le n$, Lemma 4.2.1 implies that $l(\theta_i, \frac{2\pi j}{n}) \le h_{K^*}(\theta_i) \le u(\theta_i, \frac{2\pi j}{n})$ for every $0 \le j < \lfloor n/4 \rfloor$. Note that when $j = 0$, we have $l(\theta_i, 0) = h_{K^*}(\theta_i) = u(\theta_i, 0)$. Averaging these inequalities for $j = 0, 1, \dots, k$ where k is a fixed integer with $0 \le k < \lfloor n/4 \rfloor$, we obtain

$$L_k(\theta_i) \le h_{K^*}(\theta_i) \le U_k(\theta_i) \qquad \text{for every } 0 \le k < \lfloor n/4 \rfloor \tag{4.9}$$

where

$$L_k(\theta_i) := \frac{1}{k+1} \sum_{j=0}^{k} l\left(\theta_i, \frac{2\pi j}{n}\right) \quad \text{and} \quad U_k(\theta_i) := \frac{1}{k+1} \sum_{j=0}^{k} u\left(\theta_i, \frac{2\pi j}{n}\right).$$

We are now ready to describe our estimator. Fix $1 \le i \le n$. Inequality (4.9) says that the quantity of interest, $h_{K^*}(\theta_i)$, is sandwiched between $L_k(\theta_i)$ and $U_k(\theta_i)$ for every k. Both $L_k(\theta_i)$ and $U_k(\theta_i)$ can naturally be estimated by unbiased estimators. Indeed, let

$$\hat{l}(\theta_i, 2j\pi/n) := \cos(2j\pi/n)(Y_{i+j} + Y_{i-j}) - \frac{Y_{i+2j} + Y_{i-2j}}{2}, \quad \hat{u}(\theta_i, 2j\pi/n) := \frac{Y_{i+j} + Y_{i-j}}{2\cos(2j\pi/n)}$$

and take

$$\hat{L}_k(\theta_i) := \frac{1}{k+1} \sum_{j=0}^{k} \hat{l}\left(\theta_i, 2j\pi/n\right), \quad \hat{U}_k(\theta_i) := \frac{1}{k+1} \sum_{j=0}^{k} \hat{u}\left(\theta_i, 2j\pi/n\right). \tag{4.10}$$

Obviously, in order for the above to be meaningful, we need to define Y_i even for $i \notin \{1, \ldots, n\}$. This is easily done in the following way: for any $i \in \mathbb{Z}$, let $s \in \mathbb{Z}$ be such that $i - sn \in \{1, \ldots, n\}$ and take $Y_i := Y_{i-sn}$.

As k increases, one averages more terms in (4.10) and hence the estimators $\hat{L}_k(\theta_i)$ and $\hat{U}_k(\theta_i)$ become more accurate. Let $\hat{\Delta}_k(\theta_i) := \hat{U}_k(\theta_i) - \hat{L}_k(\theta_i)$ which is the same as

$$\hat{\Delta}_k(\theta_i) = \frac{1}{k+1} \sum_{j=0}^{k} \left(\frac{Y_{i+2j} + Y_{i-2j}}{2} - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} \frac{Y_{i+j} + Y_{i-j}}{2} \right). \tag{4.11}$$

Because of (4.9), a natural strategy for estimating $h_{K^*}(\theta_i)$ is to choose k for which $\hat{\Delta}_k(\theta_i)$ is the smallest and then use either $\hat{L}_k(\theta_i)$ or $\hat{U}_k(\theta_i)$ at that k as the estimator. This is essentially our estimator with one small difference in that we also take into account the noise present in $\hat{\Delta}_k(\theta_i)$. Formally, our estimator for $h_{K^*}(\theta_i)$ is given by:

$$\hat{h}_i = \hat{U}_{\hat{k}(i)}(\theta_i), \text{ where } \hat{k}(i) := \operatorname*{argmin}_{k \in \mathcal{I}} \left\{ \left(\hat{\Delta}_k(\theta_i) \right)_+ + \frac{2\sigma}{\sqrt{k+1}} \right\} \tag{4.12}$$

and $\mathcal{I} := \{0\} \cup \{2^j : j \geq 0 \text{ and } 2^j \leq \lfloor n/16 \rfloor\}$.

Our estimator $\hat{h}_i$ can be viewed as an angle-adjusted local averaging estimator. It is inspired by the estimator of Cai and Low [29] for convex regression. The number of terms averaged equals $\hat{k}(i) + 1$ and this is analogous to the bandwidth in kernel-based smoothing methods. Our $\hat{k}(i)$ is determined from an optimization scheme. Notice that unlike the least squares estimator $h_{\hat{K}_{\text{ls}}}(\theta_i)$, the construction of $\hat{h}_i$ for a fixed i does not depend on the construction of $\hat{h}_j$ for $j \neq i$.

4.2.2 Set estimators for K^*

We next present estimators for the set K^*. The point estimators $\hat{h}_1, \ldots, \hat{h}_n$ do not directly give an estimator for K^* because $(\hat{h}_1, \ldots, \hat{h}_n)$ is not necessarily a valid support vector i.e., $(\hat{h}_1, \ldots, \hat{h}_n)$ does not always belong to the following set:

$$\mathcal{H} := \left\{ (h_K(\theta_1), \ldots, h_K(\theta_n)) : K \subseteq \mathbb{R}^2 \text{ is compact and convex} \right\}.$$

To get a valid support vector from $(\hat{h}_1, \ldots, \hat{h}_n)$, we need to project it onto $\mathcal{H}$ to obtain:

$$\hat{h}^P := (\hat{h}_1^P, \ldots, \hat{h}_n^P) := \arg \min_{(h_1, \ldots, h_n) \in \mathcal{H}} \sum_{i=1}^{n} \left(\hat{h}_i - h_i \right)^2 \tag{4.13}$$

The superscript P here stands for projection. An estimator for the set K^* can now be constructed immediately from $\hat{h}_1^P, \ldots, \hat{h}_n^P$ via

$$\hat{K} := \left\{ (x_1, x_2) : x_1 \cos\theta_i + x_2 \sin\theta_i \leq \hat{h}_i^P \text{ for all } i = 1, \ldots, n \right\}. \tag{4.14}$$

In Theorems 4.3.3 and 4.3.4, we prove upper bounds on the accuracy of $\hat{K}$ under the loss function L_f given in (4.3).

There is another reasonable way of constructing a set estimator for K^* based on the point estimators $\hat{h}_1, \dots, \hat{h}_n$. We first interpolate $\hat{h}_1, \dots, \hat{h}_n$ to define a function $\hat{h}' : (-\pi, \pi] \to \mathbb{R}$ as follows:

$$\hat{h}'(\theta) := \frac{\sin(\theta_{i+1} - \theta)}{\sin(\theta_{i+1} - \theta_i)} \hat{h}_i + \frac{\sin(\theta - \theta_i)}{\sin(\theta_{i+1} - \theta_i)} \hat{h}_{i+1} \qquad \text{for } \theta_i \leq \theta \leq \theta_{i+1}. \tag{4.15}$$

Here i ranges over $1, \dots, n$ with the convention that $\theta_{n+1} = \theta_1 + 2\pi$ (and $\theta_n \leq \theta \leq \theta_{n+1}$ should be identified with $-\pi \leq \theta \leq -\pi + 2\pi/n$). Based on this function $\hat{h}'$, we can define our estimator $\hat{K}'$ of K^* by

$$\hat{K}' := \operatorname*{argmin}_K \int_{-\pi}^{\pi} \left(\hat{h}'(\theta) - h_K(\theta) \right)^2 d\theta. \tag{4.16}$$

The existence and uniqueness of $\hat{K}'$ can be justified in the usual way by the Hilbert space projection theorem. In Theorem 4.3.5, we prove bounds on the accuracy of $\hat{K}'$ as an estimator for K^* under the integral loss L given in (4.7).

Let us now briefly comment on the algorithms for computing our set estimators $\hat{K}$ and $\hat{K}'$. The expression (4.14) shows how to write $\hat{K}$ in terms of $\hat{h}_i^P, i = 1, \dots, n$ and therefore, we only need to be able to compute $\hat{h}_i^P, i = 1, \dots, n$ for computing $\hat{K}$. This can be done via quadratic programming because the set $\mathcal{H}$ can explicitly written as $\{h \in \mathbb{R}^n : a_i^T h \leq 0, i = 1, \dots, n\}$ for some collection of vectors $a_1, \dots, a_n$ in $\mathbb{R}^n$ (see, for example, Prince and Willsky [115, Theorem1]). To compute $\hat{K}'$, we take a fine uniform grid of points $\alpha_1, \dots, \alpha_M$ in $(-\pi, \pi]$ for a large value of M and approximate $\hat{K}'$ via

$$\operatorname*{argmin}_K \sum_{i=1}^{M} \left(\hat{h}'(\alpha_i) - h_K(\alpha_i) \right)^2.$$

More precisely, one can take $\hat{K}' := \left\{ (x_1, x_2) : x_1 \cos \alpha_i + x_2 \sin \alpha_i \leq \tilde{h}_i \text{ for all } i = 1, \dots, M \right\}$ where

$$(\tilde{h}_1, \dots, \tilde{h}_M) := \arg \min_{(h_1, \dots, h_M) \in \mathcal{H}^M} \sum_{i=1}^{M} \left(\hat{h}'(\alpha_i) - h_i \right)^2$$

with $\mathcal{H}^M := \{(h_K(\alpha_1), \dots, h_K(\alpha_M)) : K \subseteq \mathbb{R}^2 \text{ is compact and convex}\}$. This estimator can then be computed in an analogous way as $\hat{K}$ by quadratic programming. We present simulation examples in Section 4.5 where one can see that there is often not much difference between $\hat{K}$ and $\hat{K}'$ in practice.

4.3 Main results

We now investigate the accuracy of the proposed point and set estimators. The proofs of these results are given in Section 4.7.

4.3.1 Accuracy of the point estimator

As mentioned in the introduction, we evaluate the performance of the point estimator $\hat{h}_i$ at individual functions, not the worst case over a large parameter space. This provides a much more precise characterization of the accuracy of the estimator. Let us first recall inequality (4.9) where $h_{K^*}(\theta_i)$ is sandwiched between $L_k(\theta_i)$ and $U_k(\theta_i)$. Define $\Delta_k(\theta_i) := U_k(\theta_i) - L_k(\theta_i)$.

Theorem 4.3.1. *Fix $i \in \{1, \dots, n\}$. There exists a universal constant $C > 0$ such that the risk of $\hat{h}_i$ as an estimator of $h_{K^*}(\theta_i)$ satisfies the inequality,*

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C \cdot \frac{\sigma^2}{k_*(i)+1} \tag{4.17}$$

where

$$k_*(i) := \operatorname*{argmin}_{k \in \mathcal{I}} \left(\Delta_k(\theta_i) + \frac{2\sigma}{\sqrt{k+1}}\right). \tag{4.18}$$

Remark 4.3.1. It turns out that the bound in (4.17) is linked to the level of smoothness of the function h_{K^*} at θ_i. However for this interpretation to be correct, one needs to regard h_{K^*} as a function on $\mathbb{R}^2$ instead of a subset of $\mathbb{R}$. This is further explained in Remark 4.4.1.

Theorem 4.3.1 gives an explicit bound on the risk of $\hat{h}_i$ in terms of the quantity $k_*(i)$ defined in (4.18). It is important to keep in mind that $k_*(i)$ depends on K^* even though this is suppressed in the notation. In the next theorem, we show that $\sigma^2/(k_*(i)+1)$ also presents a lower bound on the accuracy of every estimator for $h_{K^*}(\theta_i)$. This implies, in particular, optimality of $\hat{h}_i$ as an estimator of $h_{K^*}(\theta_i)$.

One needs to be careful in formulating the lower bound in this setting. A first attempt might perhaps be to prove that, for a universal constant $c > 0$,

$$\inf_{\tilde{h}} \mathbb{E}_{K^*}\left(\tilde{h} - h_{K^*}(\theta_i)\right)^2 \geq c \cdot \frac{\sigma^2}{k_*(i)+1}$$

where the infimum is over all possible estimators $\tilde{h}$. This, of course, would not be possible because one can take $\tilde{h} = h_{K^*}(\theta_i)$ which would make the left hand side zero. A formulation of the lower bound which avoids this difficulty was proposed by [29] in the context of convex function estimation. Their idea, translated to our setting of estimating the support function h_{K^*} at a point θ_i, is to consider, instead of the risk at K^*, the maximum of the risk at K^* and the risk at L^* which is most difficult to distinguish from K^* in term of estimating $h_{K^*}(\theta_i)$. This leads to the benchmark $R_n(K^*, \theta_i)$ defined in (4.4).

Theorem 4.3.2. *For any fixed $i \in \{1, \dots, n\}$, we have*

$$R_n(K^*, \theta_i) \geq c \cdot \frac{\sigma^2}{k_*(i)+1} \tag{4.19}$$

for a universal constant $c > 0$.

Theorems 4.3.1 and 4.3.2 together imply that $\sigma^2/(k_*(i)+1)$ is the optimal rate of estimation of $h_{K^*}(\theta_i)$ for a given compact, convex set K^*. The results show that our data driven estimator $\hat{h}_i$ for $h_{K^*}(\theta_i)$ performs uniformly within a constant factor of the ideal benchmark $R_n(K^*, \theta_i)$ for every i. This means that $\hat{h}_i$ adapts to every unknown set K^* instead of a collection of large parameter spaces as in the conventional minimax theory commonly used in nonparametric literature.

Remark 4.3.2 (A stronger upper bound on the risk of $\hat{h}_i$)**.** From the proof of Theorem 4.3.2, it can be seen that the following statement is true: there exists a compact, convex set L^* such that

$$\inf_{\tilde{h}} \max\left(\mathbb{E}_{K^*}(\tilde{h} - h_{K^*}(\theta_i))^2, \mathbb{E}_{L^*}(\tilde{h} - h_{L^*}(\theta_i))^2\right) \geq \frac{c\sigma^2}{k_*(i)+1} \tag{4.20}$$

the infimum above being over all estimators $\tilde{h}$ of $h_{K^*}(\theta_i)$. In light of this, it is natural to ask whether the following inequality

$$\max\left(\mathbb{E}_{K^*}(\hat{h}_i - h_{K^*}(\theta_i))^2, \mathbb{E}_{L^*}(\hat{h}_i - h_{L^*}(\theta_i))^2\right) \leq \frac{C\sigma^2}{k_*(i)+1} \tag{4.21}$$

holds for the same L^* where $\hat{h}_i$ refers to our estimator defined in (4.12) and C represents a universal constant. Note that this is a stronger inequality than (4.17). It turns out that (4.21) is indeed a true inequality and we provide a proof in Section B.1.3.

Given a specific set K^* and $1 \leq i \leq n$, the quantity $k_*(i)$ is often straightforward to compute up to constant multiplicative factors. Several examples are provided in Section 4.4. From these examples, it will be clear that the size of $\sigma^2/(k_*(i)+1)$ is linked to the level of smoothness of the function h_{K^*} at θ_i. However for this interpretation to be correct, one needs to regard h_{K^*} as a function on $\mathbb{R}^2$ instead of a subset of $\mathbb{R}$. This is explained in Remark 4.4.1.

The following corollaries shed more light on the quantity $\sigma^2/(k_*(i)+1)$. The proofs of these corollaries are given in Section B.1.4. The first corollary below shows that $\sigma^2/(k_*(i)+1)$ is at most $C(\sigma^2 R/n)^{-2/3}$ for every i and K^* (C is a universal constant) provided K^* is contained in a ball of radius R. In Example 4.4.3, we provide an explicit choice of i and K^* for which $\sigma^2/(k_*(i)+1) \geq c(\sigma^2 R/n)^{-2/3}$ (c is a universal constant). This implies that the conclusion of the following corollary cannot in general be improved.

Corollary 4.3.1. *Suppose K^* is contained in some closed ball of radius R. Then for every $i \in \{1, \dots, n\}$, we have, for a universal constant $C > 0$,*

$$\frac{\sigma^2}{k_*(i)+1} \leq C\left\{\left(\frac{\sigma^2 R}{n}\right)^{2/3} + \frac{\sigma^2}{n}\right\} \tag{4.22}$$

and

$$\mathbb{E}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C\left\{\left(\frac{\sigma^2 R}{n}\right)^{2/3} + \frac{\sigma^2}{n}\right\}. \tag{4.23}$$

Note that the above corollary implies the consistency of $\hat{h}_i$ as an estimator for $h_{K^*}(\theta_i)$ for every i and K^*. It turns out that $\hat{h}_i$ is a minimax optimal estimator of $h_{K^*}(\theta_i)$ over the class of all compact convex sets K^* contained in some closed ball of radius R. This is proved in the next result.

Proposition 4.3.1. *For $R \geq 0$, let $\mathcal{K}(R)$ denote the class of all compact, convex sets that are contained in some fixed closed ball of radius R. Then for every $i \in \{1, \dots, n\}$, we have*

$$\sup_{K^* \in \mathcal{K}(R)} \mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C \left\{\frac{\sigma^2}{n} + \left(\frac{\sigma^2 R}{n}\right)^{2/3}\right\} \tag{4.24}$$

for a universal constant C. We further have

$$\inf_{\tilde{h}} \sup_{K^* \in \mathcal{K}(R)} \mathbb{E}_{K^*} \left(\tilde{h} - h_{K^*}(\theta_i)\right)^2 \geq c \left\{\frac{\sigma^2}{n} + \left(\frac{\sigma^2 R}{n}\right)^{2/3}\right\} \tag{4.25}$$

for a universal constant $c > 0$ where the infimum is taken over all possible estimators $\tilde{h}$ of $h_{K^*}(\theta_i)$.

It is clear from the definition (4.18) that $k_*(i) \leq n$ for all i and K_*. In the next corollary, we prove that there exist sets K_* and i for which $k_*(i) \geq cn$ for a constant c. For these sets, the optimal rate of estimating $h_{K^*}(\theta_i)$ is therefore parametric.

For a fixed i and K^*, let $\phi_1(i)$ and $\phi_2(i)$ be such that $\phi_1(i) \leq \theta_i \leq \phi_2(i)$ and such that there exists a single point $(x_1, x_2) \in K^*$ with

$$h_{K^*}(\theta) = x_1 \cos\theta + x_2 \sin\theta \qquad \text{for all } \theta \in [\phi_1(i), \phi_2(i)]. \tag{4.26}$$

The following corollary says that if the distance of θ_i to its nearest end-point in the interval $[\phi_1(i), \phi_2(i)]$ is large (i.e., of constant order), then the optimal rate of estimation of $h_{K^*}(\theta_i)$ is parametric. This situation happens usually for polytopes (polytopes are compact, convex sets with finitely many vertices); see Examples 4.4.1 and 4.4.3 for specific instances of this phenomenon. For non-polytopes, it can often happen that $\phi_1(i) = \phi_2(i) = \theta_i$ in which case the conclusion of the next corollary is not useful.

Corollary 4.3.2. *For every $i \in \{1, \dots, n\}$, we have*

$$k_*(i) \geq c\, n \min\left(\theta_i - \phi_1(i), \phi_2(i) - \theta_i, \pi\right) \tag{4.27}$$

for a universal constant $c > 0$. Consequently

$$\mathbb{E}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq \frac{C\sigma^2}{1 + n\min(\theta_i - \phi_1(i), \phi_2(i) - \theta_i, \pi)} \tag{4.28}$$

for a universal constant $C > 0$.

From the above two corollaries, it is clear that the optimal rate of estimation of $h_{K^*}(\theta_i)$ can be as large as $n^{-2/3}$ and as small as the parametric rate n^{-1}. The rate $n^{-2/3}$ is achieved, for example, in the setting given in Example 4.4.3 while the parametric rate is achieved, for example, for polytopes.

The next corollary argues that bounding $k_*(i)$ in specific examples requires only bounding the quantity $\Delta_k(\theta_i)$ from above and below. This corollary will be useful in Section 4.4 while working out $k_*(i)$ in specific examples.

Corollary 4.3.3. *Fix $1 \leq i \leq n$. Let $\{f_k(\theta_i), k \in \mathcal{I}\}$ and $\{g_k(\theta_i), k \in \mathcal{I}\}$ be two sequences which satisfy $g_k(\theta_i) \leq \Delta_k(\theta_i) \leq f_k(\theta_i)$ for all $k \in \mathcal{I}$. Also let*

$$\breve{k}(i) := \max\left\{k \in \mathcal{I} : f_k(\theta_i) < \frac{(\sqrt{6}-2)\sigma}{\sqrt{k+1}}\right\} \tag{4.29}$$

and

$$\tilde{k}(i) := \min\left\{k \in \mathcal{I} : g_k(\theta_i) > \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k+1}}\right\} \tag{4.30}$$

as long as there is some $k \in \mathcal{I}$ for which $g_k(\theta_i) > 6(\sqrt{2}-1)\sigma/\sqrt{k+1}$; otherwise take $\tilde{k}(i) := \max_{k\in\mathcal{I}} k$. We then have $\breve{k}(i) \leq k_*(i) \leq \tilde{k}(i)$ and

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C\frac{\sigma^2}{\breve{k}(i)+1} \tag{4.31}$$

for a universal constant $C > 0$.

4.3.2 Accuracy of set estimators

We now turn to study the accuracy of the set estimators $\hat{K}$ (defined in (4.14)) and $\hat{K}'$ (defined in (4.16)). The accuracy of $\hat{K}$ will be investigated under the loss function L_f (defined in (4.3)) while the accuracy of $\hat{K}'$ will be studied under the loss function L (defined in (4.7)).

In Theorem 4.3.3 below, we prove that $\mathbb{E}_{K^*}L_f(K^*, \hat{K})$ is bounded from above by a constant multiple of $n^{-4/5}$ as long as K^* is contained in a ball of radius R. The discussions following the theorem shed more light on its implications.

Theorem 4.3.3. *If K^* is contained in some closed ball of radius $R \geq 0$, then*

$$\mathbb{E}_{K^*}L_f\left(K^*, \hat{K}\right) \leq C\left\{\frac{\sigma^2}{n} + \left(\frac{\sigma^2\sqrt{R}}{n}\right)^{4/5}\right\} \tag{4.32}$$

for a universal constant $C > 0$. Note here that $R = 0$ is allowed (in which case K^* is a singleton).

Note that as long as $R > 0$, the right hand side in (4.32) will be dominated by the $(\sigma^2\sqrt{R}/n)^{-4/5}$ term for all large n. This would mean that

$$\sup_{K^* \in \mathcal{K}(R)} \mathbb{E}_{K^*} L_f(K^*, \hat{K}) \leq C \left(\frac{\sigma^2 \sqrt{R}}{n} \right)^{4/5} \tag{4.33}$$

where $\mathcal{K}(R)$ denotes the set of all compact convex sets contained in some fixed closed ball of radius R.

The minimax rate of estimation over the class $\mathcal{K}(R)$ was studied in Guntuboyina [74]. In Theorems 3.1 and 3.2 [74], it was proved that

$$\inf_{\tilde{K}} \sup_{K^* \in \mathcal{K}(R)} \mathbb{E}_{K^*} L_f(K^*, \hat{K}) \asymp \left(\frac{\sigma^2 \sqrt{R}}{n} \right)^{4/5} \tag{4.34}$$

where $\asymp$ denotes equality upto constant multiplicative factors. From (4.33) and (4.34), it follows that $\hat{K}$ is a minimax optimal estimator of K^*. We should mention here that an inequality of the form (4.33) was proved for the least squares estimator $\hat{K}_{\mathrm{ls}}$ by Gardner et al. [59] which implies that $\hat{K}_{\mathrm{ls}}$ is also a minimax optimal estimator of K^*.

The $n^{-4/5}$ minimax rate here is quite natural in connection with estimation of smooth functions. Indeed, this is the minimax rate for estimating twice differentiable one-dimensional functions. Although we have not made any smoothness assumptions here, we are working under a convexity-based constraint and convexity is associated, in a broad sense, with twice smoothness (see, for example, Alexandrov [2]).

Remark 4.3.3. Because of the formula (4.3) for the loss function L_f, the risk $\mathbb{E}_{K^*} L_f(K^*, \hat{K})$ can be seen as the average of the risk of $\hat{K}$ for estimating $h_{K^*}(\theta_i)$ over $i = 1, \dots, n$. We have seen in Section 4.3.1 that the optimal rate of estimating $h_{K^*}(\theta_i)$ can be as high as $n^{-2/3}$. Theorem 4.3.3, on the other hand, can be interpreted as saying that, on average over $i = 1, \dots, n$, the optimal rate of estimating $h_{K^*}(\theta_i)$ is at most $n^{-4/5}$. Indeed, the key to proving Theorem 4.3.3 is to establish the following inequality:

$$\frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{k_*(i) + 1} \leq C \left\{ \frac{\sigma^2}{n} + \left(\frac{\sigma^2 \sqrt{R}}{n} \right)^{4/5} \right\}.$$

under the assumption that K^* is contained in a ball of radius R. Therefore, even though each term $\sigma^2/(k_*(i) + 1)$ can be as large as $n^{-2/3}$, on average, their size is at most $n^{-4/5}$.

Remark 4.3.4. Theorem 4.3.3 provides different qualitative conclusions when K^* is a singleton. In this case, one can take $R = 0$ in (4.32) to get the parametric bound $C\sigma^2/n$ for $\mathbb{E}_{K^*} L_f(K^*, \hat{K})$. Because this is smaller than the nonparametric $n^{-4/5}$ rate, it means that $\hat{K}$ adapts to singletons. Singletons are simple examples of polytopes and one naturally wonders here if $\hat{K}$ also adapts to other polytopes as well. This is however not implied by inequality

(4.32) which gives the rate $n^{-4/5}$ for every K^* that is not a singleton. It turns out that $\hat{K}$ indeed adapts to other polytopes and we prove this in the next theorem. In fact, we prove that $\hat{K}$ adapts to any K^* that is well-approximated by a polytope with not too many vertices. It is currently not known if the least squares estimator $\hat{K}_{\text{ls}}$ has such adaptivity.

We next prove another bound for $\mathbb{E}_{K^*} L_f(K^*, \hat{K})$. This bound demonstrates adaptivity of $\hat{K}$ as described in the previous remark. Recall that polytopes are compact, convex sets with finitely many extreme points (or vertices). The space of all polytopes in $\mathbb{R}^n$ will be denoted by $\mathcal{P}$. For a polytope $P \in \mathcal{P}$, we denote by v_P, the number of extreme points of P. Also recall the notion of Hausdorff distance between two compact, convex sets K and L defined by

$$\ell_H(K, L) := \sup_{\theta \in \mathbb{R}} \left| h_K(\theta) - h_L(\theta) \right|. \tag{4.35}$$

This is not the usual way of defining the Hausdorff distance. For an explanation of the connection between this and the usual definition, see, for example, Schneider [127, Theorem 1.8.11].

Theorem 4.3.4. *There exists a universal constant $C > 0$ such that*

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) \le C \inf_{P \in \mathcal{P}} \left[\frac{\sigma^2 v_P}{n} \log\left(\frac{en}{v_P} \right) + \ell_H^2(K^*, P) \right]. \tag{4.36}$$

Remark 4.3.5 (Near-parametric rates for polytopes)**.** The bound (4.36) implies that $\hat{h}$ has the parametric rate (upto a logarithmic factor of n) for estimating polytopes. Indeed, suppose that K^* is a polytope with v vertices. Then using $P = K^*$ in the infimum in (4.36), we have the risk bound

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) \le \frac{C\sigma^2 v}{n} \log\left(\frac{en}{v} \right). \tag{4.37}$$

This is the parametric rate $\sigma^2 v/n$ up to logarithmic factors and is smaller than the nonparametric rate $n^{-4/5}$ given in (4.32).

Remark 4.3.6. When $v = 1$, inequality (4.37) has a redundant logarithmic factor. Indeed, when $v = 1$, we can use (4.32) with $R = 0$ which gives (4.37) without the additional logarithmic factor. We do not know if the logarithmic factor in (4.37) can be removed for values of v larger than one as well.

Now consider the second set estimator $\hat{K}'$. The next theorem gives an upper bound on its accuracy under the integral loss function L (defined in (4.7)).

Theorem 4.3.5. *Suppose K^* is contained in some closed ball of radius $R \ge 0$. The risk $\mathbb{E}_{K^*} L(K^*, \hat{K}')$ satisfies both the following inequalities:*

$$\mathbb{E}_{K^*} L(K^*, \hat{K}') \le C \left\{ \frac{\sigma^2}{n} + \left(\frac{\sigma^2 \sqrt{R}}{n} \right)^{4/5} + \frac{R^2}{n^2} \right\} \tag{4.38}$$

and

$$\mathbb{E}_{K^*} L(K^*, \hat{K}') \leq C \inf_{P \in \mathcal{P}} \left[\frac{\sigma^2 v_P}{n} \log \left(\frac{en}{v_P} \right) + \ell_H^2(K^*, P) + \frac{R^2}{n^2} \right]. \tag{4.39}$$

The only difference between the inequalities (4.38) and (4.39) on one hand and (4.32) and (4.36) on the other is the presence of the R^2/n^2 term. This term is usually very small and does not change the qualitative behavior of the bounds. However note that inequality (4.36) did not require any assumption on K^* being in a ball of radius R while this assumption is necessary for (4.39).

Remark 4.3.7. The rate $(\sigma^2 \sqrt{R}/n)^{4/5}$ is the minimax rate for this problem under the loss function L. Although this has not been proved explicitly anywhere, it can be shown by modifying the proof of Guntuboyina [74, Theorem 3.2] appropriately. Theorem 4.3.5 therefore shows that $\hat{K}'$ is a minimax optimal estimator of K^* under the loss function L.

4.4 Examples

We now investigate the results given in the last section for specific choices of K^*. It is useful here to note that $\Delta_k(\theta_i) = U_k(\theta_i) - L_k(\theta_i)$ has the following alternative expression:

$$\frac{1}{k+1} \sum_{j=0}^{k} \left(h_{K^*}(\theta_i \pm 4j\pi/n) - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} h_{K^*}(\theta_i \pm 2j\pi/n) \right). \tag{4.40}$$

where we write $h_{K^*}(\theta_i \pm \phi)$ for $\left(h_{K^*}(\theta_i + \phi) + h_{K^*}(\theta_i - \phi)\right)/2$ with $\phi = 2j\pi/n, 4j\pi/n$.

Example 4.4.1 (Single point)**.** Suppose $K^* := \{(x_1, x_2)\}$ for a fixed point $(x_1, x_2) \in \mathbb{R}^2$. In this case

$$h_{K^*}(\theta) = x_1 \cos\theta + x_2 \sin\theta \qquad \text{for all } \theta. \tag{4.41}$$

It can then be directly checked from (4.40) that $\Delta_k(\theta_i) = 0$ for every $k \in \mathcal{I}$ and $i \in \{1, \dots, n\}$. As a result, it follows that $k_*(i) = \max_{k \in \mathcal{I}} k \geq cn$ for a constant $c > 0$. Theorem 4.3.1 then says that the point estimator $\hat{h}_i$ satisfies

$$\mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i) \right)^2 \leq \frac{C\sigma^2}{n} \tag{4.42}$$

for a universal constant $C > 0$. One therefore gets the parametric rate here.

Also, Theorem 4.3.3 and inequality (4.38) in Theorem 4.3.5 can both be used here with $R = 0$. This implies that the set estimators $\hat{K}$ and $\hat{K}'$ both converge to K^* at the parametric rate under the loss functions L_f and L respectively.

Example 4.4.2 (Ball)**.** Suppose K^* is a ball centered at (x_1, x_2) with radius $R > 0$. It is then easy to verify that

$$h_{K^*}(\theta) = x_1 \cos\theta + x_2 \sin\theta + R \qquad \text{for all } \theta. \tag{4.43}$$

As a result, for every $k \in \mathcal{I}$ and $i \in \{1, \dots, n\}$, we have

$$\Delta_k(\theta_i) = \frac{R}{k+1} \sum_{j=0}^{k} \left(1 - \frac{\cos \frac{4\pi j}{n}}{\cos \frac{2\pi j}{n}}\right) \le R \left(1 - \frac{\cos 4\pi k/n}{\cos 2\pi k/n}\right). \tag{4.44}$$

Because $k \le n/16$ for all $k \in \mathcal{I}$, it is easy to verify that $\Delta_k(\theta_i) \le 8R \sin^2(\pi k/n) \le 8R\pi^2 k^2/n^2$. Taking $f_k(\theta_i) = 8R\pi^2 k^2/n^2$ in Corollary 4.3.3, we obtain that $k_*(i) \ge c \min(n, (n^2\sigma/R)^{2/5})$ for a constant c. Also since the function $1 - \cos(2x)/\cos(x)$ is a strongly convex function on $[-\pi/4, \pi/4]$ with second derivative lower bounded by 3, we have

$$\Delta_k(\theta_i) = \frac{R}{k+1} \sum_{j=0}^{k} \left(1 - \frac{\cos \frac{4\pi j}{n}}{\cos \frac{2\pi j}{n}}\right) \ge \frac{R}{k+1} \sum_{j=0}^{k} \frac{3}{2} \left(\frac{2\pi j}{n}\right)^2 = \frac{R\pi^2 k(2k+1)}{n^2}.$$

This gives $k_*(i) \le C \min(n, (n^2\sigma/R)^{2/5})$ as well for a constant C. We thus have $k_*(i) \asymp (n^2\sigma/R)^{2/5}$ for every i. Theorem 4.3.1 then gives

$$\mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \le C \left\{ \frac{\sigma^2}{n} + \left(\frac{\sigma^2 \sqrt{R}}{n}\right)^{4/5} \right\} \tag{4.45}$$

for every $i \in \{1, \dots, n\}$. Theorem 4.3.3 and inequality (4.38) prove that the set estimators $\hat{K}$ and $\hat{K}'$ also converge to K^* at the $n^{-4/5}$ rate.

In the preceding examples, we saw that the optimal rate $\sigma^2/(k_*(i) + 1)$ for estimating $h_{K^*}(\theta_i)$ did not depend on i. Next, we consider *asymmetric* examples where the rate changes with i.

Example 4.4.3 (Segment)**.** Let K^* be the vertical line segment joining $(0, R)$ and $(0, -R)$ for a fixed $R > 0$. Then $h_{K^*}(\theta) = R|\sin\theta|$ for all θ. Assume that n is even and consider $i = n/2$ so that $\theta_{n/2} = 0$. It can be verified that

$$\Delta_k(\theta_{n/2}) = \Delta_k(0) = \frac{R}{k+1} \sum_{j=0}^{k} \tan \frac{2\pi j}{n} \qquad \text{for every } k \in \mathcal{I}.$$

Because $j \mapsto \tan(2\pi j/n)$ is increasing, it is straightforward to deduce from above that $3\pi Rk/(8n) \le \Delta_k(0) \le 4\pi Rk/n$. Corollary 4.3.3 then gives

$$\frac{\sigma^2}{k_*(n/2) + 1} \asymp \frac{\sigma^2}{n} + \left(\frac{\sigma^2 R}{n}\right)^{2/3}. \tag{4.46}$$

It was shown in Corollary 4.3.1 that the right hand side above represents the maximum possible value of $\sigma^2/(k_*(i) + 1)$ when K^* lies in a closed ball of radius R. Therefore this example presents the situation where estimation of $h_{K^*}(\theta_i)$ is the most difficult. See Remark 4.4.1 for the connection to smoothness of $h_{K^*}(\cdot)$ at θ_i.

Now suppose that $i = 3n/4$ (assume that $n/4$ is an integer for simplicity) so that $\theta_i = \pi/2$. Observe then that $h_{K^*}(\theta) = R \sin\theta$ (without the modulus) for $\theta = \theta_i \pm 4j\pi/n$ for every $0 \leq j \leq k, k \in \mathcal{I}$. Using (4.40), we have $\Delta_k(\theta_i) = 0$ for every $k \in \mathcal{I}$. This immediately gives $k_*(i) = \lfloor n/16 \rfloor$ and hence

$$\frac{\sigma^2}{k_*(3n/4)+1} \asymp \frac{\sigma^2}{n}. \tag{4.47}$$

In this example, the risk for estimating $h_{K^*}(\theta_i)$ changes with i. For $i = n/2$, we get the $n^{-2/3}$ rate while for $i = 3n/4$, we get the parametric rate. For other values of i, one gets a range of rates between $n^{-2/3}$ and n^{-1}.

Because K^* is a polytope with 2 vertices, Theorem 4.3.4 and inequality (4.39) imply that the set estimators $\hat{K}$ and $\hat{K}'$ converge at the near parametric rate $\sigma^2 \log n/n$. It is interesting to note here that even though for some θ_i, the optimal rate of estimation of $h_{K^*}(\theta_i)$ is $n^{-2/3}$, the entire set can be estimated at the near parametric rate.

Example 4.4.4 (Half-ball)**.** Suppose $K^* := \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1, x_2 \leq 0\}$. One then has $h_K(\theta) = 1$ for $-\pi \leq \theta \leq 0$ and $h_K(\theta) = |\cos\theta|$ for $0 < \theta \leq \pi$. Assume n is even and take $i = n/2$ so that $\theta_i = 0$. It can be checked that

$$\Delta_k(0) = \frac{1}{2(k+1)} \sum_{j=0}^{k} \left(1 - \frac{\cos 4\pi j/n}{\cos 2\pi j/n}\right).$$

This is exactly as in (4.44) with $R = 1$ and an additional factor of $1/2$. Arguing as in Example 4.4.2, we obtain that

$$\frac{\sigma^2}{k_*(n/2)+1} \asymp \frac{\sigma^2}{n} + \left(\frac{\sigma^2}{n}\right)^{4/5}.$$

Now take $i = 3n/4$ (assume $n/4$ is an integer) so that $\theta_i = \pi/2$. Observe then that $h_{K^*}(\theta) = |\cos\theta|$ for $\theta = \theta_i \pm 4j\pi/n$ for every $0 \leq j \leq k, k \in \mathcal{I}$. The situation is therefore similar to (4.46) and we obtain

$$\frac{\sigma^2}{k_*(3n/4)+1} \asymp \frac{\sigma^2}{n} + \left(\frac{\sigma^2}{n}\right)^{2/3}.$$

Similar to the previous example, the risk for estimating $h_{K^*}(\theta_i)$ changes with i and varies from $n^{-2/3}$ to $n^{-4/5}$. On the other hand, Theorem 4.3.3 states that the set estimator $\hat{K}$ still estimates K^* at the rate $n^{-4/5}$.

Remark 4.4.1 (Connection between risk and smoothness)**.** The reader may observe that the support functions (4.41) and (4.43) in the two examples above differ only by the constant R. It might then seem strange that only the addition of a non-zero constant changes the risk of estimating $h_{K^*}(\theta_i)$ from n^{-1} to $n^{-4/5}$. It turns out that the function (4.41) is much more smoother than the function (4.43); the right way to view smoothness of $h_{K^*}(\cdot)$ is to regard it as a function on $\mathbb{R}^2$. This is done in the following way. Define, for each $z = (z_1, z_2) \in \mathbb{R}^2$,

$$h_{K^*}(z) = \max_{(x_1,x_2)\in K^*} (x_1 z_1 + x_2 z_2) \,.$$

When $z = (\cos\theta, \sin\theta)$ for some $\theta \in \mathbb{R}$, this definition coincides with our definition of $h_{K^*}(\theta)$. A standard result (see for example Corollary 1.7.3 and Theorem 1.7.4 in [127]) states that the subdifferential of $z \mapsto h_{K^*}(z)$ exists at every $z = (z_1, z_2) \in \mathbb{R}^2$ and is given by

$$F(K^*, z) := \{(x_1, x_2) \in K^* : h_{K^*}(z) = x_1 z_1 + x_2 z_2\} \,.$$

In particular, $z \mapsto h_{K^*}(z)$ is differentiable at z if and only if $F(K^*, z)$ is a singleton.

Studying h_{K^*} as a function on $\mathbb{R}^2$ sheds qualitative light on the risk bounds obtained in the examples. In the case of Example 4.4.1 when $K^* = \{(x_1, x_2)\}$, it is clear that $F(K^*, z) = \{(x_1, x_2)\}$ for all z. Because this set does not change with z, this provides the case of maximum smoothness (because the derivative is constant) and thus we get the n^{-1} rate.

In Example 4.4.2 when K^* is a ball centered at $x = (x_1, x_2)$ with radius R, it can be checked that $F(K^*, z) = \{x + Rz/\|z\|\}$ for every $z \neq 0$. Since $F(K^*, z)$ is a singleton for each $z \neq 0$, it follows that $z \mapsto h_{K^*}(z)$ is differentiable for every z. For $R \neq 0$, the set $F(K^*, z)$ changes with z and thus here h_{K^*} is not as smooth as in Example 4.4.1. This explains the slower rate in Example 4.4.2 compared to 4.4.1.

Finally in Example 4.4.3, when K^* is the vertical segment joining $(0, R)$ and $(0, -R)$, it is easy to see that $F(K^*, z) = K^*$ when $z = (1, 0)$. Here $F(K^*, z)$ is not a singleton which implies that $h_{K^*}(z)$ is non-differentiable at $z = (1, 0)$. This is why one gets the slow rate $n^{-2/3}$ for estimating $h_{K^*}(\theta_{n/2})$ in Example 4.4.3.

4.5 Numerical results

In this section, we compare the performance of our estimators to other existing estimators for both the pointwise estimation and set estimation problems. We shall refer to our estimator $\hat{h}_i$ (defined in (4.12)) as the local averaging estimator (*LAE*). The set estimator $\hat{K}$ (defined in (4.14)) will be referred to as *LAE with projection* and the set estimator $\hat{K}'$ (defined in (4.16)) will be referred to as *LAE with infinite projection*.

Note that our estimators require knowledge of the noise level σ (which we have assumed to be known for our theoretical analysis). In practice, σ is typically unknown and needs to be estimated. Under the setting of the present chapter, σ is easily estimable by using the median of the consecutive differences. Let $\delta_i = Y_{2i} - Y_{2i-1},\ i = 1, \ldots, \lfloor \frac{n}{2} \rfloor$. A simple robust estimator of the noise level σ is the following median absolute deviation (MAD) estimator:

$$\hat{\sigma} = \frac{\text{median}_i |\delta_i - \text{median}_j(\delta_j)|}{\sqrt{2}\Phi^{-1}(0.75)} \approx 1.05 \times \text{median}_i |\delta_i - \text{median}_j(\delta_j)|. \tag{4.48}$$

We use this estimate of σ in our simulations.

Let us now briefly describe the other estimators to which our estimators will be compared. The first of these is the least squares estimator [115] which we have already described in this chapter. The other estimators come from Fisher et al. [52, Section 2] where the authors propose four different estimators for K^*. These are: (A) a second-order local linear method; (B) a second-order Nadaraya-Watson kernel method; (C) a third-order local quadratic estimator, and (D) a fourth-order Nadaraya-Watson kernel method. As remarked in [52, Section 3], their method (D) is always inferior to (C) (even when the smoothing parameters for (D) were chosen optimally). Therefore, we only compare our estimators with the first three methods from [52]. We shall denote these estimators by *FHTW-A*, *FHTW-B* and *FHTW-C* respectively (*FHTW* is an acronym for the author names of [52]). In our simulations, we allow these three estimators to have knowledge of the true noise level σ.

In total therefore, we evaluate the performance of seven estimators in this section: three estimators proposed in this chapter (*LAE*, *LAE with projection* and *LAE with infinite projection*), the least squares estimator (*LSE*) and the three estimators from [52] (*FHTW-A*, *FHTW-B* and *FHTW-C*).

In the interest of space, we present simulation results here for only two cases: K^* is (a) the unit ball, and (b) the segment joining $(0, -3)$ to $(0, +3)$. Simulation results for other choices of K^* including square, ellipsoid and random polytope are given in the Section B.2.

4.5.1 Pointwise estimation

In this section, we evaluate the performances of the seven pointwise estimators $h_{K^*}(\theta_i)$ for fixed $1 \leq i \leq n$. We measure the performance of each estimator $\tilde{h}$ by the mean squared error (MSE) $\mathbb{E}_{K^*}(\tilde{h} - h_{K^*}(\theta_i))^2$. For every fixed n, we simulate 200 random ensembles according to the model (4.2) and then approximate the expectation by the average of error $(\tilde{h} - h_{K^*}(\theta_i))^2$. In simulations, $\sigma = 0.5$ and n ranges over $\{20, 50, 100, 200, 300, 500\}$. We plot the risk as a function of n.

Ball: We start with the case when K^* is a ball. Without loss of generality then, we can assume that the ball is the standard unit ball whose support function always equals one. By rotation invariance of the ball, it is enough to study the case when $\theta_i = 0$. In the following plot, we draw the mean squared errors of all the estimators against the sample size n.

From Figure 4.1, it is clear that the behaviors of *LSE* and both the *LAE projection* estimators (*LAE with projection* and *LAE with infinite projection*) are almost the same, while the performance of *LAE* is quite comparable. When n is large, the performance of *LAE* is as good as that of *LSE* and the *LAE projection* estimators i.e., in this case, projecting the *LAE* onto the support function space is unnecessary. Here the *LAE*, which only uses local information, is quite similar to that of the *LSE*. Also note that the best performance in this setting is achieved by the three *FHTW* estimators.

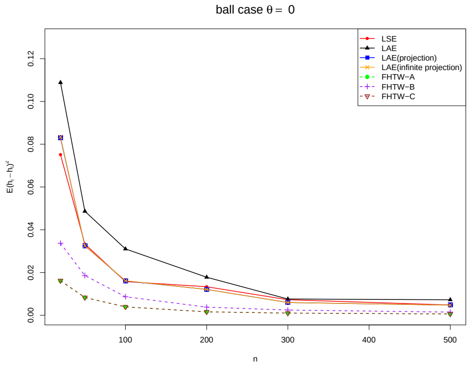


Figure 4.1: Point estimation error when K^* is a ball

Segment: Our second example is when K^* is the segment from $(0,-3)$ to $(0,+3)$ and we study the MSE when θ_i equal to $0, \pi/4, \pi/2$ (in this example, the performance of various estimators will vary with θ_i). The support function of K^* here equals $3|\sin\theta|$ (this function is plotted in the first plot of Figure 4.2); the three choices of θ_i are indicated in this plot in red. The mean squared errors of all estimators against n are plotted in the last three subplots of Figure 4.2 for each of the three choices of θ_i.

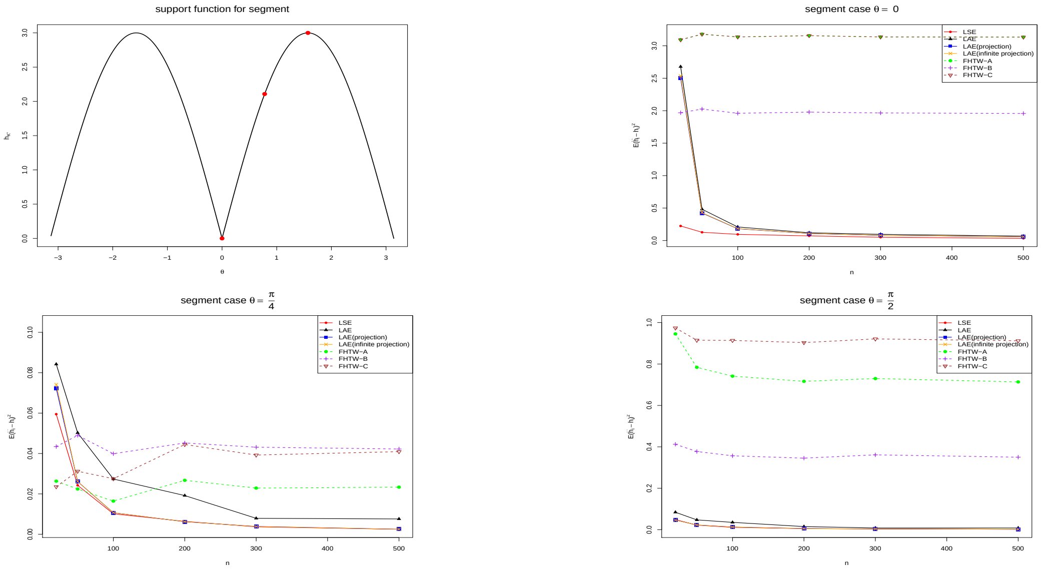


Figure 4.2: Point estimation error when K^* is a segment

Observe that similar to the case of the ball, the behaviors of *LSE* and both the *LAE projection* estimators are almost the same. The *LAE* has comparable performance. An interesting fact is that if one looks at the range of y-axis in the last three subplots of Figure 4.2, although the mean squared error is decreasing at each θ_i, the rate of decay varies with

θ_i. It may be noted that this phenomenon is predicted in our theoretical analysis because the benchmark $R_n(K^*, \theta_i)$ is adaptive to the structure of h_{K^*} at θ_i.

Note that in this example, the *FHTW* estimators perform poorly unlike the case of the ball. The reason is that in [52], the support function is assumed to be twice differentiable and so is the fitted $\hat{h}$. On the other hand, in this example, the true support function is non-differentiable which explains their poor performance. Note that in contrast, our local averaging estimator requires no assumptions on the local smoothness and as we have seen, the estimator actually adapts to local smoothness.

Analogous plots for other choices of K^* are given in Appendix B.2.1. These plots reveal the same story as the previous two settings.

4.5.2 Set estimation

We now turn to set estimation. Recall that we proposed two estimators for set estimation: the *LAE with projection* estimator $\hat{K}$ (defined in (4.14)) and the *LAE with infinite projection* estimator $\hat{K}'$ (defined in (4.16)). We compare these two estimators to the *LSE* and the *FHTW* estimators from [52]. In our simulations, we found that *FHTW-B* works much better compared to *FHTW-A* and *FHTW-C*, which can also be seen from the simulations for point estimation above. So we only present the results for *FHTW-B* among all the three *FHTW* estimators.

For a set of specific choices of K^* and n, we compute the expected squared errors $\mathbb{E}_{K^*} L_f(\hat{K}, K^*)$ and $\mathbb{E}_{K^*} L(\hat{K}, K^*)$ for each of the estimators, where L_f and L are defined in (4.3) and (4.7) respectively. Similar to the point estimation case, these two expectations are approximated by the empirical average of 200 random ensembles according to the model (4.2). For our *LAE projection* estimators which require the value of σ, we estimate σ via (4.48). For the *FHTW-B* estimator which also requires σ, we take σ to be its true value.

We plot $\mathbb{E}_{K^*} L_f(\hat{K}, K^*)$ and $\mathbb{E}_{K^*} L(\hat{K}, K^*)$ for each estimator $\hat{K}$ as a function of n. For visualizing the set estimator, we picked an ensemble randomly from the 200 ensembles and plotted each estimator. Note that for the LAE with infinite projection, as we mentioned before, we take a finer uniform grid of points $\alpha_1, \dots, \alpha_M$ on $(-\pi, \pi]$ for a large value of M and approximate the set by the intersection of M hyperplanes. In this case, M is set to be 1000.

Ball: Figure 4.3 presents the simulation results when K^* is the unit ball. It shows that the performance of the *LAE projection* estimator is almost identical to the that of the *LSE*. The three set estimators *LSE*, *LAE with projection* and *LAE with infinite projection* all look alike in the last subplot. Observe that for the *LAE with infinite projection* estimator, there are many more support lines compared to the *LAE with projection* estimator. This is because of the infinite nature of the projection that is used to define the *LAE with infinite projection* estimator. The best estimator in this example is the *FHTW-B* estimator because it captures the geometry of K^* exactly.

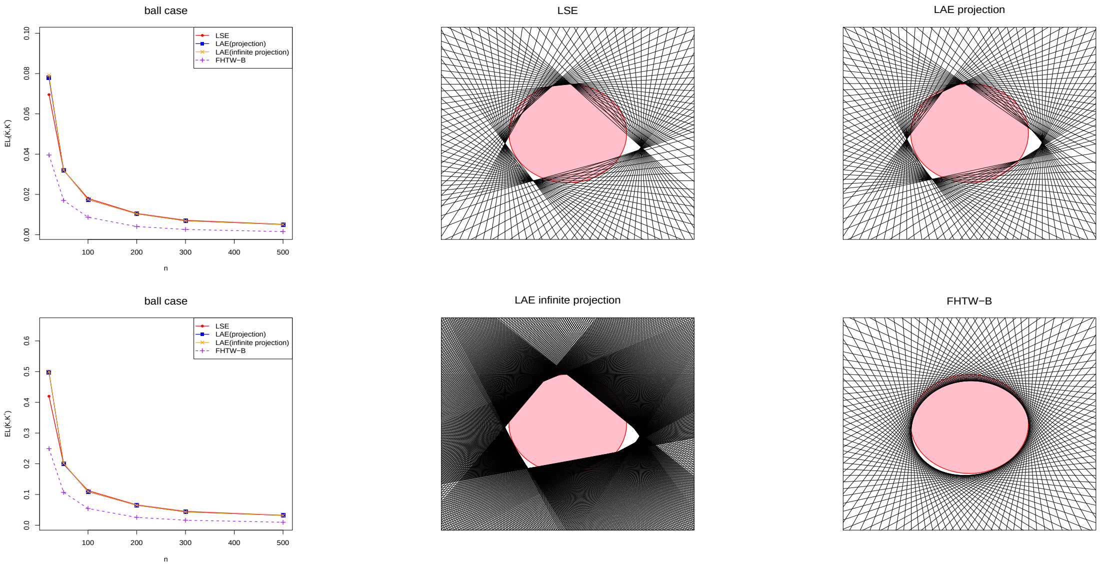


Figure 4.3: Set estimation when K^* is a ball

Segment : Our second example takes K^* to be the segment from $(0, -3)$ to $(0, +3)$. The plots are given in Figure 4.4. Similar to the ball case, our *LAE projection* estimators are comparable to that of the *LSE*. Note that the *FHTW-B* estimator which assumes smoothness of the support function becomes quite off (much higher risk) in this case.

From both these figures (as well as other set estimation figures in [27]), it is clear that both our set estimators ($\hat{K}$ and $\hat{K}'$) look quite similar and have near identical performance.

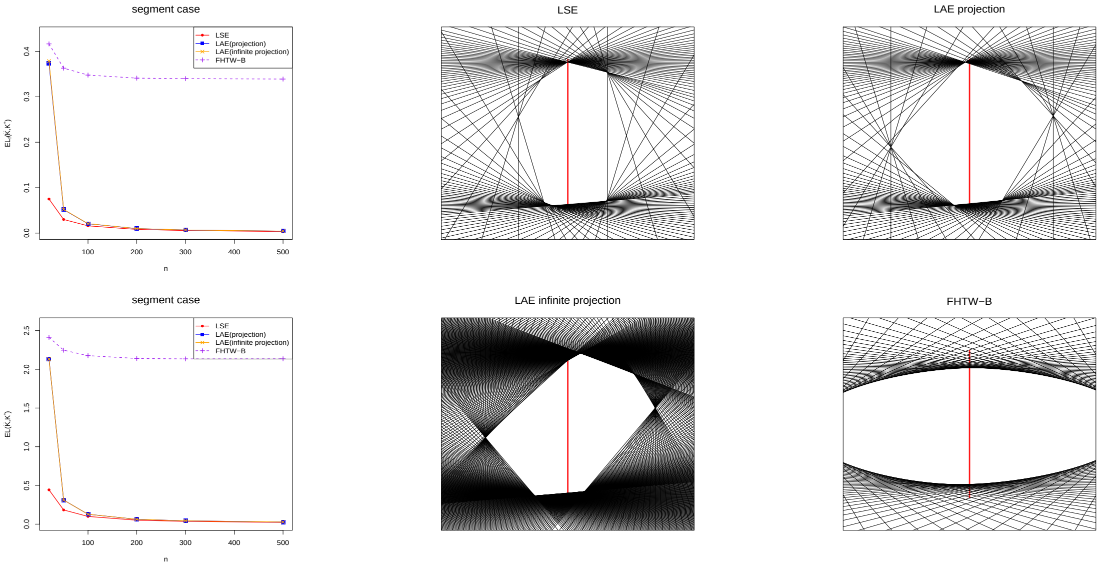


Figure 4.4: Set estimation when K^* is a segment

4.6 Discussion

In this chapter, we study the problems of estimating both the support function at a point, $h_{K^*}(\theta_i)$, and the whole convex set K^*. Data-driven adaptive estimators are constructed and their optimality is established. For pointwise estimation, the quantity $k_*(i)$, which appears in both the upper bound (4.17) and the lower bound (4.19), is related to the smoothness of $h_{K^*}(\theta)$ at $\theta = \theta_i$. The construction of $\hat{h}_i$ is based on local smoothing together with an optimization algorithm for choosing the bandwidth. Smoothing methods for estimating the support function have previously been studied by Fisher et al. [52]. Specifically, working under certain smoothness assumptions on the true support function $h_{K^*}(\theta)$, Fisher et al. [52] estimated it using periodic versions of standard nonparametric regression techniques such as local regression, kernel smoothing and splines. They evade the problem of bandwidth selection however by assuming that the true support function is sufficiently smooth. Our estimator comes with a data-driven method for choosing the bandwidth automatically and we do not need any smoothness assumptions on the true convex set. The fact that our pointwise estimator uses only local information (i.e., for computing $\hat{h}_i$, we only use information on Y_j corresponding to θ_j near θ_i) is quite advantageous in that the computational complexity can be substantially reduced by parallelizing the computation.

It was noted that the construction of our estimators $\hat{K}$ and $\hat{K}'$ given in Section 4.2.2 does not involve any special treatment for polytopes; yet we obtain faster rates for polytopes. Such automatic adaptation to polytopes has been observed in other contexts: isotonic regression where one gets automatic adaptation for piecewise constant monotone functions (see Sabyasachi et al. [38]) and convex regression where one gets automatic adaptation for piecewise affine convex functions (see Guntuboyina and Sen [75]).

Finally, we note that because $\sigma^2/(k_*(i)+1)$ gives the optimal rate in pointwise estimation, it can potentially be used as a benchmark to evaluate other estimators for $h_{K^*}(\theta_i)$ such as the least squares estimator $h_{\hat{K}_{\text{ls}}}(\theta_i)$. From our simulations in Section 4.5, it seems that the least squares estimator is also optimal in our strong sense for pointwise estimation. It is however difficult to prove accuracy results for the least squares estimator for pointwise estimation. The main difficulty comes from the fact that the least squares estimator is technically a non-local estimator (meaning that $h_{\hat{K}_{ls}}(\theta_i)$ can depend on the values of Y_j for θ_j far from θ_i). This and the other fact that there is no closed form expression for the least squares estimator makes it very hard to study its pointwise estimation properties. In the related problem of convex function estimation, pointwise properties of the least squares estimator have been studied in Groeneboom et al. [71]. But their results are asymptotic in nature and, more importantly, they make certain smoothness assumptions on the true function. In the generality considered in this chapter, studying the least squares estimator seems difficult; it will probably require new techniques which are beyond the scope of this chapter. This is an interesting topic for future research.

4.7 Proofs of the main results

This section contains the proofs of the main theorems stated in Section 4.3. The proofs of the corollaries of Subsection 4.3.1 are given in the Section B.1.4. Some technical lemmas are required for the proofs given below. These lemmas are also given in the Section B.1.6.

Please note that because of space constraints, for the first three proofs given below (those of Theorem 4.3.1, Theorem 4.3.2 and Theorem 4.3.3), we only give a few details here and relegate the complete argument to the appendix.

4.7.1 Proof of Theorem 4.3.1

We provide the proof of Theorem 4.3.1 here. The proof uses three simple lemmas: Lemma B.1.2, B.1.3 and B.1.4 which are stated and proved in the Section B.1.6. Due to space constraints, we only provide the initial part of the proof here moving the rest to Section B.1.1.

Fix $i = 1, \dots, n$. Because $\hat{h}_i = \hat{U}_{\hat{k}(i)}(\theta_i)$, we write

$$\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 = \sum_{k \in \mathcal{I}} \left(\hat{U}_k(\theta_i) - h_{K^*}(\theta_i)\right)^2 I\left\{\hat{k}(i) = k\right\}$$

where $I(\cdot)$ denotes the indicator function. Taking expectations on both sides and using Cauchy-Schwartz inequality, we obtain

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq \sum_{k \in \mathcal{I}} \sqrt{\mathbb{E}(\hat{U}_k(\theta_i) - h_{K^*}(\theta_i))^4} \sqrt{\mathbb{P}_{K^*}\left\{\hat{k}(i) = k\right\}}.$$

The random variable $\hat{U}_k - h_{K^*}(0)$ is normally distributed and we know that $\mathbb{E}Z^4 \leq 3(\mathbb{E}Z^2)^2$ for every gaussian random variable Z. We therefore have

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq \sqrt{3} \sum_{k \in \mathcal{I}} \mathbb{E}(\hat{U}_k(\theta_i) - h_{K^*}(\theta_i))^2 \sqrt{\mathbb{P}_{K^*}\left\{\hat{k}(i) = k\right\}}.$$

Because $\mathbb{E}_{K^*}\hat{U}_k(\theta_i) = U_k(\theta_i)$ (defined in (4.9)), we have

$$\mathbb{E}_{K^*}(\hat{U}_k(\theta_i) - h_{K^*}(\theta_i))^2 = (U_k(\theta_i) - h_{K^*}(\theta_i))^2 + \operatorname{var}(\hat{U}_k(\theta_i)).$$

Because $L_k(\theta_i) \leq h_{K^*}(\theta_i) \leq U_k(\theta_i)$, it is clear that $U_k(\theta_i) - h_{K^*}(\theta_i) \leq U_k(\theta) - L_k(\theta_i) = \Delta_k(\theta_i)$. Also, Lemma B.1.4 states that the variance of $\hat{U}_k$ is at most $\sigma^2/(k+1)$. Putting these together, we obtain

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq \sqrt{3} \sum_{k \in \mathcal{I}} \left(\Delta_k^2(\theta_i) + \frac{\sigma^2}{k+1}\right) \sqrt{\mathbb{P}_{K^*}\left\{\hat{k}(i) = k\right\}}.$$

The proof of (4.17) will therefore be complete if we show that

$$\sum_{k\in\mathcal{I}}\left(\Delta_k^2(\theta_i)+\frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}_{K^*}\left\{\hat{k}(i)=k\right\}}\le C\frac{\sigma^2}{k_*(i)+1} \tag{4.49}$$

for a universal positive constant C. The proof of this inequality is technical and we have moved it to the Section B.1.1.

4.7.2 Proof of Theorem 4.3.2

This subsection is dedicated to the proof of Theorem 4.3.2. The proof is again long and we have moved most of the Section B.1.2. The basic idea is presented below and is based on a classical inequality due to Le Cam [90] which states that for every estimator $\tilde{h}$ and compact, convex set L^*, the quantity

$$\max\left[\mathbb{E}_{K^*}\left(\tilde{h}-h_{K^*}(\theta_i)\right)^2,\mathbb{E}_{L^*}\left(\tilde{h}-h_{L^*}(\theta_i)\right)^2\right]$$

is bounded from above by

$$\ge\frac{1}{4}\left(h_{K^*}(\theta_i)-h_{L^*}(\theta_i)\right)^2\left(1-\|P_{K^*}-P_{L^*}\|_{TV}\right). \tag{4.50}$$

Here P_{L^*} is the product of the Gaussian probability measures with mean $h_{L^*}(\theta_i)$ and variance σ^2 for $i=1,\dots,n$. Also $\|P-Q\|_{TV}$ denotes the total variation distance between P and Q.

For ease of notation, we assume, without loss of generality, that $\theta_i=0$. We also write Δ_k for $\Delta_k(\theta_i)$ and k_* for $k_*(i)$.

Suppose first that K^* satisfies the following condition: There exists some $\alpha\in(0,\pi/4)$ such that

$$\frac{h_{K^*}(\alpha)+h_{K^*}(-\alpha)}{2\cos\alpha}-h_{K^*}(0)>\frac{\sigma}{\sqrt{n_\alpha}} \tag{4.51}$$

where n_α denotes the number of integers i for which $-\alpha<2i\pi/n<\alpha$. This condition will not be satisfied, for example, when K^* is a singleton. We shall handle such K^* later. Observe that $n_\alpha\ge 1$ for all $0<\alpha<\pi/4$ because we can take $i=0$.

Let us define, for each $\alpha\in(0,\pi/4)$,

$$a_{K^*}(\alpha):=\left(\frac{h_{K^*}(\alpha)+h_{K^*}(-\alpha)}{2\cos\alpha},\frac{h_{K^*}(\alpha)-h_{K^*}(-\alpha)}{2\sin\alpha}\right). \tag{4.52}$$

and let $L^*=L^*(\alpha)$ be defined as the smallest convex set that contains both K^* and the point $a_{K^*}(\alpha)$. In other words, L^* is the convex hull of $K^*\cup\{a_{K^*}(\alpha)\}$.

We now use Le Cam's bound (4.50) with this choice of L^*. Details are given in [27, Subsection B.1.2].

4.7.3 Proof of Theorem 4.3.3

Recall the definition of $\tilde{h}^P$ in (4.13) and the definition of the estimator $\hat{K}$ in (4.14). The first thing to note is that

$$h_{\hat{K}}(\theta_i) = \hat{h}_i^P \qquad \text{for every } i = 1, \dots, n. \tag{4.53}$$

To see this, observe first that, because $\hat{h}^P = (\hat{h}_1^P, \dots, \hat{h}_n^P)$ is a valid support vector, there exists a set $\tilde{K}$ with $h_{\tilde{K}}(\theta_i) = \hat{h}_i^P$ for every i. It is now trivial (from the definition of $\hat{K}$) to see that $\tilde{K} \subseteq \hat{K}$ which implies that $h_{\hat{K}(\theta_i)} \geq h_{\tilde{K}}(\theta_i) = \hat{h}_i^P$. On the other hand, the definition of $\hat{K}$ immediately gives $h_{\hat{K}}(\theta_i) \leq \hat{h}_i^P$.

The observation (4.53) immediately gives

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) = \mathbb{E}_{K^*} \frac{1}{n} \sum_{i=1}^{n} \left(h_{K^*}(\theta_i) - \hat{h}_i^P \right)^2$$

It will be convenient here to introduce the following notation. Let $h_{K^*}^{vec}$ denote the vector $(h_{K^*}(\theta_1), \dots, h_{K^*}(\theta_n))$. Also, for $u, v \in \mathbb{R}^n$, let $\ell(u, v)$ denote the scaled Euclidean distance defined by $\ell^2(u, v) := \sum_{i=1}^n (u_i - v_i)^2 / n$. With this notation, we have

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) = \mathbb{E}_{K^*} \ell^2(h_{K^*}^{vec}, \hat{h}^P). \tag{4.54}$$

Recall that $\hat{h}^P$ is the projection of $\hat{h} := (\hat{h}_1, \dots, \hat{h}_n)$ onto $\mathcal{H}$. Because $\mathcal{H}$ is a closed convex subset of $\mathbb{R}^n$, it follows that (see, for example, [133])

$$\ell^2(h, \hat{h}) \geq \ell^2(\hat{h}, \hat{h}^P) + \ell^2(h, \hat{h}^P) \qquad \text{for every } h \in \mathcal{H}.$$

In particular, with $h = h_{K^*}^{vec}$, we obtain $\ell^2(h_{K^*}^{vec}, \hat{h}^P) \leq \ell^2(h_{K^*}^{vec}, \hat{h})$. Combining this with (4.54), we obtain

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) \leq \mathbb{E}_{K^*} \ell^2(h_{K^*}^{vec}, \hat{h}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i) \right)^2. \tag{4.55}$$

In Theorem 4.3.1, we proved that

$$\mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i) \right)^2 \leq \frac{C\sigma^2}{k_*(i) + 1} \qquad \text{for every } i = 1, \dots, n.$$

This implies that

$$\mathbb{E}_{K^*} L_f(K^*, \hat{K}) \leq \frac{C\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{k_*(i) + 1}.$$

For inequality (4.32), it is therefore enough to prove that

$$\sum_{i=1}^{n} \frac{1}{k_*(i) + 1} \leq C \left\{ 1 + \left(\frac{R\sqrt{n}}{\sigma} \right)^{2/5} \right\}. \tag{4.56}$$

Proving the above inequality is the main part of the proof of Theorem 4.3.3. Because of space constraints, we have moved this proof to [27, Subsection B.1.5]. Our proof of (4.56) is inspired by an argument due to Zhang [159, Proof of Theorem 2.1] in a very different context.

4.7.4 Proof of Theorem 4.3.4

Let us start with some notation. For every compact, convex set P and $i = 1, \dots, n$, let $k_*^P(i)$ denote the quantity $k_*(i)$ with K^* replaced by P. More precisely,

$$k_*^P(i) := \operatorname*{argmin}_{k \in \mathcal{I}} \left(\Delta_k^P(\theta_i) + \frac{2\sigma}{\sqrt{k+1}} \right) \tag{4.57}$$

where $\Delta_k^P(\theta_i)$ is defined as in (4.40) with K^* replaced by P. Lemma B.1.6 (stated and proved in Section B.1.6 will be used crucially in the proof below. This lemma states that for every $i = 1, \dots, n$, the risk $\mathbb{E}_{K^*}(\hat{h}_i - h_{K^*}(\theta_i))^2$ can be bounded from above by a combination of $k_*^P(i)$ and how well K^* can be approximated by P. This result holds for every P. The approximation of K^* by P is measured in terms of the Hausdorff distance (defined in (4.35)).

We are now ready to prove Theorem 4.3.4. We first use inequality (4.55) from the proof of Theorem 4.3.3 which states

$$\mathbb{E}_{K^*} L_f \left(K^*, \hat{K} \right) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{K^*} \left(\hat{h}_i - h_{K^*}(\theta_i) \right)^2 .$$

An application of Lemma B.1.6, specifically inequality (B.47) for $i = 1, \dots, n$, now implies the existence of a universal positive constant C such that

$$\mathbb{E}_{K^*} L_f \left(K^*, \hat{K} \right) \leq C \left(\frac{\sigma^2}{n} \sum_{i=1}^n \frac{1}{k_*^P(i) + 1} + \ell_H^2(K^*, P) \right)$$

for every compact, convex set P. By restricting P to be in the class of polytopes, we get

$$\mathbb{E}_{K^*} L_f \left(K^*, \hat{K} \right) \leq C \inf_{P \in \mathcal{P}} \left(\frac{\sigma^2}{n} \sum_{i=1}^n \frac{1}{k_*^P(i) + 1} + \ell_H^2(K^*, P) \right) .$$

For the proof of (4.36), it is therefore enough to show that

$$\sum_{i=1}^n \frac{1}{k_*^P(i) + 1} \leq C v_P \log \frac{en}{v_P} \qquad \text{for every } P \in \mathcal{P} \tag{4.58}$$

where v_P denotes the number of extreme points of P and C is a universal positive constant. Fix a polytope P with $v_P = k$. Let the extreme points of P be $z_1, \dots, z_k$. Let $S_1, \dots, S_k$ denote a partition of $\{\theta_1, \dots, \theta_n\}$ into k nonempty sets such that for each $j = 1, \dots, m$, we have

$$h_P(\theta_i) = z_j(1) \cos \theta_i + z_j(2) \sin \theta_i \qquad \text{for all } \theta_i \in S_j$$

where $z_j = (z_j(1), z_j(2))$. For (4.58), it is enough to prove that

$$\sum_{i : \theta_i \in S_j} \frac{1}{k_*^P(i) + 1} \leq C \log(en_j) \qquad \text{for every } j = 1, \dots, k \tag{4.59}$$

where n_j is the cardinality of S_j. This is because we can write

$$\sum_{i=1}^{n} \frac{1}{k_*^P(i)+1} = \sum_{j=1}^{k} \sum_{i:\theta_i \in S_j} \frac{1}{k_*^P(i)+1} \leq C \sum_{j=1}^{k} \log(en_j) \leq Ck \log \frac{en}{k}.$$

where we used the concavity of $x \mapsto \log(ex)$. We prove (4.59) below. Fix $1 \leq j \leq k$. The inequality is obvious if S_j is a singleton because $k_*^P(i) \geq 0$. So suppose that $n_j = m \geq 2$. Without loss of generality assume that $S_j = \{\theta_{u+1}, \dots, \theta_{u+m}\}$ where $0 \leq u \leq n-m$. The definition of S_j implies that

$$h_P(\theta) = z_j(1)\cos\theta + z_j(2)\sin\theta \qquad \text{for all } \theta \in [\theta_{u+1}, \theta_{u+m}].$$

We can therefore apply inequality (4.27) to claim the existence of a positive constant c such that

$$k_*^P(i) \geq c\, n \min\left(\theta_i - \theta_{u+1}, \theta_{u+m} - \theta_i\right) \qquad \text{for all } u+1 \leq i \leq u+m.$$

The minimum with π in (4.27) is redundant here because $\theta_{u+m} - \theta_{u+1} < 2\pi$. Because $\theta_i = 2\pi i/n - \pi$, we get

$$k_*^P(i) \geq 2\pi c \min\left(i-u-1, u+m-i\right) \qquad \text{for all } u+1 \leq i \leq u+m.$$

Therefore, there exists a universal constant C such that

$$\sum_{i:\theta_i \in S_j} \frac{1}{k_*^P(i)+1} \leq C \sum_{i=1}^{m} \frac{1}{1+\min(i-1, m-i)} \leq C \sum_{i=1}^{m} \frac{1}{i} \leq C \log(em).$$

This proves (4.59) thereby completing the proof of Theorem 4.3.4.

4.7.5 Proof of Theorem 4.3.5

Recall the definition (4.16) of the estimator $\hat{K}'$ and that of the interpolating function (4.15). Following an argument similar to that used at the beginning of the proof of Theorem 4.3.3, we observe that

$$\mathbb{E}_{K^*} L(K^*, \hat{K}') \leq \int_{-\pi}^{\pi} \mathbb{E}_{K^*} \left(h_{K^*}(\theta) - \hat{h}'(\theta)\right)^2 d\theta = \sum_{i=1}^{n} \int_{\theta_i}^{\theta_{i+1}} \mathbb{E}_{K^*} \left(h_{K^*}(\theta) - \hat{h}'(\theta)\right)^2 d\theta \tag{4.60}$$

Now fix $1 \leq i \leq n$, $\theta_i \leq \theta \leq \theta_{i+1}$ and let $u(\theta) := \mathbb{E}_{K^*}\left(h_{K^*}(\theta) - \hat{h}'(\theta)\right)^2$. Using the expression (4.15) for $\hat{h}'(\theta)$, we get that

$$u(\theta) = \mathbb{E}_{K^*} \left(h_{K^*}(\theta) - \frac{\sin(\theta_{i+1} - \theta)}{\sin(\theta_{i+1} - \theta_i)} \hat{h}_i - \frac{\sin(\theta - \theta_i)}{\sin(\theta_{i+1} - \theta_i)} \hat{h}_{i+1}\right)^2.$$

We now write $\hat{h}_i = \hat{h}_i - h_{K^*}(\theta_i) + h_{K^*}(\theta_i)$ and a similar expression for $\hat{h}_{i+1}$. The elementary inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ along with $\max(\sin(\theta-\theta_i), \sin(\theta_{i+1}-\theta)) \leq \sin(\theta_{i+1}-\theta_i)$ then imply that

$$u(\theta) \leq 3\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 + 3\mathbb{E}_{K^*}\left(\hat{h}_{i+1} - h_{K^*}(\theta_{i+1})\right)^2 + 3b^2(\theta)$$

where

$$b(\theta) := h_{K^*}(\theta) - \frac{\sin(\theta_{i+1}-\theta)}{\sin(\theta_{i+1}-\theta_i)} h_{K^*}(\theta_i) - \frac{\sin(\theta-\theta_i)}{\sin(\theta_{i+1}-\theta_i)} h_{K^*}(\theta_{i+1})$$

Therefore from (4.60) (remember that $|\theta_{i+1}-\theta_i| = 2\pi/n$), we deduce

$$\mathbb{E}_{K^*} L(K^*, \hat{K}') \leq \frac{12\pi}{n}\sum_{i=1}^{n} \mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 + 3\int_{-\pi}^{\pi} b^2(\theta)d\theta.$$

Now to bound $\sum_{i=1}^n \mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2$, we can simply use the arguments from the proofs of Theorems 4.3.3 and 4.3.4. Therefore, to complete the proof of Theorem 4.3.5, we only need to show that

$$|b(\theta)| \leq \frac{CR}{n} \qquad \text{for every } \theta \in (-\pi, \pi] \tag{4.61}$$

for some universal constant C. For this, we use the hypothesis that K^* is contained in a ball of radius R. Suppose that the center of the ball is (x_1, x_2). Define $K' := K^* - \{(x_1,x_2)\} := \{(y_1,y_2) - (x_1,x_2) : (y_1,y_2) \in K^*\}$ and note that $h_{K'}(\theta) = h_{K^*}(\theta) - x_1\cos\theta - x_2\sin\theta$. It is then easy to see that $b(\theta)$ is the same for both K^* and K'. It is therefore enough to prove (4.61) assuming that $(x_1,x_2) = (0,0)$. In this case, it is straightforward to see that $|h_{K^*}(\theta)| \leq R$ for all θ and also that h_{K^*} is Lipschitz with constant R. Now, because $\max(\sin(\theta-\theta_i), \sin(\theta_{i+1}-\theta)) \leq \sin(\theta_{i+1}-\theta_i)$, it can be checked that $|b(\theta)|$ is bounded from above by

$$|h_{K^*}(\theta)|\left|1 - \frac{\sin(\theta_{i+1}-\theta)}{\sin(\theta_{i+1}-\theta_i)} - \frac{\sin(\theta-\theta_i)}{\sin(\theta_{i+1}-\theta_i)}\right| + \sum_{j=i}^{i+1}|h_{K^*}(\theta_j) - h_{K^*}(\theta)|.$$

Because h_{K^*} is R-Lipschitz and bounded by R, it is clear that we only need to show

$$\left|1 - \frac{\sin(\theta_{i+1}-\theta)}{\sin(\theta_{i+1}-\theta_i)} - \frac{\sin(\theta-\theta_i)}{\sin(\theta_{i+1}-\theta_i)}\right| \leq \frac{C}{n}$$

in order to prove (4.61). For this, write $\alpha = \theta_{i+1} - \theta$ and $\beta = \theta - \theta_i$ so that the above expression becomes

$$\left|1 - \frac{\sin\alpha + \sin\beta}{\sin(\alpha+\beta)}\right| \leq |1-\cos\alpha| + |1-\cos\beta| \leq \frac{\alpha^2+\beta^2}{2} \leq \frac{C}{n^2} \leq \frac{C}{n}.$$

This completes the proof of Theorem 4.3.5.

Part III

Optimization

Chapter 5

Early stopping for kernel boosting algorithms

5.1 Introduction

While non-parametric models offer great flexibility, they can also lead to overfitting, and thus poor generalization performance. For this reason, it is well-understood that procedures for fitting non-parametric models must involve some form of regularization. When models are fit via a form of empirical risk minimization, the most classical form of regularization is based on adding some type of penalty to the objective function. An alternative form of regularization is based on the principle of *early stopping*, in which an iterative algorithm is run for a pre-specified number of steps, and terminated prior to convergence.

While the basic idea of early stopping is fairly old (e.g., [134, 4, 142]), recent years have witnessed renewed interests in its properties, especially in the context of boosting algorithms and neural network training (e.g., [114, 35]). Over the past decade, a line of work has yielded some theoretical insight into early stopping, including works on classification error for boosting algorithms [15, 53, 84, 101, 155, 160], L^2-boosting algorithms for regression [26, 25], and similar gradient algorithms in reproducing kernel Hilbert spaces (e.g. [33, 32, 141, 155, 116]). A number of these papers establish consistency results for particular forms of early stopping, guaranteeing that the procedure outputs a function with statistical error that converges to zero as the sample size increases. On the other hand, there are relatively few results that actually establish *rate optimality* of an early stopping procedure, meaning that the achieved error matches known statistical minimax lower bounds. To the best of our knowledge, Bühlmann and Yu [26] were the first to prove optimality for early stopping of L^2-boosting as applied to spline classes, albeit with a rule that was not computable from the data. Subsequent work by Raskutti et al. [116] refined this analysis of L^2-boosting for kernel classes and first established an important connection to the localized Rademacher complexity; see also the related work [155, 123, 31] with rates for particular kernel classes.

More broadly, relative to our rich and detailed understanding of regularization via penalization (e.g., see the books [76, 138, 136, 144] and papers [13, 88] for details), our understanding of early stopping regularization is not as well developed. Intuitively, early stopping should depend on the same bias-variance tradeoffs that control estimators based on penalization. In particular, for penalized estimators, it is now well-understood that complexity measures such as the *localized Gaussian width*, or its Rademacher analogue, can be used to characterize their achievable rates [13, 88, 136, 144]. Is such a general and sharp characterization also possible in the context of early stopping?

The main intention of this chapter is to answer this question in the affirmative for the early stopping of boosting algorithms for a certain class of regression and classification problems involving functions in reproducing kernel Hilbert spaces (RKHS). A standard way to obtain a good estimator or classifier is through minimizing some penalized form of loss functions of which the method of kernel ridge regression [143] is a popular choice. Instead, we consider an iterative update involving the kernel that is derived from a greedy update. Borrowing tools from empirical process theory, we are able to characterize the "size" of the effective function space explored by taking T steps, and then to connect the resulting estimation error naturally to the notion of localized Gaussian width defined with respect to this effective function space. This leads to a principled analysis for a broad class of loss functions used in practice, including the loss functions that underlie the L^2-boost, LogitBoost and AdaBoost algorithms, among other procedures.

The remainder of this chapter is organized as follows. In Section 5.2, we provide background on boosting methods and reproducing kernel Hilbert spaces, and then introduce the updates studied in this chapter. Section 5.3 is devoted to statements of our main results, followed by a discussion of their consequences for particular function classes in Section 5.4. We provide simulations that confirm the practical effectiveness of our stopping rules, and show close agreement with our theoretical predictions. In Section 5.6, we provide the proofs of our main results, with certain more technical aspects deferred to the appendices.

5.2 Background and problem formulation

The goal of prediction is to learn a function that maps *covariates* $x \in \mathcal{X}$ to *responses* $y \in \mathcal{Y}$. In a regression problem, the responses are typically real-valued, whereas in a classification problem, the responses take values in a finite set. In this chapter, we study both regression ($\mathcal{Y} = \mathbb{R}$) and classification problems (e.g., $\mathcal{Y} = \{-1, +1\}$ in the binary case). Our primary focus is on the case of *fixed design*, in which we observe a collection of n pairs of the form $\{(x_i, Y_i)\}_{i=1}^n$, where each $x_i \in \mathcal{X}$ is a fixed covariate, whereas $Y_i \in \mathcal{Y}$ is a random response drawn independently from a distribution $\mathbb{P}_{Y|x_i}$ which depends on x_i. Later in the chapter, we also discuss the consequences of our results for the case of random design, where the (X_i, Y_i) pairs are drawn in an i.i.d. fashion from the joint distribution $\mathbb{P} = \mathbb{P}_X \mathbb{P}_{Y|X}$ for some distribution $\mathbb{P}_X$ on the covariates.

In this section, we provide some necessary background on a gradient-type algorithm which is often referred to as *boosting* algorithm. We also discuss briefly about the reproducing kernel

Hilbert spaces before turning to a precise formulation of the problem that is studied in this chapter.

5.2.1 Boosting and early stopping

Consider a cost function $\phi : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, where the non-negative scalar $\phi(y, \theta)$ denotes the cost associated with predicting θ when the true response is y. Some common examples of loss functions ϕ that we consider in later sections include:

- the *least-squares loss* $\phi(y, \theta) := \frac{1}{2}(y - \theta)^2$ that underlies L^2-boosting [26],
- the *logistic regression loss* $\phi(y, \theta) = \ln(1 + e^{-y\theta})$ that underlies the LogitBoost algorithm [55, 56], and
- the *exponential loss* $\phi(y, \theta) = \exp(-y\theta)$ that underlies the AdaBoost algorithm [53].

The least-squares loss is typically used for regression problems (e.g., [26, 33, 32, 141, 155, 116]), whereas the latter two losses are frequently used in the setting of binary classification (e.g., [53, 101, 56]).

We have set up the non-parametric estimation problem in our Section 2.2. To recall, we define the *population cost functional* $f \mapsto \mathcal{L}(f)$ via

$$\mathcal{L}(f) := \mathbb{E}_{Y_1^n}\Big[\frac{1}{n}\sum_{i=1}^{n} \phi\big(Y_i, f(x_i)\big)\Big]. \tag{5.1}$$

Note that with the covariates $\{x_i\}_{i=1}^n$ fixed, the functional $\mathcal{L}$ is a non-random object. Given some function space $\mathcal{F}$, the optimal function[*] minimizes the population cost functional—that is

$$f^* := \arg\min_{f \in \mathcal{F}} \mathcal{L}(f). \tag{5.2}$$

As a standard example, when we adopt the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, the population minimizer f^* corresponds to the conditional expectation $x \mapsto \mathbb{E}[Y \mid x]$.

Since we do not have access to the population distribution of the responses however, the computation of f^* is impossible. Given our samples $\{Y_i\}_{i=1}^n$, we consider instead some procedure applied to the *empirical loss*

$$\mathcal{L}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, f(x_i)), \tag{5.3}$$

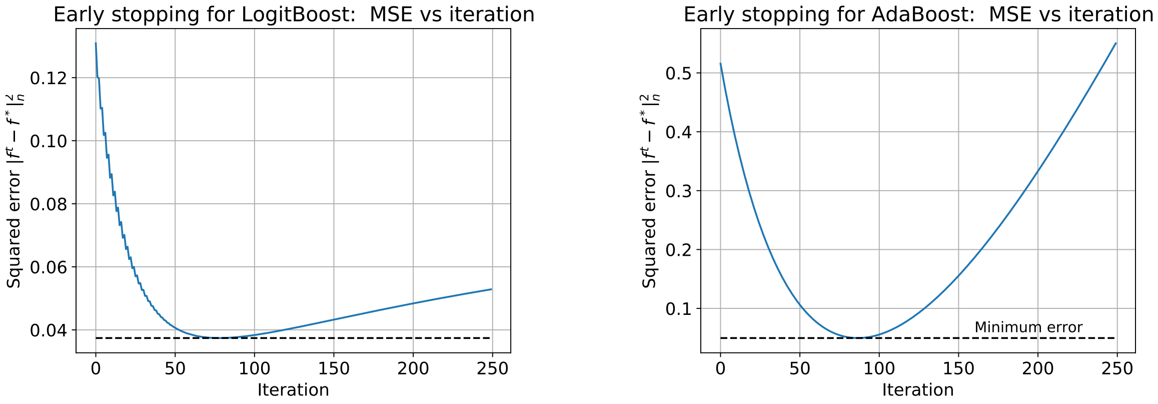
where the population expectation has been replaced by an empirical expectation. For example, when $\mathcal{L}_n$ corresponds to the log likelihood of the samples with $\phi(Y_i, f(x_i)) =$

[*] As clarified in the sequel, our assumptions guarantee uniqueness of f^*.

$\log[\mathbb{P}(Y_i; f(x_i))]$, direct unconstrained minimization of $\mathcal{L}_n$ would yield the maximum likelihood estimator.

It is well-known that direct minimization of $\mathcal{L}_n$ over a sufficiently rich function class $\mathcal{F}$ may lead to overfitting. There are various ways to mitigate this phenomenon, among which the most classical method is to minimize the sum of the empirical loss with a penalty regularization term. Adjusting the weight on the regularization term allows for trade-off between fit to the data, and some form of regularity or smoothness in the fit. The behavior of such penalized of regularized estimation methods is now quite well understood (for instance, see the books [76, 138, 136, 144] and papers [13, 88] for more details).

In this chapter, we study a form of *algorithmic regularization*, based on applying a gradient-type algorithm to $\mathcal{L}_n$ but then stopping it "early"—that is, after some fixed number of steps. Such methods are often referred to as *boosting algorithms*, since they involve "boosting" or improve the fit of a function via a sequence of additive updates (see e.g. [124, 53, 21, 20, 125]). Many boosting algorithms, among them AdaBoost [53], L^2-boosting [26] and LogitBoost [55, 56], can be understood as forms of functional gradient methods [101, 56]; see the survey paper [25] for further background on boosting. The way in which the number of steps is chosen is referred to as a stopping rule, and the overall procedure is referred to as *early stopping* of a boosting algorithm.



(a) (b)

Figure 5.1. Plots of the squared error $\|f^t - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f^t(x_i) - f^*(x_i))^2$ versus the iteration number t for (a) LogitBoost using a first-order Sobolev kernel (b) AdaBoost using the same first-order Sobolev kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$ which generates a class of Lipschitz functions (splines of order one). Both plots correspond to a sample size $n = 100$.

In more detail, a broad class of boosting algorithms [101] generate a sequence $\{f^t\}_{t=0}^\infty$ via updates of the form

$$f^{t+1} = f^t - \alpha^t g^t \quad \text{with} \quad g^t \propto \arg\max_{\|d\|_{\mathcal{F}} \leq 1} \langle \nabla \mathcal{L}_n(f^t),\, d(x_1^n) \rangle, \tag{5.4}$$

where the scalar $\{\alpha^t\}_{t=0}^{\infty}$ is a sequence of step sizes chosen by the user, the constraint $\|d\|_{\mathcal{F}} \leq 1$ defines the unit ball in a given function class $\mathcal{F}$, $\nabla \mathcal{L}_n(f) \in \mathbb{R}^n$ denotes the gradient taken at the vector $\big(f(x_1), \ldots, f(x_n)\big)$, and $\langle h, g \rangle$ is the usual inner product between vectors $h, g \in \mathbb{R}^n$. For non-decaying step sizes and a convex objective $\mathcal{L}_n$, running this procedure for an infinite number of iterations will lead to a minimizer of the empirical loss, thus causing overfitting. In order to illustrate this phenomenon, Figure 5.1 provides plots of the squared error $\|f^t - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \big(f^t(x_i) - f^*(x_i)\big)^2$ versus the iteration number, for LogitBoost in panel (a) and AdaBoost in panel (b). See Section 5.4.2 for more details on exactly how these experiments were conducted.

In the plots in Figure 5.1, the dotted line indicates the minimum mean-squared error ρ_n^2 over all iterates of that particular run of the algorithm. Both plots are qualitatively similar, illustrating the existence of a "good" number of iterations to take, after which the MSE greatly increases. Hence a natural problem is to decide at what iteration T to stop such that the iterate f^T satisfies bounds of the form

$$\mathcal{L}(f^T) - \mathcal{L}(f^*) \precsim \rho_n^2 \quad \text{and} \quad \|f^T - f^*\|_n^2 \precsim \rho_n^2 \tag{5.5}$$

with high probability. Here $f(n) \precsim g(n)$ indicates that $f(n) \leq cg(n)$ for some universal constant $c \in (0, \infty)$. The main results of this part provide a stopping rule T for which bounds of the form (5.5) do in fact hold with high probability over the randomness in the observed responses.

5.2.2 Reproducing Kernel Hilbert Spaces

The analysis of this chapter focuses on algorithms with the update (5.4) when the function class $\mathcal{F}$ is a reproducing kernel Hilbert space $\mathscr{H}$. Several important properties of this space is summarized in our Section 2.2.2. To recall, a reproducing kernel Hilbert space $\mathscr{H}$ (short as RKHS), consisting of functions mapping a domain $\mathcal{X}$ to the real line $\mathbb{R}$. Any RKHS is defined by a bivariate symmetric *kernel function* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which is required to be positive semidefinite, i.e. for any integer $N \geq 1$ and a collection of points $\{x_j\}_{j=1}^N$ in $\mathcal{X}$, the matrix $[\mathbb{K}(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$ is positive semidefinite.

Throughout this chapter, we assume that the kernel function is uniformly bounded, meaning that there is a constant L such that $\sup_{x \in \mathcal{X}} \mathbb{K}(x, x) \leq L$. Such a boundedness condition holds for many kernels used in practice, including the Gaussian, Laplacian, Sobolev, other types of spline kernels, as well as any trace class kernel with trigonometric eigenfunctions. By rescaling the kernel as necessary, we may assume without loss of generality that $L = 1$. As a consequence, for any function f such that $\|f\|_{\mathscr{H}} \leq r$, we have by the reproducing relation (2.13) that

$$\|f\|_\infty = \sup_x \langle f, \mathbb{K}(\cdot, x) \rangle_{\mathscr{H}} \leq \|f\|_{\mathscr{H}} \sup_x \|\mathbb{K}(\cdot, x)\|_{\mathscr{H}} \leq r.$$

Given samples $\{(x_i, y_i)\}_{i=1}^n$, by the representer theorem [86], it is sufficient to restrict ourselves to the linear subspace $\mathscr{H}_n = \overline{\operatorname{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$, for which all $f \in \mathscr{H}_n$ can be expressed as

$$f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i \mathbb{K}(\cdot, x_i) \tag{5.6}$$

for some coefficient vector $\omega \in \mathbb{R}^n$. Among those functions which achieve the infimum in expression (5.1), let us define f^* as the one with the minimum Hilbert norm. This definition is equivalent to restricting f^* to be in the linear subspace $\mathscr{H}_n$.

5.2.3 Boosting in kernel spaces

For a finite number of covariates x_i from $i = 1 \dots n$, let us define the *normalized kernel matrix* $K \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathbb{K}(x_i, x_j)/n$. Since we can restrict the minimization of $\mathcal{L}_n$ and $\mathcal{L}$ from $\mathscr{H}$ to the subspace $\mathscr{H}_n$ w.l.o.g., using expression (5.6) we can then write the function value vectors $f(x_1^n) := (f(x_1), \dots, f(x_n))$ as $f(x_1^n) = \sqrt{n} K \omega$. As there is a one-to-one correspondence between the n-dimensional vectors $f(x_1^n) \in \mathbb{R}^n$ and the corresponding function $f \in \mathscr{H}_n$ in $\mathscr{H}$ by the representer theorem, minimization of an empirical loss in the subspace $\mathscr{H}_n$ essentially becomes the n-dimensional problem of fitting a response vector y over the set range(K). In the sequel, all updates will thus be performed on the function value vectors $f(x_1^n)$.

With a change of variable $d(x_1^n) = \sqrt{n}\sqrt{K} z$ we then have

$$d^t(x_1^n) := \arg \max_{\substack{\|d\|_{\mathscr{H}} \leq 1 \\ d \in \text{range}(K)}} \langle \nabla \mathcal{L}_n(f^t),\, d(x_1^n) \rangle = \frac{\sqrt{n} K \nabla \mathcal{L}_n(f^t)}{\sqrt{\nabla \mathcal{L}_n(f^t) K \nabla \mathcal{L}_n(f^t)}}.$$

In this chapter, we study the choice $g^t = \langle \nabla \mathcal{L}_n(f^t),\, d^t(x_1^n) \rangle d^t$ in the boosting update (5.4), so that the function value iterates take the form

$$f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha n K \nabla \mathcal{L}_n(f^t), \tag{5.7}$$

where $\alpha > 0$ is a constant stepsize choice. Choosing $f^0(x_1^n) = 0$ ensures that all iterates $f^t(x_1^n)$ remain in the range space of K.

In this chapter, we consider the following three error measures for an estimator $\widehat{f}$:

$$\begin{aligned} L^2(\mathbb{P}_n) \text{ norm:} \quad & \|\widehat{f} - f^*\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{f}(x_i) - f^*(x_i)\big)^2, \\ L^2(\mathbb{P}_X) \text{ norm:} \quad & \|\widehat{f} - f^*\|_2^2 := \mathbb{E}\big(\widehat{f}(X) - f^*(X)\big)^2, \\ \text{Excess risk:} \quad & \mathcal{L}(\widehat{f}) - \mathcal{L}(f^*), \end{aligned}$$

where the expectation in the $L^2(\mathbb{P}_X)$-norm is taken over random covariates X which are independent of the samples (X_i, Y_i) used to form the estimate $\widehat{f}$. Our goal is to propose a stopping time T such that the averaged function $\widehat{f} = \frac{1}{T}\sum_{t=1}^{T} f^t$ satisfies bounds of the type (5.5). We begin our analysis by focusing on the empirical $L^2(\mathbb{P}_n)$ error, but as we will see in Corollary 3, bounds on the empirical error are easily transformed to bounds on the population $L^2(\mathbb{P}_X)$ error. Importantly, we exhibit such bounds with a statistical error term δ_n that is specified by the *localized Gaussian complexity* of the kernel class.

5.3 Main results

We now turn to the statement of our main results, beginning with the introduction of some regularity assumptions.

5.3.1 Assumptions

Recall from our earlier set-up that we differentiate between the empirical loss function $\mathcal{L}_n$ in expression (5.3), and the population loss $\mathcal{L}$ in expression (5.1). Apart from assuming differentiability of both functions, all of our remaining conditions are imposed on the population loss. Such conditions at the population level are weaker than their analogues at the empirical level.

For a given radius $r > 0$, let us define the Hilbert ball around the optimal function f^* as

$$\mathbb{B}_{\mathscr{H}}(f^*, r) := \{f \in \mathscr{H} \mid \|f - f^*\|_{\mathscr{H}} \le r\}. \tag{5.8}$$

Our analysis makes particular use of this ball defined for the radius $C_{\mathscr{H}}^2 := 2\max\{\|f^*\|_{\mathscr{H}}^2,\ 32,\sigma^2\}$ where the effective noise level σ is defined in the sequel.

We assume that the population loss is m-strongly convex and M-smooth over $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, meaning that the

$$\textbf{m-M-condition:} \qquad \frac{m}{2}\|f-g\|_n^2 \le \mathcal{L}(f) - \mathcal{L}(g) - \langle \nabla\mathcal{L}(g),\, f(x_1^n) - g(x_1^n)\rangle \le \frac{M}{2}\|f-g\|_n^2$$

holds for all $f, g \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$ and all design points $\{x_i\}_{i=1}^n$. In addition, we assume that the function ϕ is M-Lipschitz in its second argument over the interval $\theta \in [\min_{i\in[n]} f^*(x_i) - 2C_{\mathscr{H}}, \max_{i\in[n]} f^*(x_i) + 2C_{\mathscr{H}}]$. To be clear, here $\nabla\mathcal{L}(g)$ denotes the vector in $\mathbb{R}^n$ obtained by taking the gradient of $\mathcal{L}$ with respect to the vector $g(x_1^n)$. It can be verified by a straightforward computation that when $\mathcal{L}$ is induced by the least-squares cost $\phi(y,\theta) = \frac{1}{2}(y-\theta)^2$, the m-M-condition holds for $m = M = 1$. The logistic and exponential loss satisfy this condition (see supp. material), where it is key that we have imposed the condition *only locally* on the ball $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$.

In addition to the least-squares cost, our theory also applies to losses $\mathcal{L}$ induced by scalar functions ϕ that satisfy the

$$\textbf{ϕ'-boundedness:} \qquad \max_{i=1,\ldots,n} \left|\frac{\partial\phi(y,\theta)}{\partial\theta}\right|_{\theta=f(x_i)} \le B, \qquad \text{for all } f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}}) \text{ and } y \in \mathcal{Y}.$$

This condition holds with $B = 1$ for the logistic loss for all $\mathcal{Y}$, and $B = \exp(2.5 C_{\mathscr{H}})$ for the exponential loss for binary classification with $\mathcal{Y} = \{-1, 1\}$, using our kernel boundedness condition. Note that whenever this condition holds with some finite B, we can always rescale the scalar loss ϕ by $1/B$ so that it holds with $B = 1$, and we do so in order to simplify the statement of our results.

5.3.2 Upper bound in terms of localized Gaussian width

Our upper bounds involve a complexity measure known as the localized Gaussian width. In general, Gaussian widths are widely used to obtain risk bounds for least-squares and other types of M-estimators. In our case, we consider Gaussian complexities for "localized" sets of the form

$$\mathcal{E}_n(\delta, 1) := \Big\{ f - g \mid \|f - g\|_{\mathscr{H}} \leq 1, \ \|f - g\|_n \leq \delta \Big\} \tag{5.9}$$

with $f, g \in \mathscr{H}$. The Gaussian complexity localized at scale δ is given by

$$\mathcal{G}_n\big(\mathcal{E}_n(\delta, 1)\big) := \mathbb{E}\Big[\sup_{g \in \mathcal{E}_n(\delta,1)} \frac{1}{n} \sum_{i=1}^{n} w_i g(x_i)\Big], \tag{5.10}$$

where $(w_1, \ldots, w_n)$ denotes an i.i.d. sequence of standard Gaussian variables.

An essential quantity in our theory is specified by a certain fixed point equation that is now standard in empirical process theory [136, 13, 88, 116]. Let us define the *effective noise level*

$$\sigma := \begin{cases} \min\Big\{ t \mid \max\limits_{i=1,\ldots,n} \mathbb{E}[e^{((Y_i - f^*(x_i))^2/t^2)}] < \infty \Big\} & \text{for L.S.} \\ 4\,(2M+1)(1 + 2C_{\mathscr{H}}) & \text{for } \phi'\text{-bounded losses.} \end{cases} \tag{5.11}$$

The *critical radius* δ_n is the smallest positive scalar such that

$$\frac{\mathcal{G}_n(\mathcal{E}_n(\delta, 1))}{\delta} \leq \frac{\delta}{\sigma}. \tag{5.12}$$

We note that past work on localized Rademacher and Gaussian complexity [105, 13] guarantees that there exists a unique $\delta_n > 0$ that satisfies this condition, so that our definition is sensible.

5.3.2.1 Upper bounds on excess risk and empirical $L^2(\mathbb{P}_n)$-error

With this set-up, we are now equipped to state our main theorem. It provides high-probability bounds on the excess risk and $L^2(\mathbb{P}_n)$-error of the estimator $\bar{f}^T := \frac{1}{T}\sum_{t=1}^{T} f^t$ defined by averaging the T iterates of the algorithm. It applies to both the least-squares cost function, and more generally, to any loss function satisfying the m-M-condition and the ϕ'-boundedness condition.

Theorem 1. *Suppose that the sample size n large enough such that $\delta_n \leq \frac{M}{m}$, and we compute the sequence $\{f^t\}_{t=0}^{\infty}$ using the update (5.7) with initialization $f^0 = 0$ and any step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$. Then for any iteration $T \in \big\{0, 1, \dots \lfloor \frac{m}{8M\delta_n^2} \rfloor\big\}$, the averaged function estimate $\bar{f}^T$ satisfies the bounds*

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \leq CM\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big), \quad \textit{and} \tag{5.13a}$$

$$\|\bar{f}^T - f^*\|_n^2 \leq C\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big), \tag{5.13b}$$

where both inequalities hold with probability at least $1 - c_1 \exp(-C_2 \frac{m^2 n \delta_n^2}{\sigma^2})$.

We prove Theorem 1 in Section 5.6.1.

A few comments about the constants in our statement: in all cases, constants of the form c_j are universal, whereas the capital C_j may depend on parameters of the joint distribution and population loss $\mathcal{L}$. In Theorem 1, we have the explicit value $C_2 = \{\frac{m^2}{\sigma^2}, 1\}$ and C^2 is proportional to the quantity $2\max\{\|f^*\|_{\mathscr{H}}^2,\ 32,\ \sigma^2\}$. While inequalities (5.13a) and (5.13b) are stated as high probability results, similar bounds for expected loss (over the response y_i, with the design fixed) can be obtained by a simple integration argument.

In order to gain intuition for the claims in the theorem, note that apart from factors depending on (m, M), the first term $\frac{1}{\alpha m T}$ dominates the second term $\frac{\delta_n^2}{m^2}$ whenever $T \lesssim 1/\delta_n^2$. Consequently, up to this point, taking further iterations reduces the upper bound on the error. This reduction continues until we have taken of the order $1/\delta_n^2$ many steps, at which point the upper bound is of the order δ_n^2.

More precisely, suppose that we perform the updates with step size $\alpha = \frac{m}{M}$; then, after a total number of $\tau := \frac{1}{\delta_n^2 \max\{8, M\}}$ many iterations, the extension of Theorem 1 to expectations guarantees that the mean squared error is bounded as

$$\mathbb{E}\|\bar{f}^\tau - f^*\|_n^2 \leq C' \frac{\delta_n^2}{m^2}, \tag{5.14}$$

where C' is another constant depending on $C_{\mathscr{H}}$. Here we have used the fact that $M \geq m$ in simplifying the expression. It is worth noting that guarantee (5.14) matches the best known upper bounds for kernel ridge regression (KRR)—indeed, this must be the case, since a sharp analysis of KRR is based on the same notion of localized Gaussian complexity (e.g. [12, 13]) . Thus, our results establish a strong parallel between the *algorithmic regularization* of early stopping, and the *penalized regularization* of kernel ridge regression. Moreover, as will be clarified in Section 5.3.3, under suitable regularity conditions on the RKHS, the critical squared radius δ_n^2 also acts as a lower bound for the expected risk, meaning that our upper bounds are not improvable in general.

Note that the critical radius δ_n^2 only depends on our observations $\{(x_i, y_i)\}_{i=1}^n$ through the solution of inequality (5.12). In many cases, it is possible to compute and/or upper bound this critical radius, so that a concrete and valid stopping rule can indeed by calculated in advance. In Section 5.4, we provide a number of settings in which this can be done in terms of the eigenvalues $\{\mu_j\}_{j=1}^n$ of the normalized kernel matrix.

5.3.2.2 Consequences for random design regression

Thus far, our analysis has focused purely on the case of fixed design, in which the sequence of covariates $\{x_i\}_{i=1}^n$ is viewed as fixed. If we instead view the covariates as being sampled in an i.i.d. manner from some distribution $\mathbb{P}_X$ over $\mathcal{X}$, then the empirical error $\|\widehat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(f(x_i) - f^*(x_i)\big)^2$ of a given estimate $\widehat{f}$ is a random quantity, and it is interesting to relate it to the squared population $L^2(\mathbb{P}_X)$-norm $\|\widehat{f} - f^*\|_2^2 = \mathbb{E}\big[(\widehat{f}(X) - f^*(X))^2\big]$.

In order to state an upper bound on this error, we introduce a population analogue of the critical radius δ_n, which we denote by $\bar{\delta}_n$. Consider the set

$$\bar{\mathcal{E}}(\delta, 1) := \Big\{ f - g \mid f, g \in \mathscr{H}, \ \|f - g\|_{\mathscr{H}} \leq 1, \ \|f - g\|_2 \leq \delta \Big\}. \tag{5.15}$$

It is analogous to the previously defined set $\mathcal{E}(\delta, 1)$, except that the empirical norm $\|\cdot\|_n$ has been replaced by the population version. The population Gaussian complexity localized at scale δ is given by

$$\bar{\mathcal{G}}_n\big(\bar{\mathcal{E}}(\delta,1)\big) := \mathbb{E}_{w,X}\Big[\sup_{g \in \bar{\mathcal{E}}(\delta,1)} \frac{1}{n}\sum_{i=1}^n w_i g(X_i) \Big], \tag{5.16}$$

where $\{w_i\}_{i=1}^n$ are an i.i.d. sequence of standard normal variates, and $\{X_i\}_{i=1}^n$ is a second i.i.d. sequence, independent of the normal variates, drawn according to $\mathbb{P}_X$. Finally, the population critical radius $\bar{\delta}_n$ is defined by equation (5.10), in which $\mathcal{G}_n$ is replaced by $\bar{\mathcal{G}}_n$.

Corollary 3. *In addition to the conditions of Theorem 1, suppose that the sequence $\{(X_i, Y_i)\}_{i=1}^n$ of covariate-response pairs are drawn i.i.d. from some joint distribution $\mathbb{P}$, and we compute the boosting updates with step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$ and initialization $f^0 = 0$. Then the averaged function estimate $\bar{f}^T$ at time $T := \lfloor \frac{1}{\bar{\delta}_n^2 \max\{8,M\}} \rfloor$ satisfies the bound*

$$\mathbb{E}_X\big(\bar{f}^T(X) - f^*(X)\big)^2 \ = \ \|\bar{f}^T - f^*\|_2^2 \leq \tilde{c}\, \bar{\delta}_n^2$$

with probability at least $1 - c_1 \exp(-C_2 \frac{m^2 n \bar{\delta}_n^2}{\sigma^2})$ over the random samples.

The proof of Corollary 3 follows directly from standard empirical process theory bounds [13, 116] on the difference between empirical risk $\|\bar{f}^T - f^*\|_n^2$ and population risk $\|\bar{f}^T - f^*\|_2^2$. In particular, it can be shown that $\|\cdot\|_2$ and $\|\cdot\|_n$ norms differ only by a factor proportion to $\bar{\delta}_n$. Furthermore, one can show that the empirical critical quantity δ_n is bounded by the population $\bar{\delta}_n$. By combining both arguments the corollary follows. We refer the reader to the papers [13, 116] for further details on such equivalences.

It is worth comparing this guarantee with the past work of Raskutti et al. [116], who analyzed the kernel boosting iterates of the form (5.7), but with attention restricted to the special case of the least-squares loss. Their analysis was based on first decomposing the squared error into bias and variance terms, then carefully relating the combination of these terms to a particular bound on the localized Gaussian complexity (see equation (5.17) below). In contrast, our theory more directly analyzes the effective function class that is explored by taking T steps, so that the localized Gaussian width (5.10) appears more naturally. In addition, our analysis applies to a broader class of loss functions.

In the case of reproducing kernel Hilbert spaces, it is possible to sandwich the localized Gaussian complexity by a function of the eigenvalues of the kernel matrix. Mendelson [105] provides this argument in the case of the localized Rademacher complexity, but similar arguments apply to the localized Gaussian complexity. Letting $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ denote the ordered eigenvalues of the normalized kernel matrix K, define the function

$$\mathcal{R}(\delta) = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{n} \min\{\delta^2, \mu_j\}}. \tag{5.17}$$

Up to a universal constant, this function is an upper bound on the Gaussian width $\mathcal{G}_n\big(\mathcal{E}(\delta, 1)\big)$ for all $\delta \geq 0$, and up to another universal constant, it is also a lower bound for all $\delta \geq \frac{1}{\sqrt{n}}$.

5.3.3 Achieving minimax lower bounds

In this section, we show that the upper bound (5.14) matches known minimax lower bounds on the error, so that our results are unimprovable in general. We establish this result for the class of *regular kernels*, as previously defined by Yang et al. [154], which includes the Gaussian and Sobolev kernels as special cases.

The class of regular kernels is defined as follows. Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ denote the ordered eigenvalues of the normalized kernel matrix K, and define the quantity $d_n := \operatorname{argmin}_{j=1,\ldots,n}\{\mu_j \leq \delta_n^2\}$. A kernel is called *regular* whenever there is a universal constant c such that the tail sum satisfies $\sum_{j=d_n+1}^{n} \mu_j \leq c\, d_n \delta_n^2$. In words, the tail sum of the eigenvalues for regular kernels is roughly on the same or smaller scale as the sum of the eigenvalues bigger than δ_n^2.

For such kernels and under the Gaussian observation model ($Y_i \sim N(f^*(x_i), \sigma^2)$), Yang et al. [154] prove a minimax lower bound involving δ_n. In particular, they show that the minimax risk over the unit ball of the Hilbert space is lower bounded as

$$\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c_\ell \delta_n^2. \tag{5.18}$$

Comparing the lower bound (5.18) with upper bound (5.14) for our estimator $\bar{f}^T$ stopped after $O(1/\delta_n^2)$ many steps, it follows that the bounds proven in Theorem 1 are unimprovable apart from constant factors.

We now state a generalization of this minimax lower bound, one which applies to a sub-class of *generalized linear models*, or GLM for short. In these models, the conditional distribution of the observed vector $Y = (Y_1, \ldots, Y_n)$ given $\big(f^*(x_1), \ldots, f^*(x_n)\big)$ takes the form

$$\mathbb{P}_\theta(y) = \prod_{i=1}^{n} \Big[h(y_i) \exp\big(\frac{y_i f^*(x_i) - \Phi(f^*(x_i))}{s(\sigma)}\big)\Big], \tag{5.19}$$

where $s(\sigma)$ is a known scale factor and $\Phi : \mathbb{R} \to \mathbb{R}$ is the cumulant function of the generalized linear model. As some concrete examples:

- The linear Gaussian model is recovered by setting $s(\sigma) = \sigma^2$ and $\Phi(t) = t^2/2$.
- The logistic model for binary responses $y \in \{-1, 1\}$ is recovered by setting $s(\sigma) = 1$ and $\Phi(t) = \log(1 + \exp(t))$.

Our minimax lower bound applies to the class of GLMs for which the cumulant function Φ is differentiable and has uniformly bounded second derivative $|\Phi''| \leq L$. This class includes the linear, logistic, multinomial families, among others, but excludes (for instance) the Poisson family. Under this condition, we have the following:

Corollary 4. *Suppose that we are given i.i.d. samples $\{y_i\}_{i=1}^n$ from a GLM (5.19) for some function f^* in a regular kernel class with $\|f^*\|_{\mathscr{H}} \leq 1$. Then running $T := \lfloor \frac{1}{\delta_n^2 \max\{8, M\}} \rfloor$ iterations with step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$ and $f^0 = 0$ yields an estimate $\bar{f}^T$ such that*

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \asymp \inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2. \tag{5.20}$$

Here $f(n) \asymp g(n)$ means $f(n) = cg(n)$ up to a universal constant $c \in (0, \infty)$. As always, in the minimax claim (5.20), the infimum is taken over all measurable functions of the input data and the expectation is taken over the randomness of the response variables $\{Y_i\}_{i=1}^n$. Since we know that $\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \precsim \delta_n^2$, the way to prove bound (5.20) is by establishing $\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \succsim \delta_n^2$. See Section 5.6.2 for the proof of this result.

At a high level, the statement in Corollary 4 shows that early stopping prevents us from overfitting to the data; in particular, using the stopping time T yields an estimate that attains the optimal balance between bias and variance.

5.4 Consequences for various kernel classes

In this section, we apply Theorem 1 to derive some concrete rates for different kernel spaces and then illustrate them with some numerical experiments. It is known that the complexity of an RKHS in association with a distribution over the covariates $\mathbb{P}_X$ can be characterized by the decay rate (2.14) of the eigenvalues of the kernel function. In the finite sample setting, the analogous quantities are the eigenvalues $\{\mu_j\}_{j=1}^n$ of the normalized kernel matrix K. The representation power of a kernel class is directly correlated with the eigen-decay: the faster the decay, the smaller the function class. When the covariates are drawn from the distribution $\mathbb{P}_X$, empirical process theory guarantees that the empirical and population eigenvalues are close.

5.4.1 Theoretical predictions as a function of decay

In this section, let us consider two broad types of eigen-decay:

- **γ-exponential decay**: For some $\gamma > 0$, the kernel matrix eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2 j^\gamma)$, where c_1, c_2 are universal constants. Examples of kernels in this class include the Gaussian kernel, which for the Lebesgue measure satisfies such a bound with $\gamma = 2$ (real line) or $\gamma = 1$ (compact domain).

- **β-polynomial decay**: For some $\beta > 1/2$, the kernel matrix eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 j^{-2\beta}$, where c_1 is a universal constant. Examples of kernels in this class include the k^{th}-order Sobolev spaces for some fixed integer $k \geq 1$ with Lebesgue measure on a bounded domain. We consider Sobolev spaces that consist of functions that have k^{th}-order weak derivatives $f^{(k)}$ being Lebesgue integrable and $f(0) = f^{(1)}(0) = \cdots = f^{(k-1)}(0) = 0$. For such classes, the β-polynomial decay condition holds with $\beta = k$.

Given eigendecay conditions of these types, it is possible to compute an upper bound on the critical radius δ_n. In particular, using the fact that the function $\mathcal{R}$ from equation (5.17) is an upper bound on the function $\mathcal{G}_n\big(\mathcal{E}(\delta, 1)\big)$, we can show that for γ-exponentially decaying kernels, we have $\delta_n^2 \precsim \frac{(\log n)^{1/\gamma}}{n}$, whereas for β-polynomial kernels, we have $\delta_n^2 \precsim n^{-\frac{2\beta}{2\beta+1}}$ up to universal constants. Combining with our Theorem 1, we obtain the following result:

Corollary 5 (Bounds based on eigendecay)**.** *Under the conditions of Theorem 1:*

(a) For kernels with γ-exponential eigen-decay, we have

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq c\,\frac{\log^{1/\gamma} n}{n} \quad \textit{at } T \asymp \tfrac{n}{\log^{1/\gamma} n} \textit{ steps.} \tag{5.21a}$$

(b) For kernels with β-polynomial eigen-decay, we have

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq c\, n^{-2\beta/(2\beta+1)} \quad \textit{at } T \asymp n^{2\beta/(2\beta+1)} \textit{ steps.} \tag{5.21b}$$

See Section 5.6.3 for the proof of Corollary 5.

In particular, these bounds hold for LogitBoost and AdaBoost. We note that similar bounds can also be derived with regard to risk in $L^2(\mathbb{P}_n)$ norm as well as the excess risk $\mathcal{L}(f^T) - \mathcal{L}(f^*)$.

To the best of our knowledge, this result is the first to show non-asymptotic and optimal statistical rates for the $\| \cdot \|_n^2$-error when early stopping LogitBoost or AdaBoost with an explicit dependence of the stopping rule on n. Our results also yield similar guarantees for L^2-boosting, as has been established in past work [116]. Note that we can observe a similar trade-off between computational efficiency and statistical accuracy as in the case of kernel least-squares regression [155, 116]: although larger kernel classes (e.g. Sobolev classes) yield higher estimation errors, boosting updates reach the optimum faster than for a smaller kernel class (e.g. Gaussian kernels).

5.4.2 Numerical experiments

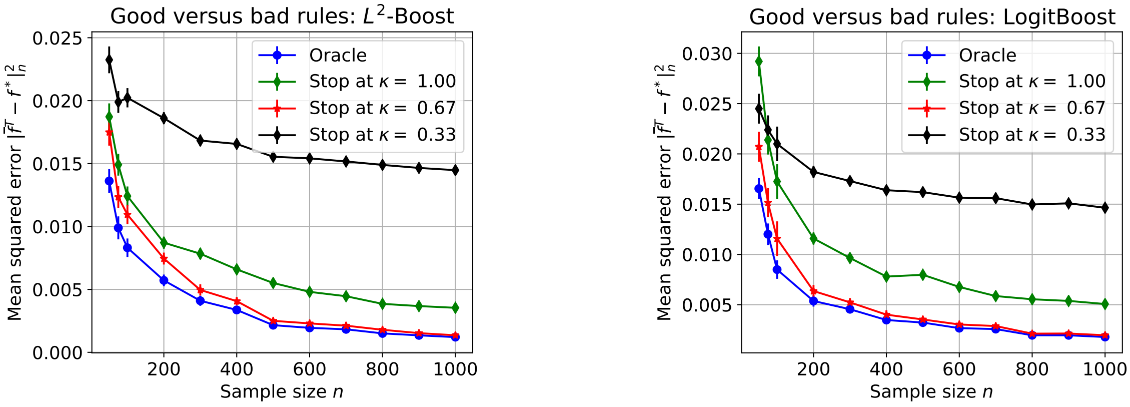
We now describe some numerical experiments that provide illustrative confirmations of our theoretical predictions. While we have applied our methods to various kernel classes, in this section, we present numerical results for the first-order Sobolev kernel as two typical examples for exponential and polynomial eigen-decay kernel classes.

Let us start with the first-order Sobolev space of Lipschitz functions on the unit interval $[0, 1]$. This function space is defined by the kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$, and with the design points $\{x_i\}_{i=1}^n$ set equidistantly over $[0, 1]$. Note that the equidistant design yields β-polynomial decay of the eigenvalues of K with $\beta = 1$ as in the case when x_i are drawn i.i.d. from the uniform measure on $[0, 1]$. Consequently we have that $\delta_n^2 \asymp n^{-2/3}$. Accordingly, our theory predicts that the stopping time $T = (cn)^{2/3}$ should lead to an estimate $\bar{f}^T$ such that $\|\bar{f}^T - f^*\|_n^2 \precsim n^{-2/3}$.

In our experiments for L^2-Boost, we sampled Y_i according to $Y_i = f^*(x_i) + w_i$ with $w_i \sim N(0, 0.5)$, which corresponds to the probability distribution $\mathbb{P}(Y \mid x_i) = N(f^*(x_i); 0.5)$, where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$ is defined on the unit interval $[0, 1]$. By construction, the function f^* belongs to the first-order Sobolev space with $\|f^*\|_{\mathscr{H}} = 1$. For LogitBoost, we sampled Y_i according to $\text{Bin}(p(x_i), 5)$ where $p(x) = \frac{\exp(f^*(x))}{1+\exp(f^*(x))}$. In all cases, we fixed the initialization $f^0 = 0$, and ran the updates (5.7) for L^2-Boost and LogitBoost with the constant step size $\alpha = 0.75$. We compared various stopping rules to the *oracle gold standard* G, meaning the procedure that examines all iterates $\{f^t\}$, and chooses the stopping time $G = \arg\min_{t \geq 1} \|f^t - f^*\|_n^2$ that yields the minimum prediction error. Of course, this procedure is unimplementable in practice, but it serves as a convenient lower bound with which to compare.

Figure 5.2 shows plots of the mean-squared error $\|\bar{f}^T - f^*\|_n^2$ over the sample size n averaged over 40 trials, for the gold standard $T = G$ and stopping rules based on $T = (7n)^\kappa$ for different choices of κ. Error bars correspond to the standard errors computed from our simulations. Panel (a) shows the behavior for L^2-boosting, whereas panel (b) shows the behavior for LogitBoost.

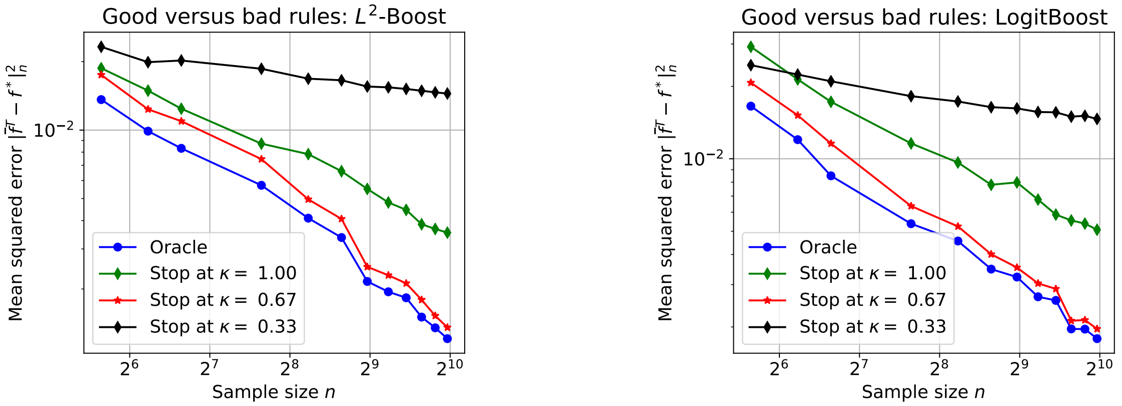
Note that both plots are qualitatively similar and that the theoretically derived stopping rule $T = (7n)^\kappa$ with $\kappa^* = 2/3 = 0.67$, while slightly worse than the Gold standard, tracks its performance closely. We also performed simulations for some "bad" stopping rules, in particular for an exponent κ *not equal* to $\kappa^* = 2/3$, indicated by the green and black curves.



(a)

(b)

Figure 5.2. The mean-squared errors for the stopped iterates $\bar{f}^T$ at the Gold standard, i.e. iterate with the minimum error among all unstopped updates (blue) and at $T = (7n)^\kappa$ (with the theoretically optimal $\kappa = 0.67$ in red, $\kappa = 0.33$ in black and $\kappa = 1$ in green) for (a) L^2-Boost and (b) LogitBoost.



(a)

(b)

Figure 5.3. Logarithmic plots of the mean-squared errors at the Gold standard in blue and at $T = (7n)^\kappa$ (with the theoretically optimal rule for $\kappa = 0.67$ in red, $\kappa = 0.33$ in black and $\kappa = 1$ in green) for (a) L^2-Boost and (b) LogitBoost.

In the log scale plots in Figure 5.3 we can clearly see that for $\kappa \in \{0.33, 1\}$ the performance is indeed much worse, with the difference in slope even suggesting a different scaling of the error with the number of observations n. Recalling our discussion for Figure 5.1, this phenomenon likely occurs due to underfitting and overfitting effects. These qualitative shifts are consistent with our theory.

5.5 Discussion

In this chapter, we have proven non-asymptotic bounds for early stopping of kernel boosting for a relatively broad class of loss functions. These bounds allowed us to propose simple stopping rules which, for the class of regular kernel functions [154], yield minimax optimal rates of estimation. Although the connection between early stopping and regularization has long been studied and explored in the theoretical literature and applications alike, to the best of our knowledge, these results are the first one to establish a general relationship between the statistical optimality of stopped iterates and the localized Gaussian complexity. This connection is important, because this localized Gaussian complexity measure, as well as its Rademacher analogue, are now well-understood to play a central role in controlling the behavior of estimators based on regularization [136, 13, 88, 144].

There are various open questions suggested by our results. The stopping rules in this chapter depend on the eigenvalues of the empirical kernel matrix; for this reason, they are data-dependent and computable given the data. However, in practice, it would be desirable to avoid the cost of computing all the empirical eigenvalues. Can fast approximation techniques for kernels be used to approximately compute our optimal stopping rules? Second, our current theoretical results apply to the averaged estimator $\bar{f}^T$. We strongly suspect that the same results apply to the stopped estimator f^T, but some new ingredients are required to extend our proofs.

5.6 Proof of main results

In this section, we present the proofs of our main results. The technical details are deferred to Appendix C.

In the following, recalling the discussion in Section 5.2.3, we denote the vector of function values of a function $f \in \mathscr{H}$ evaluated at $(x_1, x_2, \ldots, x_n)$ as $\theta_f := f(x_1^n) = (f(x_1), f(x_2), \ldots f(x_n)) \in \mathbb{R}^n$, where we omit the subscript f when it is clear from the context. As mentioned in the main text, updates on the function value vectors $\theta^t \in \mathbb{R}^n$ correspond uniquely to updates of the functions $f^t \in \mathscr{H}$. In the following we repeatedly abuse notation by defining the Hilbert norm and empirical norm on vectors in $\Delta \in \text{range}(K)$ as

$$\|\Delta\|_{\mathscr{H}}^2 = \frac{1}{n}\Delta^T K^{\dagger}\Delta \quad \text{and} \quad \|\Delta\|_n^2 = \frac{1}{n}\|\Delta\|_2^2,$$

where $K^{\dagger}$ is the pseudoinverse of K. We also use $\mathbb{B}_{\mathscr{H}}(\theta, r)$ to denote the ball with respect to the $\|\cdot\|_{\mathscr{H}}$-norm in range(K).

5.6.1 Proof of Theorem 1

The proof of our main theorem is based on a sequence of lemmas, all of which are stated with the assumptions of Theorem 1 in force. The first lemma establishes a bound on the empirical norm $\|\cdot\|_n$ of the error $\Delta^{t+1} := \theta^{t+1} - \theta^*$, provided that its Hilbert norm is suitably controlled.

Lemma 1. *For any stepsize $\alpha \in (0, \frac{1}{M}]$ and any iteration t we have*

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle.$$

See Section C.1 for the proof of this claim.

The second term on the right-hand side of the bound (5.22) involves the difference between the population and empirical gradient operators. Since this difference is being evaluated at the random points Δ^t and Δ^{t+1}, the following lemma establishes a form of uniform control on this term.

Let us define the set

$$\mathbb{S} := \Bigg\{\Delta, \widetilde{\delta} \in \mathbb{R}^n \mid \|\Delta\|_{\mathscr{H}} \geq 1, \text{ and } \Delta,\, \widetilde{\delta} \in \mathbb{B}_{\mathscr{H}}(0, 2C_{\mathscr{H}})\Bigg\}, \tag{5.22}$$

and consider the uniform bound

$$\begin{aligned}\langle \nabla\mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\delta}),\, \Delta\rangle &\leq 2\delta_n\|\Delta\|_n \\ &\quad + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2 \quad \text{for all } \Delta, \widetilde{\delta} \in \mathbb{S}.\end{aligned} \tag{5.23}$$

Lemma 2. *Let $\mathcal{E}$ be the event that bound (5.23) holds. There are universal constants (c_1, c_2) such that $\mathbb{P}[\mathcal{E}] \geq 1 - c_1 \exp(-c_2 \frac{m^2 n \delta_n^2}{\sigma^2})$.*

See Section C.2 for the proof of Lemma 2.

Note that Lemma 1 applies only to error iterates with a bounded Hilbert norm. Our last lemma provides this control for some number of iterations:

Lemma 3. *There are constants (C_1, C_2) independent of n such that for any step size $\alpha \in \big(0, \min\{M, \frac{1}{M}\}\big]$, we have*

$$\|\Delta^t\|_{\mathscr{H}} \leq C_{\mathscr{H}} \qquad \textit{for all iterations } t \leq \tfrac{m}{8M\delta_n^2} \tag{5.24}$$

with probability at least $1 - C_1 \exp(-C_2 n \delta_n^2)$, where $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

See Section C.3 for the proof of this lemma which also uses Lemma 2.

Taking these lemmas as given, we now complete the proof of the theorem. We first condition on the event $\mathcal{E}$ from Lemma 2, so that we may apply the bound (5.23). We then fix some iterate t such that $t < \frac{m}{8M\delta_n^2} - 1$, and condition on the event that the bound (5.24) in Lemma 3 holds, so that we are guaranteed that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq C_{\mathscr{H}}$. We then split the analysis into two cases:

Case 1 First, suppose that $\|\Delta^{t+1}\|_n \leq \delta_n C_{\mathscr{H}}$. In this case, inequality (5.13b) holds directly.

Case 2 Otherwise, we may assume that $\|\Delta^{t+1}\|_n > \delta_n \|\Delta^{t+1}\|_{\mathscr{H}}$. Applying the bound (5.23) with the choice $(\widetilde{\delta}, \Delta) = (\Delta^t, \Delta^{t+1})$ yields

$$\langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1} \rangle \leq 4\delta_n \|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2. \tag{5.25}$$

Substituting inequality (5.25) back into equation (5.22) yields

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + 4\delta_n\|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2.$$

Re-arranging terms yields the bound

$$\gamma m \|\Delta^{t+1}\|_n^2 \leq D^t + 4\delta_n \|\Delta^{t+1}\|_n, \tag{5.26}$$

where we have introduced the shorthand notation $D^t := \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\}$, as well as $\gamma = \frac{1}{2} - \frac{1}{c_3}$

Equation (5.26) defines a quadratic inequality with respect to $\|\Delta^{t+1}\|_n$; solving it and making use of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ yields the bound

$$\|\Delta^{t+1}\|_n^2 \leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{2D^t}{\gamma m}, \tag{5.27}$$

for some universal constant c. By telescoping inequality (5.27), we find that

$$\frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2 \leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{T}\sum_{t=1}^{T}\frac{2D^t}{\gamma m} \tag{5.28}$$

$$\leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{\alpha \gamma m T}[\|\Delta^0\|_{\mathscr{H}}^2 - \|\Delta^T\|_{\mathscr{H}}^2]. \tag{5.29}$$

By Jensen's inequality, we have

$$\|\bar{f}^T - f^*\|_n^2 = \|\frac{1}{T}\sum_{t=1}^{T}\Delta^t\|_n^2 \;\leq\; \frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2,$$

so that inequality (5.13b) follows from the bound (5.28).

On the other hand, by the smoothness assumption, we have

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \leq \frac{M}{2}\|\bar{f}^T - f^*\|_n^2,$$

from which inequality (5.13a) follows.

5.6.2 Proof of Corollary 4

Similar to the proof of Theorem 1 in Yang et al. [154], a generalization can be shown using a standard argument of Fanos inequality. By definition of the transformed parameter $\theta = DU\alpha$ with $K = U^T DU$, we have for any estimator $\widehat{f} = \sqrt{n}U^T\theta$ that $\|\widehat{f} - f^*\|_n^2 = \|\theta - \theta^*\|_2^2$. Therefore our goal is to lower bound the Euclidean error $\|\theta - \theta^*\|_2$ of any estimator of θ^*. Borrowing Lemma 4 in Yang et al. [154], there exists $\delta/2$-packing of the set $B = \{\theta \in \mathbb{R}^n \mid \|D^{-1/2}\theta\|_2 \leq 1\}$ of cardinality $M = e^{d_n/64}$ with $d_n := \arg\min_{j=1,\ldots,n}\{\mu_j \leq \delta_n^2\}$. This is done through packing the following subset of B

$$\mathcal{E}(\delta) := \Big\{\theta \in \mathbb{R}^n \ \mid \ \sum_{j=1}^n \frac{\theta_j^2}{\min\{\delta^2, \mu_j\}} \leq 1\Big\}.$$

Let us denote the packing set by $\{\theta^1, \ldots, \theta^M\}$. Since $\theta \in \mathcal{E}(\delta)$, by simple calculation, we have $\|\theta^i\|_2 \leq \delta$.

By considering the random ensemble of regression problem in which we first draw at index Z at random from the index set $[M]$ and then condition on $Z = z$, we observe n i.i.d samples $y_1^n := \{y_1, \ldots, y_n\}$ from $\mathbb{P}_{\theta^z}$, Fano's inequality implies that

$$\mathbb{P}(\|\widehat{\theta} - \theta^*\|_2 \geq \frac{\delta^2}{4}) \geq 1 - \frac{I(y_1^n; Z) + \log 2}{\log M}.$$

where $I(y_1^n; Z)$ is the mutual information between the samples Y and the random index Z.

So it is only left for us to control the mutual information $I(y_1^n; Z)$. Using the mixture representation, $\bar{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^M \mathbb{P}_{\theta^i}$ and the convexity of the KullbackLeibler divergence, we have

$$I(y_1^n; Z) = \frac{1}{M}\sum_{j=1}^M \|\mathbb{P}_{\theta^j},\ \bar{\mathbb{P}}\|_{\mathrm{KL}} \leq \frac{1}{M^2}\sum_{i,j} \|\mathbb{P}_{\theta^i},\ \mathbb{P}_{\theta^j}\|_{\mathrm{KL}}.$$

We now claim that

$$\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} \leq \frac{nL\|\theta - \theta'\|_2^2}{s(\sigma)}. \tag{5.30}$$

Since each $\|\theta^i\|_2 \leq \delta$, triangle inequality yields $\|\theta_i - \theta_j\|_2 \leq 2\delta$ for all $i \neq j$. It is therefore guaranteed that

$$I(y_1^n; Z) \leq \frac{4nL\delta^2}{s(\sigma)}.$$

Therefore, similar to Yang et al. [154], following by the fact that the kernel is regular and hence $s(\sigma)d_n \geq cn\delta_n^2$, any estimator $\widehat{f}$ has prediction error lower bounded as

$$\sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c_l\delta_n^2.$$

Corollary 4 thus follows using the upper bound in Theorem 1.

Proof of inequality (5.30) Direct calculations of the KL-divergence yield

$$\begin{aligned}\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} =& \int \log(\frac{\mathbb{P}_\theta(y)}{\mathbb{P}_{\theta'}(y)})\mathbb{P}_\theta(y)dy\\ =&\frac{1}{s(\sigma)}\sum_{i=1}^n \Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle)\\ &+ \frac{\sqrt{n}}{s(\sigma)}\int\sum_{i=1}^n \big[y_i\langle u_i,\, \theta-\theta'\rangle\big]\mathbb{P}_\theta dy. \end{aligned} \tag{5.31}$$

To further control the right hand side of expression (5.31), we concentrate on expressing $\int\sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy$ differently. Leibniz's rule allow us to inter-change the order of integral and derivative, so that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{d}{d\theta}\int P_\theta dy = 0. \tag{5.32}$$

Observe that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{\sqrt{n}}{s(\sigma)}\int P_\theta \cdot \sum_{i=1}^n u_i\big(y_i - \Phi'(\sqrt{n}\langle u_i,\, \theta'\rangle)\big)dy$$

so that equality (5.32) yields

$$\int \sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy = \sum_{i=1}^n u_i \Phi'(\sqrt{n}\langle u_i,\, \theta\rangle).$$

Combining the above inequality with expression (5.31), the KL divergence between two generalized linear models $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$ can thus be written as

$$\begin{aligned}\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} = \frac{1}{s(\sigma)}\sum_{i=1}^n &\Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle)\\ &-\sqrt{n}\langle u_i,\, \theta'-\theta\rangle\Phi'(\sqrt{n}\langle u_i,\, \theta\rangle). \end{aligned} \tag{5.33}$$

Together with the fact that

$$\begin{aligned}|\Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle) - \sqrt{n}\langle u_i,\, \theta'-\theta\rangle\Phi'(\sqrt{n}\langle u_i,\, \theta\rangle)|&\\ \leq nL\|\theta-\theta'\|_2^2.&\end{aligned}$$

which follows by assumption on Φ having a uniformly bounded second derivative. Putting the above inequality with inequality (5.33) establishes our claim (5.30).

5.6.3 Proof of Corollary 5

The general statement follows directly from Theorem 1. In order to invoke Theorem 1 for the particular cases of LogitBoost and AdaBoost, we need to verify the conditions, i.e. that the m-M-condition and ϕ'-boundedness conditions hold for the respective loss function over the ball $\mathbb{B}_{\mathscr{H}}(\theta^*, 2C_{\mathscr{H}})$. The following lemma provides such a guarantee:

Lemma 4. *With $D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$, the logistic regression cost function satisfies the m-M-condition with parameters*

$$m = \frac{1}{e^{-D} + e^{D} + 2}, \quad M = \frac{1}{4}, \quad \textit{and} \quad B = 1.$$

The AdaBoost cost function satisfies the m-M-condition with parameters

$$m = \mathbb{E}^{-D}, \quad M = \mathbb{E}^{D}, \quad \textit{and} \quad B = \mathbb{E}^{D}.$$

See Section C.4 for the proof of Lemma 4.

γ-exponential decay If the kernel eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2 j^\gamma)$, where c_1, c_2 are universal constants, the function $\mathcal{R}$ from equation (5.17) can be upper bounded as

$$\mathcal{R}(\delta) = \sqrt{\frac{2}{n}}\sqrt{\sum_{i=1}^{n} \min\{\delta^2, \mu_j\}} \leq \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \sum_{j=k+1}^{n} c_1 e^{-c_2 j^2}},$$

where k is the smallest integer such that $c_1 \exp(-c_2 k^\gamma) < \delta^2$. Since the localized Gaussian width $\mathcal{G}_n\big(\mathcal{E}_n(\delta, 1)\big)$ can be sandwiched above and below by multiples of $\mathcal{R}(\delta)$, some algebra shows that the critical radius scales as $\delta_n^2 \asymp \frac{n}{\log(n)^{1/\gamma}\sigma^2}$.

Consequently, if we take $T \asymp \frac{\log(n)^{1/\gamma}\sigma^2}{n}$ steps, then Theorem 1 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \lesssim \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right) \frac{\log^{1/\gamma} n}{n} \sigma^2,$$

with probability $1 - c_1 \text{exp}(-c_2 m^2 \log^{1/\gamma} n)$.

β-polynomial decay Now suppose that the kernel eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 j^{-2\beta}$ for some $\beta > 1/2$ and constant c_1. In this case, a direct calculation yields the bound

$$\mathcal{R}(\delta) \leq \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + c_2 \sum_{j=k+1}^{n} j^{-2}},$$

where k is the smallest integer such that $c_2 k^{-2} < \delta^2$. Combined with upper bound $c_2 \sum_{j=k+1}^{n} j^{-2} \leq c_2 \int_{k+1} j^{-2} \leq k\delta^2$, we find that the critical radius scales as $\delta_n^2 \asymp n^{-2\beta/(1+2\beta)}$.

Consequently, if we take $T \asymp n^{-2\beta/(1+2\beta)}$ many steps, then Theorem 1 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \;\leq\; \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right)\left(\frac{\sigma^2}{n}\right)^{2\beta/(2\beta+1)},$$

with probability at least $1 - c_1 \exp(-c_2 m^2 (\frac{n}{\sigma^2})^{1/(2\beta+1)})$.

Chapter 6

Future directions

In this thesis, a range of different problems are described ranging from hypothesis testing, non-parametric estimation to optimization algorithms. A common theme underlying much of this work is the underlying geometric structure of the problem. For example, Chapter 3 on cone testing showed that in addition to the Gaussian complexity, other geometric quantities play a role in determining the difficulty of testing; the convex set estimation project showed that polytopes with a controlled number of vertices are significantly easier to estimate.

It is interesting to see whether these results can provide some insights to other questions in statistics and optimization that have a geometric flavor, such as manifold structure. In the area of covariance estimation, Wiesel [151] noted that if covariance matrices are regarded as elements of a Riemannian manifold, then maximum likelihood estimation of these covariance matrices is a convex problem under the notion of geodesic convexity. This perspective opens up a variety of new questions and methods for matrix estimation. In addition, manifold learning is an area of active research in machine learning, applied mathematics, and statistics. In recent years, researchers have established a number of theoretical guarantees for such methods (e.g., [64, 85]). However, it remains unclear how to optimally extract the features of a manifold that are sufficient for subsequent clustering and/or classification tasks under minimal assumptions. It is my intention to tackle some of these interesting and fundamental problems in my future career, using the skills that I have developed thus far.

In addition, statistical inference has long been one of the most important topics in statistics. Compared to its estimation analogue, there are many interesting problems still remain to be open. One general question that interests me a lot is how to do inference on structured data. As one concrete instance, there is an evolving line of work on testing problems involving complicated structures such as communities in network data and trees (e.g., [1, 5]). Such structures arise frequently in applications such as genetics, neuroscience, and the social sciences, and the corresponding theory for testing methods is relatively undeveloped. Moreover, I am also interested in the problem of detecting multi-scaled signals and change-points from plain background. This problem is one of the key problems in applied mathematics and signal processing, and although some relevant results are known in different contexts (e.g., [46, 54, 6, 146]), several issues are not yet resolved, including fundamental limits for high-dimensional problems, behaviors of different non-parametric function classes, and efficient algorithms.

Another direction that I am interested in understanding is the role of regularization in fitting complex models. A phenomenon that has been observed over this decade, is the great generalization performance of deep neural nets despite the fact that it is highly overparametrized. To shed lights on this mystery, recent couple of years have witnessed many brand-new ideas from statistics and optimization community to reach a better understanding of non-convex problems. A line of work focus on studying the landscape of particular classes of non-convex objective functions such any stationary point of the non-convex objective is close to global optima, so it suffices to find a locally optimal solution (see e.g. [99, 78, 61, 62, 60]) Another line of work concentrated on analyzing the local convergence for various algorithms and problems (e.g. [41, 98, 79] and showed that given a good initialization, many simple local search algorithms including gradient descent succeed. However, the work listed so far are of a case-to-case flavor mostly, namely each analysis is highly dependent on the particular structure of individual problem. One interesting open problem is that can we obtain a more general way of analyzing these optimization landscapes and understand the role of generalization better.

Appendix A

Proofs for Chapter 3

This chapter is organized as follows. In Section A.1, we first explain the intuition behind the example in Section 3.3.2.3 where the GLRT is shown to be sub-optimal, and construct a series of other cases where this sub-optimality is observed. We then provide the proofs of Propositions 3.3.1 and 3.3.2 in Sections A.3.1 and A.3.2, respectively. It follows by some background on distance metrics and their properties in Section A.2. The proofs of Theorem 3.3.1 (a) and (b) are completed in Section A.4 and A.5 respectively. The proofs of the lemmas for Theorem 3.3.2 are collected in Section A.6. Finally, the technical lemmas which were crucially used in the proofs of the Proposition 3.3.2 and the monotone cone example are proved in Section A.7.

A.1 The GLRT sub-optimality

In this appendix, we first try to understand why the GLRT is sub-optimal for the Cartesian product cone $K_\times = \mathrm{Circ}_{d-1}(\alpha) \times \mathbb{R}$, and use this intuition to construct a more general class of problems for which a similar sub-optimality is witnessed.

A.1.1 Why is the GLRT sub-optimal?

Let us consider tests with null $C_1 = \{0\}$ and a general product alternative of the form $C_2 = K_\times = K \times \mathbb{R}$, where $K \subseteq \mathbb{R}^{d-1}$ is a base cone. Note that $K = \mathrm{Circ}_{d-1}$ in our previous example.

Now recall the decomposition (3.22) of the statistic T that underlies the GLRT. By the product nature of the cone, we have

$$T(y) = \|\Pi_{K_\times} y\|_2 = \|(\Pi_K(y_{-d}),\ y_d)\|_2 = \sqrt{\|\Pi_K(y_{-d})\|_2^2 + \|y_d\|_2^2},$$

where $y_{-d} := (y_1, \ldots, y_{d-1}) \in \mathbb{R}^{d-1}$ is formed from the first $d-1$ coordinates of y. Suppose that we are interested in testing between the zero vector and a vector $\theta^* = (0, \ldots, 0, \theta_d^*)$, non-zero only in the last coordinate, which belongs to the alternative. With this particular choice, under the null distribution, we have $y = \sigma g$ whereas under the alternative, we have $y = \theta^* + \sigma g$. Letting $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectations under these two Gaussian distributions, the performance of the GLRT in this direction is governed by the difference

$$\frac{1}{\sigma}\big\{\mathbb{E}_1[T(y)] - \mathbb{E}_0[T(y)]\big\} = \mathbb{E}_1\sqrt{\|\Pi_K(g_{-d})\|_2^2 + \|\frac{\theta_d^*}{\sigma} + g_d\|_2^2} \\ -\mathbb{E}_0\sqrt{\|\Pi_K(g_{-d})\|_2^2 + \|g_d\|_2^2}.$$

Note both terms in this difference involve a $(d-1)$-dimensional "pure noise" component—namely, the quantity $\|\Pi_K(g_{-d})\|_2^2$ defined by the sub-vector $g_{-d} := (g_1, \ldots, g_{d-1})$—with the only signal lying the last coordinate. For many choices of cone K, the pure noise component acts as a strong mask for the signal component, so that the GLRT is poor at detecting differences in the direction θ^*. Since the vector θ^* belongs to the alternative, this leads to sub-optimality in its overall behavior. Guided by this idea, we can construct a series of other cases where the GLRT is sub-optimal. See Appendix A.1.2 for details.

A.1.2 More examples on the GLRT sub-optimality

Now let us construct a larger class of product cones for which the GLRT is sub-optimal. For a given subset $S \subseteq \{1, \ldots, d\}$, define the subvectors $\theta_S = (\theta_i, i \in S)$ and $\theta_{S^c} = (\theta_j, j \in S^c\}$, where S^c denotes the complement of S. For an integer $\ell \geq 1$, consider any cone $K_\ell \subset \mathbb{R}^d$ with the following two properties:

- its Gaussian width scales as $\mathbb{E}\mathbb{W}(K_\ell \cap \mathbb{B}(1)) \asymp \sqrt{d}$, and
- for some fixed subset $\{1, 2 \ldots, d\}$ of cardinality ℓ, there is a scalar $\gamma > 0$ such that

$$\|\theta_S\|_2 \geq \gamma \|\theta_{S^c}\|_2 \quad \text{for all } \theta \in K_\ell.$$

As one concrete example, it is easy to check that the circular cone is a special example with $\ell = 1$ and $\gamma = 1/\tan(\alpha)$. The following result applies to the GLRT when applied to testing the null $C_1 = \{0\}$ versus the alternative $C_2 = K_\times^s = K \times \mathbb{R}$.

Proposition A.1.1. *For the previously described cone testing problem, the GLRT testing radius is sandwiched as*

$$\epsilon_{GLR}^2 \asymp \sqrt{d}\sigma^2,$$

whereas a truncation test can succeed at radius $\epsilon^2 \asymp \sqrt{\ell}\sigma^2$.

Proof. The claimed scaling of the GLRT testing radius follows as a corollary of Theorem 3.3.1 after a direct evaluation of $\delta_{\mathrm{LR}}^2(C_1, C_2)$. In order to do so, we begin by observing that

$$\inf_{\eta \in C_2 \times S^{-1}} \langle \eta, \, \mathbb{E}\Pi_{C_2} g\rangle \leq \langle e_d, \, \mathbb{E}\Pi_{c_2} g\rangle \; = \; 0, \quad \text{and}$$

$$\mathbb{E}\mathbb{W}(C_2 \cap \mathbb{B}(1)) = \mathbb{E}\|\Pi_{C_2} g\|_2 \asymp \sqrt{d}$$

which implies that $\delta^2_{\mathrm{LR}}(C_1, C_2) \asymp \sqrt{d}$, and hence implies the sandwich claim on the GLRT via Theorem 3.3.1.

On the other hand, for some pre-selected $\beta > 0$, consider the truncation test

$$\varphi(y) := \mathbb{I}\big[\|y_S\|_2 \geq \beta\big],$$

This test can be viewed as a GLRT for testing the zero null against the alternative $\mathbb{R}^\ell$, and hence it will succeed with separation $\epsilon^2 \asymp \sigma^2\sqrt{\ell}$. Putting these pieces together, we conclude that the GLRT is sub-optimal whenever ℓ is of lower order than d.

□

A.2 Distances and their properties

Here we collect some background on distances between probability measures that are useful in analyzing testing error. Suppose $\mathbb{P}_1$ and $\mathbb{P}_2$ are two probability measures on Euclidean space $(\mathbb{R}^d, \mathcal{B})$ equipped with Lebesgue measure. For the purpose of this paper, we assume $\mathbb{P}_1 \ll \mathbb{P}_2$. The *total variation* (TV) distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as

$$\|\mathbb{P}_1 - \mathbb{P}_2\|_{\mathrm{TV}} := \sup_{B \in \mathcal{B}} |\mathbb{P}_1(B) - \mathbb{P}_2(B)| = \frac{1}{2}\int |d\mathbb{P}_1 - d\mathbb{P}_2|. \tag{A.1a}$$

A closely related measure of distance is the χ^2 *distance* given by

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) := \int (\frac{d\mathbb{P}_1}{d\mathbb{P}_2} - 1)^2 d\mathbb{P}_2. \tag{A.1b}$$

For future reference, we note that the TV distance and χ^2 distance are related via the inequality

$$\|\mathbb{P}_1 - \mathbb{P}_2\|_{\mathrm{TV}} \leq \frac{1}{2}\sqrt{\chi^2(\mathbb{P}_1, \mathbb{P}_2)}. \tag{A.1c}$$

A.3 Proofs for Proposition 3.3.1 and 3.3.2

In this section, we complete the proofs of Propositions 3.3.1 and 3.3.2 in Sections A.3.1 and A.3.2, respectively.

A.3.1 Proof of Proposition 3.3.1

As in the proof of Theorem 3.3.1 and Theorem 3.3.2, we can assume without loss of generality that $\sigma = 1$ since K_+ is invariant under rescaling by positive numbers. We split our proof into two cases, depending on whether or not the dimension d is less than 81.

Case 1: First suppose that $d < 81$. If the separation is upper bounded as $\epsilon^2 \leq \kappa_\rho \sqrt{d}$, then setting $\kappa_\rho = 1/18$ yields

$$\epsilon^2 \leq \kappa_\rho \sqrt{d} < 1/2.$$

Similar to our proof for Theorem 3.3.1(b) Case 1, if $\epsilon^2 < 1/2$, every test yields testing error no smaller than $1/2$. It is seen by considering a simple verses simple testing problem (3.58a). So our lower bound directly holds for the case when $d < 81$ satisfies.

Case 2: Let us consider the case when dimension $d \geq 81$. The idea is to make use of our Lemma 3.5.3 to show that the testing error is at least ρ whenever $\epsilon^2 \leq \kappa_\rho \sqrt{d}$. In order to apply Lemma 3.5.3, the key is to construct a probability measure $\mathbb{Q}$ supported on set $K \cap B^c(1)$ such that for i.i.d. pair η, η' drawn from $\mathbb{Q}$, quantity $\mathbb{E}e^{\lambda\langle\eta,\,\eta'\rangle}$ can be well controlled. We claim that there exists such a probability measure $\mathbb{Q}$ that

$$\mathbb{E}_{\eta,\eta'} e^{\lambda\langle\eta,\,\eta'\rangle} \leq \exp\left(\exp\left(\frac{2+\lambda}{\sqrt{d}-1}\right) - \left(1 - \frac{1}{\sqrt{d}}\right)^2\right) \qquad \text{where } \lambda := \epsilon^2. \tag{A.2}$$

Taking inequality (A.2) as given for now, letting $\kappa_\rho = 1/8$, we have $\lambda = \epsilon^2 \leq \sqrt{d}/8$. So the right hand side in expression (A.2) can be further upper bounded as

$$\begin{aligned}\exp\left(\exp\left(\frac{2}{\sqrt{d}-1} + \frac{\sqrt{d}}{\sqrt{d}-1}\frac{\lambda}{\sqrt{d}}\right) - \left(1-\frac{1}{\sqrt{d}}\right)^2\right) &\leq \exp\left(\exp\left(\frac{1}{4} + \frac{9}{8}\cdot\frac{1}{8}\right) - \left(1 - \frac{1}{9}\right)^2\right) \\ &< 2,\end{aligned}$$

where we use the fact that $d \geq 81$. As a consequence of Lemma 3.5.3, the testing error of every test satisfies

$$\inf_{\psi} \mathcal{E}(\psi; \{0\}, K_+, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle\eta,\,\eta'\rangle) - 1} > \frac{1}{2} \geq \rho.$$

Putting these two cases together, our lower bound holds for any dimension thus we complete the proof of Proposition 3.3.1.

So it only remains to construct a probability measure $\mathbb{Q}$ such that the inequality (A.2) holds. We begin by introducing some helpful notation. For an integer s to be specified, consider a collection of vectors $\mathcal{S}$ containing all d-dimensional vectors with exactly s non-zero entries and each non-zero entry equals to $1/\sqrt{s}$. Note that there are in total $M := \binom{d}{s}$ vectors of this type. Letting $\mathbb{Q}$ be the uniform distribution over this set of vectors namely

$$\mathbb{Q}(\{\eta\}) := \frac{1}{M}, \qquad \eta \in \mathcal{S}. \tag{A.3}$$

Then we can write the expectation as

$$\mathbb{E}e^{\lambda\langle\eta,\,\eta'\rangle} = \frac{1}{M^2}\sum_{\eta,\eta'\in\mathcal{S}} e^{\lambda\langle\eta,\,\eta'\rangle}.$$

Note that the inner product $\langle\eta,\,\eta'\rangle$ takes values i/s, for integer $i \in \{0, 1, \dots, s\}$ and given every vector η and integer $i \in \{0, 1, \dots, s\}$, the number of η' such that $\langle\eta,\,\eta'\rangle = i/s$ equals to $\binom{s}{i}\binom{d-s}{s-i}$. Consequently, we obtain

$$\mathbb{E}e^{\lambda\langle\eta,\,\eta'\rangle} = \binom{d}{s}^{-1}\sum_{i=0}^{s}\binom{s}{i}\binom{d-s}{s-i}e^{\lambda i/s} = \sum_{i=0}^{s}\frac{A_i z^i}{i!}, \tag{A.4}$$

where

$$z := e^{\lambda/s} \text{ and } A_i := \frac{(s!(d-s)!)^2}{((s-i)!)^2 d!(d-2s+i)!}.$$

Let us set integer $s := \lfloor\sqrt{d}\rfloor$. We claim quantity A_i satisfies the following bound

$$A_i \le \exp\Big(-(1-\frac{1}{\sqrt{d}})^2 + \frac{2i}{\sqrt{d}-1}\Big) \qquad \text{for all } i \in \{0, 1, \dots, s\}. \tag{A.5}$$

Taking expression (A.5) as given for now and plugging into inequality (A.4), we have

$$\begin{aligned}
\mathbb{E}e^{\lambda\langle\eta,\,\eta'\rangle} &\le \exp\Big(-(1-\frac{1}{\sqrt{d}})^2\Big)\sum_{i=0}^{s}\frac{(z\exp(\frac{2}{\sqrt{d}-1}))^i}{i!} \\
&\overset{(a)}{\le} \exp\Big(-(1-\frac{1}{\sqrt{d}})^2\Big)\exp\Big(z\exp(\frac{2}{\sqrt{d}-1})\Big) \\
&\overset{(b)}{\le} \exp\left(-\left(1-\frac{1}{\sqrt{d}}\right)^2 + \exp\left(\frac{2+\lambda}{\sqrt{d}-1}\right)\right),
\end{aligned}$$

where step (a) follows from the standard power series expansion $e^x = \sum_{i=0}^{\infty}\frac{x^i}{i!}$ and step (b) follows by $z = e^{\lambda/s}$ and $s = \lfloor\sqrt{d}\rfloor > \sqrt{d}-1$. Therefore it verifies inequality (A.2) and complete our argument.

It is only left for us to check inequality (A.5) for A_i. Using the fact that $1 - x \le e^{-x}$, it is guaranteed that

$$A_0 = \frac{((d-s)!)^2}{d!(d-2s)!} = (1-\frac{s}{d})(1-\frac{s}{d-1})\cdots(1-\frac{s}{d-s+1}) \le \exp(-s\sum_{i=1}^{s}\frac{1}{d-s+i}). \tag{A.6a}$$

Recall that integer $s = \lfloor\sqrt{d}\rfloor$, then we can bound the sum in expression (A.6a) as

$$s\sum_{i=1}^{s}\frac{1}{d-s+i} \geq s\sum_{i=1}^{s}\frac{1}{d} = \frac{s^2}{d} \geq (1-\frac{1}{\sqrt{d}})^2,$$

which, when combined with inequality (A.6a), implies that $A_0 \leq \exp(-(1-\frac{1}{\sqrt{d}})^2)$.

Moreover, direct calculations yield

$$\frac{A_i}{A_{i-1}} = \frac{(s-i+1)^2}{d-2s+i}, \qquad 1 \leq i \leq s. \tag{A.6b}$$

This ratio is decreasing with index i as $1 \leq i \leq s$, thus is upper bounded by A_1/A_0, which implies that

$$\frac{A_i}{A_{i-1}} \leq \frac{d}{d-2\sqrt{d}+1} = (1+\frac{1}{\sqrt{d}-1})^2 \leq \exp(\frac{2}{\sqrt{d}-1}),$$

where the last inequality follows from $1+x \leq e^x$. Putting pieces together validates bound (A.5) thus finishing the proof of Proposition 3.3.1.

A.3.2 Proof of Proposition 3.3.2

As in the proof of Theorem 3.3.1 and Theorem 3.3.2, we can assume without loss of generality that $\sigma = 1$ since L and M are both invariant under rescaling by positive numbers.

We split our proof into two cases, depending on whether or not $\sqrt{\log(ed)} < 14$.

Case 1: First suppose $\sqrt{\log(ed)} < 14$, so that the choice $\kappa_\rho = 1/28$ yields the upper bound

$$\epsilon^2 \leq \kappa_\rho\sqrt{\log(ed)} < 1/2.$$

Similar to our proof of the lower bound in Theorem 3.3.1, by reducing to a simple testing problem (3.58a), any test yields testing error no smaller than 1/2 if $\epsilon^2 < 1/2$. Thus, we conclude that the stated lower bound holds when $\sqrt{\log(ed)} < 14$.

Case 2: Otherwise, we may assume that $\sqrt{\log(ed)} \geq 14$. In this case, we exploit Lemma 3.5.3 in order to show that the testing error is at least ρ whenever $\epsilon^2 \leq \kappa_\rho\sqrt{\log(ed)}$. Doing so requires constructing a probability measure $\mathbb{Q}_L$ supported on $M \cap L^\perp \cap B^c(1)$ such that the expectation $\mathbb{E}e^{\epsilon^2\langle\eta,\eta'\rangle}$ can be well controlled, where (η,η') are drawn i.i.d according to $\mathbb{Q}_L$. Note that L can be either $\{0\}$ or $\text{span}(\mathbf{1})$.

Before doing that, let us first introduce some notation. Let $\delta := 9$ and $r := 1/3$ (note that $\delta = r^{-2}$). Let

$$m := \max\left\{n \;\middle|\; \sum_{i=1}^{n}\lfloor\frac{\delta-1}{\delta^i}(d+\log_\delta d+3)\rfloor \;<\; d\right\}. \tag{A.7}$$

We claim that the integer m defined above satisfies:

$$\lceil \frac{3}{4}\log_\delta(d)\rceil + 1 \le m \le \lceil \log_\delta d\rceil, \tag{A.8}$$

where $\lceil x\rceil$ denotes the smallest integer that is greater than or equal to x. To see this, notice that for $t = \lceil \frac{3}{4}\log_\delta(d)\rceil + 1$, we have

$$\begin{aligned}\sum_{i=1}^{t}\lfloor \frac{\delta-1}{\delta^i}(d+\log_\delta d+3)\rfloor \le \sum_{i=1}^{t}\frac{\delta-1}{\delta^i}(d+\log_\delta d+3) = (1-\frac{1}{\delta^t})(d+\alpha)\\ \overset{(i)}{\le} d+\alpha-\frac{d+\alpha}{\delta^2 d^{3/4}} \overset{(ii)}{<} d,\end{aligned}$$

where we denote $\alpha := \log_\delta d + 3$. The step (i) follows by definition that $t = \lceil \frac{3}{4}\log_\delta(d)\rceil + 1$ while step (ii) holds because as $\sqrt{\log(ed)} \ge 14$, we have $\alpha = \log_\delta d + 3 < d^{1/4}/\delta^2$. On the other hand, for $t = \lceil \log_\delta d\rceil$, we have

$$\begin{aligned}\sum_{i=1}^{t}\lfloor \frac{\delta-1}{\delta^i}(d+\log_\delta d+3)\rfloor &\ge \sum_{i=1}^{t}\frac{\delta-1}{\delta^i}(d+\alpha) - t\\ &= (1-\frac{1}{\delta^t})(d+\alpha) - t\\ &> d+\alpha-\frac{d+\alpha}{d} - (\log_\delta d + 1),\end{aligned}$$

where the last step uses fact $t = \lceil \log_\delta d\rceil$. Since when $\sqrt{\log(ed)} \ge 14$, we have $\alpha = \log_\delta d + 3 < d$, therefore $(d+\alpha)/d + \log_\delta d + 1 \le 2 + \log_\delta d + 1 = \alpha$, which guarantees that

$$\sum_{i=1}^{t}\lfloor \frac{\delta-1}{\delta^i}(d+\log_\delta d+3)\rfloor > d.$$

We thereby established inequality (A.8).

We now claim that there exists a probability measure $\mathbb{Q}_L$ supported on $M \cap L^\perp \cap B^c(1)$ such that

$$\mathbb{E}_{\eta,\eta'\sim\mathbb{Q}_L} e^{\lambda\langle\eta,\eta'\rangle} \le \exp\left(\exp\left(\frac{9\lambda/4+2}{\sqrt{m}-1}\right) - \left(1-\frac{1}{\sqrt{m}}\right)^2 + \frac{27\lambda}{32(\sqrt{m}-1)}\right), \quad \text{where } \lambda := \epsilon^2. \tag{A.9}$$

Recall that we showed in inequality (A.8) that $m \ge \lceil \frac{3}{4}\log_\delta(d)\rceil + 1$. Setting $\kappa_\rho = 1/62$ implies that whenever $\epsilon^2 \le \kappa_\rho\sqrt{\log(ed)}$, we have

$$\epsilon^2 \le \frac{1}{62}\sqrt{\log(ed)} = \frac{1}{62}\sqrt{1+\frac{4}{3}\log\delta\cdot\frac{3}{4}\log_\delta d} \le \frac{1}{62}\sqrt{\frac{4}{3}\log\delta\left(1+\frac{3}{4}\log_\delta d\right)} \le \frac{1}{36}\sqrt{m}. \tag{A.10}$$

So the right hand side in expression (A.9) can be made less than 2 by

$$\begin{aligned}
&\exp\left(\frac{9\lambda/4+2}{\sqrt{m}-1}\right) - \left(1-\frac{1}{\sqrt{m}}\right)^2 + \frac{27\lambda}{32(\sqrt{m}-1)} \\
\leq &\exp\left(\frac{9\lambda}{4\sqrt{m}}\frac{\sqrt{m}}{\sqrt{m}-1} + \frac{2}{7}\right) - \left(1-\frac{1}{8}\right)^2 + \frac{27\lambda}{32\sqrt{m}}\frac{\sqrt{m}}{\sqrt{m}-1} \\
\leq &\exp\left(\frac{9}{4\cdot 36}\frac{8}{7} + \frac{2}{7}\right) - \left(1-\frac{1}{8}\right)^2 + \frac{27}{32\cdot 36}\frac{8}{7} < \log 2,
\end{aligned}$$

where we use the fact that $\sqrt{m} \geq \sqrt{1+\frac{3}{4}\log_\delta d} \geq 8$. Lemma 3.5.3 thus guarantees the testing error to be no less than

$$\inf_\psi \mathcal{E}(\psi; L, M, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'}\, \exp(\epsilon^2\langle\eta,\, \eta'\rangle) - 1} > \frac{1}{2} \geq \rho,$$

which leads to our result in Proposition 3.3.2.

Now it only remains to construct a probability measure $\mathbb{Q}_L$ with the right support such that inequality (A.9) holds. To do this, we make use of a fact from the proof of Proposition 3.3.1 for the orthant cone $K_+ \subset \mathbb{R}^m$. Recall that to establish Proposition 3.3.1, we constructed a probability measure $\mathbb{D}$ supported on $K_+ \cap S^{m-1} \subset \mathbb{R}^m$ such that if b, b' are an i.i.d pair drawn from $\mathbb{D}$, we have

$$\mathbb{E}_{b,b'\sim\mathbb{D}} e^{\lambda\langle b,\, b'\rangle} \leq \exp\left(\exp\left(\frac{2+\lambda}{\sqrt{m}-1}\right) - \left(1-\frac{1}{\sqrt{m}}\right)^2\right). \tag{A.11}$$

By construction, $\mathbb{D}$ is a uniform probability measure on the finite set $\mathcal{S}$ which consists of all vectors in $\mathbb{R}^m$ which have s non-zero entries which are all equal to $1/\sqrt{s}$ where $s = \lfloor\sqrt{m}\rfloor$.

Based on this measure $\mathbb{D}$, let us define $\mathbb{Q}_L$ as in the following lemma and establish some of its properties under the assumption that $\sqrt{\log(ed)} \geq 14$.

Lemma A.3.1. *Let G be the $m \times m$ lower triangular matrix given by*

$$G := \begin{pmatrix} 1 & & & & \\ r & 1 & & & \\ r^2 & r & 1 & & \\ \vdots & \vdots & & \ddots & \\ r^{m-1} & r^{m-2} & \cdots & & 1 \end{pmatrix}. \tag{A.12a}$$

There exists an $d \times m$ matrix F such that

$$F^T F = \mathbb{I}_m \tag{A.12b}$$

and such that for every $b \in \mathcal{S}$ and $\eta := FGb$, we have

1. *$\eta \in M \cap L^{\perp} \cap B^c(1)$ if $L = \{0\}$, and*
2. *$\eta - \bar{\eta}\mathbf{1} \in M \cap L^{\perp} \cap B^c(1)$ if $L = span(\mathbf{1})$, where $\bar{\eta} = \sum_{i=1}^{d} \eta_i/d$ denotes the mean of the vector η.*

See Appendix A.7.2 for the proof of this claim.

If $L = \{0\}$, let probability measure $\mathbb{Q}_L$ be defined as the distribution of $\eta := FGb$ where $b \sim \mathbb{D}$. Otherwise if $L = \text{span}(\mathbf{1})$, let $\mathbb{Q}_L$ be the distribution of $\eta - \bar{\eta}\mathbf{1}$ where again $\eta := FGb$ and $b \sim \mathbb{D}$. From Lemma A.3.1 we know that $\mathbb{Q}_L$ is supported on $M \cap L^{\perp} \cap B^c(1)$. It only remains to verify the critical inequality (A.9) to complete the proof of Proposition 3.3.2. Let $\eta = FGb$ and $\eta' = FGb'$ with b, b' being i.i.d having distribution $\mathbb{D}$. Using the fact that $F^T F = \mathbb{I}_m$, we can write the inner product of η, η' as

$$\langle \eta,\, \eta' \rangle = b^T G^T F^T F G b' = \langle Gb,\, Gb' \rangle.$$

The following lemma relates inner product $\langle \eta,\, \eta' \rangle$ to $\langle b,\, b' \rangle$, and thereby allows us to derive inequality (A.9) based on inequality (A.11). Recall that $\mathcal{S}$ consists of all vectors in $\mathbb{R}^m$ which have s non-zero entries which are all equal to $1/\sqrt{s}$ where $s = \lfloor \sqrt{m} \rfloor$.

Lemma A.3.2. *For every $b, b' \in \mathcal{S}$, we have*

$$\langle Gb,\, Gb' \rangle \;\leq\; \frac{\langle b,\, b' \rangle}{(1-r)^2} + \frac{r}{s(1-r)^2(1-r^2)}, \tag{A.13a}$$

$$\|Gb\|_2^2 \;\geq\; \frac{1}{(1-r)^2} - \frac{2r + r^2}{s(1-r^2)(1-r)^2}. \tag{A.13b}$$

See Appendix A.7.3 for the proof of this claim.

We are now ready to prove inequality (A.9). We consider the two cases $L = \{0\}$ and $L = \text{span}(\mathbf{1})$ separately.

For $L = \{0\}$, recall that $r = 1/3$ and $s = \lfloor \sqrt{m} \rfloor \geq \sqrt{m} - 1$. Therefore as a direct consequence of inequality (A.13a), we have

$$\mathbb{E}_{\eta, \eta \sim \mathbb{Q}} e^{\lambda \langle \eta,\, \eta' \rangle} \leq \mathbb{E}_{b, b' \sim \mathbb{D}} \exp\left(\frac{9\lambda}{4} \langle b,\, b' \rangle + \frac{27\lambda}{32(\sqrt{m}-1)} \right). \tag{A.14}$$

Combining inequality (A.14) with (A.11) completes the proof of inequality (A.9).

Let us now turn to the case when $L = \text{span}(\mathbf{1})$. The proof is essentially the same as for $L = \{0\}$ with only some minor changes. Again our goal is to check inequality (A.9). For this, we write

$$\mathbb{E}_{\eta, \eta' \sim \mathbb{Q}_L} e^{\lambda \langle \eta,\, \eta' \rangle} = \mathbb{E}_{\eta, \eta' \sim \mathbb{Q}_{\{0\}}} e^{\lambda \langle \eta - \bar{\eta}\mathbf{1},\, \eta' - \bar{\eta}'\mathbf{1} \rangle} \;\leq\; \mathbb{E}_{\eta, \eta' \sim \mathbb{Q}_{\{0\}}} e^{\lambda \langle \eta,\, \eta' \rangle}.$$

Here the last step use the fact that $\langle \eta - \bar{\eta}\mathbf{1},\, \eta' - \bar{\eta'}\mathbf{1} \rangle = \langle \eta,\, \eta' \rangle - d\bar{\eta}\bar{\eta'} \leq \langle \eta,\, \eta' \rangle$ where the last inequality follows from the non-negativity of every entry of vectors η and η' (this non-negativity is a consequence of the non-negativity of F and G from Lemma A.3.1 and non-negativity of vectors in $\mathcal{S}$).

Thus, we have completed the proof of Proposition 3.3.2.

A.4 Completion of the proof of Theorem 3.3.1(a)

In this appendix, we collect the proofs of lemmas involved in the proof of Theorem 3.3.1(a).

A.4.1 Proof of Lemma A.4.1

Let us start with the statement with this lemma.

Lemma A.4.1. *For a standard Gaussian random vector $g \sim N(0, I_d)$, closed convex cone $K \in \mathbb{R}^d$ and vector $\theta \in \mathbb{R}^d$, we have*

$$\mathbb{P}\Big(\pm \left(Z(\theta) - \mathbb{E}[Z(\theta)]\right) \geq t \Big) \leq \exp\Big(- \frac{t^2}{2} \Big), \qquad \textit{and} \tag{A.15a}$$

$$\mathbb{P}\Big(\pm \left(\langle \theta,\, \Pi_K g \rangle - \mathbb{E}\langle \theta,\, \Pi_K g \rangle\right) \geq t \Big) \leq \exp\Big(- \frac{t^2}{2\|\theta\|_2^2} \Big), \tag{A.15b}$$

where both inequalities hold for all $t \geq 0$.

For future reference, we also note that tail bound (A.15a) implies that the variance is bounded as

$$\mathrm{var}(Z(\theta)) = \int_0^\infty \mathbb{P}\Big(\big|Z(\theta) - \mathbb{E}[Z(\theta)]\big| \geq \sqrt{u} \Big) du \;\leq\; 2 \int_0^\infty e^{-u/2} du \;=\; 4. \tag{A.16}$$

To prove Lemma A.4.1, given every vector θ, we claim that the function $g \mapsto \|\Pi_K(\theta + g)\|_2$ is 1-Lipschitz, whereas the function $g \mapsto \langle \theta,\, \Pi_K g \rangle$ is a $\|\theta\|_2$-Lipschitz function. From these claims, the concentration results then follow from Borell's theorem [19].

In order to establish the Lipschitz property, consider two vectors $g, g' \in \mathbb{R}^d$. By the triangle inequaliuty non-expansiveness of Euclidean projection, we have

$$\Big| \|\Pi_K(\theta + g)\|_2 - \|\Pi_K(\theta + g')\|_2 \Big| \leq \|\Pi_K(\theta + g) - \Pi_K(\theta + g')\|_2 \;\leq\; \|g - g'\|_2.$$

Combined with the Cauchy-Schwarz inequality, we conclude that

$$\big| \langle \theta,\, \Pi_K g \rangle - \langle \theta,\, \Pi_K g' \rangle \big| \leq \|\theta\|_2 \, \|\Pi_K g - \Pi_K g'\|_2 \leq \|\theta\|_2 \, \|g - g'\|_2,$$

which completes the proof of Lemma A.4.1.

A.4.2 Proof of Lemma 3.5.1

We define the random variable $Z(\theta) := \|\Pi_K(\theta + g)\|_2 - \|\Pi_K g\|_2$, as well as its positive and negative parts $Z^+(\theta) = \max\{0, Z(\theta)\}$ and $Z^-(\theta) = \max\{0, -Z(\theta)\}$, so that $\Gamma(\theta) = \mathbb{E}Z(\theta) = \mathbb{E}Z^+(\theta) - \mathbb{E}Z^-(\theta)$. Our strategy is to bound $\mathbb{E}Z^-(\theta)$ from above and then bound $\mathbb{E}Z^+(\theta)$ from below. The following auxiliary lemma is useful for these purposes:

Lemma A.4.2. *For every closed convex cone $K \subset \mathbb{R}^d$ and vectors $x \in K$ and $y \in \mathbb{R}^d$, we have:*

$$\Big| \|\Pi_K(x+y)\|_2 - \|\Pi_K(y)\|_2 \Big| \leq \|x\|_2, \qquad \textit{and} \tag{A.17}$$

$$\max\left\{2\langle x,\, y\rangle + \|x\|_2^2,\, 2\langle x,\, \Pi_K y\rangle - \|x\|_2^2\right\} \overset{(i)}{\leq} \|\Pi_K(x+y)\|_2^2 - \|\Pi_K(y)\|_2^2 \overset{(ii)}{\leq} 2\langle x,\, \Pi_K y\rangle + \|x\|_2^2. \tag{A.18}$$

We return to prove this claim in Appendix A.4.3.

Inequality (A.17) implies that $Z(\theta) \geq -\|\theta\|_2$ and thus $\mathbb{E}Z^-(\theta) \leq \|\theta\|_2 \mathbb{P}\{Z(\theta) \leq 0\}$. The lower bound in inequality (A.18) then implies that $\mathbb{P}\{Z(\theta) \leq 0\} \leq \mathbb{P}\{\langle\theta,\, g\rangle \leq -\|\theta\|_2^2/2\} \leq \exp\big(-\frac{\|\theta\|_2^2}{8}\big)$, whence

$$\mathbb{E}Z^-(\theta) \leq \|\theta\|_2 \exp\left(\frac{-\|\theta\|_2^2}{8}\right) \leq \sup_{u>0}\left(ue^{-u^2/8}\right) = \frac{2}{\sqrt{e}}.$$

Putting together the pieces, we have established the lower bound

$$\mathbb{E}Z(\theta) \;=\; \mathbb{E}Z^+(\theta) - \mathbb{E}Z^-(\theta) \;\geq\; \mathbb{E}Z^+(\theta) - \frac{2}{\sqrt{e}}. \tag{A.19}$$

The next task is to lower bound the expectation $\mathbb{E}Z^+(\theta)$. By the triangle inequality, we have

$$\begin{aligned} \|\Pi_K(\theta+g)\|_2 &\leq \|\Pi_K(\theta+g) - \Pi_K(g)\|_2 + \|\Pi_K(g)\|_2 \\ &\leq \|\theta\|_2 + \|\Pi_K(g)\|_2, \end{aligned}$$

where the second inequality uses non-expansiveness of the projection. Consequently, we have the lower bound

$$\mathbb{E}Z^+(\theta) = \mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+}{\|\Pi_K(\theta+g)\|_2 + \|\Pi_K g\|_2} \geq \mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+}{\|\theta\|_2 + 2\|\Pi_K g\|_2}. \tag{A.20}$$

Note that inequality (A.18)(i) implies two lower bounds on the difference $\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2$. We treat each of these lower bounds in turn, and show how they lead to inequalities (3.55a) and (3.55b).

Proof of inequality (3.55a)**:** Inequality (A.20) and the first lower bound term from inequality (A.18)(i) imply that

$$\mathbb{E}Z^+(\theta) \geq \mathbb{E}\frac{(2\langle\theta,\, g\rangle + \|\theta\|_2^2)^+}{\|\theta\|_2 + 2\|\Pi_K g\|_2} \geq \mathbb{E}\frac{\|\theta\|_2^2}{\|\theta\|_2 + 2\|\Pi_K g\|_2}\mathbb{I}\{\langle\theta,\, g\rangle \geq 0\}.$$

Jensen's inequality (and the fact that $\mathbb{P}\{\langle\theta,\, g\rangle \geq 0\} = 1/2$) now allow us to deduce

$$\mathbb{E}Z^+(\theta) \geq \mathbb{P}\left\{\langle\theta,\, g\rangle \geq 0\right\} \|\theta\|_2^2 \left(\|\theta\|_2 + \frac{2\mathbb{E}\|\Pi_K g\|_2}{P\left\{\langle\theta,\, g\rangle \geq 0\right\}}\right)^{-1} = \frac{\|\theta\|_2^2}{2\|\theta\|_2 + 8\mathbb{E}\|\Pi_K g\|_2}$$

and this gives inequality (3.55a).

Proof of inequality (3.55b)**:** Putting inequality (A.20), the second term on the left hand side of inequality (A.18)(i), along with the fact that $\langle \theta, \, \mathbb{E}\Pi_K g\rangle \geq \|\theta\|_2^2$ together guarantees that

$$\mathbb{E}Z^+(\theta) \geq \mathbb{E}\frac{(2\langle \theta, \, \Pi_K g\rangle - \|\theta\|_2^2)^+}{\|\theta\|_2 + 2\|\Pi_K g\|_2} \geq \mathbb{E}\frac{\langle \theta, \, \mathbb{E}\Pi_K g\rangle - \|\theta\|_2^2}{\|\theta\|_2 + 2\|\Pi_K g\|_2} \; \mathbb{I}\left\{\langle \theta, \, \Pi_K g\rangle > \frac{1}{2}\langle \theta, \, \mathbb{E}\Pi_K g\rangle\right\}.$$

Now introducing the event $\mathcal{D} := \big\{\langle \theta, \, \Pi_K g\rangle > \langle \theta, \, \mathbb{E}\Pi_K g\rangle/2\big\}$, Jensen's inequality implies that

$$\mathbb{E}Z^+(\theta) \geq \mathbb{P}(\mathcal{D}) \, \mathbb{E}\frac{\langle \theta, \, \mathbb{E}\Pi_K g\rangle - \|\theta\|_2^2}{\|\theta\|_2 + 2\frac{\mathbb{E}\|\Pi_K g\|_2}{\mathbb{P}(\mathcal{D})}}. \tag{A.21}$$

The concentration inequality (A.15b) from Lemma A.4.1 gives us that

$$\mathbb{P}(\mathcal{D}) \geq \mathbb{P}\left\{\langle \theta, \, \Pi_K g\rangle > \frac{1}{2}\langle \theta, \, \mathbb{E}\Pi_K g\rangle\right\} \geq 1 - \exp\left(-\frac{\langle \theta, \, \mathbb{E}\Pi_K g\rangle^2}{8\|\theta\|_2^2}\right). \tag{A.22}$$

Inequality (3.55b) now follows by combining inequalities (A.19), (A.21) and (A.22).

A.4.3 Proof of Lemma A.4.2

Let us turn to prove Lemma A.4.2. Inequality (A.17) is a standard Lipschitz property of projection onto a closed convex cone. Turning to inequality (A.18), recall the polar cone $K^* := \{z \mid \langle z, \, \theta\rangle \leq 0, \; \forall \; \theta \in K\}$, as well as the Moreau decomposition (3.18)—namely, $z = \Pi_K(z) + \Pi_{K^*}(z)$. Using this notation, we have

$$\begin{aligned}\|\Pi_K(x+y)\|_2^2 - \|\Pi_K y\|_2^2 &= \|x + y - \Pi_{K^*}(x+y)\|_2^2 - \|y - \Pi_{K^*}y\|_2^2 \\ &= \|x\|_2^2 + 2\langle x, \, y - \Pi_{K^*}(x+y)\rangle + \|y - \Pi_{K^*}(x+y)\|_2^2 - \|y - \Pi_{K^*}y\|_2^2.\end{aligned}$$

Since $\Pi_{K^*}(y)$ is the closest point in K^* to y, we have $\|y - \Pi_{K^*}(x+y)\|_2 \geq \|y - \Pi_{K^*}(y)\|_2$, and hence

$$\|\Pi_K(x+y)\|_2^2 - \|\Pi_K y\|_2^2 \geq \|x\|_2^2 + 2\langle x, \, y - \Pi_{K^*}(x+y)\rangle. \tag{A.23}$$

Since $x \in K$ and $\Pi_{K^*}(x+y) \in K^*$, we have $\langle x, \, \Pi_{K^*}(x+y)\rangle \leq 0$, and hence, inequality (A.23) leads to the bound (i) in equation (A.18). In order to establish inequality (ii) in equation (A.18), we begin by rewriting expression (A.23) as

$$\|\Pi_K(x+y)\|_2^2 - \|\Pi_K y\|_2^2 \geq \|x\|_2^2 + 2\langle x, \, y - \Pi_{K^*}y\rangle + 2\langle x, \, \Pi_{K^*}y - \Pi_{K^*}(x+y)\rangle.$$

Applying the Cauchy-Schwarz inequality to the final term above and using the 1-Lipschitz property of $z \mapsto \Pi_{K^*}z$, we obtain:

$$\langle x, \, \Pi_{K^*}y - \Pi_{K^*}(x+y)\rangle \geq -\|x\|_2\|\Pi_{K^*}y - \Pi_{K^*}(x+y)\|_2 \geq -\|x\|_2^2,$$

which establishes the upper bound of inequality (A.18).

Finally, in order to prove the lower bound in inequality (A.18), we write

$$\begin{aligned}
&\|\Pi_K(x+y)\|_2^2 - \|\Pi_K y\|_2^2 \\
=&\|x+y-\Pi_{K^*}(x+y)\|_2^2 - \|x+y-\Pi_{K^*}y - x\|_2^2 \\
=&\|x+y-\Pi_{K^*}(x+y)\|_2^2 - \|x+y-\Pi_{K^*}y\|_2^2 + 2\langle x,\, x+y-\Pi_{K^*}y\rangle - \|x\|_2^2.
\end{aligned}$$

Since the vector $\Pi_{K^*}(x+y)$ corresponds to the projection of $x+y$ onto K^*, we have $\|x+y-\Pi_{K^*}(x+y)\|_2 \leq \|x+y-\Pi_{K^*}y\|_2$ and thus

$$\|\Pi_K(x+y)\|_2^2 - \|\Pi_K y\|_2^2 \leq \|x\|_2^2 + 2\langle x,\, \Pi_K y\rangle,$$

which completes the proof of inequality (A.18).

A.5 Completion of the proof of Theorem 3.3.1(b)

In this appendix, we collect the proofs of lemmas involved in the proof of Theorem 3.3.1(b), corresponding to the lower bound on the GLRT performance.

A.5.1 Proof of Lemma A.5.1

Let us first state Lemma A.5.1 and give a proof of it.

Lemma A.5.1. *For any constant $a \geq 1$ and for every closed convex cone $K \neq \{0\}$, we have*

$$0 \leq \Gamma(\theta) \leq \frac{2a\|\theta\|_2^2 + 4\langle\theta,\, \mathbb{E}\Pi_K g\rangle}{\mathbb{E}\|\Pi_K g\|_2} + b\|\theta\|_2 \qquad \text{for all } \theta \in K, \tag{A.24a}$$

where

$$b := 3\exp\Big(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}\Big) + 24\exp\Big(-\frac{a^2\|\theta\|_2^2}{16}\Big). \tag{A.24b}$$

In order to prove that $\Gamma(\theta) \geq 0$, we first introduce the convenient shorthand notation $v_1 := \Pi_{K^*}(\theta+g)$ and $v_2 := \Pi_{K^*}g$. Recall that K^* denotes the polar cone of K defined in expression (3.17). With this notation, the the Moreau decomposition (3.18) then implies that

$$\begin{aligned}
\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2 &= \|\theta+g-v_1\|_2^2 - \|g-v_2\|_2^2 \\
&= \|\theta\|_2^2 + 2\langle\theta,\, g-v_1\rangle + \|g-v_1\|_2^2 - \|g-v_2\|_2^2.
\end{aligned}$$

The right hand side above is greater than $\|\theta\|_2^2 + 2\langle\theta,\, g-v_1\rangle$ because $\|g-v_1\|_2^2 \geq \min_{v\in K^*}\|g-v\|_2^2 = \|g-v_2\|_2^2$. From the fact that $\mathbb{E}\langle\theta,\, g\rangle = 0$ and $\langle\theta,\, v\rangle \leq 0$ for all $v \in K^*$, we have $\Gamma(\theta) \geq 0$.

Now let us prove the upper bound for expected difference $\Gamma(\theta)$. Using the convenient shorthand notation $Z(\theta) := \|\Pi_K(\theta+g)\|_2 - \|\Pi_K g\|_2$, we define the event

$$\mathcal{B} := \{\|\Pi_K g\|_2 \geq \frac{1}{2}\mathbb{E}\|\Pi_K g\|_2\}, \qquad \text{where } g \sim N(0, I_d).$$

Our proof is then based on the decomposition $\Gamma(\theta) = \mathbb{E}Z(\theta) = \mathbb{E}Z(\theta)\mathbb{I}(\mathcal{B}^c) + \mathbb{E}Z(\theta)\mathbb{I}(\mathcal{B})$. In particular, we upper bound each of these two terms separately.

Bounding $\mathbb{E}[Z(\theta)\mathbb{I}(\mathcal{B}^c)]$: The analysis of this term is straightforward: inequality (A.17) from Lemma A.4.2 guarantees that $Z(\theta) \leq \|\theta\|_2$, whence

$$\mathbb{E}Z(\theta)\mathbb{I}(\mathcal{B}^c) \leq \|\theta\|_2\mathbb{P}(\mathcal{B}^c). \tag{A.25}$$

Bounding $\mathbb{E}[Z(\theta)\mathbb{I}(\mathcal{B})]$: Turning to the second term, we have

$$\begin{aligned}\mathbb{E}Z(\theta)\mathbb{I}(\mathcal{B}) &\leq \mathbb{E}Z^+(\theta)\mathbb{I}(\mathcal{B}) \\ &= \mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+}{\|\Pi_K(\theta+g)\|_2 + \|\Pi_K g\|_2}\mathbb{I}(\mathcal{B}) \leq \mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+}{\|\Pi_K g\|_2}\mathbb{I}(\mathcal{B}).\end{aligned}$$

On event $\mathcal{B}$, we can lower bound quantity $\|\Pi_K g\|_2$ with $\mathbb{E}\|\Pi_K g\|_2/2$ thus

$$\mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+}{\|\Pi_K g\|_2}\mathbb{I}(\mathcal{B}) \leq \underbrace{\mathbb{E}\frac{(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2)^+\,\mathbb{I}(\mathcal{B})}{\mathbb{E}\|\Pi_K g\|_2/2}}_{:=T_1}. \tag{A.26}$$

Next we use inequality (A.18) to bound the numerator of the quantity T_1, namely

$$\begin{aligned}\mathbb{E}\left(\|\Pi_K(\theta+g)\|_2^2 - \|\Pi_K g\|_2^2\right)^+\mathbb{I}(\mathcal{B}) &\leq \mathbb{E}\left(2\langle\theta,\, \Pi_K g\rangle + \|\theta\|_2^2\right)^+\mathbb{I}(\mathcal{B}) \\ &\leq \mathbb{E}\left(2\langle\theta,\, \Pi_K g\rangle + a\|\theta\|_2^2\right)^+\mathbb{I}(\mathcal{B}),\end{aligned}$$

for every constant $a \geq 1$. To further simplify notation, introduce event $\mathcal{C} := \{\theta^T\Pi_K g \geq -a\|\theta\|_2^2/2\}$ and by definition, we obtain

$$\begin{aligned}\mathbb{E}\left(2\langle\theta,\, \Pi_K g\rangle + a\|\theta\|_2^2\right)^+\mathbb{I}(\mathcal{B}) &= \mathbb{E}\left(2\langle\theta,\, \Pi_K g\rangle + a\|\theta\|_2^2\right)\mathbb{I}(\mathcal{B}\cap\mathcal{C}) \\ &\leq a\|\theta\|_2^2 + 2\mathbb{E}[\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{B}\cap\mathcal{C})]. \end{aligned}\tag{A.27}$$

The right hand side of inequality (A.27) consists of two terms. The first term $a\|\theta\|_2^2$ is a constant, so that we only need to further bound the second term $2\mathbb{E}\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{B}\cap\mathcal{C})$. We claim that

$$\mathbb{E}[\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{B}\cap\mathcal{C})] \leq \mathbb{E}\langle\theta,\, \Pi_K g\rangle + \|\theta\|_2\mathbb{E}\|\Pi_K g\|_2(6\sqrt{\mathbb{P}(\mathcal{C}^c)} + \mathbb{P}(\mathcal{B}^c)/2). \tag{A.28}$$

Taking inequality (A.28) as given for the moment, combining inequalities (A.26), (A.27) and (A.28) yields

$$\mathbb{E}Z^{+}(\theta)\mathbb{I}(\mathcal{B}) \le T_1 \le \frac{2a\|\theta\|_2^2 + 4\mathbb{E}\langle\theta,\, \Pi_K g\rangle}{\mathbb{E}\|\Pi_K g\|_2} + \|\theta\|_2(24\sqrt{\mathbb{P}(\mathcal{C}^c)} + 2\mathbb{P}(\mathcal{B}^c)). \tag{A.29}$$

As a summary of the above two parts—namely inequalities (A.25) and (A.29), if we assume inequality (A.28), we have

$$\Gamma(\theta) \le \frac{2a\|\theta\|_2^2 + 4\mathbb{E}\langle\theta,\, \Pi_K g\rangle}{\mathbb{E}\|\Pi_K g\|_2} + \|\theta\|_2(24\sqrt{\mathbb{P}(\mathcal{C}^c)} + 3\mathbb{P}(\mathcal{B}^c)). \tag{A.30}$$

Based on expression (A.30), the last step in proving Lemma A.5.1 is to control the probabilities $\mathbb{P}(\mathcal{C}^c)$ and $\mathbb{P}(\mathcal{B}^c)$ respectively. Using the fact that $\langle\theta,\, \Pi_K g\rangle = \langle\theta,\, (g - \Pi_{K^*} g)\rangle \ge \langle\theta,\, g\rangle$ and the concentration of $\langle\theta,\, g\rangle$, we have

$$\begin{aligned} &\mathbb{P}(\mathcal{C}^c) = \mathbb{P}(\langle\theta,\, \Pi_K g\rangle < -\frac{a}{2}\|\theta\|_2^2) \le \mathbb{P}(\langle\theta,\, g\rangle < -\frac{a}{2}\|\theta\|_2^2) \le \exp(-\frac{a^2\|\theta\|_2^2}{8}),\\ \text{and}\quad &\mathbb{P}(\mathcal{B}^c) = \mathbb{P}(\|\Pi_K g\|_2 < \frac{1}{2}\mathbb{E}\|\Pi_K g\|_2) \le \exp(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}). \end{aligned}$$

where the second inequality follows directly from concentration result in Lemma A.4.1 (A.15a). Substituting the above two inequalities into expression (A.30) yields Lemma A.5.1.

So it is only left for us to show inequality (A.28). To see this, first notice that

$$\mathbb{E}[\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{B}\cap\mathcal{C})] = \mathbb{E}\langle\theta,\, \Pi_K g\rangle - \mathbb{E}\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{C}^c\cup\mathcal{B}^c). \tag{A.31}$$

The Cauchy-Schwarz inequality and triangle inequality allow us to deduce

$$\begin{aligned} -\mathbb{E}\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{C}^c\cup\mathcal{B}^c) &= \langle\theta,\, -\mathbb{E}[\Pi_K g\mathbb{I}(\mathcal{C}^c\cup\mathcal{B}^c)]\rangle\\ &\le \|\theta\|_2\|\mathbb{E}[\Pi_K g\mathbb{I}(\mathcal{C}^c\cup\mathcal{B}^c)]\|_2\\ &\le \|\theta\|_2\Big\{\|\mathbb{E}\Pi_K g\mathbb{I}(\mathcal{C}^c)\|_2 + \|\mathbb{E}\Pi_K g\mathbb{I}(\mathcal{B}^c)\|_2\Big\}. \end{aligned}$$

Jensen's inequality further guarantees that

$$-\mathbb{E}\langle\theta,\, \Pi_K g\rangle\mathbb{I}(\mathcal{C}^c\cup\mathcal{B}^c) \le \|\theta\|_2\Big\{\underbrace{\mathbb{E}[\|\Pi_K g\|_2\mathbb{I}(\mathcal{C}^c)]}_{:=T_2} + \underbrace{\mathbb{E}[\|\Pi_K g\|_2\mathbb{I}(\mathcal{B}^c)]}_{:=T_3}\Big\}, \tag{A.32}$$

By definition, on event $\mathcal{B}^c$, we have $\|\Pi_K g\|_2 \le \mathbb{E}\|\Pi_K g\|_2/2$, and consequently

$$T_3 \le \frac{\mathbb{E}\|\Pi_K g\|_2\mathbb{P}(\mathcal{B}^c)}{2}. \tag{A.33}$$

Turning to the quantity T_2, applying Cauchy-Schwartz inequality yields

$$T_2 \le \sqrt{\mathbb{E}\|\Pi_K g\|_2^2}\sqrt{\mathbb{E}\mathbb{I}(\mathcal{C}^c)} = \sqrt{(\mathbb{E}\|\Pi_K g\|_2)^2 + \mathrm{var}(\|\Pi_K g\|_2)}\sqrt{\mathbb{P}(\mathcal{C}^c)}.$$

The variance term can be bounded as in inequality (A.16) which says that $\mathrm{var}(\|\Pi_K g\|_2) \leq 4$.

From inequality (3.21), for every non-trivial cone ($K \neq \{0\}$), we are guaranteed that $\mathbb{E}\|\Pi_K g\|_2 \geq 1/\sqrt{2\pi}$, and hence $\mathrm{var}(\|\Pi_K g\|_2) \leq 8\pi(\mathbb{E}\|\Pi_K g\|_2)^2$. Consequently, the quantity T_2 can be further bounded as

$$T_2 \leq \sqrt{1+8\pi}\mathbb{E}\|\Pi_K g\|_2 \sqrt{\mathbb{P}(\mathcal{C}^c)} \leq 6\mathbb{E}\|\Pi_K g\|_2 \sqrt{\mathbb{P}(\mathcal{C}^c)}. \tag{A.34}$$

Putting together inequalities (A.33), (A.34) and (A.32) yields

$$-\mathbb{E}[\langle\theta,\, \Pi_K g\rangle \mathbb{I}(\mathcal{C}^c \cup (\mathcal{C} \cap \mathcal{B}^c))] \leq \|\theta\|_2 \mathbb{E}\|\Pi_K g\|_2 (6\sqrt{\mathbb{P}(\mathcal{C}^c)} + \mathbb{P}(\mathcal{B}^c)/2),$$

which validates claim (A.28) when combined with inequality (A.31). We finish the proof of Lemma A.5.1.

A.5.2 Proof of inequality (3.59)

Now let us turn to the proof of inequality (3.59). First notice that if the radius satisfies $\epsilon^2 \leq b_\rho \delta^2_{\mathrm{LR}}(\{0\}, K)$, then there exists some $\theta \in \mathcal{H}_1$ with $\|\theta\|_2 = \epsilon$ that satisfies

$$\|\theta\|_2^2 \leq b_\rho \mathbb{E}\|\Pi_K g\|_2 \text{ and } \langle\theta,\, \mathbb{E}\Pi_K g\rangle \leq \sqrt{b_\rho}\mathbb{E}\|\Pi_K g\|_2. \tag{A.35}$$

Setting $a = 4/\sqrt{b_\rho} \geq 1$ in inequality (A.24a) yields

$$\Gamma(\theta) \leq \frac{8\|\theta\|_2^2/\sqrt{b_\rho} + 4\langle\theta,\, \mathbb{E}\Pi_K g\rangle}{\mathbb{E}\|\Pi_K g\|_2} + b\|\theta\|_2$$

where $b := 3\exp(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}) + 24\exp(-\frac{\|\theta\|_2^2}{b_\rho})$. Now we only need to bound the two terms in the upper bound separately. First, note that inequality (A.35) yields

$$\frac{8\|\theta\|_2^2/\sqrt{b_\rho} + 4\langle\theta,\, \mathbb{E}\Pi_K g\rangle}{\mathbb{E}\|\Pi_K g\|_2} \leq 12\sqrt{b_\rho}. \tag{A.36}$$

On the other hand, again by applying inequality (A.35), it is straightforward to verify the following two facts that

$$\begin{aligned}\|\theta\|_2 \exp(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}) &\leq \sqrt{b_\rho \mathbb{E}\|\Pi_K g\|_2}\exp(-\frac{(\mathbb{E}\|\Pi_K g\|_2)^2}{8}) \\ &\leq \sqrt{b_\rho}\max_{x\in(0,\infty)}\sqrt{x}\exp(-\frac{x^2}{8}) = \sqrt{b_\rho}\left(\frac{2}{e}\right)^{1/4},\end{aligned}$$

$$\text{and} \quad \|\theta\|_2 \exp(-\frac{\|\theta\|_2^2}{b_\rho}) \leq \sup_{x\in(0,\infty)} x\exp(-\frac{x^2}{b_\rho}) = \sqrt{\frac{b_\rho}{2e}}.$$

Combining the above two inequalities ensures an upper bound for product $b\|\theta\|_2$ and directly leads to upper bound of quantity $\Gamma(\theta)$, namely

$$\Gamma(\theta) \leq 12\sqrt{b_\rho} + 3\sqrt{b_\rho}\left(\frac{2}{e}\right)^{1/4} + 24\sqrt{\frac{b_\rho}{2e}},$$

With the choice of b_ρ, we established inequality (3.59).

A.5.3 Proof of Lemma 3.5.2

In order to prove this result, we first define random variable $F := \|\Pi_K g\|_2^2 - m$, where $m := \mathbb{E}\|\Pi_K g\|_2^2$ and $\tilde{\sigma}^2 := \mathrm{var}(F)$. We make use of the Theorem 2.1 in Goldstein et al. [65] which shows that the distribution of F and Gaussian distribution $Z \sim N(0, \tilde{\sigma}^2)$ are very close, more specifically, the Theorem says

$$\|F - Z\|_{\mathrm{TV}} \leq \frac{16}{\tilde{\sigma}^2}\sqrt{m} \leq \frac{8}{\mathbb{E}\|\Pi_K g\|_2}. \tag{A.37}$$

In the last inequality, we use the facts that $\tilde{\sigma}^2 \geq 2m$ and $\sqrt{\mathbb{E}\|\Pi_K g\|_2^2} \geq \mathbb{E}\|\Pi_K g\|_2$.

It is known that the quantity $\|\Pi_K g\|_2^2$ is distributed as a mixture of χ^2 distributions(see e.g., [117, 65])—in particular, we can write

$$\|\Pi_K g\|_2^2 \overset{\text{law}}{=} \sum_{i=1}^{V_K} X_i = W_K + V_K, \qquad W_K = \sum_{i=1}^{V_K}(X_i - 1),$$

where each $\{X_i\}_{i\geq 1}$ is an i.i.d. sequence χ_1^2 variables, independent of V_K. Applying the decomposition of variance yields

$$\tilde{\sigma}^2 = \mathrm{var}(V_K) + 2\mathbb{E}\|\Pi_K g\|_2^2 \geq 2m.$$

We can write the probability $\mathbb{P}(\|\Pi_K g\|_2 > \mathbb{E}\|\Pi_K g\|_2)$ as

$$\mathbb{P}(\|\Pi_K g\|_2 > \mathbb{E}\|\Pi_K g\|_2) = \mathbb{P}(\|\Pi_K g\|_2^2 - \mathbb{E}\|\Pi_K g\|_2^2 > (\mathbb{E}\|\Pi_K g\|_2)^2 - \mathbb{E}\|\Pi_K g\|_2^2) \geq \mathbb{P}(F > 0).$$

So if $\mathbb{E}\|\Pi_K g\|_2 \geq 128$, then inequality (A.37) ensures that $d_{TV}(F, N) \leq 1/16$, and hence

$$\mathbb{P}(F > 0) \geq \mathbb{P}(Z > 0) - \|F - Z\|_{\mathrm{TV}} \geq \frac{7}{16}.$$

We finish the proof of Lemma 3.5.2.

A.6 Completion of the proof of Theorem 3.3.2

In this appendix, we collect the proofs of various lemmas used in the proof of Theorem 3.3.2.

A.6.1 Proof of Lemma 3.5.3

For every probability measure $\mathbb{Q}$ supported on $K \cap B^c(1)$, let vector θ be distributed accordingly to measure $\epsilon\mathbb{Q}$ then it is supported on $K \cap B^c(\epsilon)$. Consider a mixture of distributions,

$$\mathbb{P}_1(y) = \mathbb{E}_\theta\ (2\pi)^{-d/2}\exp(-\frac{\|y-\theta\|_2^2}{2}). \tag{A.38}$$

Let us first control the χ^2 distance between distributions $\mathbb{P}_1$ and $\mathbb{P}_0 := N(0, I_d)$. Direct calculations yield

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) + 1 = \mathbb{E}_{\mathbb{P}_0}\left(\frac{\mathbb{P}_1}{\mathbb{P}_0}\right)^2 = \mathbb{E}_{\mathbb{P}_0}\left(\mathbb{E}_\theta \exp\{-\frac{\|y-\theta\|_2^2}{2} + \frac{\|y\|_2^2}{2}\}\right)^2$$
$$= \mathbb{E}_{\mathbb{P}_0}\left(\mathbb{E}_\theta \exp\{\langle y,\, \theta\rangle - \frac{\|\theta\|_2^2}{2}\}\right)^2.$$

Suppose random vector θ' is an independent copy of random vector θ, then

$$\begin{aligned}\chi^2(\mathbb{P}_1, \mathbb{P}_0) + 1 &= \mathbb{E}_{\mathbb{P}_0}\mathbb{E}_{\theta,\theta'} \exp\{\langle y,\, \theta+\theta'\rangle - \frac{\|\theta\|_2^2 + \|\theta'\|_2^2}{2}\} \\ &= \mathbb{E}_{\theta,\theta'} \exp\{\frac{\|\theta+\theta'\|_2^2}{2} - \frac{\|\theta\|_2^2 + \|\theta'\|_2^2}{2}\} \\ &= \mathbb{E}_{\theta,\theta'} \exp(\langle \theta,\, \theta'\rangle) \\ &= \mathbb{E}\exp(\epsilon^2\langle \eta,\, \eta'\rangle), \end{aligned} \tag{A.39}$$

where the second step uses the fact the moment generating function of multivariate normal distribution. As we know, the testing error is always bounded below by $1 - \|\mathbb{P}_1, \mathbb{P}_0\|_{\mathrm{TV}}$, so by the relation between the χ^2 distance and TV distance, we have:

$$\text{testing error } \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}\exp\left(\epsilon^2\langle \eta,\, \eta'\rangle\right) - 1},$$

which completes our proof.

A.6.2 Proof of Lemma A.6.1

Let us first provide a formal statement of Lemma A.6.1 and then prove it.

Lemma A.6.1. *Letting η and η' denote an i.i.d pair of random variables drawn from the distribution $\mathbb{Q}$ defined in equation (3.62), we have*

$$\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle \eta,\, \eta'\rangle) \leq \frac{1}{a^2}\exp\left(\frac{5\epsilon^2\|\mathbb{E}\Pi_K g\|_2^2}{(\mathbb{E}\|\Pi_K g\|_2)^2} + \frac{40\epsilon^4\mathbb{E}(\|\Pi_K g\|_2^2)}{(\mathbb{E}\|\Pi_K g\|_2)^4}\right), \tag{A.40}$$

where $a := \mathbb{P}(\|\Pi_K g\|_2 \geq \frac{1}{2}\mathbb{E}\|\Pi_K g\|_2)$ and $\epsilon > 0$ satisfies the inequality $\epsilon^2 \leq (\mathbb{E}\|\Pi_K g\|_2)^2/32$.

To prove this result, we use Borell's lemma [19] which states that for a standard Gaussian vector $Z \sim N(0, I_d)$ and a function $f : \mathbb{R}^d \to \mathbb{R}$ which is L-Lipschitz, we have

$$\mathbb{E}\exp(af(Z)) \leq \exp(a\mathbb{E}f(Z) + a^2L^2/2) \tag{A.41}$$

for every $a \geq 0$.

Let g, g' be i.i.d standard normal vectors in $\mathbb{R}^d$. Let

$$\mathcal{A}(g) := \{\|\Pi_K g\|_2 > \frac{1}{2}\mathbb{E}\|\Pi_K g\|_2\} \text{ and } \mathcal{A}(g') := \{\|\Pi_K g'\|_2 > \frac{1}{2}\mathbb{E}\|\Pi_K g'\|_2\}$$

By definition of the probability measure $\mathbb{Q}$ in expression (3.62), we have

$$\begin{aligned}\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle\eta,\, \eta'\rangle) &= \mathbb{E}_{g,g'}\left[\exp\left(\frac{4\epsilon^2\langle\Pi_K g,\, \Pi_K g'\rangle}{\mathbb{E}\|\Pi_K g\|_2\mathbb{E}\|\Pi_K g'\|_2}\right) \,\middle|\, \mathcal{A}(g)\cap\mathcal{A}(g')\right] \\ &= \frac{1}{\mathbb{P}(\mathcal{A}(g)\cap\mathcal{A}(g'))}\mathbb{E}_{g,g'} \exp\left(\frac{4\epsilon^2\langle\Pi_K g,\, \Pi_K g'\rangle}{\mathbb{E}\|\Pi_K g\|_2\mathbb{E}\|\Pi_K g'\|_2}\right)\mathbb{I}(\mathcal{A}(g)\cap\mathcal{A}(g')).\end{aligned}$$

Using the independence of g, g' and non-negativity of the exponential function, we have

$$\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle\eta,\, \eta'\rangle) \le \frac{1}{\mathbb{P}(\mathcal{A}(g))^2}\underbrace{\mathbb{E}_{g,g'} \exp\left(\frac{4\epsilon^2\langle\Pi_K g,\, \Pi_K g'\rangle}{\mathbb{E}\|\Pi_K g\|_2\mathbb{E}\|\Pi_K g'\|_2}\right)}_{:=T_1}. \tag{A.42}$$

To simplify the notation, we write $\lambda := 4\epsilon^2/(\mathbb{E}\|\Pi_K g\|_2)^2$ so that

$$T_1 = \mathbb{E}_{g,g'} \exp\left(\lambda\langle\Pi_K g,\, \Pi_K g'\rangle\right). \tag{A.43}$$

Now for every fixed value of g, the function $h \mapsto \langle\Pi_K g,\, \Pi_K h\rangle$ is Lipschitz with Lipschitz constant equal to $\|\Pi_K g\|_2$. This is because

$$|\langle\Pi_K g,\, \Pi_K h\rangle - \langle\Pi_K g,\, \Pi_K h'\rangle| \le \|\Pi_K g\|_2\|\Pi_K h - \Pi_K h'\|_2 \le \|\Pi_K g\|_2\|h - h'\|_2,$$

where we used Cauchy-Schwartz inequality and the non-expansive property of convex projection. As a consequence of inequality (A.41) and Cauchy-Schwartz inequality, the term T_1 can be upper bounded as

$$\begin{aligned}T_1 &\le \mathbb{E}_g \exp\left(\lambda\langle\Pi_K g,\, \mathbb{E}\Pi_K g'\rangle + \frac{\lambda^2\|\Pi_K g\|_2^2}{2}\right) \\ &\le \underbrace{\sqrt{\mathbb{E}_g \exp\left(2\lambda\langle\Pi_K g,\, \mathbb{E}\Pi_K g'\rangle\right)}}_{:=T_2}\underbrace{\sqrt{\mathbb{E}_g \exp\left(\lambda^2\|\Pi_K g\|_2^2\right)}}_{:=T_3}.\end{aligned} \tag{A.44}$$

We now control T_2, T_3 separately. For T_2, note again that $h \mapsto \langle\Pi_K h,\, \mathbb{E}\Pi_K g'\rangle$ is a Lipschitz function with Lipschitz constant equal to $\|\mathbb{E}\Pi_K g'\|_2$. Inequality (A.41) implies therefore that

$$T_2 \le \sqrt{\exp\left(2\lambda\langle\mathbb{E}\Pi_K g,\, \mathbb{E}\Pi_K g'\rangle + 2\lambda^2\|\mathbb{E}\Pi_K g'\|_2^2\right)}. \tag{A.45}$$

To control quantity T_3, we use a result from [3, Sublemma E.3] on the moment generating function of $\|\Pi_K g\|^2$ which gives

$$T_3 \le \sqrt{\exp\left(\lambda^2\mathbb{E}(\|\Pi_K g\|_2^2) + \frac{2\lambda^4\mathbb{E}(\|\Pi_K g\|_2^2)}{1-4\lambda^2}\right)}, \qquad \text{whenever } \lambda < 1/4. \tag{A.46}$$

Because of the assumption that $\epsilon^2 \le (\mathbb{E}\|\Pi_K g\|_2)^2/32$, we have $\lambda \le 1/8 < 1/4$. Therefore putting all the pieces together as above, we obtain

$$\begin{aligned}\mathbb{E}_{\eta,\eta'} \exp(\epsilon^2\langle \eta,\, \eta'\rangle) &\le \frac{1}{\mathbb{P}(\mathcal{A}(g))^2} \exp\Bigg((\lambda+\lambda^2)\|\mathbb{E}\Pi_K g\|_2^2 + \frac{\lambda^2\mathbb{E}(\|\Pi_K g\|_2^2)}{2} + \frac{\lambda^4\mathbb{E}(\|\Pi_K g\|_2^2)}{1-4\lambda^2} \Bigg) \\ &\le \frac{1}{\mathbb{P}(\mathcal{A}(g))^2} \exp\big(1.25\lambda\|\mathbb{E}\Pi_K g\|_2^2 + 2.5\lambda^2\mathbb{E}(\|\Pi_K g\|_2^2)\big) \\ &= \frac{1}{\mathbb{P}(\mathcal{A}(g))^2} \exp\Bigg(\frac{5\epsilon^2\|\mathbb{E}\Pi_K g\|_2^2}{(\mathbb{E}(\|\Pi_K g\|_2^2)} + \frac{40\epsilon^4\mathbb{E}(\|\Pi_K g\|_2^2)}{(\mathbb{E}\|\Pi_K g\|_2)^4}) \Bigg) .\end{aligned}$$

This completes the proof of inequality (A.40).

A.7 Completion of the proof of Proposition 3.3.2 and the monotone cone

In this appendix, we collect various results related to the monotone cone, and the proof of Proposition 3.3.2.

A.7.1 Proof of Lemma 3.3.1

So as to simplify notation, we define $\xi = \Pi_K g$, with j^{th} coordinate denoted as ξ_j. Moreover, for a given vector $g \in \mathbb{R}^d$ and integers $1 \le u < v \le d$, we define the u to v average as

$$\bar{g}_{uv} := \frac{1}{v-u+1} \sum_{j=u}^{v} g_j .$$

To demonstrate an upper bound for the inner product $\inf_{\eta \in K \cap S^{d-1}} \langle \eta,\, \mathbb{E}\Pi_K g\rangle$, it turns out that it is enough to take $\eta = \frac{1}{\sqrt{2}}(-1, 1, 0, \ldots, 0) \in K \cap S^{d-1}$ and uses the fact that

$$\inf_{\eta \in K \cap S^{d-1}} \langle \eta,\, \mathbb{E}\Pi_K g\rangle \le \frac{1}{\sqrt{2}} \mathbb{E}(\xi_2 - \xi_1). \tag{A.47}$$

So it is only left for us to analyze $\mathbb{E}(\xi_2 - \xi_1)$ which actually has an explicit form based on the explicit representation of projection to the monotone cone (see Robertson et al. [121], Chapter 1) where

$$\xi_i = \lambda_i - \bar{\lambda}, \qquad \lambda_i = \max_{u \le j} \min_{v \ge j} \bar{g}_{uv}. \tag{A.48}$$

This is true because projecting to cone $K = M \cap L^{\perp}$ can be written into two steps $\Pi_K g = \Pi_{L^{\perp}}(\Pi_M g)$ and projecting to subspace $L^{\perp}$ only shifts the vector to be mean zero.

We claim that the difference satisfies

$$\xi_2 - \xi_1 \leq \max_{v \geq 2} |\bar{g}_{2v}| + \max_{v \geq 1} |\bar{g}_{1v}|. \tag{A.49}$$

To see this, as a consequence of expression (A.48), we have

$$\xi_2 - \xi_1 = \max\{\min_{v\geq 2} \bar{g}_{1v}, \ \min_{v \geq 2} \bar{g}_{2v}\} - \min_{v \geq 1} \bar{g}_{1v}.$$

The right hand side above only takes value in set $\{\min_{v\geq 2} \bar{g}_{1v} - g_1, \ 0, \ \min_{v\geq 2} \bar{g}_{2v} - \min_{v \geq 1} \bar{g}_{1v}\}$ where the last two values agree with bound (A.49) obviously while the first value can be written as

$$\min_{v \geq 2} \bar{g}_{1v} - g_1 = \min_{v \geq 2} \left(\frac{1}{v} \sum_{i=2}^{v} g_i - (1 - \frac{1}{v}) g_1 \right) = \min_{v \geq 2} (1 - \frac{1}{v})(\bar{g}_{2v} - g_1) \leq |\bar{g}_{2v}| + |g_1|,$$

which also agrees with inequality (A.49).

Next let us prove that for every $j = 1, 2$, we have

$$\mathbb{E} \max_{v \geq j} |\bar{g}_{jv}| < 20\sqrt{2}, \tag{A.50}$$

and combine this fact with expressions (A.49) and (A.47) gives us $\inf\limits_{\eta \in K \cap S^{d-1}} \langle \eta, \, \mathbb{E}\Pi_K g \rangle \leq 40$ which validates the conclusion in Lemma 3.3.1.

It is only left for us to verify inequality (A.50). First as we can partition the interval $[j, d]$ into k smaller intervals where each smaller interval is of length 2^m except the last one, then

$$\mathbb{E} \max_{j \leq v \leq d} |\bar{g}_{jv}| = \mathbb{E} \max_{1 \leq m \leq k} \max_{v \in I_m} |\bar{g}_{jv}| \leq \sum_{m=1}^{k} \mathbb{E} \max_{v \in I_k} |\bar{g}_{jv}|, \tag{A.51}$$

where $I_m = [2^m + j - 2, 2^{m+1} + j - 3], \ 1 \leq m < k$, the number of intervals k and length of I_k are chosen to make those intervals sum up to d.

Given index $2^m + j - 2 \leq v \leq 2^{m+1} + j - 3$, random variables $\bar{g}_{jv}$ are Gaussian distributed with mean zero and variance $1/(v - j + 1)$. Suppose we have Gaussian random variable X_v with mean zero and variance $\sigma_m^2 = 1/(2^m - 1)$ and the covariance satisfies $\text{cov}(X_v, X_{v'}) = \text{cov}(\bar{g}_{jv}, \bar{g}_{jv'})$. Since $\sigma_m^2 \geq 1/(v - j + 1)$, the variable $\max_{v \in I_m} |\bar{g}_{jv}|$ is stochastically dominated by the maximum $\max_{2^m \leq v \leq 2^{m+1}-1} |X_v|$, and therefore

$$\sum_{m=1}^{k} \mathbb{E} \max_{v \in I_m} |\bar{g}_{jv}| \leq \sum_{m=1}^{k} \mathbb{E} \max_{2^m \leq v \leq 2^{m+1}-1} |X_v|.$$

Applying the fact that for $t \geq 2$ number of Gaussian random variable $\epsilon_i \sim N(0, \sigma^2)$, we have $\mathbb{E} \max_{1 \leq i \leq t} |\epsilon_i| \leq 4\sigma\sqrt{2 \log t}$ which gives

$$\sum_{m=1}^{k} \mathbb{E} \max_{v \in I_m} |\bar{g}_{jv}| \leq \sum_{m=1}^{k} 4\sigma_m \sqrt{2 \log(2^m)} = 4\sqrt{2 \log 2} \left(\sum_{m=1}^{k} \sqrt{\frac{m}{2^m - 1}} \right). \tag{A.52}$$

The last step is to control the sum $\sum_{m=1}^{k}\sqrt{\frac{m}{2^m-1}}$. There are many ways to show that it is upper bounded by some constant. One crude way is use the fact that $\frac{\sqrt{m}}{2^m-1} \leq 2^{m/4}$ whenever $m \geq 5$, therefore we have

$$\begin{aligned}\sum_{m=1}^{k}\sqrt{\frac{m}{2^m-1}} = \sum_{m=1}^{4}\sqrt{\frac{m}{2^m-1}} + \sum_{m=5}^{k}\sqrt{\frac{m}{2^m-1}} &< \sum_{m=1}^{4}\sqrt{\frac{m}{2^m-1}} + \sum_{m=5}^{k}\frac{1}{2^{m/4}} \\ &< \sum_{m=1}^{4}\sqrt{\frac{m}{2^m-1}} + \frac{2^{-5/4}}{1-2^{-1/4}} < 6,\end{aligned}$$

which validates inequality (A.50) when combined with inequalities (A.51) and (A.52). This completes the proof of Lemma 3.3.1.

A.7.2 Proof of Lemma A.3.1

The proof of Lemma A.3.1 involves two parts. First, we define the matrices G, F. Then we prove that the distribution of η has the right support where we make use of Lemma A.3.2.

As stated, matrix G is a lower triangular matrix satisfying (A.12a). Let us now specify the matrix F. Recall that we denote $\delta := r^{-2}$ and $r := 1/3$. To define matrix F, let us first define a partition of $[d]$ into m consecutive intervals $\{I_1, \ldots, I_m\}$ with m specified in expression (A.7) and the length of each interval $|I_i| = \ell_i$ where ℓ_i is defined as

$$\ell_i := \lfloor \frac{\delta-1}{\delta^i}(d + \log_\delta d + 3)\rfloor, \qquad 1 \leq i \leq m-1, \tag{A.53}$$

and $\ell_m := d - \sum_{i=1}^{m-1}\ell_i$.

Following directly from the definition (A.53), each length $\ell_i \geq 1$ and ℓ_i is a decreasing sequence with regard to i. Also ℓ_i satisfies the following

$$\ell_1 = \lfloor \frac{\delta-1}{\delta}(d + \log_\delta d + 3)\rfloor < d \qquad \text{and} \quad \ell_i \geq \delta\ell_{i+1},\ \text{for } 1 \leq i \leq m-1, \tag{A.54}$$

where the first inequality holds since as $\sqrt{\log(ed)} \geq 14$, we have $(\delta-1)(\log_\delta d + 3) \leq d$ and the last inequality follows from the fact that $\lfloor ab \rfloor \geq a\lfloor b \rfloor$ for positive integer a and $b \geq 0$ (because $a\lfloor b \rfloor$ is an integer that is smaller than ab).

We are now ready to define the $d \times m$ matrix F. We take

$$F(i,j) = \begin{cases} \frac{1}{\sqrt{\ell_j}} & i \in I_j, \\ 0 & \text{otherwise.} \end{cases} \tag{A.55}$$

It is easy to check that matrix F satisfies $F^TF = \mathbb{I}_m$ which validates inequality (A.12b).

First we show that both $\eta = FGb$ and $\eta - \bar{\eta}\mathbf{1}$ belong to M. The i-th coordinate of η can be written as

$$\eta_i = \frac{1}{\sqrt{\ell_j}} \sum_{t=1}^{j} r^{j-t} b_t, \qquad \forall\ i \in I_j.$$

Therefore we can denote u_j as the value of η_i for $i \in I_j$. To establish monotonicity, we only need to compare the value in the consecutive blocks. Direct calculation of the consecutive ratio yields

$$\frac{u_{j+1}}{u_j} = \frac{r(\sum_{t=1}^{j} r^{j-t} b_t) + b_{j+1}}{\sqrt{\ell_{j+1}}} \frac{\sqrt{\ell_j}}{\sum_{t=1}^{j} r^{j-t} b_t} \geq r\sqrt{\frac{\ell_j}{\ell_{j+1}}} \geq 1,$$

where we used the non-negativity of coordinates of vector b and the last inequality follows from inequality (A.54) and $\delta = r^{-2}$. The monotonicity of $\eta - \bar{\eta}\mathbf{1}$ thus inherits directly from the monotonicity of η.

To complete the proof of Lemma A.3.1, we only need to prove lower bounds on $\|\eta\|_2$ and $\|\eta - \bar{\eta}\|_2$. For these, we shall use inequality (A.13b) of Lemma A.3.2.

Proof of the bound $\|\eta\|_2 \geq 1$: Recall that $r = 1/3$ and as a direct consequence of inequality (A.13b) in Lemma A.3.2, we have

$$\langle \eta, \eta \rangle = \|Gb\|_2^2 \ \geq \ \frac{9}{4} - \frac{63}{32s} > 1.96, \tag{A.56}$$

where the last step follows form the fact that $s = \lfloor \sqrt{m} \rfloor \geq 7$. Therefore, the norm condition holds so η is supported on $M \cap L^T \cap B^c(1)$.

Proof of the bound $\|\eta - \bar{\eta}\mathbf{1}\|_2 \geq 1$: The norm $\|\eta - \bar{\eta}\mathbf{1}\|_2^2$ has the following decomposition where

$$\|\eta - \bar{\eta}\mathbf{1}\|_2^2 = \|\eta\|_2^2 - d(\bar{\eta})^2.$$

We claim that $d(\bar{\eta})^2 \leq 0.2$. If we take this for now, combining with inequality (A.56) which says $\|\eta\|_2^2$ is greater than 1.96, we can deduce that $\|\eta - \bar{\eta}\mathbf{1}\|_2^2 \geq 1$. So it suffices to verify the claim $d(\bar{\eta})^2 \leq 0.2$. Recall that $\eta = FGb$. Direct calculation yields

$$d\bar{\eta} = \langle \mathbf{1}, \eta \rangle = \mathbf{1}^T \cdot FGb = \sum_{k=1}^{m} b_k \underbrace{\sum_{i=k}^{m} \sqrt{\ell_i} r^{i-k}}_{:=a_k}.$$

Plugging into the definitions of r and ℓ_i guarantees that

$$a_k \le \sum_{i=k}^{m} \sqrt{\frac{(\delta-1)(d+\log_\delta d+3)}{\delta^i}} \frac{1}{\delta^{(i-k)/2}} = \sqrt{(\delta-1)(d+\log_\delta d+3)\delta^k} \sum_{i=k}^{m} \delta^{-i}$$
$$\le \sqrt{\frac{(d+\log_\delta d+3)}{(\delta-1)\delta^{k-2}}},$$

where the last step uses the summability of a geometric sequence—namely $\sum_{i=k}^{m} \delta^{-i} \le \delta^{-k+1}/(\delta-1)$. Now for every vector b, our goal is to control $\sum a_k b_k$. Recall that every vector b has s non-zero entries which equal to $1/\sqrt{s}$ where $s = \lfloor \sqrt{m} \rfloor$. Since a_k decreases with k, this inner product $\sum a_k b_k$ is largest when the first s coordinates of b are non-zero, therefore

$$d\bar{\eta} \le \sum_{k=1}^{s} a_k \frac{1}{\sqrt{s}} \le \frac{1}{\sqrt{s}} \sqrt{\frac{\delta^2(d+\log_\delta d+3)}{\delta-1}} \sum_{k=1}^{s} \frac{1}{\delta^{k/2}} \le \frac{1}{\sqrt{s}} \sqrt{\frac{\delta^2(d+\log_\delta d+3)}{\delta-1}} \frac{1}{\sqrt{\delta}-1},$$

and thus we have

$$d(\bar{\eta})^2 \le \frac{1}{\sqrt{m}-1} \frac{(d+\log_\delta d+3)}{d} \frac{\delta^2}{(\delta-1)(\sqrt{\delta}-1)^2} \le \frac{81(d+\log_\delta d+3)}{32d(\sqrt{m}-1)} < 0.2,$$

where the last step uses $\sqrt{m} \ge 8$. Therefore, the norm condition also holds so $\eta - \bar{\eta}\mathbf{1}$ is supported on $M \cap L^T \cap B^c(1)$.

Thus, we have completed the proof of Lemma A.3.1.

A.7.3 Proof of Lemma A.3.2

By definition of the matrix G, we have

$$\langle Gb,\, Gb' \rangle = \sum_{t=1}^{m} (Gb)_t (Gb')_t = \sum_{t=1}^{m} (b_t + rb_{t-1} + \cdots + r^{t-1}b_1)(b'_t + rb'_{t-1} + \cdots + r^{t-1}b'_1)$$
$$= \sum_{t=1}^{m} \sum_{u=1}^{t} \sum_{v=1}^{t} r^{2t-u-v} b_u b'_v.$$

Switching the order of summation yields

$$\begin{aligned}\langle Gb,\, Gb' \rangle &= \sum_{u=1}^{m} \sum_{v=1}^{m} b_u b'_v \sum_{t=\max\{u,v\}}^{m} r^{2t-u-v} \\ &= \sum_{u=1}^{m} \sum_{v=1}^{m} \frac{b_u b'_v}{r^{u+v}} \frac{r^{2\max\{u,v\}} - r^{2m+2}}{1-r^2} \\ &= \underbrace{\frac{1}{1-r^2} \sum_{u=1}^{m} \sum_{v=1}^{m} b_u b'_v r^{|u-v|}}_{:=\Delta_1} - \underbrace{\frac{1}{1-r^2} \sum_{u=1}^{m} \sum_{v=1}^{m} b_u b'_v r^{2m+2-u-v}}_{:=\Delta_2}. \end{aligned} \tag{A.57}$$

We bound the two terms Δ_1 and Δ_2 separately.

Recall the fact that b, b' belong to $\mathcal{S}$, so there are exactly $s = \lfloor\sqrt{m}\rfloor$ non-zero entry in both b and b' and these entries equal to $1/\sqrt{s}$. The summation defining Δ_1 is not affected by the permutation of coordinates, so that we can assume without loss of generality that the indices of non-zero entries in b are indexed by $\{1, \ldots, s\}$, and that the indices of non-zero entries in b' are indexed by $\{k, k+1, \ldots, k+s-1\}$ for some $1 \le k \le m+1-s$.

We split our proof into two cases depending on whether $k \le s$ or $k > s$.

Case 1 ($k \le s$): The summation Δ_1 can be written as

$$s(1-r^2)\Delta_1 = s\sum_{u=1}^{m}\sum_{v=1}^{m} b_u b'_v r^{|u-v|} = \sum_{u=1}^{s}\sum_{v=k}^{k+s-1} r^{|u-v|}.$$

Direct calculation yields

$$\begin{aligned}
s(1-r^2)\Delta_1 &= \sum_{u=1}^{k-1}\sum_{v=k}^{k+s-1} r^{v-u} + \sum_{u=k}^{s}\sum_{v=k}^{u} r^{u-v} + \sum_{u=k}^{s}\sum_{v=u+1}^{k+s-1} r^{v-u} \\
&= \frac{(1-r^s)(r-r^k)}{(1-r)^2} + \frac{s-k+1}{1-r} - \frac{r}{(1-r)^2}(1-r^{s-k+1}) + \frac{r(s-k+1)}{1-r} - \frac{r^k - r^{s+1}}{(1-r)^2} \\
&= \frac{1+r}{1-r}(s-k+1) + \frac{r^k(r^s + r^{s+2} - 2)}{(1-r)^2}.
\end{aligned}$$

Notice the following two facts that

$$\langle b,\, b'\rangle = \frac{s-k+1}{s} \qquad \text{and} \quad \frac{-2r}{(1-r)^2} \le \frac{r^k(r^s + r^{s+2} - 2)}{(1-r)^2} < 0,$$

so that

$$\frac{1}{(1-r)^2}\langle b,\, b'\rangle + \frac{-2r}{s(1-r^2)(1-r)^2} \le \Delta_1 \le \frac{1}{(1-r)^2}\langle b,\, b'\rangle. \tag{A.58}$$

Case 2 ($k > s$): The summation Δ_1 satisfies the bounds

$$s(1-r^2)\Delta_1 = s\sum_{u=1}^{m}\sum_{v=1}^{m} b_u b'_v r^{|u-v|} = \sum_{u=1}^{s}\sum_{v=k}^{k+s-1} r^{v-u} = \frac{r^{k-s}(1-r^s)^2}{(1-r)^2}.$$

Since $k - s \ge 1$, we have $\langle b,\, b'\rangle = 0$ and consequently

$$\Delta_1 \le \frac{1}{(1-r)^2}\langle b,\, b'\rangle + \frac{r}{s(1-r^2)(1-r)^2}. \tag{A.59}$$

Combining inequalities (A.57), (A.58) and (A.59), we can deduce that

$$\langle Gb,\, Gb' \rangle \;\le\; \Delta_1 \;\le\; \frac{1}{(1-r)^2}\langle b,\, b' \rangle + \frac{r}{s(1-r^2)(1-r)^2},$$

which validates inequality (A.13a).

On the other hand, when $b = b'$, the summation Δ_2 is the largest when the non-zero entries of b lie on coordinates $m-s+1, \ldots, m$. Thus we have

$$s(1-r^2)\Delta_2 \le \sum_{u=m-s+1}^{m} \sum_{v=m-s+1}^{m} r^{2m+2-u-v} = \frac{r^2(1-r^s)^2}{(1-r)^2} < \frac{r^2}{(1-r)^2}. \tag{A.60}$$

Combining decomposition (A.57) with the inequalities (A.58), we can deduce that

$$\langle Gb,\, Gb \rangle \le \frac{1}{(1-r)^2} - \frac{2r}{s(1-r^2)(1-r)^2} - \frac{r^2}{s(1-r^2)(1-r)^2},$$

where we use the fact that $\langle b,\, b \rangle = 1$. This completes the proof of inequality (A.13b).

Appendix B

Proofs for Chapter 4

This chapter is organized as follows. We complete the proofs of Theorems 4.3.1 and 4.3.2 in Subsections B.1.1 and B.1.2 respectively. The proof of Inequality (4.21) in Remark 4.3.2 is given in Subsection B.1.3. The proofs of the corollaries of Section 4.3.1 are given in Subsection B.1.4. The proof of Theorem 4.3.3 is completed in Subsection B.1.5. Technical lemmas which were crucially used in the proofs of the main results are stated and proved in Subsection B.1.6.

Finally, note that additional simulations (similar to those in the main text) are presented in Section B.2.

B.1 Additional proofs and technical results

B.1.1 Completion of the proof of Theorem 4.3.1

We use the same notation as in the proof of Theorem 4.3.1 in the main text. To complete the proof, we need to prove inequality (4.49).

Below, we write $\Delta_k, \hat{k}$ and k_* for $\Delta_k(\theta_i), \hat{k}(i)$ and $k_*(i)$ respectively for ease of notation. We also write $\mathbb{P}$ for $\mathbb{P}_{K^*}$.

We prove (4.49) by considering the two cases: $k \le k_*, k \in \mathcal{I}$ and $k > k_*, k \in \mathcal{I}$ separately.

The first case is $k \le k_*, k \in \mathcal{I}$. By Lemma B.1.2 and (B.44), we get

$$\Delta_k \le \Delta_{k_*} \le \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k_*+1}} \le \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k+1}}$$

and consequently

$$\Delta_k^2 + \frac{\sigma^2}{k+1} \le \frac{\sigma^2}{k+1}\left(36(\sqrt{2}-1)^2+1\right) \qquad \text{for all } k \le k_*, k \in \mathcal{I}. \tag{B.1}$$

We bound $\mathbb{P}\{\hat{k} = k\}$ from above by

$$\mathbb{P}\left\{\left(\hat{\Delta}_k\right)_+ + \frac{2\sigma}{\sqrt{k+1}} \le \left(\hat{\Delta}_{k_*}\right)_+ + \frac{2\sigma}{\sqrt{k_*+1}}\right\} \le \mathbb{P}\left\{\left(\hat{\Delta}_{k_*}\right)_+ \ge \frac{2\sigma}{\sqrt{k+1}} - \frac{2\sigma}{\sqrt{k_*+1}}\right\}.$$

Because $k \leq k_*$, the positive part above can be dropped and we obtain

$$\mathbb{P}\{\hat{k} = k\} \leq \mathbb{P}\left\{\hat{\Delta}_{k_*} \geq \frac{2\sigma}{\sqrt{k+1}} - \frac{2\sigma}{\sqrt{k_*+1}}\right\}.$$

Because $\hat{\Delta}_{k_*}$ is normally distributed with mean Δ_{k_*}, we have

$$\mathbb{P}\{\hat{k} = k\} \leq \mathbb{P}\left\{Z \geq \frac{2\sigma(k+1)^{-1/2} - 2\sigma(k_*+1)^{-1/2} - \Delta_{k_*}}{\sqrt{\mathrm{var}(\hat{\Delta}_{k_*})}}\right\},$$

where Z is a standard normal random variable. From (B.44), we have

$$\frac{2\sigma}{\sqrt{k+1}} - \frac{2\sigma}{\sqrt{k_*+1}} - \Delta_{k_*} \geq \frac{2\sigma}{\sqrt{k+1}}\left(1 - \sqrt{\frac{k+1}{k_*+1}}\left(3\sqrt{2}-2\right)\right).$$

As a result,

$$\mathbb{P}\{\hat{k} = k\} \leq \mathbb{P}\left\{Z \geq \frac{2\sigma}{\sqrt{(k+1)\mathrm{var}(\hat{\Delta}_{k_*})}}\left(1 - \sqrt{\frac{k+1}{k_*+1}}\left(3\sqrt{2}-2\right)\right)\right\}.$$

Suppose $\tilde{k} := (k_*+1)\left(3\sqrt{2}-2\right)^{-2} - 1$. For $k < \tilde{k}$, we use the bound given by Lemma B.1.4 on the variance of $\hat{\Delta}_{k^*}$ to obtain

$$\mathbb{P}\{\hat{k} = k\} \leq \mathbb{P}\left\{Z \geq 2\left(\sqrt{\frac{k_*+1}{k+1}} - 3\sqrt{2} + 2\right)\right\} \leq \exp\left(-2\left[\sqrt{\frac{k_*+1}{k+1}} - 3\sqrt{2} + 2\right]^2\right).$$

Using this and (B.1), we see that the quantity

$$\sum_{k<\tilde{k}, k\in\mathcal{I}}\left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k} = k\}}$$

is bounded from above by

$$\frac{\sigma^2}{k_*+1}\left(36(\sqrt{2}-1)^2 + 1\right)\sum_{k<\tilde{k}, k\in\mathcal{I}}\frac{k_*+1}{k+1}\exp\left(-\left[\sqrt{\frac{k_*+1}{k+1}} - 3\sqrt{2} + 2\right]^2\right).$$

Because $\mathcal{I}$ consists of integers of the form 2^j, it follows that for any two successive integers k_1 and k_2 in $\mathcal{I}$, we have $3/2 \leq (k_1+1)/(k_2+1) \leq 2$. Using this, it is easily seen that

$$\sum_{k<\tilde{k}, k\in\mathcal{I}}\frac{k_*+1}{k+1}\exp\left(-\left[\sqrt{\frac{k_*+1}{k+1}} - 3\sqrt{2} + 2\right]^2\right)$$

is bounded from above by

$$\sum_{j\geq 4} 2^j \exp\left(-\left[(3/2)^{j/2} - 3\sqrt{2} + 2\right]^2\right) + \sum_{0\leq j\leq 3} 2^j,$$

which is just a universal positive constant. We have proved therefore that

$$\sum_{k<\tilde{k}, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \frac{C_1\sigma^2}{k_*+1}, \tag{B.2}$$

for a positive constant C_1.

For $\tilde{k} \leq k \leq k_*$, we simply use (B.1) along with the trivial bound $\mathbb{P}\{\hat{k}=k\} \leq 1$ to get

$$\sum_{\tilde{k}\leq k\leq k_*, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \left(36(\sqrt{2}-1)^2+1\right)\frac{\sigma^2}{k_*+1}\sum_{\tilde{k}\leq k<k_*, k\in\mathcal{I}} \frac{k_*+1}{k+1}.$$

Once again because $\mathcal{I}$ consists of integers of the form 2^j, we get

$$\sum_{\tilde{k}\leq k\leq k_*, k\in\mathcal{I}} \frac{k_*+1}{k+1} \leq \sum_{j\geq 0} 2^j \left\{(3/2)^j \leq \left(3\sqrt{2}-2\right)^2\right\}.$$

The right hand side above is just a constant. It follows therefore that

$$\sum_{\tilde{k}\leq k\leq k_*, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \frac{C_2\sigma^2}{k_*+1}, \tag{B.3}$$

for a positive constant C_2. Combining (B.2) and (B.3), we deduce that

$$\sum_{k\leq k_*, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \frac{C\sigma^2}{k_*+1} \tag{B.4}$$

where $C := C_1 + C_2$ is a universal positive constant.

To complete the proof of Theorem 4.3.1, we need to deal with the case $k > k_*, k \in \mathcal{I}$ and prove that

$$\sum_{k>k_*, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \frac{C\sigma^2}{k_*+1} \tag{B.5}$$

for a constant C. Assume that $\{k \in \mathcal{I} : k > k_*\}$ is non-empty for otherwise there is nothing to prove. By the first part of (B.45) in Lemma B.1.3, we get

$$\sum_{k>k_*, k\in\mathcal{I}} \left(\Delta_k^2 + \frac{\sigma^2}{k+1}\right)\sqrt{\mathbb{P}\{\hat{k}=k\}} \leq \left(1 + \frac{1}{(\sqrt{6}-2)^2}\right)\sum_{k>k_*, k\in\mathcal{I}} \Delta_k^2\sqrt{\mathbb{P}\{\hat{k}=k\}}. \tag{B.6}$$

We first bound $\mathbb{P}\{\hat{k}=k\}$ for $k>k_*, k\in\mathcal{I}$. We proceed by writing

$$\begin{aligned}
\mathbb{P}\{\hat{k}=k\} &\leq \mathbb{P}\left\{\hat{\Delta}_k^+ + \frac{2\sigma}{\sqrt{k+1}} \leq \hat{\Delta}_{k_*}^+ + \frac{2\sigma}{\sqrt{k_*+1}}\right\} \\
&\leq \mathbb{P}\left\{\hat{\Delta}_k + \frac{2\sigma}{\sqrt{k+1}} \leq \hat{\Delta}_{k_*}^+ + \frac{2\sigma}{\sqrt{k_*+1}}\right\} \qquad \text{(because } x\leq x^+\text{)} \\
&\leq \mathbb{P}\left\{\hat{\Delta}_k + \frac{2\sigma}{\sqrt{k+1}} \leq \hat{\Delta}_{k_*} + \frac{2\sigma}{\sqrt{k_*+1}}\right\} + \mathbb{P}_K\left\{\hat{\Delta}_k + \frac{2\sigma}{\sqrt{k+1}} \leq \frac{2\sigma}{\sqrt{k_*+1}}\right\} \\
&\leq \mathbb{P}\left\{\hat{\Delta}_k \leq \hat{\Delta}_{k_*} + \frac{2\sigma}{\sqrt{k_*+1}}\right\} + \mathbb{P}_K\left\{\hat{\Delta}_k \leq \frac{2\sigma}{\sqrt{k_*+1}}\right\} \\
&\leq \mathbb{P}\left\{\hat{\Delta}_{k_*} - \hat{\Delta}_k \geq -\frac{2\sigma}{\sqrt{k_*+1}}\right\} + \mathbb{P}\left\{-\hat{\Delta}_k \geq -\frac{2\sigma}{\sqrt{k_*+1}}\right\}
\end{aligned}$$

Both $\hat{\Delta}_{k_*}-\hat{\Delta}_k$ and $\hat{\Delta}_k$ are normally distributed with means $\Delta_{k_*}-\Delta_k$ and Δ_k respectively. As a result

$$\mathbb{P}\{\hat{k}=k\} \leq \mathbb{P}\left\{Z \geq \frac{\Delta_k - \Delta_{k_*} - 2\sigma(k_*+1)^{-1/2}}{\sqrt{\operatorname{var}(\hat{\Delta}_{k_*}-\hat{\Delta}_k)}}\right\} + \mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}}{\sqrt{\operatorname{var}(\hat{\Delta}_k)}}\right\}$$

where Z is a standard normal random variable. Using (B.44) in Lemma B.1.3, we obtain

$$\mathbb{P}\{\hat{k}=k\} \leq \mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}\left(3\sqrt{2}-2\right)}{\sqrt{\operatorname{var}(\hat{\Delta}_{k_*}-\hat{\Delta}_k)}}\right\} + \mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}}{\sqrt{\operatorname{var}(\hat{\Delta}_k)}}\right\}.$$

By the Cauchy-Schwarz inequality and Lemma B.1.4, we get, for $k>k_*$,

$$\sqrt{\operatorname{var}(\hat{\Delta}_{k_*}-\hat{\Delta}_k)} \leq \sqrt{\operatorname{var}(\hat{\Delta}_{k_*})} + \sqrt{\operatorname{var}(\hat{\Delta}_k)} \leq \frac{\sigma}{\sqrt{k+1}} + \frac{\sigma}{\sqrt{k_*+1}} \leq \frac{2\sigma}{\sqrt{k_*+1}}$$

Also $\operatorname{var}(\hat{\Delta}_k) \leq \sigma^2/(k+1) \leq \sigma^2/(k_*+1)$. Therefore if $k>k_*, k\in\mathcal{I}$ is such that

$$\Delta_k \geq 2\sigma(k_*+1)^{-1/2}\left(3\sqrt{2}-2\right), \tag{B.7}$$

we obtain

$$\begin{aligned}
\mathbb{P}\{\hat{k}=k\} &\leq \mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}\left(3\sqrt{2}-2\right)}{\sigma\sqrt{2}(k_*+1)^{-1/2}}\right\} + \mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}}{\sigma(k_*+1)^{-1/2}}\right\} \\
&\leq 2\mathbb{P}\left\{Z \geq \frac{\Delta_k - 2\sigma(k_*+1)^{-1/2}\left(3\sqrt{2}-2\right)}{\sigma\sqrt{2}(k_*+1)^{-1/2}}\right\} \\
&\leq 2\exp\left(-\frac{k_*+1}{2\sigma^2}\left(\Delta_k - 2\sigma(k_*+1)^{-1/2}(3\sqrt{2}-2)\right)^2\right).
\end{aligned}$$

Using the inequality $(x-y)^2 \geq x^2/2 - y^2$ with $x = \Delta_k$ and $y = 2\sigma(k_*+1)^{-1/2}(3\sqrt{2}-2)$, we obtain

$$\mathbb{P}\{\hat{k} = k\} \leq 2\exp\left(2(3\sqrt{2}-2)^2\right)\exp\left(-\frac{(k_*+1)\Delta_k^2}{4\sigma^2}\right) \tag{B.8}$$

whenever $k \in I, k > k_*$ satisfies (B.7). It is easy to see that when (B.7) is not satisfied, the right hand side above is larger than 2. Thus, inequality (B.8) is true for all $k \in \mathcal{I}, k > k_*$. As a result,

$$\Delta_k^2\sqrt{\mathbb{P}\{\hat{k} = k\}} \leq \sqrt{2}\exp\left((3\sqrt{2}-2)^2\right)\xi\left(\Delta_k^2\right) \qquad \text{for all } k \in \mathcal{I}, k > k_*. \tag{B.9}$$

where

$$\xi(z) := z\exp\left(-\frac{(k_*+1)z}{8\sigma^2}\right) \qquad \text{for } z > 0.$$

By (B.6) and (B.9), the proof would therefore be complete if we show that $\sum_{k\in\mathcal{I}:k>k_*}\xi\left(\Delta_k^2\right)$ is bounded from above by a universal positive constant. For this, note first that the function $\xi(z)$ is decreasing for $z \geq \breve{z} := 8\sigma^2/(k_*+1)$ and attains its maximum over $z > 0$ at $z = \breve{z}$. Note also the second part of inequality (B.45) gives $\Delta_k^2 \geq z_k$ for all $k \in \mathcal{I}, k > k_*$ where

$$z_k := \frac{(\sqrt{6}-2)^2\sigma^2(k+1)}{4(k_*+1)^2}$$

We therefore get

$$\begin{aligned}\xi\left(\Delta_k^2\right) \leq \xi(\max(z_k,\breve{z})) &= \max(z_k,\breve{z})\exp\left(\frac{-(k_*+1)\max(z_k,\breve{z})}{8\sigma^2}\right)\\ &\leq \max(z_k,\breve{z})\exp\left(\frac{-(k_*+1)z_k}{8\sigma^2}\right) \leq (z_k+\breve{z})\exp\left(\frac{-(k_*+1)z_k}{8\sigma^2}\right).\end{aligned}$$

Because $k > k_*$, it is easy to see that

$$\breve{z} = \frac{8\sigma^2}{k_*+1} \leq \frac{8\sigma^2(k+1)}{(k_*+1)^2}.$$

We deduce that

$$\xi\left(\Delta_k^2\right) \leq \left[\frac{(\sqrt{6}-2)^2}{4} + 8\right]\frac{\sigma^2(k+1)}{(k_*+1)^2}\exp\left(-\frac{(\sqrt{6}-2)^2}{32}\frac{k+1}{k_*+1}\right).$$

Denoting the constants above by c_1 and c_2, we can write

$$\sum_{k\in\mathcal{I}:k>k_*}\xi\left(\Delta_k^2\right) \leq \frac{c_1\sigma^2}{k_*+1}\sum_{k\in\mathcal{I}:k>k_*}\frac{k+1}{k_*+1}\exp\left(-\frac{k+1}{c_2(k_*+1)}\right).$$

The sum in the right hand side above is easily seen to be bounded from above by

$$\sum_{j \geq 0} 2^j \exp\left(-\frac{1}{c_2}\left(\frac{3}{2}\right)^j\right)$$

which is further bounded by a universal constant. This completes the proof of Theorem 4.3.1.

B.1.2 Completion of the proof of Theorem 4.3.2

We continue from where we left off in the proof of chapter 4. We first work with the case when K^* satisfies the condition (4.51). The idea here is to use Le Cam's bound (4.50) with the choice of L^* given in the proof in the chapter 4. In the remainder of the proof, we use Lemma B.1.5 which is stated and proved in Section B.1.

To control the total variation distance in (4.50), we use Pinsker's inequality:

$$||P_{K^*} - P_{L^*}||_{TV} \leq \sqrt{\frac{1}{2} D(P_{K^*}||P_{L^*})},$$

and the fact that (note that $\theta_i = 2\pi i/n - \pi$)

$$D(P_{K^*}||P_{L^*}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(h_{K^*}(2i\pi/n - \pi) - h_{L^*}(2i\pi/n - \pi)\right)^2$$

where $D(P_{K^*}\|P_{L^*})$ denotes the Kullback-Leibler divergence between the probability measures P_{K^*} and P_{L^*}.

The support function of L^* is easily seen to be the maximum of the support functions of K^* and the singleton $\{a_{K^*}(\alpha)\}$. Therefore,

$$\begin{aligned} h_{L^*}(\theta) &:= \max\left(h_{K^*}(\theta), \frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha}\cos\theta + \frac{h_{K^*}(\alpha) - h_{K^*}(-\alpha)}{2\sin\alpha}\sin\theta\right) \\ &= \max\left(h_{K^*}(\theta), \frac{\sin(\theta+\alpha)}{\sin 2\alpha} h_{K^*}(\alpha) + \frac{\sin(\alpha-\theta)}{\sin 2\alpha} h_{K^*}(-\alpha)\right). \end{aligned}$$

Using (4.1), it can be shown that

$$h_{K^*}(\theta) \leq \frac{\sin(\theta+\alpha)}{\sin 2\alpha} h_{K^*}(\alpha) + \frac{\sin(\alpha-\theta)}{\sin 2\alpha} h_{K^*}(-\alpha) \tag{B.10}$$

for $-\alpha < \theta < \alpha$ and

$$h_{K^*}(\theta) \geq \frac{\sin(\theta+\alpha)}{\sin 2\alpha} h_{K^*}(\alpha) + \frac{\sin(\alpha-\theta)}{\sin 2\alpha} h_{K^*}(-\alpha) \tag{B.11}$$

for $\theta \in [-\pi, -\alpha] \cup [\alpha, \pi]$. To see this, assume that $\theta > 0$ without loss of generality. We then work with the two separate cases $\theta \in [0, \alpha]$ and $\theta \in [\alpha, \pi]$. In the first case, apply (4.1) with $\alpha_1 = \alpha, \alpha = \theta$ and $\alpha_2 = -\alpha$ to get (B.10). In the second case, apply (4.1) with $\alpha_1 = \theta, \alpha = \alpha$ and $\alpha_2 = -\alpha$ to get (B.11).

As a result of (B.10) and (B.11), we get that

$$h_{L^*}(\theta) = \frac{\sin(\theta + \alpha)}{\sin 2\alpha} h_{K^*}(\alpha) + \frac{\sin(\alpha - \theta)}{\sin 2\alpha} h_{K^*}(-\alpha) \tag{B.12}$$

for $-\alpha \le \theta \le \alpha$, and that $h_{L^*}(\theta)$ equals $h_{K^*}(\theta)$ for every other θ in $(-\pi, \pi]$.

We now give an upper bound on $h_{L^*}(\theta) - h_{K^*}(\theta)$ for $0 \le \theta < \alpha$. Using (4.1) with $\alpha_1 = \theta, \alpha = 0$ and $\alpha_2 = -\alpha$, we obtain

$$h_{K^*}(\theta) \ge \frac{\sin(\alpha + \theta)}{\sin \alpha} h_{K^*}(0) - \frac{\sin \theta}{\sin \alpha} h_{K^*}(-\alpha).$$

Thus for $0 \le \theta < \alpha$, we obtain the inequality

$$\begin{aligned} 0 \le h_{L^*}(\theta) - h_{K^*}(\theta) &= \frac{\sin(\theta + \alpha)}{\sin 2\alpha} h_{K^*}(\alpha) + \frac{\sin(\alpha - \theta)}{\sin 2\alpha} h_{K^*}(-\alpha) - h_{K^*}(\theta) \\ &\le \frac{\sin(\theta + \alpha)}{\sin \alpha} \left(\frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \right). \end{aligned}$$

Because $0 < \alpha < \pi/4, 0 \le \theta \le \alpha$, we use the fact that the sine function is increasing on $(0, \pi/2)$ to deduce that

$$0 \le h_{L^*}(\theta) - h_{K^*}(\theta) \le \frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \qquad \text{for all } 0 \le \theta < \alpha.$$

One can similarly deduce the same inequality for the case $-\alpha < \theta \le 0$ as well. Because of this and the fact that $h_{L^*}(\theta)$ equals $h_{K^*}(\theta)$ for all θ in $(-\pi, \pi]$ that are not in the interval $(-\alpha, \alpha)$, we obtain

$$\begin{aligned} D(P_{K^*}||P_{L^*}) &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(h_{K^*}(2i\pi/n - \pi) - h_{L^*}(2i\pi/n - \pi) \right)^2 \\ &\le \frac{n_\alpha}{2\sigma^2} \left(\frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \right)^2. \end{aligned}$$

Also because $h_{L^*}(0) = (h_{K^*}(\alpha) + h_{K^*}(-\alpha))/(2\cos\alpha)$, Le Cam's inequality gives

$$r \ge \frac{1}{4} \left(\frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \right)^2 \left(1 - \sqrt{\frac{n_\alpha}{4\sigma^2}} \left(\frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \right) \right) \tag{B.13}$$

for every $0 < \alpha < \pi/4$ where

$$r := \inf_{\tilde{h}} \max \left[\mathbb{E}_{K^*} \left(\tilde{h} - h_{K^*}(\theta_i) \right)^2, \mathbb{E}_{L^*} \left(\tilde{h} - h_{L^*}(\theta_i) \right)^2 \right] \tag{B.14}$$

where the infimum above is over all estimators $\tilde{h}$. Our strategy now is to choose an appropriate $\alpha_* \in (0, \pi/4)$ in order to prove that $r \geq c\sigma^2/(k_* + 1)$ for some positive constant c. Let us now define α_* by

$$\alpha_* := \inf\left\{0 < \alpha < \pi/4 : \frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) > \frac{\sigma}{\sqrt{n_\alpha}}\right\}.$$

Note first that $\alpha_* > 0$ because $n_\alpha \geq 1$ for all α and thus for α very small while the quantity $(h_{K^*}(\alpha) + h_{K^*}(-\alpha))/(2\cos\alpha) - h_{K^*}(0)$ becomes close to 0 for small α (by continuity of $h_{K^*}(\cdot)$).

Also because we have assumed (4.51), it follows that $0 < \alpha_* < \pi/4$. Now for each $\epsilon > 0$ sufficiently small, we have

$$\frac{h_{K^*}(\alpha_* - \epsilon) + h_{K^*}(-\alpha_* + \epsilon)}{2\cos(\alpha_* - \epsilon)} - h_{K^*}(0) \leq \frac{\sigma}{\sqrt{n_{\alpha_* - \epsilon}}}.$$

Letting $\epsilon \downarrow 0$ in the above and using the fact that $n_{\alpha_*-\epsilon} \to n_{\alpha_*}$ and the continuity of h_{K^*}, we deduce

$$\frac{h_{K^*}(\alpha_*) + h_{K^*}(-\alpha_*)}{2\cos\alpha_*} - h_{K^*}(0) \leq \frac{\sigma}{\sqrt{n_{\alpha_*}}}. \tag{B.15}$$

Because $0 < \alpha_* < \pi/4$, by the definition of the infimum, there exists a decreasing sequence $\{\alpha_k\} \in (0, \pi/4)$ converging to α_* such that

$$\frac{h_{K^*}(\alpha_k) + h_{K^*}(-\alpha_k)}{2\cos\alpha_k} - h_{K^*}(0) > \frac{\sigma}{\sqrt{n_{\alpha_k}}} \qquad \text{for all } k.$$

For k large, n_{α_k} is either n_{α_*} or $n_{\alpha_*} + 2$, and hence letting $k \to \infty$, we get

$$\frac{h_{K^*}(\alpha_*) + h_{K^*}(-\alpha_*)}{2\cos\alpha_*} - h_{K^*}(0) \geq \frac{\sigma}{\sqrt{n_{\alpha_*} + 2}} \geq \frac{1}{\sqrt{3}}\frac{\sigma}{\sqrt{n_{\alpha_*}}},$$

where we also used that $n_{\alpha_*} \geq 1$. Combining the above with (B.15), we conclude that

$$\frac{1}{\sqrt{3}}\frac{\sigma}{\sqrt{n_{\alpha_*}}} \leq \frac{h_{K^*}(\alpha_*) + h_{K^*}(-\alpha_*)}{2\cos\alpha_*} - h_{K^*}(0) \leq \frac{\sigma}{\sqrt{n_{\alpha_*}}}.$$

Using $\alpha = \alpha_*$ in (B.13), we get

$$r \geq \frac{\sigma^2}{24 n_{\alpha_*}}. \tag{B.16}$$

We shall now show that

$$\alpha_* \leq \tilde{\alpha} := \frac{8(k_* + 1)\pi}{n} \tag{B.17}$$

when $8(k_* + 1)\pi/n \leq \pi/4$ (otherwise (B.17) is obvious). This would imply, because $\alpha \mapsto n_\alpha$ is non-decreasing, that

$$n_{\alpha_*} \leq n_{\tilde{\alpha}} = \frac{n\tilde{\alpha}}{\pi} - 1 = 8k_* + 7.$$

This and (B.16) would give

$$r \geq \frac{\sigma^2}{24(8k_* + 7)} \geq \frac{c\sigma^2}{k_* + 1}$$

for a positive constant c. This would prove the theorem when assumption (4.51) is true.

To prove (B.17), we only need to show that

$$\frac{h_{K^*}(\tilde{\alpha}) + h_{K^*}(-\tilde{\alpha})}{2\cos\tilde{\alpha}} - h_{K^*}(0) > \frac{\sigma}{\sqrt{n_{\tilde{\alpha}}}} = \frac{\sigma}{\sqrt{8k_* + 7}}. \tag{B.18}$$

We verify this via Lemma B.1.5 on a case-by-case basis. When $k_* = 0$, we have $\tilde{\alpha} = 8\pi/n$ so that, by Lemma B.1.5, the left hand side above is bounded from below by Δ_2. Because k_* is zero, by definition of k_*, we have

$$\Delta_2 + \frac{2\sigma}{\sqrt{3}} \geq \Delta_0 + 2\sigma = 2\sigma.$$

This gives $\Delta_2 \geq 2\sigma(1 - (1/\sqrt{3}))$ which can be verified to be larger than $\sigma/\sqrt{8k_* + 7} = \sigma/\sqrt{7}$.

When $k_* = 1$, we have $\tilde{\alpha} = 16\pi/n$ so that, by Lemma B.1.5, the left hand side in (B.18) is bounded from below by Δ_4. Because $k_* = 1$, by definition of k_*, we have

$$\Delta_4 + \frac{2\sigma}{\sqrt{5}} \geq \Delta_1 + \frac{2\sigma}{\sqrt{2}} \geq \frac{2\sigma}{\sqrt{2}}$$

which gives $\Delta_4 \geq 2\sigma((1/\sqrt{2}) - (1/\sqrt{5}))$. This can be verified to be larger than $\sigma/\sqrt{8k_* + 7} = \sigma/\sqrt{15}$.

When $k_* \geq 2$, we again use Lemma B.1.5 to argue that the left hand side in (B.18) is bounded from below by $\Delta_{2(k_*+1)}$. Because Δ_k is increasing in k (Lemma B.1.2), we have $\Delta_{2(k_*+1)} \geq \Delta_{2k_*}$. By the definition of k_* (and the fact that $\Delta_{k_*} \geq 0$), we have

$$\Delta_{2k_*} \geq \frac{2\sigma}{k_* + 1}\left(1 - \sqrt{\frac{k_* + 1}{2k_* + 1}}\right).$$

Because $k_* \geq 2$, it can be easily checked that $(k_* + 1)/(2k_* + 1) \leq 3/5$ and $(8k_* + 7)/(k_* + 1) \geq 23/3$. These, together with the fact that $2(1 - \sqrt{3/5})\sqrt{23/3} > 1$, imply (B.18). This completes the proof of the theorem when assumption (4.51) holds.

We now deal with the simpler case when (4.51) is violated. When (4.51) is violated, we first show that

$$k_* > \frac{12n}{16(1 + 2\sqrt{3})^2} - 1. \tag{B.19}$$

To see this, note first that, because (4.51) is violated, we have

$$\frac{h_{K^*}(\alpha) + h_{K^*}(-\alpha)}{2\cos\alpha} - h_{K^*}(0) \leq \frac{\sigma}{\sqrt{n_\alpha}} \leq \sigma\left(\frac{n\alpha}{\pi} - 1\right)^{-1/2}$$

for all $\alpha \in (0, \pi/4]$. Lemma B.1.5 implies that for every $1 \leq k \leq n/16$, we get

$$\Delta_k \leq \frac{h_{K^*}(4k\pi/n) + h_{K^*}(-4k\pi/n)}{2\cos 4k\pi/n} - h_{K^*}(0) \leq \frac{\sigma}{\sqrt{4k-1}} \leq \frac{\sigma}{\sqrt{3k}}. \tag{B.20}$$

Now for every

$$k \leq \frac{12n}{16(1+2\sqrt{3})^2} - 1, \tag{B.21}$$

we have

$$\Delta_k + \frac{2\sigma}{\sqrt{k+1}} \geq \frac{2\sigma}{\sqrt{k+1}} \geq \frac{\sigma}{\sqrt{3n/16}} + \frac{2\sigma}{\sqrt{n/16}} > \Delta_{n/16} + \frac{2\sigma}{\sqrt{n/16+1}}.$$

It follows therefore that any k satisfying (B.21) cannot be a minimizer of $\Delta_k + 2\sigma(k+1)^{-1/2}$, thereby implying (B.19).

Let L^* be defined as the Minkowski sum of K^* and the closed ball with center 0 and radius $\sigma(3n/2)^{-1/2}$. In other words, $L^* := \left\{x + \sigma(3n/2)^{-1/2}y : x \in K \text{ and } ||y|| \leq 1\right\}$. The support function L^* can be checked to equal:

$$h_{L^*}(\theta) = h_{K^*}(\theta) + \sigma(3n/2)^{-1/2}. \tag{B.22}$$

Le Cam's bound again gives

$$r \geq \frac{1}{4}\left(h_{K^*}(0) - h_{L^*}(0)\right)^2 \{1 - ||P_{K^*} - P_{L^*}||_{TV}\} \tag{B.23}$$

where r is as defined in (B.14). By use of Pinsker's inequality, we have

$$||P_{K^*} - P_{L^*}||_{TV} \leq \frac{1}{2\sigma}\sqrt{\sum_{i=1}^{n}\left(h_K(2i\pi/n - \pi) - h_{\breve{K}}(2i\pi/n - \pi)\right)^2} = \frac{1}{2\sigma}\sqrt{\frac{n\sigma^2}{3n/2}} \leq \frac{1}{2}.$$

Therefore, from (B.23) and (B.19), we get that

$$r \geq \frac{\sigma^2}{12n} \geq \frac{1}{16(1+2\sqrt{3})^2}\frac{\sigma^2}{k_* + 1}.$$

This completes the proof of Theorem 4.3.2.

B.1.3 Proof of Inequality (4.21) in Remark4.3.2

Fix $i \in \{1, \ldots, n\}$ and a compact, convex set K^*. Let L^* be defined as in the proof of Theorem 4.3.2. We want to show that

$$\max\left(\mathbb{E}_{K^*}(\hat{h}_i - h_{K^*}(\theta_i))^2, \mathbb{E}_{L^*}(\hat{h}_i - h_{L^*}(\theta_i))^2\right) \leq C.\frac{\sigma^2}{k_*(i)+1} \tag{B.24}$$

for a universal constant C where $\hat{h}_i$ denotes our estimator defined in (4.12). We have already proved in Theorem 4.3.1 that

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C.\frac{\sigma^2}{k_*(i)+1}. \tag{B.25}$$

It can similarly be proved that

$$\mathbb{E}_{L^*}\left(\hat{h}_i - h_{L^*}(\theta_i)\right)^2 \leq C.\frac{\sigma^2}{k_*^{L^*}(i)+1} \tag{B.26}$$

where $k_*^{L^*}$ denotes the quantity $k_*(i)$ with K^* replaced by L^*. More precisely

$$k_*^{L^*}(i) := \operatorname*{argmin}_{k\in\mathcal{I}}\left(\Delta_k^{L^*}(\theta_i) + \frac{2\sigma}{\sqrt{k+1}}\right)$$

where $\Delta_k^{L^*}(\theta_i)$ is defined as in (4.40) with K^* replaced by L^*.

Inequalities (B.25) and (B.26) together imply that the left hand side of (B.24) is bounded from above by

$$C\sigma^2 \max\left(\frac{1}{k_*(i)+1}, \frac{1}{k_*^{L^*}(i)+1}\right). \tag{B.27}$$

We show below how to establish (B.24) from the above bound. As in the proof of Theorem 4.3.2, we shall work with two separate cases.

In the first case, we suppose that the condition (4.51) in the proof of Theorem 4.3.2 holds. In this case, recall from the proof of Theorem 4.3.2 that the set L^* is defined as the convex hull of $K^* \cup \{a_{K^*}(\alpha)\}$ where $a_{K^*}(\alpha)$ is defined as in (4.52). We show below then that

$$\Delta_k^{L^*}(\theta_i) \leq \Delta_k(\theta_i) \qquad \text{for every } k \in \mathcal{I} \tag{B.28}$$

This would immediately imply that $k_*(i) \leq k_*^{L^*}(i)$. The inequality (B.24) would then follow from the bound (B.27).

In order to prove (B.28), we first recall from (B.12) the expression for the support function of L^* i.e., $h_{L^*}(\theta)$ equals the right hand side of (B.12) when $-\alpha \leq \theta \leq \alpha$ and it equals $h_{K^*}(\theta)$ for every other $\theta \in (-\pi, \pi]$. For notational convenience, let us denote by $\delta_j^{K^*}(\theta_i)$, the quantity inside the summation in (4.40) i.e.,

$$\delta_j^{K^*}(\theta_i) = h_{K^*}(\theta_i \pm 4j\pi/n) - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} h_{K^*}(\theta_i \pm 2j\pi/n) \tag{B.29}$$

where $h_{K^*}(\theta_i \pm \phi)$ has the same meaning as in (4.40). This means that $\Delta_k(\theta_i) = \sum_{j=0}^k \delta_j^{K^*}(\theta_i)/(k+1)$. We similarly define $\delta_j^{L^*}(\theta_i)$ with L^* replacing K^* in (B.29) so that $\Delta_k^{L^*}(\theta_i) = \sum_{j=0}^k \delta_j^{L^*}(\theta_i)/(k+1)$.

We now verify (B.28) as follows. From the formula (B.12), it is easy to observe that $\delta_j^{L^*}(\theta_i)$ equals zero whenever $4j\pi/n \leq \alpha$ or, equivalently, $j \leq n\alpha/(4\pi)$. We therefore have

$$\Delta_k^{L^*}(\theta_i) = \frac{1}{k+1}\sum_{j=0}^{k} \delta_j^{L^*}(\theta_i) = \frac{1}{k+1}\sum_{j=0}^{k} \delta_j^{L^*}(\theta_i) I\{j > n\alpha/(4\pi)\}. \tag{B.30}$$

where $I\{\cdot\}$ denotes the indicator function. When $j > n\alpha/(4\pi)$, again from the form of the support function of h_{L^*} described in (B.12), it follows that $h_{L^*}(4j\pi/n) = h_{K^*}(4j\pi/n)$. On the other hand, $h_{L^*}(\theta) \geq h_{K^*}(\theta)$ for all θ simply because $K^* \subseteq L^*$. We thus have

$$\delta_j^{L^*}(\theta_i) \leq \delta_j^{K^*}(\theta_i) \qquad \text{for } j > n\alpha/(4\pi).$$

Because $\delta_j^{K^*}(\theta_i)$ is always nonnegative, inequality (B.28) is now immediate from this and (B.30).

We now turn to the case when the condition (4.51) in the proof of Theorem 4.3.2 does not hold. Observe that in this case, we proved in (B.19) and (B.20) respectively that

$$k_*(i) > \frac{12n}{16(1+2\sqrt{3})^2} - 1 \quad \text{and} \quad \Delta_k \leq \frac{\sigma}{\sqrt{3k}} \tag{B.31}$$

for every $k \in \mathcal{I}$. It may also be recalled that L^* in this case was chosen to be such that its support function satisfies the identity given in (B.22). As a result of this, it is easily seen that

$$\Delta_k^{L^*}(\theta_i) = \Delta_k(\theta_i) + \sigma\left(\frac{3n}{2}\right)^{-1/2} \frac{1}{k+1}\sum_{j=0}^{k}\left(1 - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right)$$

for every k. Now following the calculations in Example 4.4.2 (immediately after inequality (4.44)), we deduce that

$$\Delta_k^{L^*}(\theta_i) \leq \Delta_k(\theta_i) + \frac{8\sqrt{2}\sigma\pi^2}{\sqrt{3}} k^2 n^{-5/2}.$$

The second inequality in (B.31) now allows us to deduce that

$$\Delta_k^{L^*}(\theta_i) \leq \frac{\sigma}{\sqrt{3k}} + \frac{8\sqrt{2}\sigma\pi^2}{\sqrt{3}} k^2 n^{-5/2} \leq c_1\sigma\left(k^{-1/2} + k^2 n^{-5/2}\right)$$

for a universal constant c_1. Thus if $k \in \mathcal{I}$ is such that $k \geq c_2 n$ for a positive constant c_2, we have

$$\Delta_k^{L^*}(\theta_i) \leq \frac{c_1\sigma}{\sqrt{n}}\left(1 + c_2^{-1/2}\right)$$

and consequently

$$\Delta_k^{L^*}(\theta_i) + \frac{2\sigma}{\sqrt{k+1}} \leq \frac{\sigma}{\sqrt{n}}\left[c_1(1 + c_2^{-1/2}) + 2c_2^{-1/2}\right] \tag{B.32}$$

for every $k \in \mathcal{I}, k \geq c_2 n$. On the other hand for $k \leq c_3 n$, we have

$$\Delta_k^{L^*}(\theta_i) + \frac{2\sigma}{\sqrt{k+1}} \geq \frac{2\sigma}{\sqrt{k+1}} \geq \frac{2\sigma}{\sqrt{c_3 n + 1}} \geq \frac{2\sigma}{\sqrt{n(c_3+1)}}. \tag{B.33}$$

From (B.32) and (B.33), it is easy to see that if c_3 and c_2 are suitably chosen, then no k for which $k \leq c_3 n$ can minimize the left hand side of (B.33). This implies therefore that $k_*^{L^*}(i) \geq cn$ for a positive constant c. On the other hand, the first inequality in (B.31) implies that $k_*(i)$ is also at least cn for a positive constant c. This allows to deduce that (B.27) is bounded from above by a constant multiple of $\sigma^2/(k_*(i)+1)$. This completes the proof of inequality (4.21) in Remark 4.3.2.

B.1.4 Proofs of Corollaries and Proposition 4.3.1 in Section 4.3.1

The proofs of the corollaries stated in Section 4.3.1 are given here. For these proofs, we need some simple properties of the $\Delta_k(\theta_i)$ which are stated and proved in Appendix B.1.

We start with the proof of Corollary 4.3.3.

Proof of Corollary 4.3.3. Fix $1 \leq i \leq n$. We will prove that $\breve{k}(i) \leq k_*(i) \leq \tilde{k}(i)$. Inequality (4.31) would then follow from Theorem 4.3.1. For simplicity, we write Δ_k for $\Delta_k(\theta_i)$, f_k for $f_k(\theta_i)$, g_k for $g_k(\theta_i)$, k_* for $k_*(i)$, $\breve{k}$ for $\breve{k}(i)$ and $\tilde{k}$ for $\tilde{k}(i)$.

Inequality (B.45) in Lemma B.1.3 gives

$$\Delta_k \geq \frac{\sigma(\sqrt{6}-2)}{\sqrt{k+1}} \qquad \text{for all } k > k_*, k \in \mathcal{I}.$$

Thus any $k \in \mathcal{I}$ for which $f_k \leq \Delta_k < \sigma(\sqrt{6}-2)/\sqrt{k+1}$ has to satisfy $k \leq k_*$. This proves $\breve{k} \leq k_*$.

For $k_* \leq \tilde{k}$, we first inequality (B.44) in Lemma B.1.3 to obtain $\Delta_{k_*} \geq 6(\sqrt{2}-1)\sigma/\sqrt{k_*+1}$. Also Lemma B.1.2 states that $k \mapsto \Delta_k$ is non-decreasing for $k \in \mathcal{I}$. We therefore have

$$g_k \leq \Delta_k \leq \Delta_{k_*} \leq \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k_*+1}} \leq \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k+1}} \qquad \text{for all } k \leq k_*, k \in \mathcal{I}.$$

Therefore any $k \in \mathcal{I}$ for which $g_k > 6(\sqrt{2}-1)\sigma/\sqrt{k+1}$ has to be larger than k_*. This proves $\tilde{k} \geq k_*$. The proof is complete. □

We next give the proof of Corollary 4.3.1.

Proof of Corollary 4.3.1. We only need to prove (4.22). Inequality (4.23) would then follow from Theorem 4.3.1. Fix $i \in \{1, \dots, n\}$ and suppose that K^* is contained in a ball of radius R centered at (x_1, x_2). We shall prove below that $\Delta_k(\theta_i) \leq 6\pi Rk/n$ for every $k \in \mathcal{I}$ and (4.22) would then follow from Corollary 4.3.3. Without loss of generality, assume that $\theta_i = 0$.

As in the proof of Theorem 4.3.5, we may assume that K^* is contained in the ball of radius R centered at the origin. This implies that $|h_{K^*}(\theta)| \leq R$ for all θ and also that h_{K^*} is Lipschitz with constant R. Note then that for every $k \in \mathcal{I}$ and $0 \leq j \leq k$, the quantity

$$Q := \frac{h_{K^*}(4j\pi/n) + h_{K^*}(-4j\pi/n)}{2} - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} \frac{h_{K^*}(2j\pi/n) + h_{K^*}(-2j\pi/n)}{2}$$

can be bounded as

$$|Q| = \left| \frac{h_{K^*}(4j\pi/n) - h_{K^*}(2j\pi/n) + h_{K^*}(-4j\pi/n) - h_{K^*}(-2j\pi/n)}{2} \right. \\ \left. - \left(\frac{\cos(4j\pi/n) - \cos(2j\pi/n)}{\cos(2j\pi/n)}\right) \frac{h_{K^*}(2j\pi/n) + h_{K^*}(-2j\pi/n)}{2} \right| \leq \frac{6Rj\pi}{n}.$$

Here we used also the fact that $\cos(\cdot)$ is Lipschitz and $\cos(2j\pi/n) \geq 1/2$. The inequality $\Delta_k(0) \leq 6\pi Rk/n$ then immediately follows. The proof is complete. □

Proof of Proposition 4.3.1. Inequality (4.24) is clearly a direct consequence of (4.23). We therefore only prove (4.25) below. We assume without loss of generality that n is even, $i = n/2$ and that $\theta_i = 0$. Also assume that $\mathcal{K}(R)$ contains of all compact, convex sets that are contained in the ball of radius R *centered at the origin*.

Take K^* to be the vertical line segment joining the two points $(0, R)$ and $(0, -R)$ for a fixed $R > 0$ (as in Example 4.4.3). Further let L^* be as in the proof of Theorem 4.3.2. It is then easy to check that $L^* \in \mathcal{K}(R)$ and thus the minimax risk in the left hand side of (4.25) is bounded from below by

$$\inf_{\tilde{h}} \max \left(\mathbb{E}_{K^*}(\tilde{h} - h_{K^*}(\theta_i))^2, \mathbb{E}_{L^*}(\tilde{h} - h_{L^*}(\theta_i))^2 \right)$$

Inequality (4.20) then gives that

$$\inf_{\tilde{h}} \sup_{K^* \in \mathcal{K}(R)} \mathbb{E}_{K^*} \left(\tilde{h} - h_{K^*}(\theta_i) \right)^2 \geq \frac{c\sigma^2}{k_*(i) + 1}.$$

We now use inequality (4.46) which proves that the right hand side above is bounded from above by a constant multiple of $(\sigma^2/n) + (\sigma^2 R/n)^{2/3}$. This completes the proof of Proposition 4.3.1. □

We conclude this section with a proof of Corollary 4.3.2.

Proof of Corollary 4.3.2. By Theorem 4.3.1, inequality (4.28) is a direct consequence of (4.27). We therefore only need to prove (4.27). Fix $k \in \mathcal{I}$ with

$$k \leq \frac{n}{4\pi} \min(\theta_i - \phi_1(i), \phi_2(i) - \theta_i). \tag{B.34}$$

It is then clear that $\theta_i \pm 4j\pi/n \in [\phi_1(i), \phi_2(i)]$ for every $0 \leq j \leq k$. From (4.26), it follows that

$$h_{K^*}(\theta) = x_1 \cos\theta + x_2 \sin\theta \qquad \text{for all } \theta = \theta_i \pm \frac{4j\pi}{n}, 0 \leq j \leq k.$$

We now argue that $\Delta_k(\theta_i) = 0$. To see this, note first that $\Delta_k(\theta_i) = U_k(\theta_i) - L_k(\theta_i)$ has the following alternative expression (4.40). Plugging in $h_{K^*}(\theta) = x_1 \cos\theta + x_2 \sin\theta$ in (4.40), one can see by direct computation that $\Delta_k(\theta_i) = 0$ for every $k \in \mathcal{I}$ satisfying (B.34). The definition (4.18) of $k_*(i)$ now immediately implies that

$$k_*(i) \geq \min\left(\frac{n}{4\pi}\min(\theta_i - \phi_1(i), \phi_2(i) - \theta_i), cn\right)$$

for a small enough universal constant c. This proves (4.27) thereby completing the proof. □

B.1.5 Completion of the proof of Theorem 4.3.3

We complete the proof of Theorem 4.3.3 starting from where we left off in the main text. The goal is to prove inequality (4.56). The argument below is inspired by an argument due to Zhang [159, Proof of Theorem 2.1] in a very different context.

Recall that $k_*(i)$ takes values in $\mathcal{I} := \{0\} \cup \{2^j : j \geq 0, 2^j \leq \lfloor n/16 \rfloor\}$. For $k \in \mathcal{I}$, let

$$\rho(k) := \sum_{i=1}^{n} I\{k_*(i) = k\} \qquad \text{and} \qquad \ell(k) := \sum_{i=1}^{n} I\{k_*(i) < k\}$$

Note that $\ell(0) = 0, \ell(1) = \rho(0)$ and $\rho(k) = \ell(2k) - \ell(k)$ for $k \geq 1, k \in \mathcal{I}$. As a result

$$\sum_{i=1}^{n} \frac{1}{k_*(i)+1} = \sum_{k \in \mathcal{I}} \frac{\rho(k)}{k+1} = \ell(1) + \sum_{k \geq 1, k \in \mathcal{I}} \frac{\ell(2k) - \ell(k)}{k+1}.$$

Let K denote the maximum element of $\mathcal{I}$. Because $\ell(2K) = n$, we can write

$$\sum_{i=1}^{n} \frac{1}{k_*(i)+1} = \frac{n}{K+1} + \frac{\ell(1)}{2} + \sum_{k \geq 2, k \in \mathcal{I}} \frac{k\ell(k)}{(k+1)(k+2)}.$$

Using $n/(K+1) \leq C$ and loose bounds for the other terms above, we obtain

$$\sum_{i=1}^{n} \frac{1}{k_*(i)+1} \leq C + \sum_{k \geq 1, k \in \mathcal{I}} \frac{3\ell(k)}{k}. \tag{B.35}$$

We shall show below that

$$\ell(k) \leq \min\left(n, \frac{ARk^{5/2}}{\sigma n}\right) \qquad \text{for all } k \in \mathcal{I} \tag{B.36}$$

for a universal positive constant A. Before that, let us first prove (4.56) assuming (B.36). Assuming (B.36), we can write

$$\sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} = \sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k \leq \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} + \sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k > \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} \tag{B.37}$$

In the first term on the right hand side above, we use the bound $\ell(k) \leq ARk^{5/2}/(\sigma n)$. We then get

$$\sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k \leq \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} \leq \frac{AR}{\sigma n} \sum_{k\geq 1, k\in\mathcal{I}} k^{3/2} I\left\{k \leq \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\}.$$

Because $\mathcal{I}$ consists of integers of the form 2^j, the sum in the right hand side above is bounded from above by a constant multiple of the last term. This gives

$$\sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k \leq \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} \leq \frac{CR}{\sigma n} \left(\frac{\sigma n^2}{AR}\right)^{3/5} = C\left(\frac{R\sqrt{n}}{\sigma}\right)^{2/5} \tag{B.38}$$

For the second term on the right hand side in (B.37), we use the bound $\ell(k) \leq n$ which gives

$$\sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k > \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} \leq n \sum_{k\geq 1, k\in\mathcal{I}} k^{-1} I\left\{k > \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\}$$

Again, because $\mathcal{I}$ consists of integers of the form 2^j, the sum in the right hand side above is bounded from above by a constant multiple of the first term. This gives

$$\sum_{k\geq 1, k\in\mathcal{I}} \frac{\ell(k)}{k} I\left\{k > \left(\frac{\sigma n^2}{AR}\right)^{2/5}\right\} \leq Cn \left(\frac{\sigma n^2}{AR}\right)^{-2/5} = C\left(\frac{R\sqrt{n}}{\sigma}\right)^{2/5}. \tag{B.39}$$

Inequalities (B.38) and (B.39) in conjunction with (B.35) proves (4.56) which would complete the proof of (4.32).

We only need to prove (B.36). For this, observe first that when $k_*(i) < k$, Corollary 4.3.3 gives that

$$\Delta_k(\theta_i) \geq \frac{(\sqrt{6}-2)\sigma}{\sqrt{k+1}}. \tag{B.40}$$

This is because if (B.40) is violated, then Corollary 4.3.3 gives $k \leq \check{k}(i) \leq k_*(i)$. Consequently, we have

$$I\{k_*(i) < k\} \leq \frac{\Delta_k(\theta_i)\sqrt{k+1}}{(\sqrt{6}-2)\sigma}$$

and

$$\ell(k) \leq \frac{\sqrt{k+1}}{(\sqrt{6}-2)\sigma} \sum_{i=1}^{n} \Delta_k(\theta_i) \qquad \text{for every } k \in \mathcal{I}. \tag{B.41}$$

We will now prove an upper bound for $\Delta_k(\theta_i), 1 \leq i \leq n$ under the assumption that K^* is contained in a ball of radius $R \geq 0$. We may assume without loss of generality that this ball is centered at the origin because the expression for $\Delta_k(\theta_i)$ given in (4.40) remains unchanged if $h_{K^*}(\theta)$ is replaced by $h_{K^*}(\theta) - a_1 \cos\theta - a_2 \sin\theta$ for any $(a_1, a_2) \in \mathbb{R}^2$.

Now using the expression (4.40) for $\Delta_k(\theta_i)$, it is easy to see that

$$\sum_{i=1}^{n} \Delta_k(\theta_i) = \frac{1}{k+1} \sum_{j=0}^{k} \delta_j \tag{B.42}$$

where δ_j is given by

$$\delta_j = \sum_{i=1}^{n} \left(\frac{h_{K^*}(\theta_{i+2j}) + h_{K^*}(\theta_{i-2j})}{2} - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} \frac{h_{K^*}(\theta_{i+j}) + h_{K^*}(\theta_{i-j})}{2} \right).$$

with $\theta_k = 2\pi k/n - \pi$. Because $\theta \mapsto h_{K^*}(\theta)$ is a periodic function of period 2π, the above expression for δ_j only depends on $h_{K^*}(\theta_1)$, ..., $h_{K^*}(\theta_n)$. In fact, it is easy to see that

$$\delta_j = \left(1 - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right) \sum_{i=1}^{n} h_{K^*}(\theta_i).$$

Now because K^* is contained in the ball of radius R centered at the origin, it follows that $|h_{K^*}(\theta_i)| \leq R$ for each i which gives

$$\delta_j \leq nR\left(1 - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right) \leq nR\left(1 - \frac{\cos(4k\pi/n)}{\cos(2k\pi/n)}\right) = \frac{nR(1 + 2\cos 2\pi k/n)}{\cos 2\pi k/n}(1 - \cos 2\pi k/n)$$

for all $0 \leq j \leq k$. Because $k \leq n/16$ for all $k \in \mathcal{I}$, it follows that

$$\delta_j \leq 8nR\sin^2(\pi k/n) \leq \frac{8R\pi^2 k^2}{n} \qquad \text{for all } 0 \leq j \leq k.$$

The identity (B.42) therefore gives $\sum_{i=1}^{n} \Delta_k(\theta_i) \leq 8R\pi^2 k^2/n$ for all $k \in \mathcal{I}$. Consequently, from (B.41) and the trivial fact that $\ell(k) \leq n$, we obtain

$$\ell(k) \leq \min\left(n, \frac{8\pi^2}{(\sqrt{6}-2)} \frac{Rk^2\sqrt{k+1}}{\sigma n}\right) \qquad \text{for all } k \in \mathcal{I}.$$

Note that $\ell(0) = 0$ so that the above inequality only gives something useful for $k \geq 1$. Using $k + 1 \leq 2k$ for $k \geq 1$ and denoting the resulting constant by C, we obtain (B.36). This completes the proof of Theorem 4.3.3.

B.1.6 Technical Lemmas

Our first task here is to provide the proof of Lemma 4.2.1. We also restate this result here for the convenience of the reader.

Lemma B.1.1. *For every $0 < \phi < \pi/2$ and every $\theta \in (-\pi, \pi]$, we have $l(\theta, \phi) \leq h_{K^*}(\theta) \leq u(\theta, \phi)$.*

Proof. The inequality $h_{K^*}(\theta) \leq u(\theta, \phi)$ is obtained by using (4.1) with $\alpha_1 = \theta + \phi, \alpha_2 = \theta - \phi$ and $\alpha = \theta$. For $l(\theta, \phi) \leq h_{K^*}(\theta)$, we use (4.1) with $\alpha_1 = \theta + 2\phi, \alpha_2 = \theta$ and $\alpha = \theta + \phi$ to obtain

$$h_{K^*}(\theta) \geq 2h_{K^*}(\theta + \phi)\cos\phi - h_{K^*}(\theta + 2\phi).$$

One similarly has $h_{K^*}(\theta) \geq 2h_{K^*}(\theta - \phi)\cos\phi - h_{K^*}(\theta - 2\phi)$ and $l(\theta, \phi) \leq h_{K^*}(\theta)$ is deduced by averaging these two inequalities. □

We next provide three lemmas which were used in the proofs of the main results of chapter 4.

Lemma B.1.2. *Recall the quantity $\Delta_k(\theta_i)$ defined in (4.40). The inequality $\Delta_{2k}(\theta_i) \geq 1.5\Delta_k(\theta_i)$ holds for every $1 \leq i \leq n$ and $0 \leq k \leq n/16$.*

Proof. We may assume without loss of generality that $\theta_i = 0$. We will simply write Δ_k for $\Delta_k(\theta_i)$ below for notational convenience. Let us define, for $\theta \in \mathbb{R}$,

$$\delta(\theta) := \frac{h_{K^*}(2\theta) + h_{K^*}(-2\theta)}{2} - \frac{\cos 2\theta}{\cos\theta}\frac{h_{K^*}(\theta) + h_{K^*}(-\theta)}{2}.$$

Note then that $\Delta_k = \sum_{j=0}^{k} \delta(2j\pi/n)/(k+1)$. We shall first prove that

$$\delta(y) \geq \left(\frac{\tan y}{\tan x}\right)\delta(x) \qquad \text{for every } 0 < y \leq \pi/4 \text{ and } x < y \leq 2x. \tag{B.43}$$

For this, first apply (4.1) to $\alpha_1 = 2x, \alpha_2 = x$ and $\alpha = y$ to get

$$h_{K^*}(y) \leq \frac{\sin(y - x)}{\sin x}h_{K^*}(2x) + \frac{\sin(2x - y)}{\sin x}h_{K^*}(x).$$

We then apply (4.1) to $\alpha_1 = 2y, \alpha_2 = x$ and $\alpha = 2x$ to get (note that $2y - x \leq 2y < \pi/2$)

$$h_{K^*}(2y) \geq \frac{\sin(2y - x)}{\sin x}h_{K^*}(2x) - \frac{\sin(2y - 2x)}{\sin x}h_{K^*}(x).$$

Combining these two inequalities, we get (note that $2y \leq \pi/2$ which implies that $\cos 2y \geq 0$)

$$h_{K^*}(2y) - \frac{\cos 2y}{\cos y}h_{K^*}(y) \geq \alpha h_{K^*}(2x) - \beta h_{K^*}(x),$$

where

$$\alpha := \frac{\sin(2y - x)}{\sin x} - \frac{\cos 2y}{\cos y}\frac{\sin(y - x)}{\sin x}$$

and

$$\beta := \frac{\sin(2y - 2x)}{\sin x} + \frac{\cos 2y}{\cos y}\frac{\sin(2x - y)}{\sin x}.$$

It can be checked by a straightforward calculation that

$$\alpha = \frac{\tan y}{\tan x} \quad \text{and} \quad \beta = \frac{\tan y}{\tan x}\frac{\cos 2x}{\cos x}.$$

It follows therefore that

$$h_{K^*}(2y) - \frac{\cos 2y}{\cos y} h_{K^*}(y) \geq \frac{\tan y}{\tan x}\left(h_{K^*}(2x) - \frac{\cos 2x}{\cos x} h_{K^*}(x)\right).$$

We similarly obtain

$$h_{K^*}(-2y) - \frac{\cos 2y}{\cos y} h_{K^*}(-y) \geq \frac{\tan y}{\tan x}\left(h_{K^*}(-2x) - \frac{\cos 2x}{\cos x} h_{K^*}(-x)\right).$$

The required inequality (B.43) now results by adding the above two inequalities. A trivial consequence of (B.43) is that $\delta(y) \geq \delta(x)$ for $0 < y \leq \pi/4$ and $x < y \leq 2x$. Further, applying (B.43) to $y = 2x$ (assuming that $0 < x < \pi/8$), we obtain $\delta(2x) \geq 2\delta(x)$. Note that $\tan 2x = 2\tan x/(1 - \tan^2 x) \geq 2\tan x$ for $0 < x < \pi/8$.

To prove $\Delta_{2k} \geq (1.5)\Delta_k$, we fix $1 \leq k \leq n/16$ (note that the inequality is trivial when $k = 0$) and note that

$$\Delta_{2k} = \frac{1}{2k+1}\sum_{j=0}^{2k} \delta\left(\frac{2j\pi}{n}\right) = \frac{1}{2k+1}\sum_{j=1}^{k}\left(\delta\left(\frac{2(2j-1)\pi}{n}\right) + \delta\left(\frac{4j\pi}{n}\right)\right)$$

where we used the fact that $\delta(0) = 0$. Using the bounds proved for $\delta(\theta)$, we have

$$\delta\left(\frac{2(2j-1)\pi}{n}\right) \geq \delta\left(\frac{2j\pi}{n}\right) \quad \text{and} \quad \delta\left(\frac{4j\pi}{n}\right) \geq 2\delta\left(\frac{2j\pi}{n}\right).$$

Therefore

$$\Delta_{2k} \geq \frac{3}{2k+1}\sum_{j=1}^{k}\delta\left(\frac{2j\pi}{n}\right) \geq \frac{3}{2(k+1)}\sum_{j=0}^{k}\delta\left(\frac{2j\pi}{n}\right) = \frac{3}{2}\Delta_k$$

and this completes the proof. □

Lemma B.1.3. *Fix $i \in \{1, \ldots, n\}$. Consider $\Delta_k(\theta_i)$ (defined in (4.40)) and $k_*(i)$ (defined in (4.18)). We then have the following inequalities*

$$\Delta_{k_*(i)}(\theta_i) \leq \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k_*(i)+1}}. \tag{B.44}$$

and

$$\Delta_k(\theta_i) \geq \max\left(\frac{(\sqrt{6}-2)\sigma}{\sqrt{k+1}}, \frac{(\sqrt{6}-2)\sqrt{k+1}\sigma}{2(k_*+1)}\right) \tag{B.45}$$

for all $k > k_*(i), k \in \mathcal{I}$.

Proof. Fix $i \in \{1, \dots, n\}$. Below we simply denote $k_*(i)$ and $\Delta_k(\theta_i)$ by k_* and Δ_k respectively for notational convenience.

We first prove (B.44). If $k_* \geq 2$, we have

$$\Delta_{k_*} + \frac{2\sigma}{\sqrt{k_*+1}} \leq \Delta_{k_*/2} + \sqrt{2}\frac{2\sigma}{\sqrt{k_*+2}} \leq \Delta_{k_*/2} + \sqrt{2}\frac{2\sigma}{\sqrt{k_*+1}}.$$

Using Lemma B.1.2 (note that $k_* \in \mathcal{I}$ and hence $k_* \leq n/16$), we have $\Delta_{k_*/2} \leq (2/3)\Delta_{k_*}$. We therefore have

$$\Delta_{k_*} + \frac{2\sigma}{\sqrt{k_*+1}} \leq \frac{2}{3}\Delta_{k_*} + \sqrt{2}\frac{2\sigma}{\sqrt{k_*+1}}$$

which proves (B.44). Inequality (B.44) is trivial when $k_* = 0$. Finally, for $k_* = 1$, we have $\Delta_1 + \sqrt{2}\sigma \leq \Delta_0 + 2\sigma = 2\sigma$ which again implies (B.44).

We now turn to (B.45). Let k' denote the smallest $k \in \mathcal{I}$ for which $k > k_*$. We start by proving the first part of (B.45):

$$\Delta_k \geq \frac{(\sqrt{6}-2)\sigma}{\sqrt{k+1}} \qquad \text{for } k > k_*, k \in \mathcal{I}. \tag{B.46}$$

Note first that if (B.46) holds for $k = k'$, then it holds for all $k \geq k'$ as well because $\Delta_k \geq \Delta_{k'}$ (from Lemma B.1.2) and $1/\sqrt{k+1} \leq 1/\sqrt{k'+1}$. We therefore only need to verify (B.46) for $k = k'$. If $k_* = 0$, then $k' = 1$ and because

$$\Delta_1 + \frac{2\sigma}{\sqrt{2}} \geq \Delta_0 + 2\sigma = 2\sigma,$$

we obtain $\Delta_1 \geq (2-\sqrt{2})\sigma$. This implies (B.46). On the other hand, if $k_* > 0$, then $k' = 2k_*$ and we can write

$$\Delta_{2k_*} + \frac{2\sigma}{\sqrt{2k_*+1}} \geq \Delta_{k_*} + \frac{2\sigma}{\sqrt{k_*+1}} \geq \frac{2\sigma}{\sqrt{k_*+1}}.$$

This gives

$$\Delta_{2k_*} \geq \frac{2\sigma}{\sqrt{2k_*+1}}\left(\sqrt{\frac{2k_*+1}{k_*+1}} - 1\right)$$

which implies inequality (B.46) for $k = 2k_*$ because $(2k_*+1)/(k_*+1) \geq 3/2$. The proof of (B.46) is complete.

For the second part of (B.45), we use Lemma B.1.2 which states $\Delta_{2k} \geq (1.5)\Delta_k \geq \sqrt{2}\Delta_k$ for all $k \in \mathcal{I}$. By a repeated application of this inequality, we get

$$\Delta_k \geq \sqrt{\frac{k}{k'}}\Delta_{k'} \geq \sqrt{\frac{k+1}{k'+1}}\Delta_{k'} \qquad \text{for all } k \geq k'.$$

Using (B.46) for $k = k'$, we get

$$\Delta_k \geq \frac{(\sqrt{6}-2)\sigma\sqrt{k+1}}{k'+1}.$$

The proof of (B.45) is now completed by observing that $k' \leq 2k_* + 1$. □

Lemma B.1.4. *Fix $i \in \{1, \dots, n\}$. For every $0 \leq k \leq n/8$, the variance of the random variable $\hat{U}_k(\theta_i)$ (defined in (4.10)) is at most $\sigma^2/(k+1)$. Also, for every $0 \leq k \leq n/16$, the variance of the random variable $\hat{\Delta}_k(\theta_i)$ (defined in (4.11)) is at most $\sigma^2/(k+1)$.*

Proof. Fix $1 \leq i \leq n$. We shall first prove the bound for the variance of $\hat{U}_k(\theta_i)$ for a fixed $0 \leq k \leq n/8$. Note that

$$\hat{U}_k(\theta_i) = \frac{1}{k+1}\sum_{j=0}^{k} \frac{Y_{i+j} + Y_{i-j}}{2\cos(2j\pi/n)}.$$

It is therefore straightforward to see that

$$\operatorname{var}(\hat{U}_k(\theta_i)) = \frac{\sigma^2}{(k+1)^2}\left(1 + \frac{1}{2}\sum_{j=1}^{k}\sec^2(2j\pi/n)\right).$$

For $1 \leq j \leq k \leq n/8$, we have $\sec(2j\pi/n) \leq \sqrt{2}$ because $2j\pi/n \leq \pi/4$. The inequality $\operatorname{var}(\hat{U}_k(\theta_i)) \leq \sigma^2/(k+1)$ then immediately follows.

Let us now turn to the variance of $\hat{\Delta}_k(\theta_i)$. When $k = 0$, the conclusion is obvious since $\hat{\Delta}_k(\theta_i) = 0$. Otherwise, the expression (4.11) for $\hat{\Delta}_k(\theta_i)$ can be rewritten as

$$\hat{\Delta}_k(\theta_i) = S_1 + S_2 + S_3$$

where

$$S_1 = \frac{-1}{k+1}\sum_{j=1}^{k}\{j \text{ is odd}\}\frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\frac{Y_{i+j} + Y_{i-j}}{2},$$

$$S_2 = \frac{1}{k+1}\sum_{j=1}^{k}\{j \text{ is even}\}\left(1 - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right)\frac{Y_{i+j} + Y_{i-j}}{2},$$

and

$$S_3 = \frac{1}{k+1}\sum_{j=k+1}^{2k}\{j \text{ is even}\}\frac{Y_j + Y_{-j}}{2}.$$

S_1, S_2 and S_3 are clearly independent. Moreover, the different terms in each S_i are also independent. Thus

$$\mathrm{var}(S_1) = \frac{\sigma^2}{2(k+1)^2} \sum_{j=1}^{k} \{j \text{ is odd}\} \frac{\cos^2(4j\pi/n)}{\cos^2(2j\pi/n)},$$

$$\mathrm{var}(S_2) = \frac{\sigma^2}{2(k+1)^2} \sum_{j=1}^{k} \{j \text{ is even}\} \left(1 - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right)^2,$$

and

$$\mathrm{var}(S_3) = \frac{\sigma^2}{2(k+1)^2} \sum_{j=k+1}^{2k} \{j \text{ is even}\} \leq \frac{\sigma^2}{2(k+1)}.$$

Now for $k \leq n/16$ and $1 \leq j \leq k$,

$$0 \leq \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} \leq 1$$

which implies that $\mathrm{var}(S_1) + \mathrm{var}(S_2) \leq \sigma^2/2(k+1)$. Thus $\mathrm{var}(\hat{\Delta}_k(\theta_i)) \leq \sigma^2/(k+1)$. □

The next lemma was used in the proof of Theorem 4.3.2.

Lemma B.1.5. *Let Δ_k be the quantity (4.40) with $\theta_i = 0$ i.e.,*

$$\Delta_k := \frac{1}{k+1} \sum_{j=0}^{k} \left(\frac{h_{K^*}(4j\pi/n) + h_{K^*}(-4j\pi/n)}{2} - \frac{\cos(4j\pi/n)}{\cos(2j\pi/n)} \frac{h_{K^*}(2j\pi/n) + h_{K^*}(-2j\pi/n)}{2} \right).$$

Then the following inequality holds for every $k \leq n/16$:

$$\Delta_k \leq \frac{h_{K^*}(4k\pi/n) + h_{K^*}(-4k\pi/n)}{2\cos(4k\pi/n)} - h_{K^*}(0).$$

Proof. From Lemma B.1.2, it follows that $\delta(2i\pi/n) \leq \delta(2k\pi/n)$ for all $1 \leq i \leq k$ (this follows by reapplying Lemma B.1.2 to $2i\pi/n, 4i\pi/n, \dots$ until we hit $2k\pi/n$). As a consequence, we have $\Delta_k \leq \delta(2k\pi/n)$. Now, if $\theta = 2k\pi/n$ then $\theta \leq \pi/8$ and we can write

$$\begin{aligned} \delta(\theta) &= \frac{h_{K^*}(2\theta) + h_{K^*}(-2\theta)}{2} - \frac{\cos 2\theta}{\cos\theta} \frac{h_{K^*}(\theta) + h_{K^*}(-\theta)}{2} \\ &= \cos 2\theta \left(\frac{h_{K^*}(2\theta) + h_{K^*}(-2\theta)}{2\cos 2\theta} - h_{K^*}(0) \right) - \cos 2\theta \left(\frac{h_{K^*}(\theta) + h_{K^*}(-\theta)}{2\cos\theta} - h_{K^*}(0) \right). \end{aligned}$$

Because $h_{K^*}(\theta) + h_{K^*}(-\theta) \geq 2h_{K^*}(0)\cos\theta$ and $\cos 2\theta \geq 0$, we have

$$\delta(\theta) \leq \cos 2\theta \left(\frac{h_{K^*}(2\theta) + h_{K^*}(-2\theta)}{2\cos 2\theta} - h_{K^*}(0) \right) \leq \frac{h_{K^*}(2\theta) + h_{K^*}(-2\theta)}{2\cos 2\theta} - h_{K^*}(0).$$

The proof is complete. □

Lemma B.1.6 (Approximation)**.** *There exists a universal positive constant C such that for every $i = 1, \ldots, n$ and every compact, convex set P, we have*

$$\mathbb{E}_{K^*}\left(\hat{h}_i - h_{K^*}(\theta_i)\right)^2 \leq C\left(\frac{\sigma^2}{k_*^P(i)+1} + \ell_H^2(K^*, P)\right). \tag{B.47}$$

Proof. Fix $i \in \{1, \ldots, n\}$ and a compact, convex set P. For notational convenience, we write $\Delta_k, \Delta_k^P, k_*$ and k_*^P for $\Delta_k(\theta_i), \Delta_k^P(\theta_i), k_*(\theta_i)$ and $k_*^P(\theta_i)$ respectively.

We assume that the following condition holds:

$$k_*^P + 1 \geq \frac{24(\sqrt{2}-1)}{\sqrt{6}-2}(k_* + 1). \tag{B.48}$$

If this condition does not hold, we have

$$\frac{1}{k_*+1} < \frac{24(\sqrt{2}-1)}{\sqrt{6}-2}\frac{1}{k_*^P+1}$$

and then (B.1.6) immediately follows from Theorem 4.3.1.

Note that (B.48) implies, in particular, that $k_*^P > k_*$. Inequality (B.45) in Lemma B.1.3 applied to $k = k_*^P$ implies therefore that

$$\Delta_{k_*^P} \geq \frac{(\sqrt{6}-2)\sqrt{k_*^P+1}\sigma}{2(k_*+1)}.$$

Also inequality (B.44) applied to the set P instead of K^* gives

$$\Delta_{k_*^P}^P \leq \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k_*^P+1}}.$$

Combining the above pair of inequalities, we obtain

$$\Delta_{k_*^P} - \Delta_{k_*^P}^P \geq \frac{(\sqrt{6}-2)\sqrt{k_*^P+1}\sigma}{2(k_*+1)} - \frac{6(\sqrt{2}-1)\sigma}{\sqrt{k_*^P+1}}.$$

The right hand above is non-decreasing in $k_*^P + 1$ and so we can replace $k_*^P + 1$ by the lower bound in (B.48) to obtain, after some simplication,

$$\Delta_{k_*^P} - \Delta_{k_*^P}^P \geq \frac{\sigma}{4\sqrt{k_*+1}}\sqrt{24(\sqrt{2}-1)(\sqrt{6}-2)}. \tag{B.49}$$

The key now is to observe that

$$|\Delta_k - \Delta_k^P| \leq 2\ell_H(K^*, P) \qquad \text{for all } k. \tag{B.50}$$

This follows from the definition (4.35) of the Hausdorff distance which gives

$$\left|\Delta_k - \Delta_k^P\right| \leq \ell_H(K^*, P)\left(1 + \frac{1}{k+1}\sum_{j=0}^{k}\frac{\cos(4j\pi/n)}{\cos(2j\pi/n)}\right)$$

and this clearly implies (B.50) because $\cos(4j\pi/n)/\cos(2j\pi/n) \leq 1$ for all $0 \leq j \leq k$.

From (B.50) and (B.49), we deduce that

$$\ell_H(K^*, P) \geq \frac{c\sigma}{\sqrt{k_* + 1}}$$

for a universal positive constant c. This, together with inequality (4.17), clearly implies (B.47) which completes the proof. □

B.2 Additional Simulation Results

We had presented simulation results only when K^* is a ball and a segment in chapter 4. Here we present additional simulation results when K^* is a square, ellipsoid and random polytope.

B.2.1 Pointwise estimation

Here, we present plots analogous to Figure 4.2 for three additional choices of K^*:

1. K^* is the square formed by the four corner points: $\{(0,0),(0,1),(1,0),(1,1)\}$ whose support function equals $h_{K^*}(\theta) = \max\{0, \sin\theta, \cos\theta, \sin\theta + \cos\theta\}$. This function is plotted in the first subplot of Figure B.1. We study pointwise estimation here for $\theta_i = 0, \pi/8$ and $\pi/4$ (these points are indicated by the red dots in the first subplot). For each of these three values of θ_i, we calculated the mean squared error as a function of n which is plotted in Figure B.1.

2. K^* is the ellipsoid $\{(x,y) : x^2/4+y^2/2 = 1\}$ and $\theta_i = 0, \pi/4, \pi/2$. The support function equals $h_{K^*}(\theta) := \left(4\cos^2\theta + 2\sin^2\theta\right)^{1/2}$. This function is plotted in the first subplot of Figure B.2. We study pointwise estimation here for $\theta_i = 0, \pi/4$ and $\pi/2$ (these points are indicated by the red dots in the first subplot). For each of these three values of θ_i, we calculated the mean squared error as a function of n which is plotted in Figure B.2.

3. For our final example, we consider a random polytope K^* generated by sampling 10 points from the uniform distribution on the square $[-2,2] \times [-2,2]$ and taking their convex hull. The performance of the seven estimators is shown in the following plots. In the first subplot, the support function is drawn in black with points $0, \pi/8, \pi/4$ marked as our choices for θ_i. Similarly as before, the last three subplots shows how the mean squared error changes with sample size n growing.

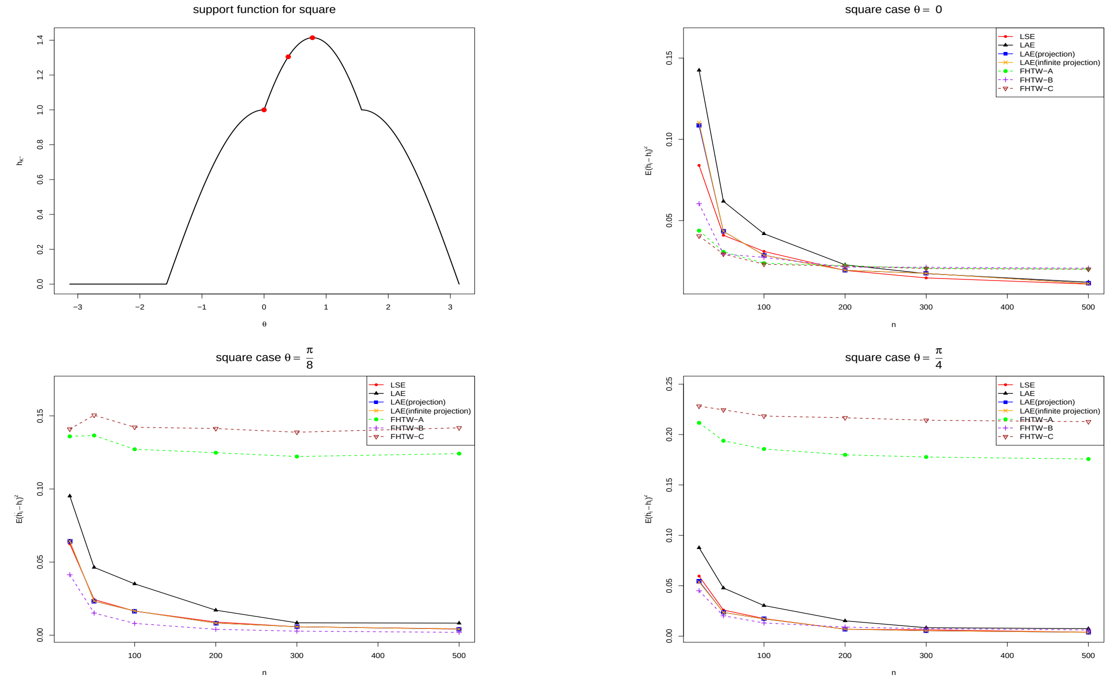


Figure B.1: Point estimation error when K^* is a square

The story in all these plots is the same. The error decays in all cases as n grows. The performance of our estimators is similar to the LSE. The performance of the *FHTW* estimators is good when smoothness assumptions are met but otherwise they can be poor.

B.2.2 Set Estimation

Here we present simulation results on set estimation for each of the three examples discussed above. The relevant plots are given in Figure B.4 (when K^* is the square), Figure B.5 (when K^* is the ellipsoid) and Figure B.6 (when K^* is the random polytope).

The conclusions are again same as before. Our estimators perform at the same level as the *LSE*. Even though, we propose two set estimators: *LAE with projection* and *LAE with infinite projection*, both of them look similar and have similar performance. The *FHTW-B* estimator seems to work well when K^* can be well-approximated by an ellipsoid.

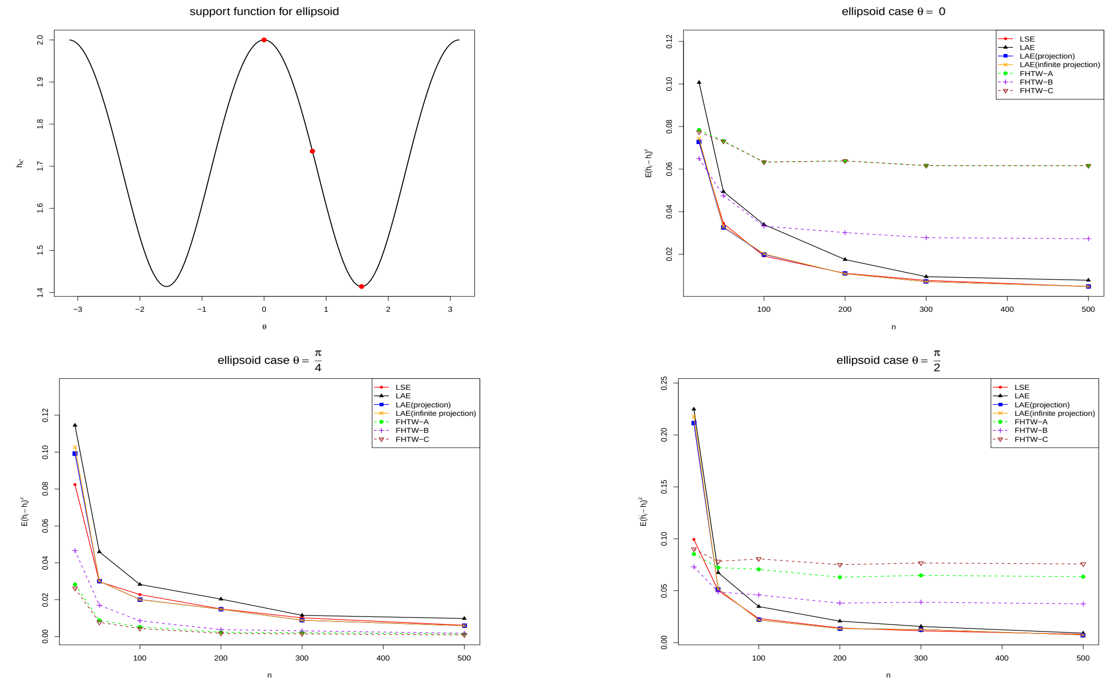


Figure B.2: Point estimation error when K^* is an ellipsoid

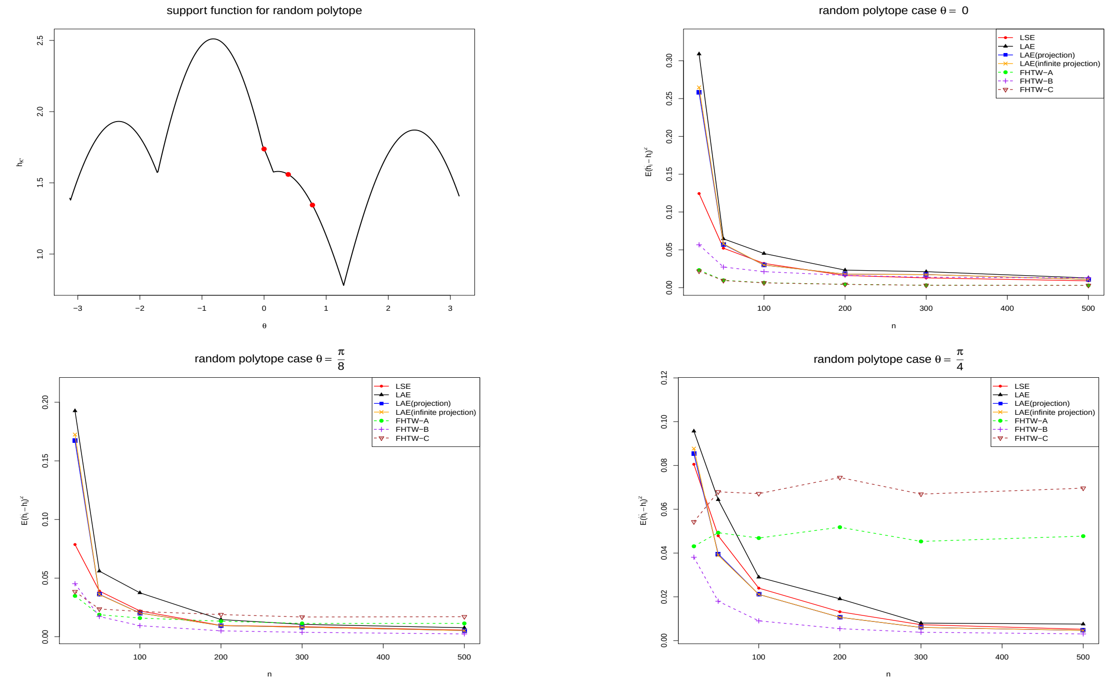


Figure B.3: Point estimation error when K^* is a random polytope

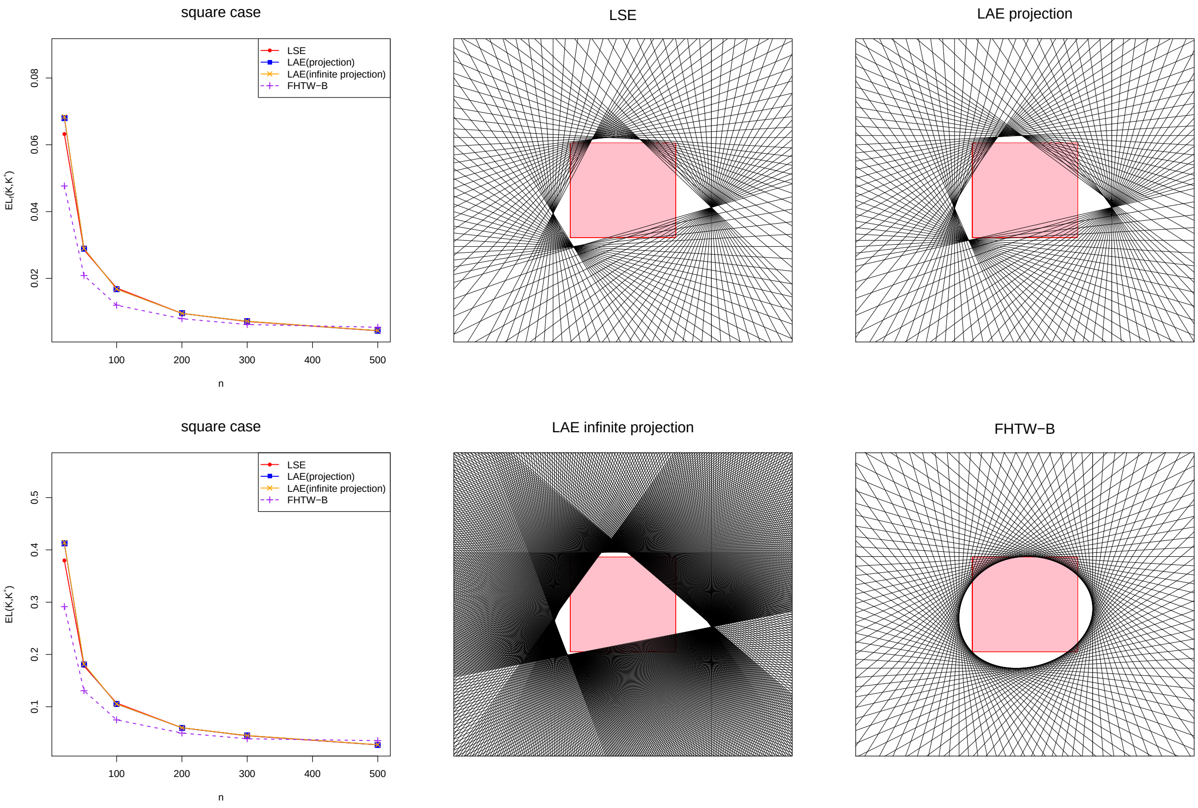


Figure B.4: Set estimation when K^* is a square

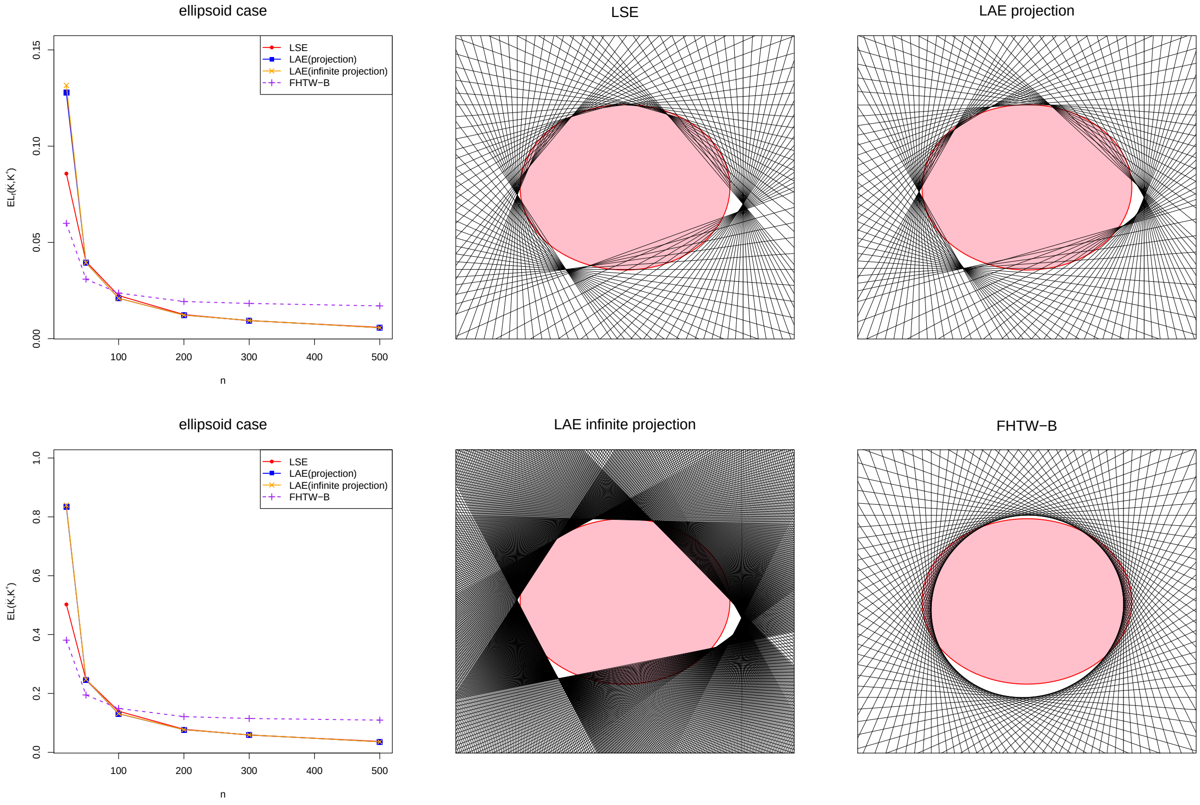


Figure B.5: Set estimation when K^* is an ellipsoid

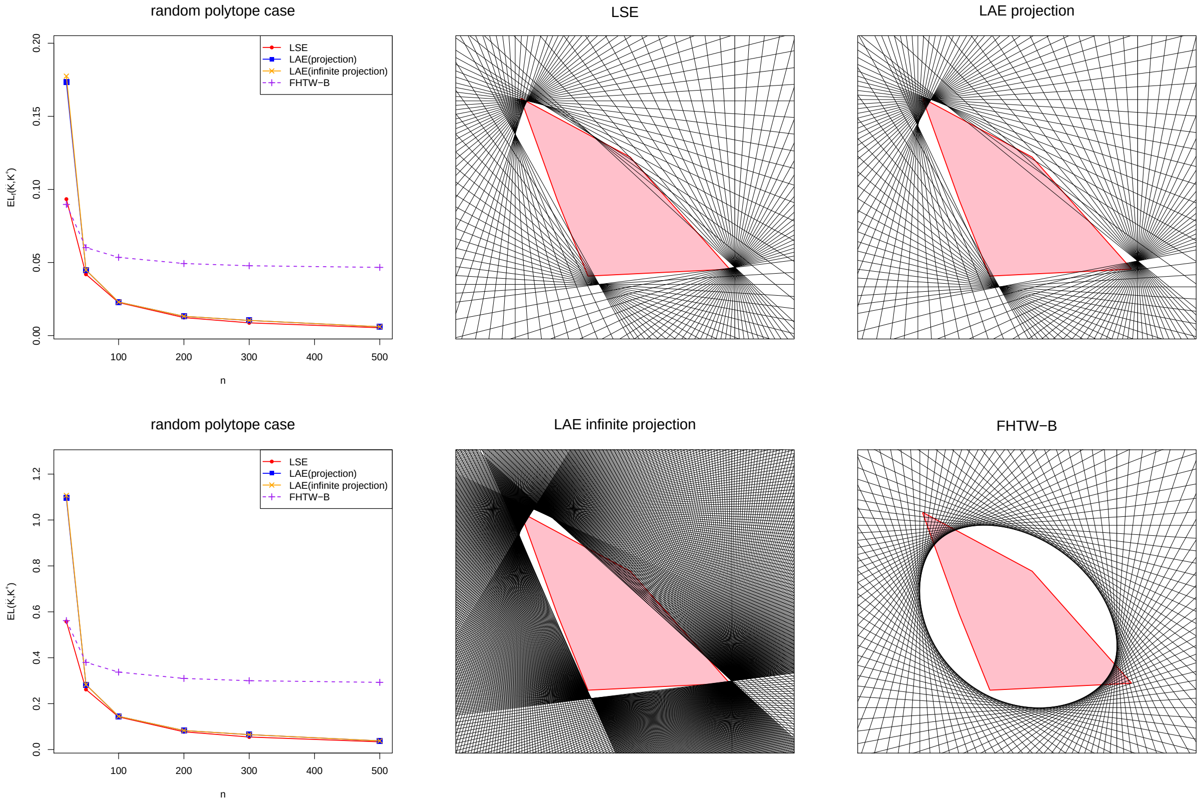


Figure B.6: Set estimation when K^* is a random polytope

Appendix C

Proofs for Chapter 5

C.1 Proof of Lemma 1

Recalling that $K^\dagger$ denotes the pseudoinverse of K, our proof is based on the linear transformation

$$z := n^{-1/2}(K^\dagger)^{1/2}\theta \quad \Longleftrightarrow \quad \theta = \sqrt{n}K^{1/2}z.$$

as well as the new function $\mathcal{J}_n(z) := \mathcal{L}_n(\sqrt{n}\sqrt{K}z)$ and its population equivalent $\mathcal{J}(z) := \mathbb{E}\mathcal{J}_n(z)$. Ordinary gradient descent on $\mathcal{J}_n$ with stepsize α takes the form

$$z^{t+1} = z^t - \alpha\nabla\mathcal{J}_n(z^t) \;=\; z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}_n(\sqrt{n}\sqrt{K}z^t). \tag{C.1}$$

If we transform this update on z back to an equivalent one on θ by multiplying both sides by $\sqrt{n}\sqrt{K}$, we see that ordinary gradient descent on $\mathcal{J}_n$ is equivalent to the kernel boosting update $\theta^{t+1} = \theta^t - \alpha nK\nabla\mathcal{L}_n(\theta^t)$.

Our goal is to analyze the behavior of the update (C.1) in terms of the population cost $\mathcal{J}(z^t)$. Thus, our problem is one of analyzing a noisy form of gradient descent on the function $\mathcal{J}$, where the noise is induced by the difference between the empirical gradient operator $\nabla\mathcal{J}_n$ and the population gradient operator $\nabla\mathcal{J}$.

Recall that the $\mathcal{L}$ is M-smooth by assumption. Since the kernel matrix K has been normalized to have largest eigenvalue at most one, the function $\mathcal{J}$ is also M-smooth, whence

$$\mathcal{J}(z^{t+1}) \leq \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, d^t\rangle + \frac{M}{2}\|d^t\|_2^2,$$
$$\text{where} \quad d^t := z^{t+1} - z^t = -\alpha\nabla\mathcal{J}_n(z^t).$$

Morever, since the function $\mathcal{J}$ is convex, we have $\mathcal{J}(z^*) \geq \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, z^* - z^t\rangle$, whence

$$\begin{aligned}\mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\leq \langle\nabla\mathcal{J}(z^t),\, d^t + z^t - z^*\rangle + \frac{M}{2}\|d^t\|_2^2\\ &= \langle\nabla\mathcal{J}(z^t),\, z^{t+1} - z^*\rangle + \frac{M}{2}\|d^t\|_2^2. \end{aligned}\tag{C.2}$$

Now define the difference of the squared errors $V^t := \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|z^{t+1} - z^*\|_2^2\Big\}$. By some simple algebra, we have

$$\begin{aligned} V^t \; &= \; \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|d^t + z^t - z^*\|_2^2\Big\} \\ &= -\,\langle d^t,\, z^t - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\,\langle d^t,\, -d^t + z^{t+1} - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\,\langle d^t,\, z^{t+1} - z^*\rangle + \frac{1}{2}\|d^t\|_2^2. \end{aligned}$$

Substituting back into equation (C.2) yields

$$\begin{aligned} \mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\leq \frac{1}{\alpha}V^t + \langle \nabla\mathcal{J}(z^t) + \frac{d^t}{\alpha},\, z^{t+1} - z^*\rangle \\ &= \; \frac{1}{\alpha}V^t + \langle \nabla\mathcal{J}(z^t) - \nabla\mathcal{J}_n(z^t),\, z^{t+1} - z^*\rangle, \end{aligned}$$

where we have used the fact that $\frac{1}{\alpha} \geq M$ by our choice of stepsize α.

Finally, we transform back to the original variables $\theta = \sqrt{n}\sqrt{K}z$, using the relation $\nabla\mathcal{J}(z) = \sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta)$, so as to obtain the bound

$$\begin{aligned} \mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*) \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} \\ + \langle \nabla\mathcal{L}(\theta^t) - \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^*\rangle. \end{aligned}$$

Note that the optimality of θ^* implies that $\nabla\mathcal{L}(\theta^*) = 0$. Combined with m-strong convexity, we are guaranteed that $\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*)$, and hence

$$\begin{aligned} \frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} \\ + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle, \end{aligned}$$

as claimed.

C.2 Proof of Lemma 2

We split our proof into two cases, depending on whether we are dealing with the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, or a classification loss with uniformly bounded gradient ($\|\phi'\|_\infty \leq 1$).

C.2.1 Least-squares case

The least-squares loss is m-strongly convex with $m = M = 1$. Moreover, the difference between the population and empirical gradients can be written as $\nabla\mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla\mathcal{L}_n(\theta^* +$

$\widetilde{\delta}) = \frac{\sigma}{n}(w_1, \ldots, w_n)$, where the random variables $\{w_i\}_{i=1}^n$ are i.i.d. and sub-Gaussian with parameter 1. Consequently, we have

$$|\langle \nabla \mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla \mathcal{L}_n(\theta^* + \widetilde{\delta}),\, \Delta\rangle| = \left|\frac{\sigma}{n}\sum_{i=1}^n w_i \Delta(x_i)\right|.$$

Under these conditions, one can show (see [144] for reference) that

$$\left|\frac{\sigma}{n}\sum_{i=1}^n w_i \Delta(x_i)\right| \leq 2\delta_n \|\Delta\|_n + 2\delta_n^2 \|\Delta\|_{\mathscr{H}} + \frac{1}{16}\|\Delta\|_n^2, \tag{C.3}$$

which implies that Lemma 2 holds with $c_3 = 16$.

C.2.2 Gradient-bounded ϕ-functions

We now turn to the proof of Lemma 2 for gradient bounded ϕ-functions. First, we claim that it suffices to prove the bound (5.23) for functions $g \in \partial\mathscr{H}$ and $\|g\|_{\mathscr{H}} = 1$ where $\partial\mathscr{H} := \{f - g \mid f, g \in \mathscr{H}\}$. Indeed, suppose that it holds for all such functions, and that we are given a function Δ with $\|\Delta\|_{\mathscr{H}} > 1$. By assumption, we can apply the inequality (5.23) to the new function $g := \Delta/\|\Delta\|_{\mathscr{H}}$, which belongs to $\partial\mathscr{H}$ by nature of the subspace $\mathscr{H} = \overline{\text{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$.

Applying the bound (5.23) to g and then multiplying both sides by $\|\Delta\|_{\mathscr{H}}$, we obtain

$$\begin{aligned}
&\langle \nabla \mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla \mathcal{L}_n(\theta^* + \widetilde{\delta}),\, \Delta\rangle \\
&\qquad \leq 2\delta_n \|\Delta\|_n + 2\delta_n^2 \|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\frac{\|\Delta\|_n^2}{\|\Delta\|_{\mathscr{H}}} \\
&\qquad \leq 2\delta_n \|\Delta\|_n + 2\delta_n^2 \|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2,
\end{aligned}$$

where the second inequality uses the fact that $\|\Delta\|_{\mathscr{H}} > 1$ by assumption.

In order to establish the bound (5.23) for functions with $\|g\|_{\mathscr{H}} = 1$, we first prove it uniformly over the set $\{g \mid \|g\|_{\mathscr{H}} = 1, \quad \|g\|_n \leq t\}$, where $t > 1$ is a fixed radius (of course, we restrict our attention to those radii t for which this set is non-empty.) We then extend the argument to one that is also uniform over the choice of t by a "peeling" argument.

Define the random variable

$$\mathcal{Z}_n(t) := \sup_{\Delta, \widetilde{\delta} \in \mathcal{E}(t,1)} \langle \nabla \mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla \mathcal{L}_n(\theta^* + \widetilde{\delta}),\, \Delta\rangle. \tag{C.4}$$

The following two lemmas, respectively, bound the mean of this random variable, and its deviations above the mean:

Lemma 5. *For any $t > 0$, the mean is upper bounded as*

$$\mathbb{E}\mathcal{Z}_n(t) \leq \sigma \mathcal{G}_n(\mathcal{E}(t, 1)), \tag{C.5}$$

where $\sigma := 2M + 4C_{\mathscr{H}}$.

Lemma 6. *There are universal constants (c_1, c_2) such that*

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq \mathbb{E}\mathcal{Z}_n(t) + \alpha\Big] \leq c_1 \exp\Big(-\frac{c_2 n \alpha^2}{t^2}\Big). \tag{C.6}$$

See Appendices C.2.3 and C.2.4 for the proofs of these two claims.

Equipped with Lemmas 5 and 6, we now prove inequality (5.23). We divide our argument into two cases:

Case $t = \delta_n$ We first prove inequality (5.23) for $t = \delta_n$. From Lemma 5, we have

$$\mathbb{E}\mathcal{Z}_n(\delta_n) \leq \sigma \mathcal{G}_n(\mathcal{E}(\delta_n, 1)) \overset{(i)}{\leq} \delta_n^2, \tag{C.7}$$

where inequality (i) follows from the definition of δ_n in inequality (5.12). Setting $\alpha = \delta_n^2$ in expression (C.6) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(\delta_n) \geq 2\delta_n^2\Big] \leq c_1 \exp\left(-c_2 n \delta_n^2\right), \tag{C.8}$$

which establishes the claim for $t = \delta_n$.

Case $t > \delta_n$ On the other hand, for any $t > \delta_n$, we have

$$\mathbb{E}\mathcal{Z}_n(t) \overset{(i)}{\leq} \sigma \mathcal{G}_n(\mathcal{E}(t, 1)) \overset{(ii)}{\leq} t\sigma \frac{\mathcal{G}_n(\mathcal{E}(t, 1))}{t} \leq t\delta_n,$$

where step (i) follows from Lemma 5, and step (ii) follows because the function $u \mapsto \frac{\mathcal{G}_n(\mathcal{E}(u,1))}{u}$ is non-increasing on the positive real line. (This non-increasing property is a direct consequence of the star-shaped nature of $\partial\mathscr{H}$.) Finally, using this upper bound on expression $\mathbb{E}\mathcal{Z}_n(\delta_n)$ and setting $\alpha = t^2 m/(4c_3)$ in the tail bound (C.6) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq t\delta_n + \frac{t^2 m}{4c_3}\Big] \leq c_1 \exp\left(-c_2 n m^2 t^2\right). \tag{C.9}$$

Note that the precise values of the universal constants c_2 may change from line to line throughout this section.

Peeling argument Equipped with the tail bounds (C.8) and (C.9), we are now ready to complete the peeling argument. Let $\mathcal{A}$ denote the event that the bound (5.23) is violated for some function $g \in \partial\mathscr{H}$ with $\|g\|_{\mathscr{H}} = 1$. For real numbers $0 \leq a < b$, let $\mathcal{A}(a, b)$ denote the event that it is violated for some function such that $\|g\|_n \in [a, b]$, and $\|g\|_{\mathscr{H}} = 1$. For $k = 0, 1, 2, \ldots$, define $t_k = 2^k \delta_n$. We then have the decomposition $\mathcal{E} = (0, t_0) \cup (\bigcup_{k=0}^{\infty} \mathcal{A}(t_k, t_{k+1}))$ and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{A}(0, \delta_n)] + \sum_{k=1}^{\infty} \mathbb{P}[\mathcal{A}(t_k, t_{k+1})]. \tag{C.10}$$

From the bound (C.8), we have $\mathbb{P}[\mathcal{A}(0, \delta_n)] \leq c_1 \exp\left(-c_2 n \delta_n^2\right)$. On the other hand, suppose that $\mathcal{A}(t_k, t_{k+1})$ holds, meaning that there exists some function g with $\|g\|_{\mathscr{H}} = 1$ and $\|g\|_n \in [t_k, t_{k+1}]$ such that

$$\begin{aligned}\langle \nabla\mathcal{L}(\theta^* + \widetilde{\delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\delta}),\, g\rangle &> 2\delta_n \|g\|_n + 2\delta_n^2 + \frac{m}{c_3}\|g\|_n^2 \\ &\overset{(i)}{\geq} 2\delta_n t_k + 2\delta_n^2 + \frac{m}{c_3} t_k^2 \\ &\overset{(ii)}{\geq} \delta_n t_{k+1} + 2\delta_n^2 + \frac{m}{4c_3} t_{k+1}^2,\end{aligned}$$

where step (i) uses the $\|g\|_n \geq t_k$ and step (ii) uses the fact that $t_{k+1} = 2t_k$. This lower bound implies that $\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4c_3}$ and applying the tail bound (C.9) yields

$$\begin{aligned}\mathbb{P}(\mathcal{A}(t_k, t_{k+1})) &\leq \mathbb{P}(\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4c_3}) \\ &\leq \exp\left(-c_2 n m^2 2^{2k+2} \delta_n^2\right).\end{aligned}$$

Substituting this inequality and our earlier bound (C.8) into equation (C.10) yields

$$\mathbb{P}(\mathcal{E}) \leq c_1 \exp(-c_2 n m^2 \delta_n^2),$$

where the reader should recall that the precise values of universal constants may change from line-to-line. This concludes the proof of Lemma 2.

C.2.3 Proof of Lemma 5

Recalling the definitions (5.1) and (5.3) of $\mathcal{L}$ and $\mathcal{L}_n$, we can write

$$\mathcal{Z}_n(t) = \sup_{\Delta, \widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} (\phi'(y_i, \theta_i^* + \widetilde{\delta}_i) - \mathbb{E}\phi'(y_i, \theta_i^* + \widetilde{\delta}_i))\Delta_i$$

Note that the vectors Δ and $\widetilde{\delta}$ contain function values of the form $f(x_i) - f^*(x_i)$ for functions $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$. Recall that the kernel function is bounded uniformly by one. Consequently, for any function $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, we have

$$\begin{aligned} |f(x) - f^*(x)| &= |\langle f - f^*, \, \mathbb{K}(\cdot, x)\rangle_{\mathscr{H}}| \\ &\leq \|f - f^*\|_{\mathscr{H}} \|\mathbb{K}(\cdot, x)\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}. \end{aligned}$$

Thus, we can restrict our attention to vectors $\Delta, \widetilde{\delta}$ with $\|\Delta\|_\infty, \|\widetilde{\delta}\|_\infty \leq 2C_{\mathscr{H}}$ from hereonwards.

Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, define the symmetrized variable

$$\tilde{\mathcal{Z}}_n(t) := \sup_{\Delta, \widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \phi'(y_i, \theta_i^* + \widetilde{\delta}_i) \; \Delta_i. \tag{C.11}$$

By a standard symmetrization argument [138], we have $\mathbb{E}_y[\mathcal{Z}_n(t)] \leq 2\mathbb{E}_{y,\epsilon}[\tilde{\mathcal{Z}}_n(t)]$. Moreover, since

$$\phi'(y_i, \theta_i^* + \widetilde{\delta}_i) \; \Delta_i \leq \frac{1}{2}\Big(\phi'(y_i, \theta_i^* + \widetilde{\delta}_i)\Big)^2 + \frac{1}{2}\Delta_i^2$$

we have

$$\begin{aligned} \mathbb{E}\mathcal{Z}_n(t) \leq & \mathbb{E} \sup_{\widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \big(\phi'(y_i, \theta_i^* + \widetilde{\delta}_i)\big)^2 + \mathbb{E} \sup_{\Delta \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i^2 \\ \leq & \underbrace{2\,\mathbb{E} \sup_{\widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \phi'(y_i, \theta_i^* + \widetilde{\delta}_i)}_{T_1} + 4C_{\mathscr{H}} \underbrace{\mathbb{E} \sup_{\Delta \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i}_{T_2}, \end{aligned}$$

where the second inequality follows by applying the Rademacher contraction inequality [92], using the fact that $\|\phi'\|_\infty \leq 1$ for the first term, and $\|\Delta\|_\infty \leq 2C_{\mathscr{H}}$ for the second term.

Focusing first on the term T_1, since $\mathbb{E}[\varepsilon_i \phi'(y_i, \theta_i^*)] = 0$, we have

$$\begin{aligned} T_1 &= \mathbb{E} \sup_{\widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \underbrace{\Big(\phi'(y_i, \theta_i^* + \widetilde{\delta}_i) - \phi'(y_i; \theta_i^*)\Big)}_{\varphi_i(\widetilde{\delta}_i)} \\ &\overset{(i)}{\leq} M\mathbb{E} \sup_{\widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \widetilde{\delta}_i \\ &\overset{(ii)}{\leq} \sqrt{\frac{\pi}{2}} M\mathcal{G}_n(\mathcal{E}(t,1)), \end{aligned}$$

where step (i) follows since each function φ_i is M-Lipschitz by assumption; and step (ii) follows since the Gaussian complexity upper bounds the Rademacher complexity up to a factor of $\sqrt{\frac{\pi}{2}}$. Similarly, we have

$$T_2 \leq \sqrt{\frac{\pi}{2}}\, \mathcal{G}_n(\mathcal{E}(t,1)),$$

and putting together the pieces yields the claim.

C.2.4 Proof of Lemma 6

Recall the definition (C.11) of the symmetrized variable $\tilde{\mathcal{Z}}_n$. By a standard symmetrization argument [138], there are universal constants c_1, c_2 such that

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq \mathbb{E}\mathcal{Z}_n[t] + c_1\alpha\Big] \leq c_2 \mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \geq \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha\Big].$$

Since $\{\varepsilon_i\}_{i=1}^n$ are $\{y_i\}_{i=1}^n$ are independent, we can study $\tilde{\mathcal{Z}}_n(t)$ conditionally on $\{y_i\}_{i=1}^n$. Viewed as a function of $\{\varepsilon_i\}_{i=1}^n$, the function $\tilde{\mathcal{Z}}_n(t)$ is convex and Lipschitz with respect to the Euclidean norm with parameter

$$L^2 := \sup_{\Delta, \widetilde{\delta} \in \mathcal{E}(t,1)} \frac{1}{n^2} \sum_{i=1}^n \Big(\phi'(y_i, \theta_i^* + \widetilde{\delta}_i)\; \Delta_i\Big)^2 \;\leq\; \frac{t^2}{n},$$

where we have used the facts that $\|\phi'\|_\infty \leq 1$ and $\|\Delta\|_n \leq t$. By Ledoux's concentration for convex and Lipschitz functions [91], we have

$$\mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \geq \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha \;\mid\; \{y_i\}_{i=1}^n\Big] \leq c_3 \exp\Big(-c_4 \frac{n\alpha^2}{t^2}\Big).$$

Since the right-hand side does not involve $\{y_i\}_{i=1}^n$, the same bound holds unconditionally over the randomness in both the Rademacher variables and the sequence $\{y_i\}_{i=1}^n$. Consequently, the claimed bound (C.6) follows, with suitable redefinitions of the universal constants.

C.3 Proof of Lemma 3

We first require an auxiliary lemma, which we state and prove in the following section. We then prove Lemma 3 in Section C.3.2.

C.3.1 An auxiliary lemma

The following result relates the Hilbert norm of the error to the difference between the empirical and population gradients:

Lemma 7. *For any convex and differentiable loss function $\mathcal{L}$, the kernel boosting error $\Delta^{t+1} := \theta^{t+1} - \theta^*$ satisfies the bound*

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}} \|\Delta^{t+1}\|_{\mathscr{H}} \\ + \alpha \langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t), \, \Delta^{t+1} \rangle. \end{aligned} \tag{C.12}$$

Proof. Recall that $\|\Delta^t\|_{\mathscr{H}}^2 = \|\theta^t - \theta^*\|_{\mathscr{H}}^2 = \|z^t - z^*\|_2^2$ by definition of the Hilbert norm. Let us define the population update operator G on the population function $\mathcal{J}$ and the empirical update operator G_n on $\mathcal{J}_n$ as

$$\begin{aligned} G(z^t) &:= z^t - \alpha \nabla \mathcal{J}(\sqrt{n}\sqrt{K} z^t), \\ \text{and} \quad z^{t+1} &:= G_n(z^t) = z^t - \alpha \nabla \mathcal{J}_n(\sqrt{n}\sqrt{K} z^t). \end{aligned} \tag{C.13}$$

Since $\mathcal{J}$ is convex and smooth, it follows from standard arguments in convex optimization that G is a non-expansive operator—viz.

$$\|G(x) - G(y)\|_2 \leq \|x - y\|_2 \qquad \text{for all } x, y \in \mathcal{C}. \tag{C.14}$$

In addition, we note that the vector z^* is a fixed point of G—that is, $G(z^*) = z^*$. From these ingredients, we have

$$\begin{aligned} &\|\Delta^{t+1}\|_{\mathscr{H}}^2 \\ &= \langle z^{t+1} - z^*, \, G_n(z^t) - G(z^t) + G(z^t) - z^* \rangle \\ &\overset{(i)}{\leq} \|z^{t+1} - z^*\|_2 \|G(z^t) - G(z^*)\|_2 \\ &\quad + \alpha \langle \sqrt{n}\sqrt{K}[\nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t)], \, z^{t+1} - z^* \rangle \\ &\overset{(ii)}{\leq} \|\Delta^{t+1}\|_{\mathscr{H}} \|\Delta^t\|_{\mathscr{H}} \\ &\quad + \alpha \langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t), \, \Delta^{t+1} \rangle \end{aligned}$$

where step (i) follows by applying the Cauchy-Schwarz to control the inner product, and step (ii) follows since $\Delta^{t+1} = \sqrt{n}\sqrt{K}(z^{t+1} - z^*)$, and the square root kernel matrix $\sqrt{K}$ is symmetric. □

C.3.2 Proof of Lemma 3

We now prove Lemma 3. The argument makes use of Lemmas 1 and 2 combined with Lemma 7.

In order to prove inequality (5.24), we follow an inductive argument. Instead of proving (5.24) directly, we prove a slightly stronger relation which implies it, namely

$$\max\{1, \|\Delta^t\|_{\mathscr{H}}^2\} \leq \max\{1, \|\Delta^0\|_{\mathscr{H}}^2\} + t\delta_n^2 \frac{4M}{\widetilde{\gamma} m}. \tag{C.15}$$

Here $\widetilde{\gamma}$ and c_3 are constants linked by the relation

$$\widetilde{\gamma} := \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2. \tag{C.16}$$

We claim that it suffices to prove that the error iterates Δ^{t+1} satisfy the inequality (C.15). Indeed, if we take inequality (C.15) as given, then we have

$$\|\Delta^t\|_{\mathscr{H}}^2 \leq \max\{1, \|\Delta^0\|_{\mathscr{H}}^2\} + \frac{1}{2\widetilde{\gamma}} \leq C_{\mathscr{H}}^2,$$

where we used the definition $C_{\mathscr{H}}^2 = 2\max\{\|\theta^*\|_{\mathscr{H}}^2, \; 32\}$. Thus, it suffices to focus our attention on proving inequality (C.15).

For $t = 0$, it is trivially true. Now let us assume inequality (C.15) holds for some $t \leq \frac{m}{8M\delta_n^2}$, and then prove that it also holds for step $t+1$.

If $\|\Delta^{t+1}\|_{\mathscr{H}} < 1$, then inequality (C.15) follows directly. Therefore, we can assume without loss of generality that $\|\Delta^{t+1}\|_{\mathscr{H}} \geq 1$.

We break down the proof of this induction into two steps:

- First, we show that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$ so that Lemma 2 is applicable.
- Second, we show that the bound (C.15) holds and thus in fact $\|\Delta^{t+1}\|_{\mathscr{H}} \leq C_{\mathscr{H}}$.

Throughout the proof, we condition on the event $\mathcal{E}$ and $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Lemma 2 guarantees that $\mathbb{P}(\mathcal{E}^c) \leq c_1 \exp(-c_2 \frac{m^2 n \delta_n^2}{\sigma^2})$ whereas $\mathbb{P}(\mathcal{E}_0) \geq 1 - \mathbb{E}^{-n}$ follows from the fact that Y^2 is sub-exponential with parameter $\sigma^2 n$ and applying Hoeffding's inequality. Putting things together yields an upper bound on the probability of the complementary event, namely

$$\mathbb{P}(\mathcal{E}^c \cup \mathcal{E}_0^c) \leq 2c_1 \exp(-C_2 n \delta_n^2)$$

with $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

Showing that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$ In this step, we assume that inequality (C.15) holds at step t, and show that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$. Recalling that $z := \frac{(K^\dagger)^{1/2}}{\sqrt{n}}\theta$, our update can be written as

$$\begin{aligned} z^{t+1} - z^* = z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta^t) - z^* \\ + \alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t)). \end{aligned}$$

Applying the triangle inequality yields the bound

$$\begin{aligned} \|z^{t+1} - z^*\|_2 \leq \| \underbrace{z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta^t)}_{G(z^t)} - z^*\|_2 \\ + \|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2 \end{aligned}$$

where the population update operator G was previously defined (C.13), and observed to be non-expansive (C.14). From this non-expansiveness, we find that

$$\|z^{t+1} - z^*\|_2 \leq \|z^t - z^*\|_2 + \|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2,$$

Note that the ℓ_2 norm of z corresponds to the Hilbert norm of θ. This implies

$$\|\Delta^{t+1}\|_{\mathscr{H}} \leq \|\Delta^t\|_{\mathscr{H}} + \underbrace{\|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2}_{:=T}$$

Observe that because of uniform boundedness of the kernel by one, the quantity T can be bounded as

$$T \leq \alpha\sqrt{n}\|\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2 = \alpha\sqrt{n}\frac{1}{n}\|v - \mathbb{E}v\|_2,$$

where we have define the vector $v \in \mathbb{R}^n$ with coordinates $v_i := \phi'(y_i, \theta_i^t)$. For functions ϕ satisfying the gradient boundedness and $m - M$ condition, since $\theta^t \in \mathbb{B}_{\mathscr{H}}(\theta^*, C_{\mathscr{H}})$, each coordinate of the vectors v and $\mathbb{E}v$ is bounded by 1 in absolute value. We consequently have

$$T \leq \alpha \leq C_{\mathscr{H}},$$

where we have used the fact that $\alpha \leq m/M < 1 \leq \frac{C_{\mathscr{H}}}{2}$. For least-squares ϕ we instead have

$$T \leq \alpha\frac{\sqrt{n}}{n}\|y - \mathbb{E}[y \mid x]\|_2 =: \frac{\alpha}{\sqrt{n}}Y \leq \sqrt{2}\sigma \leq C_{\mathscr{H}}$$

conditioned on the event $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Since Y^2 is sub-exponential with parameter $\sigma^2 n$ it follows by Hoeffding's inequality that $\mathbb{P}(\mathcal{E}_0) \geq 1 - \mathbb{E}^{-n}$.

Putting together the pieces yields that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$, as claimed.

Completing the induction step We are now ready to complete the induction step for proving inequality (C.15) using Lemma 1 and Lemma 2 since $\|\Delta^{t+1}\|_{\mathscr{H}} \geq 1$. We split the argument into two cases separately depending on whether or not $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$. In general we can assume that $\|\Delta^{t+1}\|_{\mathscr{H}} > \|\Delta^t\|_{\mathscr{H}}$, otherwise the induction inequality (C.15) satisfies trivially.

Case 1 When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$, inequality (5.23) implies that

$$\begin{aligned}\langle\nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^*+\widetilde{\Delta}),\, \Delta^{t+1}\rangle& \\ \leq 4\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2,&\end{aligned} \tag{C.17}$$

Combining Lemma 7 and inequality (C.17), we obtain

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq & \|\Delta^t\|_{\mathscr{H}}\|\Delta^{t+1}\|_{\mathscr{H}} + 4\alpha\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ \Longrightarrow \|\Delta^{t+1}\|_{\mathscr{H}} \leq & \frac{1}{1-\alpha\delta_n^2\frac{m}{c_3}}\big[\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2\big], \end{aligned} \tag{C.18}$$

where the last inequality uses the fact that $\|\Delta^{t+1}\|_n \leq \delta_n\|\Delta^{t+1}\|_{\mathscr{H}}$.

Case 2 When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n < \|\Delta^{t+1}\|_n$, we use our assumption $\|\Delta^{t+1}\|_{\mathscr{H}} \geq \|\Delta^t\|_{\mathscr{H}}$ together with Lemma 7 and inequality (5.23) which guarantee that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq & \|\Delta^t\|_{\mathscr{H}}^2 + 2\alpha\langle\nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle \\ \leq & \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\delta_n\|\Delta^{t+1}\|_n + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2.\end{aligned}$$

Using the elementary inequality $2ab \leq a^2 + b^2$, we find that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq & \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\left[m\widetilde{\gamma}\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ \leq & \|\Delta^t\|_{\mathscr{H}}^2 + \alpha\frac{m}{4}\|\Delta^{t+1}\|_n^2 + \frac{2\alpha\delta_n^2}{\widetilde{\gamma}m},\end{aligned} \tag{C.19}$$

where in the final step, we plug in the constants $\widetilde{\gamma}, c_3$ which satisfy equation (C.16).

Now Lemma 1 implies that

$$\begin{aligned}\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq & D^t + 4\|\Delta^{t+1}\|_n\delta_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ \overset{(i)}{\leq} & D^t + 4\left[\widetilde{\gamma}m\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2,\end{aligned}$$

where step (i) again uses $2ab \leq a^2 + b^2$. Thus, we have $\frac{m}{4}\|\Delta^{t+1}\|_n^2 \leq D^t + \frac{1}{\widetilde{\gamma}m}\delta_n^2$. Together with expression (C.19), we find that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq & \|\Delta^t\|_{\mathscr{H}}^2 + \frac{1}{2}(\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2) + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2 \\ \Longrightarrow \|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq & \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2.\end{aligned} \tag{C.20}$$

Combining the pieces By combining the two previous cases, we arrive at the bound

$$\begin{aligned} &\max\left\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\right\} \\ \leq &\max\left\{1, \kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2, \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2\right\}, \end{aligned} \tag{C.21}$$

where $\kappa := \frac{1}{(1-\alpha\delta_n^2\frac{m}{c_3})}$ and we used that $\alpha \leq \min\{\frac{1}{M}, M\}$.

Now it is only left for us to show that with the constant c_3 chosen such that $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, we have

$$\kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2.$$

Define the function $f : (0, C_{\mathscr{H}}] \to \mathbb{R}$ via $f(\xi) := \kappa^2(\xi + 4\alpha\delta_n^2)^2 - \xi^2 - \frac{4M}{\widetilde{\gamma}m}\delta_n^2$. Since $\kappa \geq 1$, in order to conclude that $f(\xi) < 0$ for all $\xi \in (0, C_{\mathscr{H}}]$, it suffices to show that $\operatorname{argmin}_{x\in\mathbb{R}} f(x) < 0$ and $f(C_{\mathscr{H}}) < 0$. The former is obtained by basic algebra and follows directly from $\kappa \geq 1$. For the latter, since $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, $\alpha < \frac{1}{M}$ and $\delta_n^2 \leq \frac{M^2}{m^2}$ it thus suffices to show

$$\frac{1}{(1 - \frac{M}{8m})^2} \leq \frac{4M}{m} + 1$$

Since $(4x+1)(1-\frac{x}{8})^2 \geq 1$ for all $x \leq 1$ and $\frac{m}{M} \leq 1$, we conclude that $f(C_{\mathscr{H}}) < 0$.

Now that we have established $\max\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\} \leq \max\{1, \|\Delta^t\|_{\mathscr{H}}^2\} + \frac{4M}{\widetilde{\gamma}m}\delta_n^2$, the induction step (C.15) follows. which completes the proof of Lemma 3.

C.4 Proof of Lemma 4

Recall that the LogitBoost algorithm is based on logistic loss $\phi(y,\theta) = \ln(1 + e^{-y\theta})$, whereas the AdaBoost algorithm is based on the exponential loss $\phi(y,\theta) = \exp(-y\theta)$. We now verify the m-M-condition for these two losses with the corresponding parameters specified in Lemma 4.

C.4.1 m-M-condition for logistic loss

The first and second derivatives are given by

$$\frac{\partial\phi(y,\theta)}{\partial\theta} = \frac{-ye^{-y\theta}}{1+e^{-y\theta}}, \qquad \text{and} \qquad \frac{\partial^2\phi(y,\theta)}{(\partial\theta)^2} = \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2}.$$

It is easy to check that $|\frac{\partial\phi(y,\theta)}{\partial\theta}|$ is uniformly bounded by $B = 1$.

Turning to the second derivative, recalling that $y \in \{-1, +1\}$, it is straightforward to show that

$$\max_{y \in \{-1,+1\}} \sup_{\theta} \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \leq \frac{1}{4},$$

which implies that $\frac{\partial \phi(y,\theta)}{\partial \theta}$ is a 1/4-Lipschitz function of θ, i.e. with $M = 1/4$.

Our final step is to compute a value for m by deriving a uniform lower bound on the Hessian. For this step, we need to exploit the fact that $\theta = f(x)$ must arise from a function f such that $\|f\|_{\mathscr{H}} \leq D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$. Since $\sup_x \mathbb{K}(x,x) \leq 1$ by assumption, the reproducing relation for RKHS then implies that $|f(x)| \leq D$. Combining this inequality with the fact that $y \in \{-1, 1\}$, it suffices to lower the bound the quantity

$$\begin{aligned} \min_{y \in \{-1,+1\}} \min_{|\theta| \leq D} \left| \frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} \right| &= \min_{|y| \leq 1} \min_{|\theta| \leq D} \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \\ &\geq \underbrace{\frac{1}{e^{-D} + e^{D} + 2}}_{m}, \end{aligned}$$

which completes the proof for the logistic loss.

C.4.2 m-M-condition for AdaBoost

The AdaBoost algorithm is based on the cost function $\phi(y, \theta) = e^{-y\theta}$, which has first and second derivatives (with respect to its second argument) given by

$$\frac{\partial \phi(y,\theta)}{\partial \theta} = -ye^{-y\theta}, \qquad \text{and} \qquad \frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} = e^{-y\theta}.$$

As in the preceding argument for logistic loss, we have the bound $|y| \leq 1$ and $|\theta| \leq D$. By inspection, the absolute value of the first derivative is uniformly bounded $B := e^D$, whereas the second derivative always lies in the interval $[m, M]$ with $M := e^D$ and $m := e^{-D}$, as claimed.

Moreover, as shown by our later results, under suitable regularity conditions, the expectation of the minimum squared error ρ_n^2 is proportional to the *statistical minimax risk* $\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[\mathcal{L}(\widehat{f}) - \mathcal{L}(f)]$, where the infimum is taken over all possible estimators $\widehat{f}$. Note that the minimax risk provides a fundamental lower bound on the performance of any estimator uniformly over the function space $\mathcal{F}$. Coupled with our stopping time guarantee (5.5), we are guaranteed that our estimate achieves the minimax risk up to constant factors. As a result, our bounds are unimprovable in general (see Corollary 4).

Bibliography

[1] L. Addario-Berry, N. Broutin, L. Devroye, G. Lugosi, et al. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[2] A. D. Alexandrov. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser.*, 6:3–35, 1939.

[3] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, 2014.

[4] R. S. Anderssen and P. M. Prenter. A formal comparison of methods proposed for the numerical solution of first kind integral equations. *Jour. Australian Math. Soc. (Ser. B)*, 22:488–500, 1981.

[5] E. Arias-Castro, E. Candès, and A. Durand. Detection of an abnormal cluster in a network. *The Bulleting of the Internation Statistical Association, Durban, South Africa*, 2009.

[6] E. Arias-Castro, D. L. Donoho, X. Huo, et al. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *The Annals of Statistics*, 34(1):326–349, 2006.

[7] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[8] Y. Baraud and L. Birgé. Rates of convergence of rho-estimators for sets of densities satisfying shape constraints. *arXiv preprint arXiv:1503.04427*, 2015.

[9] R. E. Barlow, D. J. Bartholomew, J. Bremner, and H. D. Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression.* Wiley New York, 1972.

[10] D. Bartholomew. A test of homogeneity for ordered alternatives. *Biometrika*, 46(1/2):36–48, 1959.

[11] D. Bartholomew. A test of homogeneity for ordered alternatives. ii. *Biometrika*, 46(3/4):328–335, 1959.

[12] P. Bartlett and S. Mendelson. Gaussian and Rademacher complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[13] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[14] P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[15] P. L. Bartlett and M. Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.

[16] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2004.

[17] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Norwell, MA, 2004.

[18] O. Besson. Adaptive detection of a signal whose signature belongs to a cone. In *Proceedings SAM Conference*, 2006.

[19] C. Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30:207–216, 1975.

[20] L. Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[21] L. Breiman et al. Arcing classifier (with discussion and a rejoinder by the author). *Annals of Statistics*, 26(3):801–849, 1998.

[22] V.-E. Brunel. *Non-parametric estimation of convex bodies and convex polytopes*. PhD thesis, Université Pierre et Marie Curie-Paris VI; University of Haifa, 2014.

[23] H. D. Brunk. Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, pages 177–197. Cambridge Univ. Press, London, 1970.

[24] P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013.

[25] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007.

[26] P. Bühlmann and B. Yu. Boosting with L^2 loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2003.

[27] T. T. Cai, A. Guntuboyina, and Y. Wei. Supplement to "adaptive estimation of planar convex sets". 2015.

[28] T. T. Cai, A. Guntuboyina, and Y. Wei. Adaptive estimation of planar convex sets. *To appear in The Annals of Statistics, arXiv:1508.03744*, 2017+.

[29] T. T. Cai and M. G. Low. A framework for estimation of convex functions. *Statistica Sinica*, 25:423–456, 2015.

[30] T. T. Cai, M. G. Low, and Y. Xia. Adaptive confidence intervals for regression functions under shape constraints. *Annals of Statistics*, 41:722–750, 2013.

[31] R. Camoriano, T. Angles, A. Rudi, and L. Rosasco. Nytro: When subsampling meets early stopping. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1403–1411, 2016.

[32] A. Caponetto and Y. Yao. Adaptation for regularization operators in learning theory. Technical Report CBCL Paper #265/AI Technical Report #063, Massachusetts Institute of Technology, September 2006.

[33] A. Caponneto. Optimal rates for regularization operators in learning theory. Technical Report CBCL Paper #264/AI Technical Report #062, Massachusetts Institute of Technology, September 2006.

[34] C. Carolan and R. Dykstra. Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.*, 27(3):557–566, 1999.

[35] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, pages 402–408, 2001.

[36] E. Cator. Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17:714–735, 2011.

[37] S. Chatterjee et al. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

[38] S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *Annals of Statistics*, 2014. to appear.

[39] S. Chatterjee, A. Guntuboyina, B. Sen, et al. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.

[40] D. Chen and R. J. Plemmons. Nonnegativity constraints in numerical analysis. *The birth of numerical analysis*, 10:109–140, 2009.

[41] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[42] D. Chetverikov. Testing regression monotonicity in econometric models. Technical report, UCLA, December 2012. arXiv:1212.6757.

[43] M. M. Deza and E. Deza. *Encyclopedia of distances*. Springer, 2009.

[44] D. L. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, December 1995.

[45] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[46] L. Dumbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, pages 124–152, 2001.

[47] R. L. Dykstra and T. Robertson. On testing monotone tendencies. *Journal of the American Statistical Association*, 78(382):342–350, 1983.

[48] A. W. F. Edwards. *Likelihood*. Cambridge University Press, Cambridge, 1972.

[49] M. S. Ermakov. Minimax detection of a signal in a gaussian white noise. *Theory of Probability & Its Applications*, 35(4):667–679, 1991.

[50] J. Fan and J. Jiang. Nonparametric inference with generalized likelihood ratio tests. *Test*, 16(3):409–444, 2007.

[51] J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and Wilk's phenomenon. *The Annals of Statistics*, pages 153–193, 2001.

[52] N. I. Fisher, P. Hall, B. A. Turlach, and G. S. Watson. On the estimation of a convex set from noisy data on its support function. *Journal of the American Statistical Association*, 92:84–91, 1997.

[53] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[54] K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[55] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, 28(2):337–407, 2000.

[56] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[57] R. J. Gardner. *Geometric Tomography.* Cambridge University Press, second edition, 2006.

[58] R. J. Gardner and M. Kiderlen. A new algorithm for 3D reconstruction from support functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:556–562, 2009.

[59] R. J. Gardner, M. Kiderlen, and P. Milanfar. Convergence of algorithms for reconstructing convex bodies and directional measures. *Annals of Statistics*, 34:1331–1374, 2006.

[60] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[61] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[62] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3656–3666, 2017.

[63] S. A. Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[64] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, et al. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.

[65] L. Goldstein, I. Nourdin, and G. Peccati. Gaussian phase transitions and conic intrinsic volumes: Steining the Steiner formula. *arXiv preprint arXiv:1411.6265*, 2014.

[66] M. Greco, F. Gini, and A. Farina. Radar detection and classification of jamming signals belonging to a cone class. *IEEE Trans. Signal Processing*, 56(5):1984–1993, May 2008.

[67] J. Gregor and F. R. Rannou. Three-dimensional support function estimation and application for projection magnetic resonance imaging. *International Journal of Imaging Systems and Technology*, 12:43–50, 2002.

[68] P. Groeneboom. The concave majorant of Brownian motion. *Ann. Probab.*, 11(4):1016–1027, 1983.

[69] P. Groeneboom. Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA, 1985. Wadsworth.

[70] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*, volume 38. Cambridge University Press, 2014.

[71] P. Groeneboom, G. Jongbloed, and J. A. Wellner. A canonical process for estimation of convex functions: The "invelope" of integrated brownian motion $+t^4$. *Annals of Statistics*, 29:1620–1652, 2001.

[72] P. Groeneboom, G. Jongbloed, and J. A. Wellner. Estimation of convex functions: characterizations and asymptotic theory. *Annals of Statistics*, 29:1653–1698, 2001.

[73] C. Gu. *Smoothing spline ANOVA models*. Springer Series in Statistics. Springer, New York, NY, 2002.

[74] A. Guntuboyina. Optimal rates of convergence for convex set estimation from support functions. *The Annals of Statistics*, 40(1):385–411, 2012.

[75] A. Guntuboyina and B. Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 2013. *To appear*, available at http://arxiv.org/abs/1305.1648.

[76] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

[77] D. L. Hanson and G. Pledger. Consistency in concave regression. *Ann. Statist.*, 4(6):1038–1050, 1976.

[78] M. Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[79] M. Hardt and M. Wootters. Fast matrix completion without the condition number. In *Conference on Learning Theory*, pages 638–678, 2014.

[80] X. Hu and F. T. Wright. Likelihood ratio tests for a class of non-oblique hypotheses. *Ann. Inst. Statist. Math.*, 46(1):137–145, 1994.

[81] Y. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the L_p-metrics. *Theory of Probability and Its Applications*, 31(2):333–337, 1987.

[82] Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science and Business Media, 2012.

[83] H. Jankowski. Convergence of linear functionals of the Grenander estimator under misspecification. *Ann. Statist.*, 42(2):625–653, 2014.

[84] W. Jiang. Process consistency for adaboost. *Annals of Statistics*, 21:13–29, 2004.

[85] A. K. Kim, H. H. Zhou, et al. Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.

[86] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[87] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[88] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

[89] A. Kudo. A multivariate analogue of the one-sided test. *Biometrika*, 50(3/4):403–418, 1963.

[90] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York, 1986.

[91] M. Ledoux. *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[92] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York, NY, 1991.

[93] E. L. Lehmann. On likelihood ratio tests. In *Selected Works of EL Lehmann*, pages 209–216. Springer, 2012.

[94] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses.* Springer Science and Business Media, 2006.

[95] A. S. Lele, S. R. Kulkarni, and A. S. Willsky. Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A*, 9:1693–1714, 1992.

[96] O. V. Lepski and V. G. Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

[97] O. V. Lepski and A. B. Tsybakov. Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*, 117(1):17–48, 2000.

[98] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

[99] P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[100] E. Mammen. Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, 19(2):741–759, 1991.

[101] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, pages 512–518, 1999.

[102] D. E. McClure and R. A. Vitale. Polygonal approximation of plane convex bodies. *Journal of Mathematical Analysis and Applications*, 51(2):326–358, 1975.

[103] M. B. McCoy and J. A. Tropp. From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete and Computational Geometry*, 51(4):926–963, 2014.

[104] N. Meinshausen. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631, 2013.

[105] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 29–43, 2002.

[106] J. Menéndez, C. Rueda, B. Salvador, et al. Dominance of likelihood ratio tests under cone constraints. *The Annals of Statistics*, 20(4):2087–2099, 1992.

[107] J. A. Menéndez and B. Salvador. Anomalies of the likelihood ratio test for testing restricted hypotheses. *Annals of Statistics*, 19(2):889–898, 1991.

[108] J. Menéndnez, C. Rueda, and B. Salvador. Testing non-oblique hypotheses. *Communications in Statistics - Theory and Methods*, 21(2):471–484, 1992.

[109] M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*, pages 1083–1104, 2000.

[110] M. C. Meyer. A test for linear versus convex regression function using shape-restricted regression. *Biometrika*, 90(1):223–232, 2003.

[111] J.-J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *CR Acad. Sci. Paris*, 255:238–240, 1962.

[112] M. D. Perlman, L. Wu, et al. The emperor's new tests. *Statistical Science*, 14(4):355–369, 1999.

[113] G. Pisier. *Probabilistic methods in the geometry of Banach spaces.* Springer, 1986.

[114] L. Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[115] J. L. Prince and A. S. Willsky. Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:377–389, 1990.

[116] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

[117] R. F. Raubertas, C.-I. Charles Lee, and E. V. Nordheim. Hypothesis tests for normal means constrained by linear inequalities. *Communications in Statistics - Theory and Methods*, 15(9):2809–2833, 1986.

[118] T. Robertson. Testing for and against an order restriction on multinomial parameters. *Journal of the American Statistical Association*, 73(361):197–202, 1978.

[119] T. Robertson. On testing symmetry and unimodality. In *Advances in Order Restricted Statistical Inference*, pages 231–248. Springer, 1986.

[120] T. Robertson and E. J. Wegman. Likelihood ratio tests for order restrictions in exponential families. *The Annals of Statistics*, pages 485–505, 1978.

[121] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference.* Wiley Series in Probability and Mathematical Statistics, 1988.

[122] W. Robertson. On measuring the conformity of a parameter set to a trend, with applications. *The Annals of Statistics*, 10(4):1234–1245, 1982.

[123] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[124] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[125] R. E. Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[126] L. L. Scharf. *Statistical signal processing: Detection, estimation and time series analysis.* Addison-Wesley, Reading, MA, 1991.

[127] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory.* Cambridge Univ. Press, Cambridge, 1993.

[128] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[129] B. Sen and M. Meyer. Testing against a linear regression model using ideas from shape-restricted estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

[130] A. Shapiro. Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review/Revue Internationale de Statistique*, pages 49–62, 1988.

[131] M. Slawski, M. Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

[132] V. G. Spokoiny. Adaptive and spatially adaptive testing a nonparametric hypothesis. *Math. Methods Statist*, 7:245–273, 1998.

[133] H. Stark and Y. Yang. Vector space projections. *John Wiley&Sons, New York*, 1998.

[134] O. N. Strand. Theory and methods related to the singular value expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Numer. Anal.*, 11:798–825, 1974.

[135] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer-Verlag, 2009.

[136] S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[137] S. A. Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.

[138] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, NY, 1996.

[139] A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

[140] R. A. Vitale. Support functions of plane convex sets. Technical report, Claremont Graduate School, Claremont, CA, 1979.

[141] E. D. Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

[142] G. Wahba. Three topics in ill-posed problems. In M. Engl and G. Groetsch, editors, *Inverse and ill-posed problems*, pages 37–50. Academic Press, 1987.

[143] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[144] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2017.

[145] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[146] G. Walther et al. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics*, 38(2):1010–1033, 2010.

[147] G. Warrack and T. Robertson. A likelihood ratio test regarding two nested but oblique order-restricted hypotheses. *Journal of the American Statistical Association*, 79(388):881–886, 1984.

[148] Y. Wei and M. J. Wainwright. Sharp minimax bounds for testing discrete monotone distributions. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2684–2688. IEEE, 2016.

[149] Y. Wei, M. J. Wainwright, and A. Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood tests and minimax radii. *arXiv preprint arXiv:1703.06810*, 2017.

[150] Y. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, pages 6067–6077, 2017.

[151] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182, 2012.

[152] F. T. Wright. The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, 9(2):443–448, 1981.

[153] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[154] Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Annals of Statistics*, 2017. To appear.

[155] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[156] B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. Springer-Verlag, New York, 1997.

[157] E. H. Zarantonello. Projections on convex sets in Hilbert spaces and spectral theory. In *Contributions to nonlinear functional analysis*, pages 237–424. Academic Press, 1971.

[158] C.-H. Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.

[159] C.-H. Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

[160] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.