Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: On Some Problems in Nonparametric and Location-Scale Estimation
Presented by: Shamindra Shrotriya
Accepted by: Department of Statistics & Data Science
Readers:

MATEY NEYKOV, ADVISOR

ARUN KUMAR KUCHIBHOTLA

YANG NING, CORNELL UNIVERSITY

ALEX REINHART

ALESSANDRO RINALDO

LARRY WASSERMAN

Approved by the Committee on Graduate Degrees:

RICHARD SCHEINES, DEAN

12/7/22
DATE

Carnegie Mellon University

On Some Problems in Nonparametric and Location-Scale Estimation

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in
Statistics

by

Shamindra Shrotriya

Department of Statistics & Data Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

Carnegie Mellon University

December 1, 2022

Dedicated to my parents, Anjali and Vijay,
and to Rev. Swami Sridharanandaji,
for their unwavering love and encouragement.

Abstract

We study three generalizations of classical nonparametric, and location-scale estimation problems.

First, we study the classical problem of deriving minimax rates for density estimation over convex density classes. Our work extends known results by demonstrating that the local metric entropy of the density class always captures the exact (up to constants) minimax optimal rates under such settings. Our bounds provide a unifying perspective across both parametric and nonparametric convex density classes, under weaker assumptions on the richness of the density class than previously considered.

Second, we consider a variation of classical isotonic regression, which we term adversarial sign-corrupted isotonic (ASCI) regression. Here, the adversary can corrupt the sign of the responses having full access to the true response terms. We formalize `ASCIFIT`, a three-step estimation procedure under this regime, and demonstrate its theoretical guarantees in the form of sharp high probability upper bounds and minimax lower bounds.

Finally, we extend classical univariate uniform location-scale estimation over an interval, to multivariate uniform location-scale estimation over general convex bodies. Unlike the univariate setting, the observations are no longer totally ordered, and previous estimation techniques prove insufficient to account for the more refined geometry of the generating process. Under fixed dimension, our proposed location estimators converge at an n^{-1} rate. Our minimax lower bounds justify the optimality of our estimators in terms of the sample complexity. We also provide practical algorithms with provable convergence rates for our estimators, over a wide class of convex bodies.

Keywords: *density estimation, minimax, metric entropy, isotonic regression, location-scale estimation, uniform distribution, convex body.*

MSC Codes: 62C20, 62F10, 62G07, 62G08.

Acknowledgments

I never expected to graduate with a PhD in Statistics and Data Science, and yet here we are. The road to this PhD has been like a tough but rewarding hike. Similar to my past favorite hikes, this journey was made possible and enjoyable by having a strong supportive team alongside me. This is my attempt to acknowledge their kind efforts thereof. Please indulge me a little, if you don't mind.

First and foremost, I'd like to thank my amazing advisor Matey Neykov, who has been my main guide and collaborator. Although Matey and I met in my third year, we hit it off almost instantly. I recall asking him in our first meeting what he was interested in, and he spontaneously jotted down five exciting and varied research directions. I was immediately drawn in by the breadth and quality of the statistical questions posed, and our collaboration began immediately thereafter. Matey is an incredibly creative statistician, who generously shares his ideas, and is always open to my input. My favorite times together were those meetings where we interactively sketched out a proof, with each one of us alternating in proposing the next line of the argument. I hope that this thesis marks the beginning of more such work together.

I'd like to thank in turn my esteemed committee members for their thoughtful mentorship over the years, namely: Arun Kumar Kuchibhotla, Yang Ning (Cornell University), Alex Reinhart, Alessandro (Ale) Rinaldo, and Larry Wasserman. Arun is impressive in the depth of his statistical knowledge, and always manages to clearly explain the intuition behind very subtle statistical concepts. I still continue to reflect on the insights from his "concentration inequalities masterclass". I was fortunate to have Yang as my external committee member. Through his gentle demeanor, Yang raised several razor-sharp questions on our work and greatly helped improve the resulting thesis exposition. I thank Alex for being a role model in both statistical methodology and computing. His ability to retain his composure whilst taking on large-scale research projects has been a vital lesson for me. Alex also developed the template for this thesis and graciously helped me to incorporate numerous additional LaTeX stylistic preferences.

Ale has been a tremendous guide throughout the program, as both a collaborator and mentor. As a collaborator, he has an uncanny knack of quickly connecting ideas from disparate statistical areas to help make progress on the problem at hand. As a mentor, I remember nervously meeting with him early in my third year and openly expressing my concern at the lack of my research output. He patiently listened to me, and told me to keep focusing on my interests with persistence, and that the research would follow in due time. I'm so grateful for this uplifting advice, which changed the course of my PhD. I'd like to thank Larry, for being one of the most knowledgeable and encouraging faculty members in the department. Prior to the PhD, I worked for eight years as a data scientist in various capacities. Larry doesn't know this, but I used to read his *All of Statistics* textbook on the Sydney trains to and from work. Although I barely understood anything at the time, I felt inspired from his book to learn more about this beautiful subject. In none of those train rides, did I ever feel that I would get to learn from the master himself one day! Moreover, I did not expect to have him on my thesis committee, since research itself was not on my radar at the time. This really has been a very positive surreal life experience for me. I look forward to the continued friendship and collaboration with my fantastic committee members.

There are also several research collaborators to whom I owe a large debt of gratitude to. I'd like to thank Riccardo Fogliato for our work together with Arun Kumar Kuchibhotla on model misspecification. Riccardo is a very focused and pragmatic researcher, whose friendship and advice means a lot to me. I thank Niccolò (Nic) Dalmasso for our work together with Alex Reinhart on wildfire prediction. I've always appreciated his frank but empathetic feedback and friendship throughout the program. I want to thank Heejong Bong and Wanshan Li for our collaborations with Ale Rinaldo on statistical ranking. Both Heejong and Wanshan have fantastic intuition, and are just plain fun to work with. I look forward to seeing what projects my collaborators and I will work on next.

In order to unwind during the research process I enjoyed many cooking sessions with several CMU students, namely: Riccardo Fogliato, Natalia Lombardi De Oliveira, Rafael Catoia Pulgrossi, Anjalie Field, Kevin Lin, Eyan Noronha, Shoham Sen, and Vanessa Vidal. Our culinary catch-ups were some of the best social highlights during my time here, and a nice way to sample various cultural cuisines. A memorable event was when we celebrated Diwali as a potluck dinner one year, and then concluded it with a viewing of the eponymous *The Office (US)* episode.

I am also grateful to have been part of such an inspiring PhD cohort. Our (pre-pandemic) bonding traditions of celebrating birthdays as a group always brought a smile to my face. I want to thank in particular Pratik Patil who is a great researcher and friend. I've thoroughly enjoyed our long and varied discussions, mostly driven just by learning for learning's sake. I eagerly look forward to seeing what exciting things my cohort achieves in the future.

More broadly, I thank the wider CMU Statistics and Data Science student body for always maintaining a friendly and approachable atmosphere. I genuinely appreciated the ad hoc and engaging discussions with Matteo Bonvini, Purvasha Chakravarti, Maria Jahja, Ilmun Kim, Jisu Kim, Yue Li, Tudor Manole, Trey McNeely, Lee Richardson, Taylor Pospisil, Ron Yurko and many others whenever I visited the student lounge at FMS.

On a similar note, the CMU Statistics and Data Science faculty have been amazing to interact with during my time here. Despite their high individual stature in the broader statistical community, I never once felt intimidated knocking on a faculty member's door to ask questions. Furthermore, the graduate courses as taught to me by Sivaraman Balakrishnan, Christopher Genovese, Edward Kennedy, Aaditya Ramdas, Ale Rinaldo, Alex Reinhart, Ryan Tibshirani, Valérie Ventura, and Larry Wasserman have all been instrumental in shaping my statistical thinking.

Our excellent managerial staff deserve my sincere praise, namely: Kira Bokalders, Laura Butler, Danielle Hamilton, Mari Alice McShane, Jessica Paschke, Carl Skipper, Margie Smykla, and Teraya White. They helped me calmly navigate through numerous administrative issues as an international student.

I'd like to thank the American Australian Association (AAA) for awarding me the 2020 Graduate Education Wildfire Research Scholarship. This has been a huge support in funding my ongoing collaborative research in wildfire prediction.

My friends outside of CMU helped me to remember that there was a bigger world outside of research. In particular I'd like to thank Aditi Gupta, Sandeep Liyanage, Nina Pak, Tracy Ong, Dean Pham, Christina Randall, Ainesh Sewak, and Chelsea Wise for keeping me motivated with their typical good cheer.

Prior to CMU, I was extremely fortunate to undertake a Masters in Statistics at UC Berkeley and meet Prof. Bin Yu. Bin generously allowed me to attend weekly meetings and work with the amazing Yu Group, thus providing me with my first taste of statistical research. I'd like to additionally thank Professors: David Aldous, Chris Paciorek, Will Fithian, and Jim Pitman for giving stunning lectures on probability and statistics that I still actively ruminate on.

Finally, and in a personal sense most importantly, I would like to thank my family and loved ones. In particular my siblings and their families, namely: Amrita, Angad, Atul, Saleel, Lauren, Seema, Praful, Shreyas, and Sonia. Although we are separated by four continents, you were always a quick phone call away and ready to help during my many moments of self-doubt. I would not be here without your continued support. Lastly, I am deeply grateful for my wonderful parents, Anjali and Vijay, and also to Rev. Swami Sridharanandaji, for all of their untiring efforts in helping me realize my ambitions. In order to pay their kindness forward, I hope that I can apply the knowledge gained here to serve humanity in a positive manner. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

One

Introduction

The motivating theme of this thesis is to generalize three classical nonparametric, and location-scale (parametric) estimation problems in statistics, and analyze in detail their theoretical properties. Here, 'classical' is simply qualitatively taken to mean those foundational canonical estimation problems that are found in most modern statistical theory reference books, e.g., Lehmann and Casella (1998); Wasserman (2004, 2006). Our main goal in analyzing such classical problems is to shed *new* inferential insights on them, in an effort to drive new research directions.

We aim to achieve this through two main perspectives. The first such perspective is to develop *multivariate generalizations* of the given classical univariate problem. Here, the dimension, $d \in \mathbb{N}$, is arbitrary but fixed, while the sample size n, increases asymptotically. The second perspective involves maintaining the classical (typically univariate) setting, but directly generalizing it in a well-motivated manner. This is done by either considering an *adversarial perturbation* of the generating process, or by simply *weakening the assumptions* upon which estimation is performed for the given problem.

1.1 Organization of the thesis

This leads us to the specific structure and content of the thesis, which is split into two parts, with each part then further subdivided into individual chapters. Each part represents a generalized analysis of nonparametric estimation (Part I), or (parametric) location-scale estimation (Part II) problems. Three core problems are studied in this thesis, with the analysis of each problem forming a separate chapter. Moreover, each chapter directly corresponds to a released (or forthcoming) preprint of the same title. In particular Chapter 2 is based on Shrotriya and Neykov (2022a), Chapter 3 is based on Shrotriya and Neykov (2022b), and finally Chapter 4 is derived from (the forthcoming preprint) Shrotriya and Neykov (2022c). Importantly, we emphasize that each of these three chapters represents work that is co-authored with Matey Neykov.

The resulting Part-Chapter content layout of this thesis can be conveniently summarized as follows.

Part I Nonparametric Estimation

Chapter 2 *Revisiting Le Cam's Equation: Exact Minimax Rates over Convex Density Classes*

Chapter 3 *Adversarial Sign-Corrupted Isotonic Regression*

Part II Location-Scale Estimation

Chapter 4 *Uniform Location Estimation on Convex Bodies*

1.2 Overview of the three core problems

Having described the Part-Chapter organization of our thesis as per Section 1.1, we now briefly describe on the core questions of interest studied in each chapter. We intentionally formalize the underlying generating processes driving each chapter, so that they are made precise up front. Our main goal is to concisely provide our driving motivation behind their study in this thesis. More detailed historical context can be found in the respective chapters in which these specific problems are studied. Before describing the detailed problem settings for Chapters 2 to 4, we note that each of these chapters can be (largely) read in a standalone manner, though they are unified by the aforementioned motivating theme of this thesis.

Part I Nonparametric Estimation

We begin with Chapter 2, where we study minimax density estimation over convex density classes. To set the stage, suppose we have constants $0 < \alpha < \beta < \infty$, for some fixed dimension $p \in \mathbb{N}$, and a common known (Borel measurable) compact support set $B \subseteq \mathbb{R}^p$ (with positive measure). We then define the *ambient* class of (probability) density functions, $\mathcal{F}_B^{[\alpha,\beta]}$, as follows:

$$\mathcal{F}_B^{[\alpha,\beta]} := \left\{ f \colon B \to [\alpha, \beta] \,\middle|\, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{1.1}$$

where μ is the dominating finite measure on B. We always take μ to be a (normalized) probability measure on B. Now, given this general bounded density class, we consider n observations, $(X_i)_{i=1}^n$, where each observation X_i

is generated from the following model:

$$
\begin{aligned}
X_i &\overset{\text{i.i.d.}}{\sim} f && (1.2)\\
\text{s.t. } f &\in \mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]} && (1.3)\\
\text{and } \mathcal{F} &\text{ is a convex set} && (1.4)
\end{aligned}
$$

This now leads to the following core questions of interest.

> ***Core questions:*** *Suppose that we observe $(X_i)_{i=1}^n$, generated according to (1.2)-(1.4). Can we propose a universal estimator for f, and derive the exact (up to constants) squared L_2-minimax rate of estimation, in expectation?*

Such a problem set up has been explicitly studied in the seminal work of Yang and Barron (1999). However, our focus here is to extend these results by establishing *exact* minimax rates over general (i.e., both nonparametric and parametric) convex density classes. We also hope to generalize these known results by considering weaker assumptions on the underlying convex density class $\mathcal{F}$.

Moving onto Chapter 3, our study here begins with the following generating process. We consider n univariate observations, $(R_i)_{i=1}^n$, where each observation R_i is generated from the following model:

$$
\begin{aligned}
R_i &= \xi_i(\mu_i + \varepsilon_i) && (1.5)\\
\text{s.t. } 0 &< \eta \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n && (1.6)\\
\text{and } \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (1.7)\\
\text{and } \xi_i &\in \{-1, 1\}. && (1.8)
\end{aligned}
$$

We note that the constant $\eta > 0$ is a known constant in the generating process.

We refer to the generating process described by (1.5)-(1.8), as adversarial sign-corrupted isotonic (ASCI) regression. This represents a partial generalization of the classical isotonic regression setup. Here, the classical isotonic regression responses, $\mu_i + \varepsilon_i$ in Equation (1.5), are *sign-corrupted* in a manner chosen by an adversary, as captured by the multiplicative ξ_i terms. Given this ASCI setting, we consider the following core questions of interest.

> ***Core questions:*** *Under* ASCI *regression setup, can we find a computationally efficient estimator for* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$*, and demonstrate its precise (non-asymptotic) statistical optimality?*

To the best of our knowledge, this is the first such appearance of this variation of classical isotonic regression. Our study is driven by three factors, which we believe make the problem challenging and thus interesting. First, the sign-corruptions can be chosen adversarially in a way that results in a strong dependence between the original isotonic responses. As such, any ASCI estimator must be able to handle arbitrary dependence structure between the sign-corrupted responses. Second, the sign-corruptions are in a sense 'extreme' in that the sign-corruptions fundamentally attack the monotonicity constraint directly. It is this convex monotone constraint which classical isotonic estimators, i.e., `PAVA`, are *designed* to exploit. Finally, we show that the ASCI setting contains interesting non-trivial special cases, for which naively applying typical least squares estimation techniques will prove insufficient.

Part II Location-scale estimation

Finally, in Chapter 4 we study multivariate uniform location estimation on convex bodies. We formalize the generating process as follows. Let $d \geq 1$ be a fixed positive integer. Now, let $K \subset \mathbb{R}^d$ also be a fixed convex body, i.e., a compact, convex set, with a non-empty interior. We assume that both d, K are known to the observer, and that centroid $(K) = \mathbf{0} \in \mathbb{R}^d$. Now, let $\mathbf{v} \in \mathbb{R}^d$ be a fixed but unknown *location* parameter. We then consider n observations, $(\mathbf{Y}_i)_{i=1}^n$, where each observation $\mathbf{Y}_i \in \mathbb{R}^d$ is generated from the following model:

$$\mathbf{Y}_i \stackrel{a.s.}{=} \mathbf{v} + \sigma \mathbf{X}_i \tag{1.9}$$

$$\text{s.t. } \mathbf{X}_i \stackrel{\text{i.i.d.}}{\sim} \mathsf{Unif}[K] \tag{1.10}$$

$$\text{and } \sigma > 0 \tag{1.11}$$

This generating process described formally by (1.9)-(1.11) is known as the multivariate uniform location-scale model. In Equation (1.11), we consider both *scaling regimes* in which the fixed *scale* parameter σ is either known or unknown to the observer. In the latter case σ is treated as a nuisance parameter. The multivariate uniform location-scale model is our proposed generalization of the well-studied case of univariate uniform location-scale estimation.

Given the above, we consider the following natural core questions of interest for multivariate uniform location estimation.

> ***Core questions:*** *How can one derive statistically optimal location estimators (i.e., for* $\mathbf{v}$*), and understand their geometry in this multivariate setting under both known and unknown scaling regimes (i.e., for* σ*)? What are practical algorithms to compute such estimators under a wide variety of use-cases with convergence rate guarantees?*

Despite the univariate uniform location-scale estimation being a well studied estimation problem, to the best of our knowledge our proposed multivariate generalization of uniform location-scale estimation has not been explicitly studied previously. A particular emphasis of this work is to consider a wide variety of location estimators under this setting and understand the statistical and computational and trade-offs that arise in the estimation process.

1.3 A note on stylistic conventions used in this thesis

We note that the thesis adopts three stylistic conventions which we state up front. First, since each chapter can be read in a standalone manner, we append the corresponding chapter appendix materials *directly* after the main chapter body. Second, for each of the appendix proofs of statements from the main chapter body, we *first restate verbatim* the relevant statement before presenting its proof. Finally, there may be slight notational differences between the chapters. To ensure notational clarity for a given chapter, we provide a *separate notation section*, and also a corresponding notational summary table at the start of each chapter appendix. All of these conventions are designed to improve the flow of readability.

Part I

Nonparametric estimation

Two

Revisiting Le Cam's Equation: Exact Minimax Rates over Convex Density Classes

Abstract: We study the classical problem of deriving minimax rates for density estimation over convex density classes. Building on the pioneering work of Le Cam (1973); Birgé (1983); Birgé (1986); Wong and Shen (1995); Yang and Barron (1999), we determine the exact (up to constants) minimax rate over any convex density class. This work thus extends these known results by demonstrating that the local metric entropy of the density class always captures the minimax optimal rates under such settings. Our bounds provide a unifying perspective across both parametric and nonparametric convex density classes, under weaker assumptions on the richness of the density class than previously considered. Our proposed 'multistage sieve' MLE applies to any such convex density class. We apply our risk bounds to rederive known minimax rates including bounded total variation, and Lipschitz density classes. We further illustrate the utility of the result by deriving upper bounds for less studied classes, e.g., convex mixture of densities.

The work in this chapter was done jointly with Matey Neykov. It is based on a preprint with the title *"Revisiting Le Cam's Equation: Exact Minimax Rates over Convex Density Classes"*.

2.1 Introduction

It is well known that (global) metric entropy often times determines the minimax rates for density estimation. Specifically, the following equation sometimes informally referred to as the 'Le Cam equation' is used to heuristically determine the minimax rate of convergence

$$\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp n\varepsilon^2,$$

where n is the sample size, $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon)$ is the *global* metric entropy of the density set $\mathcal{F}$ at a Hellinger distance ε (see Definition 2.9), and ε^2 determines the order of the minimax rate. In this paper we complement these known results, by establishing that *local* metric entropy *always* determines the minimax rate for convex density classes, where the densities are assumed to be (uniformly) bounded from above and below.

In detail, under the setting of density estimation just described, we suggest a small revision to the Le Cam equation: namely, change the global entropy to local entropy, and the Hellinger metric to the L_2-metric. Furthermore, the same result holds when the convex density class contains densities only (uniformly) bounded from above, and a single density which is bounded from below. Unlike previous known results, our result unites minimax density estimation under both parametric and nonparametric convex density classes. A further contribution is that our proposed 'multistage sieve' maximum likelihood estimator (MLE) achieves these bounds regardless of the density class (as long as it is convex).

We will now formally describe the setting we consider. To that end, we first define a general class of bounded densities, i.e., $\mathcal{F}_B^{[\alpha,\beta]}$. Later, we will assume that the true density of interest belongs to a known convex subset of this general ambient density class.

Definition 2.1 (Ambient density class $\mathcal{F}_B^{[\alpha,\beta]}$)**.** Given constants $0 < \alpha < \beta < \infty$, for some fixed dimension $p \in \mathbb{N}$, and a common known (Borel measurable) compact support set $B \subseteq \mathbb{R}^p$ (with positive measure), we then define the class of density functions, $\mathcal{F}_B^{[\alpha,\beta]}$, as follows:

$$\mathcal{F}_B^{[\alpha,\beta]} := \left\{ f\colon B \to [\alpha,\beta] \,\middle|\, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{2.1}$$

where μ is the dominating finite measure on B. We always take μ to be a (normalized) probability measure on B.

Furthermore, we can endow $\mathcal{F}_B^{[\alpha,\beta]}$ with the L_2-metric. That is, for any two densities $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$, we denote the L_2-metric between them to be

$$\|f - g\|_2 \coloneqq \left(\int_B (f-g)^2 \, \mathrm{d}\mu \right)^{\frac{1}{2}}. \tag{2.2}$$

Remark 2.2. Qualitatively, we have that $\mathcal{F}_B^{[\alpha,\beta]}$ is the class of all densities that are uniformly α-lower bounded and β-upper bounded, on a common compact support $B \subseteq \mathbb{R}^p$. Furthermore, Definition 2.1 implies that $\mathcal{F}_B^{[\alpha,\beta]}$ forms a convex set, and that the metric space $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ is complete, bounded, but may not be totally bounded[1].

In this paper we will focus on the scenario where it is known that the true density $f \in \mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$, where $\mathcal{F}$ is a known convex set. The set $\mathcal{F}$ represents our knowledge on the true density, before observing any data. With these mathematical preliminaries, we formalize our core density estimation problem of interest as follows.

> ***Core problem:*** *Suppose that we observe n observations $\mathbf{X} \coloneqq (X_1, \ldots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f$, for some (fixed but unknown) $f \in \mathcal{F}$. Here $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$ is a convex set, which is known to the observer. Can we propose a universal estimator for f, and derive the exact (up to constants) squared L_2-minimax rate of estimation, in expectation?*

For convenience, we can illustrate the generating process for a univariate example of our density estimation problem of interest in Figure 2.1. It will serve as a useful conceptual guide to later help visualize our proposed estimator over such general convex class of densities $\mathcal{F}$.

Now, without further ado, we will informally state our main result as a direct affirmative answer to our core question of interest. Namely, there does exist a likelihood-based estimator (one can think of it as a multistage sieve MLE), i.e., $\nu^*(\mathbf{X})$, which achieves the following rate of estimation error

$$\sup_{f \in \mathcal{F}} \mathbb{E}\|\nu^*(\mathbf{X}) - f\|_2^2 \lesssim \varepsilon^{*2} \wedge \mathrm{diam}_2(\mathcal{F})^2. \tag{2.3}$$

[1]These fundamental (and additional) analytic properties of $\mathcal{F}_B^{[\alpha,\beta]}$ are formally justified in Section 2.A.2.

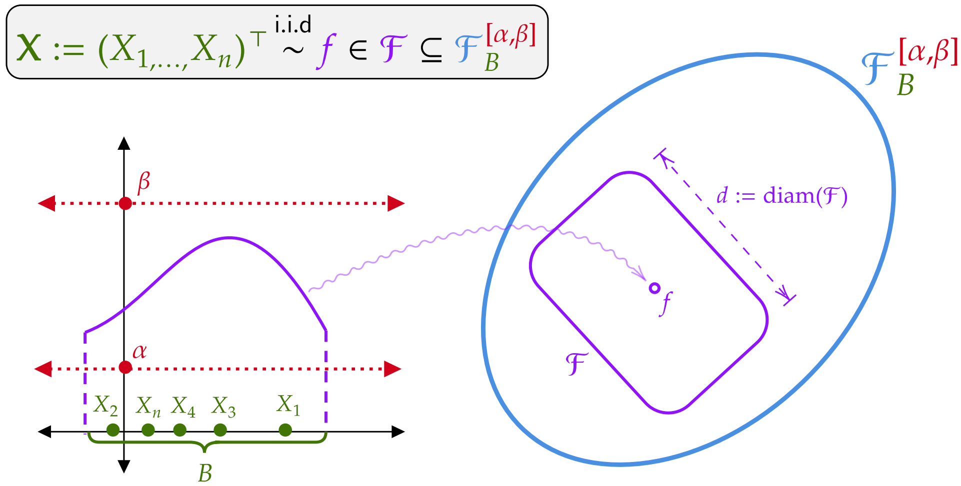

Figure 2.1: Illustrative generating process for a univariate density $f \in \mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$.

Here $\varepsilon^* := \sup\{\varepsilon : n\varepsilon^2 \leq \log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)\}$, with $\log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)$ being the L_2-*local* metric entropy of $\mathcal{F}$ (see Definition 2.10). The quantity $\text{diam}_2(\mathcal{F})$, refers to the L_2-diameter of $\mathcal{F}$, which is finite by the boundedness of $\mathcal{F}_B^{[\alpha,\beta]}$ in our setting. In addition, the rate above is minimax optimal, as there is a matching (up to constants) lower bound.

Remark 2.3. We will later see that we can largely relax the α-lower boundedness condition on $\mathcal{F}$. That is, the results we are about to derive can be readily generalized to convex subsets $\mathcal{F} \subset \mathcal{F}_B^{[0,\beta]}$. This is so as long as the class $\mathcal{F}$ contains a *single* density which is bounded away from 0.

Next, we turn our attention to reviewing some relevant literature.

2.1.1 Relevant Literature

Classical work

As noted, density estimation is a classical statistical estimation problem with a rich history. Lively accounts of the key references, particularly for nonparametric density estimation as relevant to our setting, are already covered in Yang and Barron (1999, Section 1) and Bilodeau et al. (2021, Section 6.1). We similarly begin with a brief panoramic overview of these references in regard to minimax risk bounds for density estimation, before comparing and contrasting the results from the most relevant references to our work.

In terms of minimax lower bounds on density estimation, Boyd and Steele (1978) prove a fundamental n^{-1} rate in the mean integrated p^{th} power error (with $p \geq 1$), for *any* arbitrary density estimator. Such generalized lower bounds on density estimation were also further studied in Devroye (1983). In the case of density estimation over classes with more assumed structure (e.g., smoothness, or regularity assumptions) minimax lower bounds have been developed based on hypothesis testing approaches coupled with information-theoretic techniques. We now provide brief highlights of such key works in this direction.

In Bretagnolle and Huber (1979), the authors derive sharp lower bounds for Sobolev smooth densities in $\mathbb{R}^d$ ($d \in \mathbb{N}$), with risk measured with respect to a power of the L_q-metric ($q \geq 1$). In Birgé (1986), sharp risk bounds for more general classes of such smooth families were provided using metric entropy based methods, with an emphasis on the Hellinger loss. The work of Efroĭmovich and Pinsker (1982) provided precise (asymptotic) analysis for an ellipsoidal class of densities in the L_2-metric. Across a wide-ranging series of related and collaborative efforts Has'minskiĭ (1978); Ibragimov and Has'minskiĭ (1977, 1978); Ibragimov and Khas'minskij (1980) used Fano's inequality type arguments to establish lower bounds over a variety of density estimation settings. These range from deriving lower bounds on nonparametric density estimation in the uniform metric, to minimax risk bounds for the Gaussian white noise model, for example. The authors also develop metric entropy based techniques in Has'minskiĭ and Ibragimov (1990) to derive minimax lower bounds for a wide variety of density classes defined on $\mathbb{R}^d$ ($d \in \mathbb{N}$), in L_q-loss ($q \geq 1$). Numerous applications of optimal lower bounds using both Assouad's and Fano's lemma arguments for densities on a compact support, are demonstrated in (Yu, 1997, Section 29.3). Later Yang and Barron (1999) demonstrated that *global* metric entropy bounds capture minimax risk for sufficiently rich density classes over a common compact support. Classical reference texts on minimax lower bound techniques with an emphasis on nonparametric density estimation include Devroye (1987); Devroye and Györfi (1985); Le Cam (1986). More modern such references include Tsybakov (2009) and Wainwright (2019, Chapter 15). The latter in particular, also incorporates metric entropy based lower bound techniques.

In addition, there is a large body of work in deriving upper bounds for specific density estimators using metric entropy methods. This includes Yatracos (1985); Barron and Cover (1991), who employ the minimum distance principle to derive density estimators and their metric entropy-based upper bounds in the Hellinger and L_1-metric, respectively. In a similar spirit to Birgé (1983); Birgé (1986), van de Geer (1993) is also concerned with density estimation using Hellinger loss. However, its focus is to use techniques from empirical process theory in order to specifically establish the Hellinger consistency of the nonparametric MLE, over convex density classes. Upper bounds for density estimation based on the 'sieve' MLE technique is studied in Wong and Shen (1995). Recall that a 'sieve' estimator effectively estimates the parameter of interest via an optimization procedure (e.g., maximum likelihood) over a constrained subset of the parameter space (Grenander, 1981, Chapter 8). In Birgé and Massart (1993a) the authors study 'minimum contrast estimators' (MCEs), which include the MLE, least squares estimators (LSEs) etc., and apply them to density estimation. This is further developed in Birgé and Massart (1998) where they analyze convergence of MCEs using sieve-based approaches.

Comparison to our work

By stating our main result early in the introduction, we now turn to contrasting it with the most relevant results in the literature. These include both the aforementioned classical references, and more recent work on convex density estimation, which have most directly inspired our efforts in this work.

First we would like to comment on the closely related landmark papers (Le Cam, 1973; Birgé, 1983; Birgé, 1986). These works consider very abstract settings and show upper bounds based on Hellinger ball testing. Although widely believed that they do, whether these results lead to bounds that are minimax optimal is unclear. Moreover, their estimator is quite involved and non-constructive. In contrast, in this paper we offer a simple to state, *constructive* multistage sieve MLE type of estimator, which is provably minimax optimal over any convex density class $\mathcal{F}$. A crucial difference is that we metrize the space $\mathcal{F}$ with the L_2-metric as we mentioned above. Even though in our instance the two distances are equivalent, in contrast to the Hellinger distance, the ε-*local* metric entropy of the convex density class in the L_2-metric can be shown to be monotonic in ε. This key observation enables us to match the upper and lower bounds exactly.

Next, we will compare our work with the celebrated paper of Yang and Barron (1999), who inspect a very similar problem. Yang and Barron (1999) demonstrate a lower and upper bound which need not match in general but do match under certain sufficient conditions. Notably their bounds involve only quantities depending on the global entropies of the set $\mathcal{F}$ (which is also assumed to be convex for some results of Yang and Barron (1999)). This is convenient as often times global metric entropy is easier to work with compared to local metric entropy, however under Yang and Barron (1999)'s sufficient condition it can be seen that the two notions are equivalent. Hence, our work can be thought of as removing the sufficient condition requirement from Yang and Barron (1999) and also unifying parametric and nonparametric density estimation problems (over convex classes) for which one typically needs to use different tools to obtain the accurate rates. Finally we would like to mention Wong and Shen (1995).

In that paper the authors propose a sieve MLE estimator and demonstrate that it is *nearly* minimax optimal under certain conditions. Our estimator is not the same as the one considered by Wong and Shen (1995), and we can provably match the minimax rate over whatever be the convex set $\mathcal{F}$. A notable difference is that Wong and Shen (1995) work with the Hellinger metric and KL divergence, which although equivalent to L_2-metric in our problem, are actually less practical in terms of matching the bounds exactly as we explained above. We will now turn our attention to reviewing some further relevant literature.

Recent work

Our estimator and proof techniques thereof, are inspired by the recent work of Neykov (2022) on the Gaussian sequence model. We would like to stress on the fact that the sequence model is a very distinct problem from density estimation. In particular, our underlying metric space of interest is $(\mathcal{F}, \|\cdot\|_2)$, as compared to[2] $(\mathbb{R}^n, \|\cdot\|_2)$ for the sequence model. Both of these metric spaces differ *vastly* from each other in their underlying geometric structure. Furthermore, unlike our setting, the sequence model contains additional Gaussian information on the underlying generating process, which can be directly exploited for estimation purposes. As such, given that Neykov (2022) provides a guiding template for our analysis, some resulting structural similarities to their work are to be expected. However, all corresponding proofs, and estimators thereof, have to be non-trivially adapted to our nonparametric density estimation setting. A notable example of such required modifications, is that our estimator presented in this paper does not use proximity in Euclidean norm, but is a likelihood-based estimator.

We additionally note that density estimation in both abstract and more concrete settings, continues to be an active area of research. It is not feasible to detail such a large and growing body of references. However, we provide a selective overview of some interesting recent directions in density estimation, to simply indicate the diversity of the research efforts thereof. For example, Cleanthous et al. (2020); Baldi et al. (2009) study convergence properties of density estimators using wavelet-based methods. The papers Goldenshluger and Lepski (2014); Efromovich (2008); Rigollet (2006); Rigollet and Tsybakov (2007); Samarov and Tsybakov (2007); Birgé (2014) study adaptive minimax density estimation on $\mathbb{R}^d$ $(d \geq 1)$ under L_p-loss $(p \geq 1)$. Here, 'adaptive' refers to the fact that the density class is defined by an unknown tuning hyperparameter, which must be explicitly accounted for during the estimation process. Recently Wang and Marzouk (2022) used techniques from optimal transport to study the convergence properties of various nonparametric density

[2]Note that $\|\cdot\|_2$ here is the Euclidean metric on $\mathbb{R}^n$.

estimators. Interestingly, Bilodeau et al. (2021) applied empirical (metric) entropy methods to establish minimax optimal rates in the adjacent setting of *conditional* density estimation. Although these works do not directly study our core problem of interest, we note that they represent new and important perspectives on classical minimax density estimation, and related problems.

2.1.2 Notation

We outline some commonly used notation here. We use $a \vee b$ and $a \wedge b$ for the max and min of two numbers $\{a, b\}$, respectively. Throughout the paper $\|\cdot\|_2$ denotes the L_2-metric in $\mathcal{F}$. Constants may change values from line to line. For an integer $m \in \mathbb{N}$, we use the shorthand $[m] \coloneqq \{1, \ldots, m\}$. We use $B_2(\theta, r)$ to denote a closed L_2-ball centered at the point θ with radius r. We use $\lesssim$ and $\gtrsim$ to mean $\leq$ and $\geq$ up to absolute (positive) constant factors, and for two sequences a_n and b_n we write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. Throughout the paper we use log to denote the natural logarithm, or we specify the base explicitly otherwise. Our use of $\{\alpha, \beta\}$ is *only* used to refer to the constants in Definition 2.1, of $\mathcal{F}_B^{[\alpha,\beta]}$ (and thus $\mathcal{F}$). We will introduce additional section-specific notation as needed.

2.1.3 Organization

The rest of this paper is organized as follows. In Section 2.2 we prove risk bounds for our underlying setting. We first establish the key topological equivalence between the L_2-metric and the Kullback-Leibler divergence in $\mathcal{F}_B^{[\alpha,\beta]}$. We then proceed to derive minimax lower bounds for our setting in Section 2.2.1, introducing additional relevant mathematical background as needed, e.g., local metric entropy. In Section 2.2.2 we define our likelihood-based estimator, and provide intuition behind its construction. We then derive its (matching) minimax risk upper bound. In Section 2.3, we apply our results to specific examples of commonly used convex density classes. We then conclude in Section 2.4 by summarizing our results, and discuss some future research directions.

2.2 Minimax Lower and Upper Bounds

Before establishing our main results, we establish a key technical lemma which drives much of the geometric arguments in our analysis to follow. Note that for any two densities $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$, the KL-divergence between them is defined to be

$$d_{\mathsf{KL}}(f||g) \coloneqq \int_B f \log\left(\frac{f}{g}\right) \mathrm{d}\mu =: \mathbb{E}_f \log\left(\frac{f(X)}{g(X)}\right), \tag{2.4}$$

where $X \sim f$ in (2.4).

Remark 2.4. We observe that $d_{\mathsf{KL}}(f||g)$ is well-defined in (2.4) for our setting, since $\inf_{x \in B} g(x) \geq \alpha > 0$, by Definition 2.1. We further emphasize that KL-divergence is not valid metric in general, since it is not symmetric in its arguments.

The crucial fact in the risk bounds we will soon derive, is the 'topological equivalence' of the L_2-metric and KL-divergence, on the density class $\mathcal{F}_B^{[\alpha,\beta]}$. Since it is hard to find a concrete reference for this folklore fact, we formalize this equivalence for our setting in Lemma 2.5.

Lemma 2.5 (KL-L_2 equivalence on $\mathcal{F}_B^{[\alpha,\beta]}$)**.** *For each pair of densities $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$, the following relationship holds:*

$$c(\alpha, \beta)\|f - g\|_2^2 \leq d_{\mathsf{KL}}(f||g) \leq (1/\alpha)\|f - g\|_2^2, \tag{2.5}$$

where we denote $c(\alpha, \beta) := \frac{h(\beta/\alpha)}{\beta} > 0$. Here $h : (0, \infty) \to \mathbb{R}$ is defined to be

$$h(\gamma) := \begin{cases} \frac{\gamma - 1 - \log \gamma}{(\gamma - 1)^2} & \text{if } \gamma \in (0, \infty) \setminus \{1\} \\ \frac{1}{2} = \lim_{x \to 1} \frac{x - 1 - \log x}{(x-1)^2} & \text{if } \gamma = 1, \end{cases} \tag{2.6}$$

and is positive over its entire support. It is also easily seen that on $\mathcal{F}_B^{[\alpha,\beta]}$, d_{KL} (and hence the L_2-metric) is also equivalent to the Hellinger metric. Furthermore, these properties are also inherited by $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$, which is our density class of interest.

Remark 2.6. We note that both the upper and lower bounds in (2.5) are stated without proof and without tracking constants in Klemelä (2009, Lemma 11.6). We formally prove this claim in Section 2.A. Importantly, the validity of (2.5) relies on the assumption of the boundedness of the densities, which holds in our setting.

2.2.1 Minimax Lower Bound

We will first establish a lower bound. For completeness, we need to introduce some additional relevant background and notation. We start by stating Fano's inequality for our convex density class, $\mathcal{F}$ (see Tsybakov, 2009, Lemma 2.10).

Lemma 2.7 (Fano's inequality for $\mathcal{F}$)**.** *Let $\{f^1, \ldots, f^m\} \subset \mathcal{F}$ be a collection of ε-separated densities (i.e. $\|f^i - f^j\|_2 > \varepsilon$ for $i \neq j$), in the L_2-metric. Suppose*

J is uniformly distributed over the index set $[m]$, and $(X_i | J = j) \overset{\text{i.i.d.}}{\sim} f^j$ for each $i \in [n]$. Then

$$\inf_{\widehat{\nu}} \sup_{f} \mathbb{E}\|\widehat{\nu}(\mathbf{X}) - f\|_2^2 \geq \frac{\varepsilon^2}{4}\left(1 - \frac{nI(X_1; J) + \log 2}{\log m}\right).$$

In the above $I(X_1; J) := \frac{1}{m}\sum_{j=1}^{m} d_{\mathsf{KL}}\big(f^j || \bar{f}\big)$, where $\bar{f} = \frac{1}{m}\sum_{j=1}^{m} f^j$ is the mutual information between X_1 and the randomly sampled index J. Further, the infimum is taken over all measurable functions of the data. Next, we define the important notion of a packing set for $\mathcal{F}$ (see Section 5.2 Wainwright, 2019, e.g., for more details).

Definition 2.8 (Packing sets and packing numbers of $\mathcal{F}$ in the L_2-metric)**.** Given any $\varepsilon > 0$, an ε-packing set of $\mathcal{F}$ in the L_2-metric, is a set $\{f^1, \ldots, f^m\} \subset \mathcal{F}$ of ε-separated densities (i.e., $\|f^i - f^j\|_2 > \varepsilon$ for $i \neq j$) in the L_2-metric. The corresponding ε-packing number, denoted by $M(\varepsilon, \mathcal{F})$, is the cardinality of the largest (maximal) ε-packing of $\mathcal{F}$. We refer to $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) := \log M(\varepsilon, \mathcal{F})$ as the *global metric entropy* of $\mathcal{F}$.

Remark 2.9. Note that we are not assuming here that $\mathcal{F}$ is totally bounded, hence some (or perhaps all) of the packing numbers may be infinite; this however does not cause a problem in what follows. Henceforth, all packing sets (or packing numbers) of $\mathcal{F}$, will be assumed to be with reference to the L_2-metric, unless stated otherwise. We will use the standard fact that a ε-maximal packing of $\mathcal{F}$, is also a ε-covering set of $\mathcal{F}$.

We will now define the notion of *local metric entropy*, which will play a key role in the development of our risk bounds.

Definition 2.10 (Local metric entropy of $\mathcal{F}$)**.** Let $c > 0$ be fixed, and $\theta \in \mathcal{F}$ be an arbitrary point. Consider the set[3] $\mathcal{F} \cap B_2(\theta, \varepsilon)$. Let $M(\varepsilon/c, \mathcal{F} \cap B_2(\theta, \varepsilon))$ denote the ε/c-packing number of $\mathcal{F} \cap B_2(\theta, \varepsilon)$, in the L_2-metric. Let

$$M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c) := \sup_{\theta \in \mathcal{F}} M(\varepsilon/c, \mathcal{F} \cap B_2(\theta, \varepsilon)) =: \sup_{\theta \in \mathcal{F}} M_{\mathcal{F} \cap B_2(\theta, \varepsilon)}^{\mathsf{glo}}(\varepsilon/c).$$

We refer to $\log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)$ as the *local metric entropy* of $\mathcal{F}$.

We show the following minimax lower bound for our convex density estimation setting over $\mathcal{F}$. It is a direct consequence of Fano's inequality per Lemma 2.7.

[3]Observe that this set may also fail to be totally bounded, since while the ball $B_2(\theta, \varepsilon)$ is a bounded set, it is not totally bounded.

Lemma 2.11 (Minimax lower bound)**.** *Let $c > 0$ be fixed, and independent of the data samples $\mathbf{X}$. Then the minimax rate satisfies*

$$\inf_{\widehat{\nu}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{\nu}(\mathbf{X}) - f\|_2^2 \geq \frac{\varepsilon^2}{8c^2},$$

if ε satisfies $\log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c) > 2n\varepsilon^2/\alpha + 2\log 2$.

2.2.2 Upper Bound

We now turn our attention to the upper bound. We note that our universal estimator over $\mathcal{F}$, will be a likelihood-based estimator for f. As such for any two densities $g, g' \in \mathcal{F}$, we will routinely work with the *log-likelihood difference* for the n observed samples $\mathbf{X} := (X_1, \ldots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f \in \mathcal{F}$. We will denote this by

$$\psi(g, g', \mathbf{X}) := \log\left(\prod_{i=1}^{n} \frac{g(X_i)}{g'(X_i)}\right) = \sum_{i=1}^{n} \log\left(\frac{g(X_i)}{g'(X_i)}\right) = \sum_{i=1}^{n} \log g(X_i) - \sum_{i=1}^{n} \log g'(X_i). \tag{2.7}$$

Remark 2.12. We note that the log-likelihood difference $\psi(g, g', \mathbf{X})$ in (2.7), is well-defined. This follows since for each $i \in [n]$, the individual random variables $\log g(X_i)/g'(X_i)$ are well-defined (as $\alpha > 0$), and bounded. That is, $-\infty < \log \alpha/\beta \leq \log g(X_i)/g'(X_i) \leq \log \beta/\alpha < \infty$, for each $i \in [n]$.

We will use the log-likelihood difference to help us decide which of the two densities is "more" correct, given the observed data samples $\mathbf{X}$. Given this, we will first need a concentration result on the density log-likelihood difference. We do this by establishing the following lemma.

Lemma 2.13 (Log-likelihood difference concentration in $\mathcal{F}$)**.** *Let $\delta > 0$ be arbitrary, and let $\mathbf{X} := (X_1, \ldots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, be the n observed samples. Suppose we are trying to distinguish between two densities $g, g' \in \mathcal{F}$. Let $\psi(g, g', \mathbf{X})$ denote their log-likelihood difference per (2.7). We then have*

$$\sup_{\substack{g, g' :\, \|g - g'\|_2 \geq C\delta, \\ \|g' - f\|_2 \leq \delta}} \mathbb{P}(\psi(g, g', \mathbf{X}) > 0) \leq \exp\left(-nL(\alpha, \beta, C)\delta^2\right) \tag{2.8}$$

where

$$C > 1 + \sqrt{1/(\alpha c(\alpha, \beta))} \tag{2.9}$$

$$L(\alpha, \beta, C) := \frac{\left(\sqrt{c(\alpha, \beta)}(C - 1) - \sqrt{1/\alpha}\right)^2}{2\left\{2K(\alpha, \beta) + \frac{2}{3}\log \beta/\alpha\right\}}, \tag{2.10}$$

with $K(\alpha, \beta) \coloneqq \beta/(\alpha^2 c(\alpha, \beta))$, and $c(\alpha, \beta)$ is as defined in Lemma 2.5. In the above $\mathbb{P}$ is taken with respect to the true density function f, i.e., $\mathbb{P} = \mathbb{P}_f$.

From Lemma 2.13, we derive a key concentration result concerning a packing set in $\mathcal{F}$, as summarized in Lemma 2.14. The relevance of such a result will become clearer later, when we introduce our sieve-based MLE for f. Our sieve estimator will be constructed using packing sets of $\mathcal{F}$, thus Lemma 2.14 will be an important tool to enable us to handle the concentration properties of our estimator.

Lemma 2.14 (Maximum likelihood concentration in $\mathcal{F}$)**.** *Let $\delta > 0$ be arbitrary, and let $\mathbf{X} \coloneqq (X_1, \ldots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, be the n observed samples. Suppose further that we have a maximal δ-packing set of $\mathcal{F}' \subset \mathcal{F}$, i.e., $\{g_1, \ldots, g_m\} \subset \mathcal{F}'$ such that $\|g_i - g_j\|_2 > \delta$ for all $i \neq j$, and it is known that $f \in \mathcal{F}'$. Now let $j^* \in [m]$, denote the index of a density whose likelihood is the largest. We then have*

$$\mathbb{P}(\|g_{j^*} - f\|_2 > (C+1)\delta) \leq m \exp\left(-nL(\alpha, \beta, C)\delta^2\right),$$

where C is assumed to satisfy (2.9), and $L(\alpha, \beta, C)$ is defined as per (2.10).

Next we establish that the map $\varepsilon \mapsto \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)$ is non-increasing. This lemma is made possible by the fact that the set $\mathcal{F}$ is convex by assumption, and that we are using the L_2-metric. This monotonicity property of the ε-local metric entropy in the L_2-metric is a *critical* technical ingredient used in the proofs establishing our upper bound.

Lemma 2.15 (Monotonicity of local metric entropy)**.** *The map $\varepsilon \mapsto \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)$ is non-increasing.*

We now turn our attention to describing our proposed likelihood-based estimator, i.e., $\nu^*(\mathbf{X})$, of $f \in \mathcal{F}$. In the discussion that follows we let $d \coloneqq \operatorname{diam}_2(\mathcal{F})$, which is finite by the boundedness of $\mathcal{F}$. The estimator is directly inspired by a recent construction used in Neykov (2022), who applied it to the Gaussian sequence model. However, there the underlying space used is $(\mathbb{R}^n, \|\cdot\|_2)$, whereas in our case it is $(\mathcal{F}, \|\cdot\|_2)$, which has a *vastly* different underlying geometric structure. Importantly since we are performing density estimation, our proposed estimator uses a fundamentally different *log-likelihood*-based selection criteria, compared to the *projection*-based sequence model estimator in Neykov (2022). Although our estimator can also be described constructively, it is not intended to be practically computable.

Construction of the multistage sieve MLE, $\nu^*(\mathbf{X})$, of $f \in \mathcal{F}$.

Step 1 Initialize inputs.

Let $\mathbf{X} \coloneqq (X_1, \ldots, X_n)^\top$ denote our n observed i.i.d. data samples. Fix some sufficiently large $c > 0$, and then define C such that $c \coloneqq 2(C+1)$. Importantly, the constant c should be set *without* looking at the data samples, i.e., *independently* of $\mathbf{X}$.

Step 2 Construct a maximal packing set tree of depth $\overline{J}$ *before* seeing the data.

Construct a tree of packing sets of depth $\overline{J} \in \mathbb{N}$, which is *independent* of the data samples $\mathbf{X}$. Here, $\overline{J}$ is as defined in Theorem 2.19. The explicit construction of such a packing set tree proceeds as follows. First, fix any arbitrary point $\Upsilon_1 \in \mathcal{F}$, which is the root node, i.e., the first level of the packing set tree. In the case where $\overline{J} = 1$, the tree construction stops at this single root node.

Assuming the (more interesting) case where $\overline{J} > 1$, we then let $d \coloneqq \operatorname{diam}(\mathcal{F})$, and construct a maximal $\frac{d}{2(C+1)}$-packing set of $B_2(\Upsilon_1, d) \cap \mathcal{F} = \mathcal{F}$. Denote this packing set by $P_{\Upsilon_1} \coloneqq \{m_1, m_2, m_3, \ldots, m_{|P_{\Upsilon_1}|}\}$. The set P_{Υ_1} forms the children (densities) of our root node, that is the second level of the tree[a]. Now, for *each* density in P_{Υ_1}, we again construct a maximal packing set as follows. For example, taking the density $m_3 \in P_{\Upsilon_1}$, we construct a maximal $\frac{d}{4(C+1)}$-packing set of $B_2(m_3, d/2) \cap \mathcal{F}$, which we denote as $P_{m_3} \coloneqq \{m_{3,1}, m_{3,2}, m_{3,3}, \ldots, m_{3,|P_{m_3}|}\}$. Here, the (finite) packing set P_{m_3} again forms the children of the node density m_3. Iterating this process over each density in P_{Υ_1}, forms the complete second level of the tree.

Now we can further iterate this process over each density in the second level of the tree to construct the third level of the tree. For example, taking the density $m_{3,3}$, we construct a maximal $\frac{d}{8(C+1)}$-packing set of $B_2(m_{3,3}, d/4) \cap \mathcal{F}$, which we denote as $P_{m_{3,3}} \coloneqq \{m_{3,3,1}, m_{3,3,2}, m_{3,3,3}, \ldots, m_{3,3|P_{m_{3,3}}|}\}$, which forms the children of node $m_{3,3}$. This process is iterated so that for the k^{th}-level of the tree, we construct $\frac{d}{2^k(C+1)}$-packing sets, with closed balls $B_2(\cdot, d/2^{k-1}) \cap \mathcal{F}$. In particular the packing set tree is extended for each depth level $k \in \{2, 3, \ldots, \overline{J} - 1\}$. This process results in a maximal packing set tree of depth $\overline{J}$, as claimed.

Step 3 Build a finite sequence of densities by traversing our packing set tree.

Now, *after* observing our data sample $\mathbf{X}$, we construct a *finite* sequence of densities, i.e., $\Upsilon := (\Upsilon_k)_{k=1}^{\overline{J}}$, using our packing set tree construction in **Step 2**. First, we initialize the first term of our sequence to Υ_1, i.e., the root node already chosen in **Step 2**. If $\overline{J} = 1$, then the sequence $\Upsilon := (\Upsilon_1)$. Otherwise, if $\overline{J} > 1$, we traverse down one level of our packing set tree, and assign Υ_2 to be the density from P_{Υ_1} which maximizes the log-likelihood *given* the data. That is, set $\Upsilon_2 := \arg\max_{\nu \in P_{\Upsilon_1}} \sum_{i=1}^{n} \log \nu(X_i)$. Since P_{Υ_1} is a finite set, this will be exhausted for each such iteration in finitely many steps.

Moreover, we note that when assigning Υ_2, there may be ties in children densities who all simultaneously maximize the log-likelihood. To break ties, by *convention*, we always select the *left-most* child from our packing set tree[b]. Once the Υ_2 is assigned from our packing set tree, once again assign Υ_3 from its children by again maximizing the log-likelihood. Keep iterating in this manner for each index[c] $k \in \{2, 3, \dots, \overline{J}\}$, and construct the finite, i.e., *terminating* sequence Υ.

Step 4 Output estimator as the $\overline{J}^{\text{th}}$-term of the sequence.

Finally, we note that the finite sequence $\Upsilon := (\Upsilon_k)_{k=1}^{\overline{J}}$ satisfies[d] $\|\Upsilon_J - \Upsilon_{J'}\|_2 \leq \frac{d}{2^{J'-2}}$, for each pair of positive integers $J' < J$. Our multistage sieve MLE, i.e., $\nu^*(\mathbf{X})$, can be taken as the final term of this sequence. That is $\nu^*(\mathbf{X}) := \Upsilon_{\overline{J}}$. The estimator $\nu^*(\mathbf{X})$ is readily understood by comparing[e] Figure 2.2 with the qualitative description in **Step 1**-**Step 4**.

[a] By *convention*, the children forming the packing set densities are arbitrarily indexed in an increasing alphanumeric manner, from left child node to right child node.

[b] This selection rule thus effectively assigns the child density maximizing log-likelihood with the *smallest* such alphanumeric index.

[c] Note that k here refers to index of the k^{th}-term our sequence Υ.

[d] We will formally justify this in the appendix in Lemma 2.38.

[e] We note that in Figure 2.2 if $\overline{J} = 1$, the estimator would just output Υ_1. In the case where $\overline{J} > 1$, the maximal packing sets for each level of the tree are illustrated on the left, and the corresponding constructed tree level is shown on the right. In this instance the finite sequence of $\overline{J}$ densities is given by $\Upsilon = (\Upsilon_1, m_3, m_{3,3}, m_{3,3,2}, \dots, m_{3,3,2,\dots,5})$. The estimator then takes the $\overline{J}^{\text{th}}$-term of Υ,

i.e., $\nu^*(\mathbf{X}) = m_{\underbrace{3,3,2,\ldots,5}_{(\bar{J}-1)-\text{terms}}}$.

Remark 2.16. We emphasize that Figure 2.2 is not drawn to any precise scale. In reality the L_2-balls should be much "wider" than the set $\mathcal{F}$ (and $\mathcal{F}_B^{[\alpha,\beta]}$). This is because they do not impose that their elements are proper densities, unlike the elements of the set $\mathcal{F}$ (and $\mathcal{F}_B^{[\alpha,\beta]}$) which are non-negative and integrate to 1. It is intended to be useful conceptual guide to understanding the construction of our multistage sieve MLE.

We observe that our proposed estimator $\nu^*(\mathbf{X})$ can be thought of as an "multistage sieve MLE" in the spirit of Wong and Shen (1995). Broadly speaking a 'sieve' MLE effectively takes the MLE over a strategically constrained subset of the parameter space, i.e., $\mathcal{F}$ in our setting (see Chapter 8 Grenander, 1981, e.g., for more details). Specifically, as we traverse the down the finite-depth maximal packing set tree, each group of children densities along with the MLE selection rule can be thought of as a "sieve". We note that the sieve MLE proposed in Wong and Shen (1995) is a construction which is also not practically computable for general density classes $\mathcal{F}$.

Remark 2.17 (An *online* finite packing set tree construction). We note that the finite-depth maximal packing set tree described in **Step 2**, can be replaced with a conceptually simpler *online* finite-depth maximal packing set tree construction. This proceeds as follows. Once again, as per **Step 2**, we can initialize $\Upsilon_1 \in \mathcal{F}$ to be the root node independently of the data. We then construct the second level of our packing set tree, i.e., $P_{\Upsilon_1} := \{m_1, m_2, m_3, \ldots, m_{|P_{\Upsilon_1}|}\}$, as the previously described maximal packing set. This first level is constructed without looking at the data samples $\mathbf{X}$. This time however, we can traverse down the first level of the tree and set $\Upsilon_2 := \arg\max_{\nu \in P_{\Upsilon_1}} \sum_{i=1}^n \log \nu(X_i)$, i.e., by using the data samples $\mathbf{X}$. Given Υ_2 selected in this data driven manner, we can construct the second level of the tree as the children of Υ_2, i.e., the maximal packing set P_{Υ_2} *without* using the data samples. We can then set $\Upsilon_3 := \arg\max_{\nu \in P_{\Upsilon_2}} \sum_{i=1}^n \log \nu(X_i)$, once again using the data. We can thus repeat this recursive process for $\overline{J}$ iterations, whereby the maximal packing set of children of each parent node are constructed without seeing the data. The specific child node is selected after seeing the data, *and then* the estimator can traverse to one of these children. This does not require the all possible children of all possible parent nodes of the maximal packing set tree to be constructed up front as described in **Step 2**. Instead, we only construct the children as required in a simple *sequential* manner.

We next show that our multistage sieve MLE is a measurable function of the data with respect to the Borel σ-field on $\mathcal{F}$ in L_2-metric topology. This is important, because all upper bound risk rates in expectation for $\nu^*(\mathbf{X})$ that follow, are with respect to the L_2-metric topology on $\mathcal{F}$.

Proposition 2.18 (Measurability of $\nu^*(\mathbf{X})$)**.** *The multistage sieve MLE, i.e., $\nu^*(\mathbf{X})$, is a measurable function of the data with respect to the Borel σ-field on $\mathcal{F}$ in the L_2-metric topology.*

With the measurability of $\nu^*(\mathbf{X})$ established, the main theorem establishing the performance of $\nu^*(\mathbf{X})$ is Theorem 2.19 below.

Theorem 2.19 (Upper bound rate for the multistage sieve MLE $\nu^*(\mathbf{X})$)**.** *Let, $\nu^*(\mathbf{X}) = \Upsilon_{\overline{J}}$ be the output of the multistage sieve MLE which is run for $\overline{J} \in \mathbb{N}$ steps. Here $\overline{J}$ is defined as the maximal integer $J \in \mathbb{N}$, such that $\varepsilon_J := \frac{\sqrt{L(\alpha,\beta,c/2-1)}\,d}{2^{(J-2)}c}$ satisfies*[4]

$$n\varepsilon_J^2 > 2\log M^{\mathsf{loc}}_{\mathcal{F}}\left(\varepsilon_J \frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}}, c\right) \vee \log 2, \tag{2.11}$$

or $\overline{J} = 1$ if no such J exists. Then

$$\mathbb{E}\|\nu^*(\mathbf{X}) - f\|_2^2 \le \bar{C}\varepsilon^{*2},$$

for some universal constant $\bar{C}$, and where $\varepsilon^* := \varepsilon_{\overline{J}}$. We remind the reader that $c := 2(C+1)$ is the constant from the definition of local metric entropy, which is assumed to be sufficiently large. Here C is assumed to satisfy (2.9), and $L(\alpha,\beta,C)$ is defined as per (2.10).

We will now formally illustrate that the above estimator achieves the minimax rate. The precise expression of the rate is quantified in the following result.

Theorem 2.20 (Minimax rate)**.** *Define $\varepsilon^* := \sup\{\varepsilon : n\varepsilon^2 \le \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)\}$, where c in the definition of local metric entropy is a sufficiently large absolute constant. Then the minimax rate is given by $\varepsilon^{*2} \wedge d^2$ up to absolute constant factors.*

[4] Observe that by the definition of $\varepsilon_{\overline{J}}$ and (2.11) we have that all packing sets used in the construction of the estimator must be finite, even though we are not assuming that the set $\mathcal{F}$ is totally bounded.

Remark 2.21 (Extending results to loss functions in KL-divergence and the Hellinger metric). Recall that by Lemma 2.5 we have the "topological equivalence" of the KL-divergence and squared Hellinger metric with the squared L_2-metric on $\mathcal{F}$. This means that we can readily extend our minimax risk bounds in Theorem 2.20 to loss functions measured via KL-divergence and the squared Hellinger metric. The important consideration is that (2.11) is still solved (in both cases) using the local metric entropy of $\mathcal{F}$ using the squared L_2-metric. Note that for the KL-divergence to be well-defined, we require that all densities are strictly positively lower bounded over the common compact support.

We now argue that the minimax rate for a class $\mathcal{F} \subset \mathcal{F}_B^{[0,\beta]}$ which is convex and not necessarily lower bounded by $\alpha > 0$ is given by the same equation, as long as there exists a single density in $f_\alpha \in \mathcal{F}$ which is α-lower bounded. The argument used to establish this claim essentially the same as used in Yang and Barron (1999, Lemma 1), which we formalize for our setting in Proposition 2.22. For completeness, we provide all details for our setting in the Appendix.

Proposition 2.22 (Extending results to $\mathcal{F}_B^{[0,\beta]}$)**.** *Let $\mathcal{F} \subset \mathcal{F}_B^{[0,\beta]}$ be a convex class of densities, with at least one $f_\alpha \in \mathcal{F}$ that is α-lower bounded, with $\alpha > 0$. Then the minimax rate in the squared L_2-metric is $\varepsilon^{*2} \wedge d^2$, where $\varepsilon^* := \sup\{\varepsilon : n\varepsilon^2 \leq \log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)\}$.*

2.2.3 Adaptivity

In this section we illustrate that the estimator, $\nu^*(\mathbf{X})$, as defined in Section 2.2.2 is adaptive to the true density f. Before that, similar to Neykov (2022) we re-define the notion of *adaptive L_2-local metric entropy* for any density $\theta \in \mathcal{F}$.

Definition 2.23 (Adaptive Local Entropy)**.** Let $\theta \in \mathcal{F}$ be a density. Let $M(\theta, \varepsilon, c)$ denote the maximal cardinality of a packing set of the set $B_2(\theta, \varepsilon) \cap \mathcal{F}$ at an L_2 distance ε/c.

$$M_{\mathcal{F}}^{\mathsf{adloc}}(\theta, \varepsilon, c) := M(\varepsilon/c, \mathcal{F} \cap B_2(\theta, \varepsilon)) =: M_{\mathcal{F} \cap B_2(\theta,\varepsilon)}^{\mathsf{glo}}(\varepsilon/c)\,.$$

We refer to $\log M_{\mathcal{F}}^{\mathsf{adloc}}(\theta, \varepsilon, c)$ as the *adaptive L_2-local metric entropy* of $\mathcal{F}$ at θ.

Theorem 2.24 (Adaptive upper bound rate for the multistage sieve MLE $\nu^*(\mathbf{X})$)**.** *Let, $\nu^*(\mathbf{X}) = \Upsilon_{\overline{J}}$ be the output of the multistage sieve MLE which is* *run for $\overline{J}$ iterations where $\overline{J}$ is defined as the maximal solution to*

$$n\varepsilon_J^2 > 2 \inf_{f \in \mathcal{F}} M_{\mathcal{F}}^{\mathsf{adloc}} \left(f, 2\varepsilon_J \frac{c}{\sqrt{L(\alpha, \beta, c/2-1)}}, 2c \right) \vee \log 2,$$

where $\varepsilon_J \coloneqq \frac{\sqrt{L(\alpha,\beta,c/2-1)}\,d}{2^{(J-2)}c}$ *and* $\overline{J} = 1$ *if no such* J *exists*[5]. *Let* J^* *be defined as the maximal integer* $J \in \mathbb{N}$, *such that* $\varepsilon_J \coloneqq \frac{\sqrt{L(\alpha,\beta,c/2-1)}\,d}{2^{(J-2)}c}$ *such that*[6].,

$$n\varepsilon_J^2 > 2 M_{\mathcal{F}}^{\mathsf{adloc}} \left(f, 2\varepsilon_J \frac{c}{\sqrt{L(\alpha, \beta, c/2-1)}}, 2c \right) \vee \log 2, \tag{2.12}$$

and $J^* = 1$ *if no such* J *exists. Then*

$$\mathbb{E}\|\nu^*(\mathbf{X}) - f\|_2^2 \leq \bar{C}\varepsilon^{*2},$$

for some universal constant $\bar{C}$, *and where* $\varepsilon^* \coloneqq \varepsilon_{J^*}$. *We remind the reader that* $c \coloneqq 2(C+1)$ *is the constant from the definition of local metric entropy, which is assumed to be sufficiently large. Here* C *is assumed to satisfy* (2.9), *and* $L(\alpha, \beta, C)$ *is defined as per* (2.10).

2.3 Examples

We will now apply our work to derive risk bounds for density estimation (under the squared L_2-metric) for various examples of convex density classes $\mathcal{F}$. To that end, per Proposition 2.22 our risk bounds only require us to establish that the stated class $\mathcal{F}$ is indeed convex, and importantly that there exists at least one density $f_\alpha \in \mathcal{F}$ that is positively bounded away from 0 over the entire support B. In order to establish the latter fact we can usually take $f_\alpha \sim \mathsf{Unif}[B]$, and check that it lies in our density class $\mathcal{F}$, and by suitably expanding our ambient space[7] $\mathcal{F}_B^{[\alpha,\beta]}$. We will also use the following key fact relating L_2-*local* and L_2-*global* metric entropies.

$$\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \leq \log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c) \leq \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \tag{2.13}$$

[5]Note that running the estimator with $\overline{J}$ many steps, may result into having non-finite packing sets — that is not an issue however.

[6]Observe that by the definition of $\varepsilon_{\overline{J}}$ and (2.11) we have that *some* packing sets used in the construction of the *adaptive* estimator may not be finite, but will be at most countable. This follows from the L_2-separability of $\mathcal{F}$ and is formalized in Lemma 2.34 in the appendix. We note that the measurability of the adaptive estimator still holds as per Proposition 2.18 in this case.

[7]We reiterate that our use of $\{\alpha, \beta\}$ in this section (and throughout the paper) is *only* used to refer to the constants in Definition 2.1, of $\mathcal{F}_B^{[\alpha,\beta]}$ and thus $\mathcal{F}$.

Here, (2.13) follows directly from Yang and Barron (1999, Lemma 2), where it is only proved for the case $c = 2$. However, their proof directly extends to the more general case for each $c > 0$, which is required for our setting. For the various examples of $\mathcal{F}$ that follow below, we will show that a stronger sufficient condition on global entropy is satisfied, namely

$$\log M^{\mathsf{glo}}_{\mathcal{F}}(\varepsilon/c) - \log M^{\mathsf{glo}}_{\mathcal{F}}(\varepsilon) \asymp \log M^{\mathsf{glo}}_{\mathcal{F}}(\varepsilon/c), \tag{2.14}$$

provided we take c to be sufficiently large enough, which is within our control to do, per our packing set tree construction. In short, (2.14) will enable us to bound the local metric entropy via (2.13). To illustrate this, we initially consider two examples from Yang and Barron (see 1999, Section 6). We begin with the class $\mathcal{F} \coloneqq \mathsf{Lip}_{\gamma,q}(\Psi)$, i.e., the (γ, q, Ψ)-Lipschitz density class defined as per (2.15). As noted in Yang and Barron (1999, Section 6.4), with fixed constants $\max\{1/q - 1/2, 0\} < \gamma \leq 1$, and $1 \leq q \leq \infty$, the ε-*global* metric entropy of $\mathsf{Lip}_{\gamma,q}(\Psi)$ is of the order $\varepsilon^{-1/\gamma}$ per Birman and Solomjak (1980).

Example 2.25 (Lipschitz density class $\mathcal{F}$)**.** Let $1 < \Psi < \beta < \infty$, $\max\{1/q - 1/2, 0\} < \gamma \leq 1$, and $1 \leq q \leq \infty$ be fixed constants, and $B \coloneqq [0, 1]$. Now, let $\mathcal{F} \coloneqq \mathsf{Lip}_{\gamma,q}(\Psi)$ denote the space of (γ, q, Ψ)-Lipschitz densities with total variation at most β. That is,

$$\mathsf{Lip}_{\gamma,q}(\Psi) \coloneqq \left\{ f \colon B \to [0, \Psi] \,\middle|\, \|f(x+h) - f(x)\|_q \leq \Psi h^{\gamma}, \|f\|_q \leq \Psi, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{2.15}$$

and $\|f\|_q \coloneqq \left(\int_B |f(x)|^q \,\mathrm{d}\mu\right)^{1/q}$. Note that in (2.15) we have that $x \in B$, and only consider $h > 0$ such that $x + h \in B$, so that the predicate of $\mathsf{Lip}_{\gamma,q}(\Psi)$ is well-defined. Then $\mathsf{Lip}_{\gamma,q}(\Psi)$ is a convex density class, there exists a density $f_\alpha \in \mathsf{Lip}_{\gamma,q}(\Psi)$ that is strictly positively bounded away from 0, and the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{Lip}_{\gamma,q}(\Psi)$ is of the order $n^{-\frac{2\gamma}{2\gamma+1}}$.

Another well studied density estimation problem is the case where $\mathcal{F} \coloneqq \mathsf{BV}_\zeta$ is total bounded variation at most ζ, defined as per (2.16). Importantly we note that the ε-*global* L_2-metric entropy of this well studied function class is of the order ε^{-1} (see Section 6.4 Yang and Barron, 1999, e.g., for more details).

Example 2.26 (Bounded total variation density class $\mathcal{F}$)**.** Let $1 < \zeta < \beta < \infty$ be a fixed constant, and $B \coloneqq [0, 1]$. Now, let $\mathcal{F} \coloneqq \mathsf{BV}_\zeta$ denote the space of univariate densities with total variation at most β. That is,

$$\mathsf{BV}_\zeta := \left\{ f \colon B \to [0, \zeta] \,\middle|\, \|f\|_\infty \leq \zeta, V(f) \leq \zeta, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{2.16}$$

where we define the total variation of f, i.e., $V(f)$ as

$$V(f) := \sup_{\{x_1,\ldots,x_m \,|\, 0 \leq x_1 < \cdots < x_m \leq 1, m \in \mathbb{N}\}} \sum_{i=1}^{m-1} |f(x_{i+1}) - f(x_i)|, \tag{2.17}$$

and $\|f\|_\infty := \sup_{x \in B} |f(x)|$. Then the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{BV}_\zeta$ is of the order $n^{-2/3}$.

Another interesting example illustrating the use case of our bounds is that where $\mathcal{F} := \mathsf{Quad}_\gamma$, forms the density class of γ-quadratic functionals defined as per (2.18). Importantly we note that the ε-*global* L_2-metric entropy of this well studied function class is of the order $\varepsilon^{-1/4}$ (see Example 15.8 and Example 15.22 Wainwright, 2019, e.g., for more details).

Example 2.27 (Quadratic functional density class $\mathcal{F}$)**.** Let $0 < \alpha < 1 < \beta < \infty$, and $\gamma > 1$ be fixed constants, with $B := [0, 1]$. Now, let $\mathcal{F} := \mathsf{Quad}_\gamma$ denote the space of univariate quadratic functional densities. That is,

$$\mathsf{Quad}_\gamma := \left\{ f \colon B \to [\alpha, \beta] \,\middle|\, \|f''\|_\infty \leq \gamma, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}. \tag{2.18}$$

Then Quad_γ is a convex density class, there exists a density $f_\alpha \in \mathsf{Quad}_\gamma$ that is strictly positively bounded away from 0, and the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{Quad}_\gamma$ is of the order $n^{-4/5}$.

We now turn our attention to an interesting example, which demonstrates that our results can yield useful bounds in cases where L_2-*global* metric entropy of $\mathcal{F}$ may be unknown (or difficult to compute), but the L_2-*local* metric entropy can be controlled.

Example 2.28 (Convex mixture density class $\mathcal{F}$)**.** Let $\mathcal{F} := \mathsf{Conv}_k$ where

$$\mathsf{Conv}_k := \left\{ \sum_{i=1}^{k} \alpha_i f_i \,\middle|\, \sum_{i=1}^{k} \alpha_i = 1, \alpha_i \geq 0, f_i \in \mathcal{F}_B^{[\alpha,\beta]} \right\}, \tag{2.19}$$

for some <u>fixed</u> $k \in \mathbb{N}$ and $f_i \in \mathcal{F}_B^{[\alpha,\beta]}$ for each $i \in [k]$. Further, let $\mathbf{G} = (\mathbf{G}_{ij})_{i,j \in [k]}$ denote the Gram matrix with $\mathbf{G}_{ij} := \int_B f_i f_j \mu(\mathrm{d}x)$, which we assume is positive definite, i.e., $\mathbf{G} \succ \mathbf{0}$. Then the minimax rate for estimating $f \in \mathsf{Conv}_k$ is bounded from above by $\sqrt{\frac{k}{n}}$ up to absolute constant factors.

2.4 Discussion

In this paper we derived exact minimax rates for density estimation over convex density classes. Our work builds on seminal research of Le Cam (1973); Birgé (1983); Yang and Barron (1999); Wong and Shen (1995). More directly, we non-trivially adapted the techniques of Neykov (2022), who used it for deriving exact rates for the Gaussian sequence model. Our results demonstrate that the L_2-*local* metric entropy *always* determines that minimax rate under squared L_2-loss in this setting. We thus provide a unifying perspective across parametric *and* nonparametric convex density classes, and under weaker assumptions than those used by Yang and Barron (1999).

An important open question that we would like to think further about is whether there exists a computationally tractable estimator which is also minimax optimal in our setting. We can also consider applying our techniques to the nonparametric regression setting (with Gaussian noise) where f is a uniformly bounded regression function of interest. We leave these exciting directions for future work. Finally, we hope that this research stimulates further activity in approximating L_2-*local* metric entropy for various convex density classes.

2.5 Acknowledgments

We would like to thank Arun Kumar Kuchibhotla for suggesting to us to prove that our estimator is adaptive to the true density. We would also like to thank Wanshan Li, Yang Ning, Alex Reinhart, Alessandro Rinaldo, and Larry Wasserman for providing insightful feedback and encouragement during this work. All figures in this paper were drawn using the `Mathcha`[8] editor.

[8] https://www.mathcha.io/editor

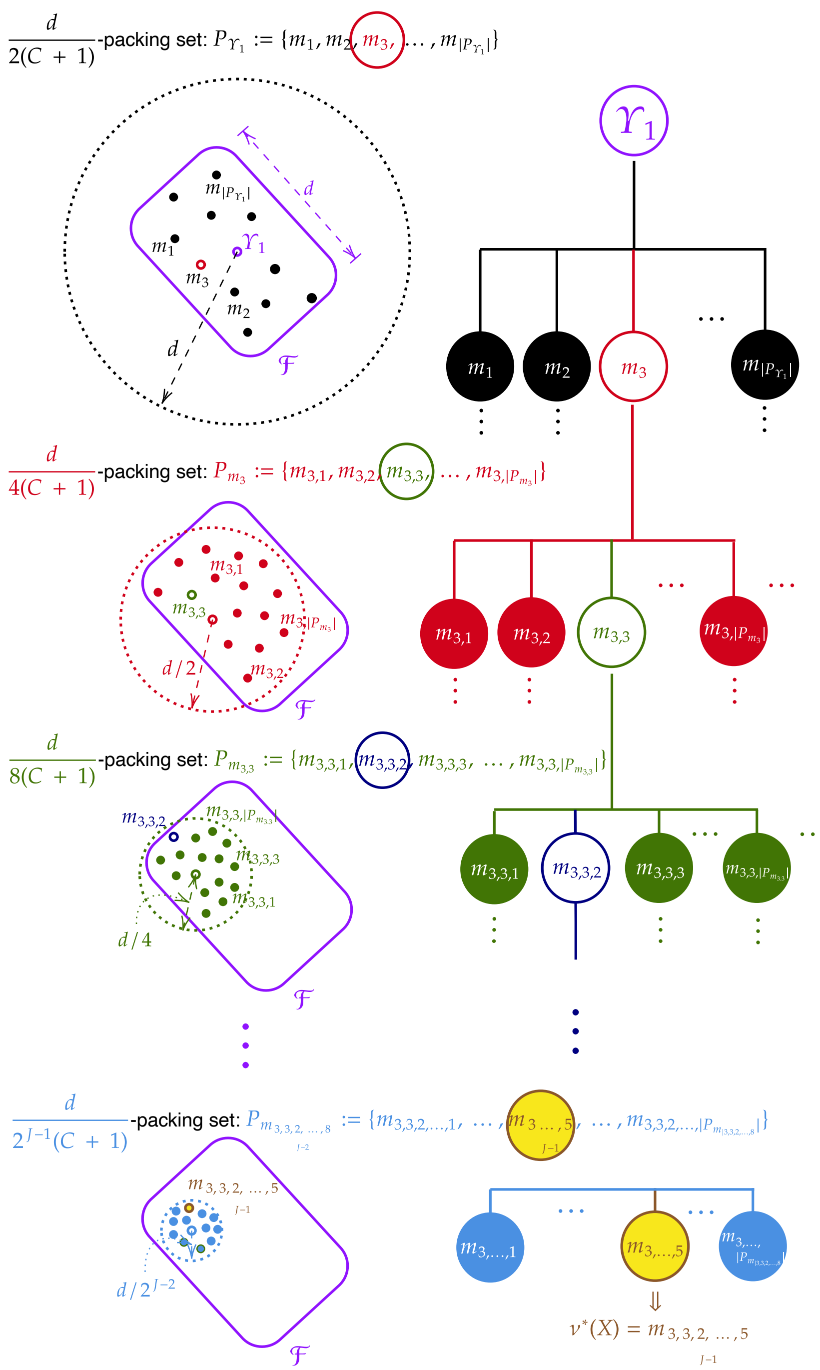

Figure 2.2: Maximal packing set tree construction in **Step 2**.

Appendix - Chapter 2

2.A Preliminary

We begin with some basic mathematical preliminaries for our work.

2.A.1 Notation Summary

To ensure that the Appendix is can be read in a standalone manner, we consolidate key notation used in the paper in Table 2.A.1.

Table 2.A.1: Notation and conventions used in this chapter

Variables and inequalities	
$a \wedge b$	$\min\{a, b\}$ for each $a, b \in \mathbb{R}$
$a \vee b$	$\max\{a, b\}$ for each $a, b \in \mathbb{R}$
$\lesssim$	$\leq$ up to positive universal constants
$\gtrsim$	$\geq$ up to positive universal constants
$\asymp$	if both $\lesssim$ and $\gtrsim$ hold
Functions and sets	
$\\|\cdot\\|_2$	the L_2-metric in $\mathcal{F}$
$[m]$	$\{1, \ldots, m\}$, for $m \in \mathbb{N}$
$B_2(\theta, r)$	closed L_2-ball centered at $\theta \in \mathcal{F}$ with radius r

2.A.2 Properties of $\mathcal{F}_B^{[\alpha,\beta]}$

Here we provide some basic analytic properties of our core density class $\mathcal{F}_B^{[\alpha,\beta]}$, as per Definition 2.1. Many of these facts will be used, sometimes implicitly, in our proofs. We hope that by documenting them rigorously, they provide the reader with a much richer understanding of the geometry of this broader density class. This may also be a useful reference for researchers in working in similar density estimation settings. We provide suitable references where the properties follow from standard real analysis theory.

Lemma 2.29 (Convexity of $\mathcal{F}_B^{[\alpha,\beta]}$)**.** *The density class $\mathcal{F}_B^{[\alpha,\beta]}$, forms a convex set, in the L_2-metric.*

Proof of Lemma 2.29. In order to show the convexity of $\mathcal{F}_B^{[\alpha,\beta]}$, Let $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$, and let $\kappa \in [0,1]$ be arbitrary. Then for each $x \in B$, we observe that

$$(\kappa f + (1-\kappa)g)(x) \coloneqq \kappa f(x) + (1-\kappa)g(x) \geq \kappa\alpha + (1-\kappa)\alpha \geq \alpha \tag{2.20}$$

$$(\kappa f + (1-\kappa)g)(x) \coloneqq \kappa f(x) + (1-\kappa)g(x) \leq \kappa\beta + (1-\kappa)\beta \leq \beta \tag{2.21}$$

From (2.20) and (2.21), it follows that $\kappa f + (1-\kappa)g \colon B \to [\alpha, \beta]$. Moreover, since $\int_B f \,\mathrm{d}\mu = \int_B g \,\mathrm{d}\mu = 1$, we have

$$\int_B (\kappa f + (1-\kappa)g) \,\mathrm{d}\mu = \kappa \int_B f \,\mathrm{d}\mu + (1-\kappa) \int_B g \,\mathrm{d}\mu = 1. \tag{2.22}$$

Since f, g are measurable functions, then so is their convex combination, i.e., $\kappa f + (1-\kappa)g$. Combining the above we have shown that $\kappa f + (1-\kappa)g \in \mathcal{F}_B^{[\alpha,\beta]}$, which proves the convexity of $\mathcal{F}_B^{[\alpha,\beta]}$, as required. □

Lemma 2.30 (Boundedness of $\mathcal{F}_B^{[\alpha,\beta]}$)**.** *The density class $\mathcal{F}_B^{[\alpha,\beta]}$, is bounded, in the L_2-metric.*

Proof of Lemma 2.30. We now show that $\mathcal{F}_B^{[\alpha,\beta]}$ is bounded in the L_2-metric. To see this observe that for any $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$:

$$\|f-g\|_2^2 \coloneqq \int_B (f-g)^2 \,\mathrm{d}\mu \leq \int_B |f-g|\, 2\beta \,\mathrm{d}\mu \leq 2\beta \left(\int_B |f| \,\mathrm{d}\mu + \int_B |g| \,\mathrm{d}\mu \right) = 4\beta. \tag{2.23}$$

It follows that $\operatorname{diam}_2\left(\mathcal{F}_B^{[\alpha,\beta]}\right) \coloneqq \sup\left\{ \|f-g\|_2 \,\middle|\, f, g \in \mathcal{F}_B^{[\alpha,\beta]} \right\} \leq 2\sqrt{\beta} < \infty$, as required. □

Lemma 2.31 ($\mathcal{F}_B^{[\alpha,\beta]}$ lies in $L^2(B)$)**.** *The density class $\mathcal{F}_B^{[\alpha,\beta]}$, satisfies $\mathcal{F}_B^{[\alpha,\beta]} \subset L^2(B)$, where*

$$L^2(B) \coloneqq \left\{ f \colon B \to \mathbb{R} \,\middle|\, \int_B f^2 \,\mathrm{d}\mu < \infty, f \text{ measurable} \right\}. \tag{2.24}$$

As such $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ is an induced metric subspace of $(L^2(B), \|\cdot\|_2)$.

Proof of Lemma 2.31. Let $f \in \mathcal{F}_B^{[\alpha,\beta]}$ be arbitrary. We then observe that

$$\int_B f^2 \,\mathrm{d}\mu \leq \int_B f\beta \,\mathrm{d}\mu = \beta < \infty, \tag{2.25}$$

since $f \leq \beta$ by definition of $\mathcal{F}_B^{[\alpha,\beta]}$. Given that $f\colon B \to [\alpha,\beta] \subset \mathbb{R}$, we have that $f \in L^2(B)$, i.e., $\mathcal{F}_B^{[\alpha,\beta]} \subset L^2(B)$ as required. □

Lemma 2.32 (Completeness and separability of $L^2(B)$)**.** *The metric space $(L^2(B), \|\cdot\|_2)$, with $L^2(B)$ defined as per Equation (2.24), is complete and separable.*

Proof of Lemma 2.32. We note that completeness of $(L^2(B), \|\cdot\|_2)$ follows directly from Brezis (2011, Theorem 4.8), and separability follows from Brezis (2011, Theorem 4.13). □

Lemma 2.33 (Completeness and separability of $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$)**.** *The metric space $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ is complete and separable.*

Proof of Lemma 2.30. Firstly we note that $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ is an induced metric subspace of $(L^2(B), \|\cdot\|_2)$ per Lemma 2.31. Now separability of $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ follows, since it is inherited from $(L^2(B), \|\cdot\|_2)$ by applying Shirali and Vasudeva (2006, Proposition 2.3.16). We now show the completeness of $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$. Take an arbitrary Cauchy sequence $(f_k)_{k=1}^{\infty}$ in $\mathcal{F}_B^{[\alpha,\beta]}$. Since $(L^2(B), \|\cdot\|_2)$ is complete per Lemma 2.32, it follows that the L_2 limit of $(f_k)_{k=1}^{\infty}$ exists in $L^2(B)$. Let f be that limit, i.e., $\lim_{k\to\infty} f_k =: f \in L^2(B)$. We will show that $f \in \mathcal{F}_B^{[\alpha,\beta]}$. First let us show that it is a density, i.e., it integrates to 1. By Cauchy-Schwartz $\int_B |f_k(x) - f(x)|\mu(\mathrm{d}x) \leq \sqrt{\int_B |f_k(x) - f(x)|^2 \mu(\mathrm{d}x)} \to 0$ so that $\int f(x)\mu(\mathrm{d}x) = 1$. Next consider the function $f' = (f \wedge \alpha) \vee \beta$. Since for any $x \in [\alpha,\beta]$ and any y we have $|x-y| \geq |x - (y\wedge\alpha)\vee\beta|$ then that implies (since $f_k \in \mathcal{F}_B^{[\alpha,\beta]}$) that $\int_B |f_k(x) - f'(x)|^2\mu(\mathrm{d}x) \leq \int_B |f_k(x) - f(x)|^2 \mu(\mathrm{d}x) \to 0$ and so f' must also be a limit of f_k. Since the limits are unique (up to considering equivalence classes modulo sets of measure 0 with respect to μ) then $f = f'$ and hence it belongs to $\mathcal{F}_B^{[\alpha,\beta]}$. □

Lemma 2.34 (Separability of $(\mathcal{F}, \|\cdot\|_2)$)**.** *The metric space $(\mathcal{F}, \|\cdot\|_2)$ is separable. Furthermore, if $A \subset \mathcal{F}$, then $(A, \|\cdot\|_2)$ is separable.*

Proof of Lemma 2.34. We first observe by Lemma 2.33 that the metric space $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$ is separable. It then follows that since $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$, that $(\mathcal{F}, \|\cdot\|_2)$ is a restriction of $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$, and thus a separable metric space by Shirali and Vasudeva (2006, Proposition 2.3.16). By a similar argument, it follows that if $A \subset \mathcal{F}$, then $(A, \|\cdot\|_2)$ is separable. □

The following lemma shows that in general $\mathcal{F}_B^{[\alpha,\beta]}$ is not totally bounded. We consider a restricted case of $B \coloneqq [0,1]$, to construct a suitable counterexample.

Lemma 2.35 (Non total boundedness of $(\mathcal{F}_{[0,1]}^{[\alpha,\beta]}, \|\cdot\|_2)$)**.** *Suppose $\beta \geq 2-\alpha$. Then the metric space $(\mathcal{F}_{[0,1]}^{[\alpha,\beta]}, \|\cdot\|_2)$ is not totally bounded, and hence not compact.*

Proof of Lemma 2.35. We note that per Shirali and Vasudeva (2006, Theorem 5.1.12) a metric space is totally bounded if and only if every sequence contains a Cauchy subsequence. We will use this characterization to construct a counterexample to demonstrate that $(\mathcal{F}_{[0,1]}^{[\alpha,\beta]}, \|\cdot\|_2)$ is not totally bounded. In particular, we will define a sequence in $(\mathcal{F}_{[0,1]}^{[\alpha,\beta]}, \|\cdot\|_2)$ which can't contain *any* Cauchy subsequence.

Specifically we consider the sequence of functions $(1, \{x \mapsto \sin(2\pi jx)\}_{j\in\mathbb{N}})$. These functions are orthonormal in $L^2([0,1])$. Construct the sequence of functions $f_j(x) = 1 + (1-\alpha)\sin(2\pi jx)$ for $j \in \mathbb{N}$. By the orthogonality of 1 and $\sin(2\pi jx)$ we have that $\int_0^1 f_j(x)dx = 1$. Furthermore, $\alpha \leq f_j(x) \leq 2-\alpha$ for all $x \in [0,1]$, hence since $\beta \geq 2-\alpha$ we have $f_j(x) \in \mathcal{F}_{[0,1]}^{[\alpha,\beta]}$. Take any two $j \neq k \in \mathbb{N}$, and consider

$$\|f_j - f_k\|_2^2 = (1-\alpha)^2\|\sin(2\pi jx) - \sin(2\pi kx)\|_2^2 = 2(1-\alpha)^2 > 0.$$

This shows that there cannot be a Cauchy subsequence and hence the set is not totally bounded. □

2.A.3 Elementary inequalities

We will state and prove Lemma 2.36, which will provide the key fact to will assist us in the proof of the lower bound in Lemma 2.5.

Lemma 2.36 (Elementary log inequality)**.** *For each $\gamma > 0$, and for any $x \in (0,\gamma]$, the following relationship holds:*

$$\log x \leq (x-1) - h(\gamma)(x-1)^2. \tag{2.26}$$

Here $h : (0,\infty) \to \mathbb{R}$ is defined as in (2.6), and is positive over its entire support.

Proof of Lemma 2.36. We first argue that $h(x) > 0$ for $x \in (0,\infty)$. This is by the elementary inequality $\log(x+1) \leq x$ for all $x \geq -1$. Next, it suffices to show that the map $x \mapsto h(x)$ is decreasing for $x > 0$ where h is defined in (2.6). This is because (2.26) holds for $x = 1$, and if $x \neq 1$ it is equivalent to

$$h(\gamma) \leq \frac{(x-1) - \log x}{(x-1)^2},$$

for $x \leq \gamma$. It is simple to verify that

$$h'(x) = \frac{-x^2 + 2x\log x + 1}{(x-1)^3 x}.$$

We will show that the above function is negative on $(0,\infty)$ which will complete the proof. First we will evaluate it at $x = 1$. By a triple application of L'Hôpital's rule it is simple to verify that $\frac{d}{dx}h(x)|_{x=1} = -\frac{1}{3} < 0$. Thus, it remains to show that for $x \neq 1$,

$$(-x^2 + 2x\log x + 1)(x-1) < 0.$$

Now let $f(x) \coloneqq \frac{x^2-1}{2x} - \log x$. We want to show that $f(x) > 0$, for each $x > 1$ and $f(x) < 0$ for $x < 1$. First observe that $f(1) = 0$. Moreover, we have that

$$f'(x) = \frac{(x-1)^2}{2x^2} > 0. \tag{2.27}$$

That is, $f(x)$ is *strictly* increasing, which implies that $f(x) > 0$ for each $x \in (1,\infty)$, and $f(x) < 0$ for each $x < 1$ as required. $\square$

2.B Proofs of Section 2.2

2.B.1 Proof of Lemma 2.5

Lemma 2.5 (KL-L_2 equivalence on $\mathcal{F}_B^{[\alpha,\beta]}$)**.** *For each pair of densities $f, g \in \mathcal{F}_B^{[\alpha,\beta]}$, the following relationship holds:*

$$c(\alpha, \beta)\|f - g\|_2^2 \leq d_{\mathsf{KL}}(f||g) \leq (1/\alpha)\|f - g\|_2^2, \tag{2.5}$$

where we denote $c(\alpha, \beta) := \frac{h(\beta/\alpha)}{\beta} > 0$. Here $h : (0, \infty) \to \mathbb{R}$ is defined to be

$$h(\gamma) := \begin{cases} \frac{\gamma - 1 - \log \gamma}{(\gamma - 1)^2} & \text{if } \gamma \in (0, \infty) \setminus \{1\} \\ \frac{1}{2} = \lim_{x \to 1} \frac{x - 1 - \log x}{(x-1)^2} & \text{if } \gamma = 1, \end{cases} \tag{2.6}$$

and is positive over its entire support. It is also easily seen that on $\mathcal{F}_B^{[\alpha,\beta]}$, d_{KL} (and hence the L_2-metric) is also equivalent to the Hellinger metric. Furthermore, these properties are also inherited by $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$, which is our density class of interest.

Proof of Lemma 2.5. We will prove the upper and lower bound in turn.

(Upper bound in (2.5)): We seek to show that $d_{\mathsf{KL}}(f||g) \leq \frac{1}{\alpha}\|f - g\|_2^2$. First, for any two densities $f, g \in \mathcal{F}$, we define the χ^2-divergence between f and g, as follows:

$$\chi^2(f||g) := \int_B \frac{(f-g)^2}{g} \, \mathrm{d}\mu. \tag{2.28}$$

Per Remark 2.4, we note that $\chi^2(f||g)$ in (2.28) is similarly well-defined. We then have:

$$\begin{aligned}
d_{\mathsf{KL}}(f||g) &\leq \chi^2(f||g) && \text{(per Gibbs and Su (2002, Theorem 5))} \\
&:= \int_B \frac{(f-g)^2}{g} \, \mathrm{d}\mu && \text{(using (2.28))} \\
&\leq \frac{1}{\alpha} \int_B (f-g)^2 \, \mathrm{d}\mu && (\text{since } \inf_{x \in B} g(x) \geq \alpha > 0) \\
&=: \frac{1}{\alpha}\|f - g\|_2^2. && \text{(by definition)}
\end{aligned}$$

As required. ■

(Lower bound in (2.5)): We seek to show that $d_{\mathsf{KL}}(f||g) \geq c(\alpha, \beta)\|f - g\|_2^2$.

We proceed as follows: First observe that for any $f, g \in \mathcal{F}$, we have that $0 < \frac{g}{f} \leq \frac{\beta}{\alpha} < \infty$

$$\begin{aligned}
d_{\mathsf{KL}}(f||g) &:= \int_B f \log\left(\frac{f}{g}\right) \mathrm{d}\mu && \text{(per (2.4))} \\
&= \int_B -f \log\left(\frac{g}{f}\right) \mathrm{d}\mu && (\text{since } \inf_{x\in B} f(x) \geq \alpha > 0) \\
&\geq \int_B -f\left(\left(\frac{g}{f} - 1\right) - h(\beta/\alpha)\left(\frac{g}{f} - 1\right)^2\right) \mathrm{d}\mu \\
& && \left(\text{using Lemma 2.36, with } C = \tfrac{\beta}{\alpha} \text{ and } x = \tfrac{g}{f}\right) \\
&= \int_B (f - g)\,\mathrm{d}\mu + h(\beta/\alpha)\int_B \frac{(g-f)^2}{f}\,\mathrm{d}\mu \\
&\geq \frac{h(\beta/\alpha)}{\beta}\int_B (g-f)^2\,\mathrm{d}\mu \\
& && \left(\text{since } \textstyle\int_B (f-g)\,\mathrm{d}\mu = 0, \text{ and } 0 < \sup_{x\in B} f(x) \leq \beta\right) \\
&=: \frac{h(\beta/\alpha)}{\beta}\|f - g\|_2^2 \\
&=: c(\alpha,\beta)\|f-g\|_2^2, && (2.29)
\end{aligned}$$

where we define $c(\alpha, \beta) := \frac{h(\beta/\alpha)}{\beta} > 0$. This proves the lower bound in (2.5), as required. ■

We now show the following equivalence between the Hellinger, i.e., d_{H}-metric, and the L_2 metric in $\mathcal{F}_B^{[\alpha,\beta]}$.

$$(1/4\beta)\|f-g\|_2^2 \leq d_{\mathsf{H}}(f||g)^2 \leq (1/\alpha)\|f-g\|_2^2. \tag{2.30}$$

To prove the upper bound in (2.30), we note that

$$\begin{aligned}
d_{\mathsf{H}}(f||g)^2 &\leq d_{\mathsf{KL}}(f||g) && \text{(from Gibbs and Su (2002))} \\
&\leq (1/\alpha)\|f-g\|_2^2, && \text{(per (2.5))}
\end{aligned}$$

as required. In order to prove the lower bound we observe that

$$\begin{aligned}
\|f-g\|_2^2 &= \int_B (f-g)^2\,\mathrm{d}\mu \\
&= \int_B (\sqrt{f} + \sqrt{g})^2(\sqrt{f} - \sqrt{g})^2\,\mathrm{d}\mu \\
&\leq 4\beta\int_B (\sqrt{f} - \sqrt{g})^2\,\mathrm{d}\mu && (\text{since } f, g \leq \beta.) \\
&=: 4\beta d_{\mathsf{H}}(f||g)^2, && \text{(by definition)}
\end{aligned}$$

which implies the required lower bound in (2.30). We have thus established the required upper and lower bounds in both (2.5) and (2.30).

Finally, we note that $(\mathcal{F}, \|\cdot\|_2)$ is metric space, since it is the restriction of the metric space $(\mathcal{F}_B^{[\alpha,\beta]}, \|\cdot\|_2)$. And so the bounds (2.5) and (2.30), are also inherited by $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$. □

2.B.2 Proof of Lemma 2.11

Lemma 2.11 (Minimax lower bound)**.** *Let $c > 0$ be fixed, and independent of the data samples $\mathbf{X}$. Then the minimax rate satisfies*

$$\inf_{\widehat{\nu}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{\nu}(\mathbf{X}) - f\|_2^2 \geq \frac{\varepsilon^2}{8c^2},$$

if ε satisfies $\log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c) > 2n\varepsilon^2/\alpha + 2\log 2$.

Proof of Lemma 2.11. Let $c > 0$ be fixed, and $\theta \in \mathcal{F}$ be an arbitrary point. Consider maximal packing the set $\{f^1, \dots, f^m\} \subset \mathcal{F} \cap B_2(\theta, \varepsilon)$ at a L_2-"distance" at least ε/c. Here $B_2(\theta, \varepsilon)$ denotes a closed L_2-ball around the point θ, with radius ε. Suppose it has m elements. Then we know that

$$I(X; J) \leq \frac{1}{m} \sum_{j=1}^{m} d_{\mathsf{KL}}\big(f^j || \theta\big) \leq \max_{j \in [m]} d_{\mathsf{KL}}\big(f^j || \theta\big) \leq \max_{j \in [m]} (1/\alpha) \|f^j - \theta\|_2^2 \leq \varepsilon^2/\alpha.$$

Here the final two inequalities follow by applying (2.5), and using the fact that $\{f^1, \dots, f^m\} \subset \mathcal{F} \cap B_2(\theta, \varepsilon)$, respectively. Hence, if the packing number satisfies $\log m \geq 2n\varepsilon^2/\alpha + 2\log 2$ we will have a lower bound proportional to ε^2 (it will be $\varepsilon^2/(8c^2)$). By taking the supremum over θ, we conclude that if $\log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c) > 2n\varepsilon^2/\alpha + 2\log 2$ we have a lower bound proportional to ε^2. □

2.B.3 Proof of Lemma 2.13

Lemma 2.13 (Log-likelihood difference concentration in $\mathcal{F}$)**.** *Let $\delta > 0$ be arbitrary, and let $\mathbf{X} \coloneqq (X_1, \dots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, be the n observed samples. Suppose we are trying to distinguish between two densities $g, g' \in \mathcal{F}$. Let $\psi(g, g', \mathbf{X})$ denote their log-likelihood difference per (2.7). We then have*

$$\sup_{\substack{g,g' : \|g-g'\|_2 \geq C\delta, \\ \|g'-f\|_2 \leq \delta}} \mathbb{P}(\psi(g, g', \mathbf{X}) > 0) \leq \exp\big(-nL(\alpha, \beta, C)\delta^2\big) \tag{2.8}$$

where

$$C > 1 + \sqrt{1/(\alpha c(\alpha,\beta))} \tag{2.9}$$

$$L(\alpha,\beta,C) \coloneqq \frac{\left(\sqrt{c(\alpha,\beta)}(C-1) - \sqrt{1/\alpha}\right)^2}{2\left\{2K(\alpha,\beta) + \frac{2}{3}\log\beta/\alpha\right\}}, \tag{2.10}$$

with $K(\alpha,\beta) \coloneqq \beta/(\alpha^2 c(\alpha,\beta))$*, and* $c(\alpha,\beta)$ *is as defined in Lemma 2.5. In the above* $\mathbb{P}$ *is taken with respect to the true density function* f*, i.e.,* $\mathbb{P} = \mathbb{P}_f$*.*

Proof of Lemma 2.13. We first observe per Remark 2.12 that the log-likelihood, $\psi(g, g', X)$, is well-defined. Next the mean of these variables, for each $i \in [n]$, is

$$\begin{aligned}
\mathbb{E}_f\left[\log\frac{g(X_i)}{g'(X_i)}\right] &= \mathbb{E}_f\left[\log\left(\frac{f(X_i)}{g'(X_i)}\Big/\frac{f(X_i)}{g(X_i)}\right)\right] \\
&\qquad\qquad\text{(which is well-defined by Remark 2.12.)} \\
&= \mathbb{E}_f\left[\log\frac{f(X_i)}{g'(X_i)}\right] - \mathbb{E}_f\left[\log\frac{f(X_i)}{g(X_i)}\right] \\
&= d_{\mathsf{KL}}(f||g') - d_{\mathsf{KL}}(f||g). \qquad (2.31)
\end{aligned}$$

Where the last line follows by definition using (2.4). We then have

$$\begin{aligned}
\mathbb{P}(\psi(g,g',X) > 0) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(X_i)}{g'(X_i)} > 0\right) \qquad \text{(using (2.7))} \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(X_i)}{g'(X_i)} - \mathbb{E}_f\log\frac{g(X_1)}{g'(X_1)} > \mathbb{E}_f\log\frac{g'(X_1)}{g(X_1)}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i}\log\frac{g(X_i)}{g'(X_i)} - \mathbb{E}_f\log\frac{g(X_1)}{g'(X_1)} > d_{\mathsf{KL}}(f||g) - d_{\mathsf{KL}}(f||g')\right) \\
&\qquad\qquad \text{(using (2.31))} \\
&\le \exp\left(-\frac{n^2t^2}{2\left\{\sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right] + \frac{1}{3}n\kappa t\right\}}\right) \\
&= \exp\left(-\frac{n^2t^2}{2\left\{n\mathbb{E}\left[Y_1^2\right] + \frac{1}{3}n\kappa t\right\}}\right) \qquad \text{(since } Y_i \text{ are i.i.d.)} \\
&= \exp\left(-\frac{nt^2}{2\left\{\mathbb{E}\left[Y_1^2\right] + \frac{1}{3}\kappa t\right\}}\right) \qquad (2.32)
\end{aligned}$$

where $\kappa \coloneqq 2\log\beta/\alpha$, $t \coloneqq d_{\mathsf{KL}}(f||g) - d_{\mathsf{KL}}(f||g')$, and $Y_i \coloneqq \log\frac{g(X_i)}{g'(X_i)} - \mathbb{E}_f\log\frac{g(X_1)}{g'(X_1)}$. This follows by the boundedness of $\log g(X_i)/g'(X_i)$, and then by applying Bernstein's inequality, provided that $t > 0$. In order to check this final positivity condition, we first note that there exists a $C > 0$ such that $\|g - g'\|_2 \geq C\delta$, and $\|g' - f\|_2 \leq \delta$ both hold. We then have

$$\|f - g\|_2 \geq (C-1)\delta. \tag{2.33}$$

To see this we observe that by assumption, and the triangle inequality respectively that $C\delta \leq \|g - g'\| \leq \|g - f\| + \|f - g'\|$. Then using $\|f - g'\| \leq \delta$ by assumption and re-arranging, we obtain (2.33) as required. As a result we obtain the following two inequalities

$$\sqrt{d_{\mathsf{KL}}(f||g)} \geq \sqrt{c(\alpha,\beta)}\|f - g\|_2 \geq \sqrt{c(\alpha,\beta)}(C-1)\delta \tag{2.34}$$

$$\sqrt{d_{\mathsf{KL}}(f||g')} \leq \sqrt{1/\alpha}\|f - g'\|_2 \leq \sqrt{1/\alpha}\,\delta, \tag{2.35}$$

where $C > 0$ is defined to be a constant satisfying $c(\alpha,\beta)(C-1)^2 > 1/\alpha$, i.e.,

$$C > 1 + \sqrt{1/(\alpha c(\alpha,\beta))}. \tag{2.36}$$

Under the condition specified by (2.36), and by squaring and subtracting (2.35) from (2.34), we obtain

$$t \coloneqq d_{\mathsf{KL}}(f||g) - d_{\mathsf{KL}}\big(f||g'\big) \geq (c(\alpha,\beta)(C-1)^2 - 1/\alpha)\delta^2 > 0 \tag{2.37}$$

Now we show that $\mathbb{E}_f\left(Y_1^2\right) \lesssim d_{\mathsf{KL}}(f||g) + d_{\mathsf{KL}}(f||g')$. To see this

$$\begin{aligned}
\mathbb{E}_f\left(Y_1^2\right) &\leq \mathbb{E}_f\left[\left(\log\frac{g(X_1)}{g'(X_1)}\right)^2\right] \\
&= \mathbb{E}_f\left[\log\left(\frac{f(X_1)}{g'(X_1)}\Big/\frac{f(X_1)}{g(X_1)}\right)\right] \\
&\qquad\qquad\qquad\qquad \text{(which is well-defined by Remark 2.12.)} \\
&= \mathbb{E}_f\left[\left(\log\frac{f(X_1)}{g'(X_1)} - \log\frac{f(X_1)}{g(X_1)}\right)^2\right] \\
&\leq 2\underbrace{\mathbb{E}_f\left[\left(\log\frac{f(X_1)}{g(X_1)}\right)^2\right]}_{=:A} + 2\underbrace{\mathbb{E}_f\left[\left(\log\frac{f(X_1)}{g'(X_1)}\right)^2\right]}_{=:B} \\
&\qquad\qquad \text{(using } (a-b)^2 \leq 2(a^2+b^2)\text{, for } a, b \geq 0.\text{)}
\end{aligned}$$

We now bound the A term above, with B handled similarly. We observe that:

$$\begin{aligned}
A &\coloneqq \mathbb{E}_f\left[\left(\log\frac{f(X_1)}{g(X_1)}\right)^2\right] && \text{(by definition)}\\
&= \int f\left(\log\frac{f}{g}\right)^2 \mathrm{d}\mu\\
&= \int_{f\leq g} f\left(\log\frac{g}{f}\right)^2 \mathrm{d}\mu + \int_{g<f} f\left(\log\frac{f}{g}\right)^2 \mathrm{d}\mu. && (2.38)
\end{aligned}$$

Now using $\log x \leq x-1$, for each $x \in \mathbb{R}_{>0}$, we have that

$$\left(\log\frac{g}{f}\right)^2 \leq \left(\frac{g-f}{f}\right)^2 \quad\text{and}\quad \left(\log\frac{f}{g}\right)^2 \leq \left(\frac{f-g}{g}\right)^2, \tag{2.39}$$

which hold for $f \leq g$ (i.e., $\frac{g}{f} \geq 1$), and $g < f$ (i.e., $\frac{f}{g} > 1$), respectively. Now we have:

$$\begin{aligned}
A &\leq \int_{f\leq g}\frac{(g-f)^2}{f}\,\mathrm{d}\mu + \int_{g<f}\frac{(f-g)^2 f}{g^2}\,\mathrm{d}\mu && \text{(using (2.38) and (2.39).)}\\
&\leq (1/\alpha)\int_{f\leq g}(g-f)^2\,\mathrm{d}\mu + (\beta/\alpha^2)\int_{g<f}(f-g)^2\,\mathrm{d}\mu\\
& && \text{(since } 0<\alpha<f,g\leq\beta)\\
&\leq (\beta/\alpha^2)\|f-g\|_2^2 && \text{(since } \beta/\alpha^2 \geq 1/\alpha.)\\
&\leq K(\alpha,\beta)d_{\mathsf{KL}}(f||g), && (2.40)
\end{aligned}$$

where $K(\alpha,\beta) \coloneqq \beta/(\alpha^2 c(\alpha,\beta))$, where $c(\alpha,\beta)$ is as defined in Lemma 2.5. By a similar argument, we also have that

$$B \leq K(\alpha,\beta)d_{\mathsf{KL}}\left(f||g'\right). \tag{2.41}$$

Let $z \coloneqq d_{\mathsf{KL}}(f||g) + d_{\mathsf{KL}}(f||g')$. Then using (2.40) and (2.41), we obtain

$$\mathbb{E}_f\left(Y_1^2\right) \leq 2K(\alpha,\beta)[d_{\mathsf{KL}}(f||g) + d_{\mathsf{KL}}\left(f||g'\right)] \eqqcolon 2zK(\alpha,\beta) \tag{2.42}$$

Now we use the basic inequality $a+b \leq \left(\sqrt{a}+\sqrt{b}\right)^2 \leq 2(a+b)$, to obtain

$$z \leq \left(\sqrt{d_{\mathsf{KL}}(f||g)} + \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2 \leq 2z. \tag{2.43}$$

Now, $t^2 := (d_{\mathsf{KL}}(f||g) - d_{\mathsf{KL}}(f||g'))^2 = \left(\sqrt{d_{\mathsf{KL}}(f||g)} - \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2 \left(\sqrt{d_{\mathsf{KL}}(f||g)} + \sqrt{d_{\mathsf{KL}}(f||g')}\right.$
we have:

$$\left(\sqrt{d_{\mathsf{KL}}(f||g)} - \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2 z \leq t^2 \leq 2\left(\sqrt{d_{\mathsf{KL}}(f||g)} - \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2 z. \tag{2.44}$$

We then conclude using (2.32),(2.37), that

$$\begin{aligned}
\mathbb{P}(\psi(g, g', X) > 0) &\leq \exp\left(-\frac{nt^2}{2\left\{\mathbb{E}\left[Y_1^2\right] + \frac{1}{3}\kappa t\right\}}\right) && \text{(per (2.32))} \\
&\leq \exp\left(-\frac{n\left(\sqrt{d_{\mathsf{KL}}(f||g)} - \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2 z}{2\left\{2zK(\alpha,\beta) + \frac{1}{3}\kappa z\right\}}\right) \\
& && \text{(since } t \leq z \text{ and (2.42))} \\
&= \exp\left(-\frac{n\left(\sqrt{d_{\mathsf{KL}}(f||g)} - \sqrt{d_{\mathsf{KL}}(f||g')}\right)^2}{2\left\{2K(\alpha,\beta) + \frac{1}{3}\kappa\right\}}\right) \\
&\leq \exp\left(-\frac{n\left(\sqrt{c(\alpha,\beta)}(C-1) - \sqrt{1/\alpha}\right)^2 \delta^2}{2\left\{2K(\alpha,\beta) + \frac{1}{3}\kappa\right\}}\right) \\
& && \text{(by subtracting (2.35) from (2.34))} \\
&=: \exp\left(-nL(\alpha,\beta,C)\delta^2\right),
\end{aligned}$$

whenever condition (2.36) holds, and $L(\alpha,\beta,C) := \frac{\left(\sqrt{c(\alpha,\beta)}(C-1) - \sqrt{1/\alpha}\right)^2}{2\left\{2K(\alpha,\beta) + \frac{2}{3}\log\beta/\alpha\right\}}$. Now, taking the supremum over all $g, g' \colon \|g - g'\|_2 \geq C\delta, \|g' - f\|_2 \leq \delta$, the required result follows. □

2.B.4 Proof of Lemma 2.14

Recall that Lemma 2.14 is concerning a packing set. Suppose we have a maximal packing set of $\mathcal{F}' \subset \mathcal{F}$, i.e., $\{g_1, \ldots, g_m\} \subset \mathcal{F}' \subset \mathcal{F}$ such that $\|g_i - g_j\|_2 > \delta$ for all $i \neq j$, and it is known that $f \in \mathcal{F}'$. We then obtain a key concentration result as per Lemma 2.14.

Lemma 2.14 (Maximum likelihood concentration in $\mathcal{F}$)**.** *Let $\delta > 0$ be arbitrary, and let $\mathbf{X} := (X_1, \ldots, X_n)^\top \overset{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, be the n observed samples. Suppose further that we have a maximal δ-packing set of $\mathcal{F}' \subset \mathcal{F}$, i.e., $\{g_1, \ldots, g_m\} \subset \mathcal{F}'$ such that $\|g_i - g_j\|_2 > \delta$ for all $i \neq j$, and it is known that $f \in \mathcal{F}'$. Now let*

$j^* \in [m]$, denote the index of a density whose likelihood is the largest. We then have

$$\mathbb{P}(\|g_{j^*} - f\|_2 > (C+1)\delta) \leq m \exp\left(-nL(\alpha, \beta, C)\delta^2\right),$$

where C is assumed to satisfy (2.9), and $L(\alpha, \beta, C)$ is defined as per (2.10).

Proof of Lemma 2.14. We first define the intermediate *thresholding* random variables

$$T_k := \begin{cases} \max_{j\in[m]} \|g_j - g_k\|_2 & \text{, s.t. } \sum_{i=1}^n \log g_j(X_i) \geq \sum_{i=1}^n \log g_k(X_i), \|g_j - g_k\|_2 > C\delta \\ 0 & \text{, otherwise,} \end{cases}$$

for each $k \in [m]$. Without loss of generality suppose that $\|g_k - f\|_2 \leq \delta$. Next

$$\begin{aligned} \mathbb{P}(\|g_{j^*} - f\|_2 > (C+1)\delta) &\leq \mathbb{P}(j^* \in \{j : \|g_j - g_k\|_2 > C\delta\}) \\ &\leq \mathbb{P}(T_k > 0). \end{aligned}$$

On the other hand

$$\begin{aligned} \mathbb{P}(T_k > 0) &= \mathbb{P}\left(\exists j \in [m] \colon \sum_{i=1}^n \log g_j(X_i) \geq \sum_{i=1}^n \log g_k(X_i), \|g_j - g_k\|_2 > C\delta\right) \\ &= \mathbb{P}\left(\bigcup_{j=1}^m \left\{\sum_{i=1}^n \log g_j(X_i) \geq \sum_{i=1}^n \log g_k(X_i), \|g_j - g_k\|_2 > C\delta\right\}\right) \\ &\leq m \exp\left(-nL(\alpha, \beta, C)\delta^2\right), \\ &\qquad\qquad \text{(using union bound and Lemma 2.13.)} \end{aligned}$$

where C is assumed to satisfy (2.9), and $L(\alpha, \beta, C)$ is defined as per (2.10). $\square$

2.B.5 Proof of Lemma 2.15

Lemma 2.15 (Monotonicity of local metric entropy)**.** *The map $\varepsilon \mapsto \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)$ is non-increasing.*

Proof of Lemma 2.15. It suffices to show that if $g_1, \ldots, g_m \in \mathcal{F} \cap B_2(\theta, \varepsilon)$ is a maximal packing set at a distance ε/c, then we can pack $B_2(\theta, \varepsilon') \cap \mathcal{F}$ at a distance ε'/c with at least m points where $\varepsilon' < \varepsilon$. Consider the points $\theta(1 - \varepsilon'/\varepsilon) + \varepsilon'/\varepsilon g_j$. These points clearly are densities since $\theta, g_j \in \mathcal{F}$. We will show that these points are an ε'/c packing of $B_2(\theta, \varepsilon') \cap \mathcal{F}$. First let us convince ourselves that the points belong to the set. We have

$$\|\theta(1 - \varepsilon'/\varepsilon) + \varepsilon'/\varepsilon g_j - \theta\|_2 = \varepsilon'/\varepsilon \|g_j - \theta\|_2 \leq \varepsilon',$$

and using the fact that $\mathcal{F}$ is convex (by assumption) grants the conclusion. Next

$$\|\theta(1-\varepsilon'/\varepsilon)+\varepsilon'/\varepsilon g_j-\theta(1-\varepsilon'/\varepsilon)-\varepsilon'/\varepsilon g_k\|_2=\varepsilon'/\varepsilon\|g_j-g_k\|_2>\varepsilon'/c,$$

which completes the proof. □

2.B.6 Proof of Proposition 2.18

Proposition 2.18 (Measurability of $\nu^*(\mathbf{X})$)**.** *The multistage sieve MLE, i.e., $\nu^*(\mathbf{X})$, is a measurable function of the data with respect to the Borel σ-field on $\mathcal{F}$ in the L_2-metric topology.*

Proof of Proposition 2.18. Recall our multistage sieve estimator $\nu^*(\mathbf{X}) \coloneqq \Upsilon_{\overline{J}}$, where $\mathbf{X} \coloneqq (X_1,\ldots,X_n)^\top$ is a fixed data sample. Here $\Upsilon_{\overline{J}}$ denotes the last term of the *finite* sequence $\Upsilon \coloneqq (\Upsilon_k)_{k=1}^{\overline{J}}$ as described in Section 2.2.2.

In order to show the measurability of $\nu^*(\mathbf{X})$ we need to formalize our setting. We note that our estimator $\nu^*\colon B^n \to \mathcal{F}$, is more precisely a map from the measurable space $(B^n, \sigma(B^n))$ to the measurable space $(\mathcal{F}, \sigma(\mathcal{F}))$. Here $\sigma(B^n)$ and $\sigma(\mathcal{F})$ denote the Borel σ-field with respect to the Euclidean and L_2-metric topologies on B^n and $\mathcal{F}$, respectively.

Our proof strategy will be to proceed by induction on $k \in [\overline{J}]$ over the sequence Υ. We will show that each k^{th}-indexed map in Υ, i.e., Υ_k, is Borel measureable, which in turn will imply the measureability of the $\nu^*(\mathbf{X})$. Following our (maximal) packing set construction as described in Section 2.2.2 and Figure 2.2, we need to consider the case where the traversal down the tree is not necessarily unique at each level, i.e., there may be collisions (ties) in the packing set children nodes, where the likelihood is equal. We do always ensure a unique path down the maximal packing set tree, by selecting the smallest alphanumerically indexed children node at each level. However, our measurability proof must account for this selection rule explicitly.

In order to proceed by induction, we consider the base case for $k=1$, i.e., $\Upsilon_1 \in \mathcal{F}$. Importantly, we note that Υ_1 is chosen arbitrarily from $\mathcal{F}$ independently of the data samples, $\mathbf{X}$. Let $A \in \sigma(\mathcal{F})$ be any Borel set. Since all samples $\mathbf{X} \in B^n$ are mapped to Υ_1 in our setting, then $\Upsilon_1^{-1}[A] = B^n$ if $\Upsilon_1 \in A \in \sigma(\mathcal{F})$, or $\Upsilon_1^{-1}[A] = \varnothing$, otherwise. In either case we have $\varnothing, B^n \in \sigma(B^n)$, which shows that Υ_1 is Borel measurable. Now consider the event $\{\Upsilon_2 = m_s\} \coloneqq \left\{(X_1,\ldots,X_n)^\top \in B^n \,\middle|\, \Upsilon_2(X_1,\ldots,X_n) = m_s\right\} \subset B^n$, for some index $s \in \mathbb{N}$. Then we have

$$\{\Upsilon_2 = m_s\} := \left\{ (X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \Upsilon_2(X_1, \ldots, X_n) = m_s \right\} \tag{2.45}$$

$$= \bigcap_{g \in P_{\Upsilon_1}} \left\{ (X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \sum_{i=1}^{n} \log(m_s(X_i)) \geq \sum_{i=1}^{n} \log(g(X_i)) \right\} \bigcap$$
$$\bigcap_{j=1}^{s-1} \left\{ (X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \sum_{i=1}^{n} \log(m_s(X_i)) > \sum_{i=1}^{n} \log(m_j(X_i)) \right\}. \tag{2.46}$$

In (2.46), we observe that $\{\Upsilon_2 = m_s\} \subset B^n$ is represented as the intersection of 2 separate (finite) set intersections. Note that the second intersection set *explicitly* accounts for our alphanumerical index selection rule in the children densities of P_{Υ_1}. Consider the first finite intersection term. Here, each $g \in P_{\Upsilon_1} \subset \mathcal{F}$ are Borel measurable by (2.1). We note that the log and the addition (i.e., "$+_{\mathbb{R}}$") functions are both continuous and measurable, and therefore, so is their composition. Thus the resulting *finite* sum, $\sum_{i=1}^{n} \log f(X_i)$, is a measurable function, for any density $f \in \mathcal{F}$ (which is always measurable). As such the Υ_2 is measurable since all these inequalities give rise to measurable sets and when one intersects them (they are finitely many) one obtains another measurable set. Once again, let $A \in \sigma(\mathcal{F})$ be any Borel set. Then such an A contains either no such densities m_s, or at most finitely many (since the number of children of our maximal packing set tree is always finite). If no such $m_s \in A$, then $\Upsilon_2^{-1}[A] = \varnothing \in \sigma(B^n)$. Thus Υ_2 is indeed Borel measurable in this case. In the case where there exist finitely many such $m_s \in A$, it follows that

$$\Upsilon_2^{-1}[A] = \bigcup_{\{s \,|\, m_s \in A\}} \{\Upsilon_2 = m_s\} =: \bigcup_{\{s \,|\, m_s \in A\}} \left\{ (X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \Upsilon_2(X_1, \ldots, X_n) = m_s \right\} \tag{2.47}$$

In (2.47) we note that $\Upsilon_2^{-1}[A]$ represents a finite union of Borel measurable sets as per (2.46), which is again Borel measurable. That is, we have shown that $\Upsilon_2^{-1}[A] \in \sigma(B^n)$, which indeed implies the Borel measurability of Υ_2, as required.

Similarly, consider the event $\{\Upsilon_3 =: m_{s,t}\} \subset B^n$, for some $t \in \mathbb{N}$ and $s \in \mathbb{N}$ taken as per (2.45). Here, the indexed density $m_{s,t}$ signifies that Υ_3 is derived from the children of the packing set of $\Upsilon_2 =: m_s$, as denoted by P_{m_s} in our work. Once again we can write this Υ_3 as

$$\begin{aligned}\{\Upsilon_3 = m_{s,t}\} &:= \left\{(X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \Upsilon_3(X_1, \ldots, X_n) = m_{s,t}\right\} \\ &= \bigcap_{g \in P_{m_s}} \left\{(X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \sum_{i=1}^n \log(m_{s,t}(X_i)) \geq \sum_{i=1}^n \log(g(X_i))\right\} \bigcap \\ &\bigcap_{j=1}^{t-1} \left\{(X_1, \ldots, X_n)^\top \in B^n \,\middle|\, \sum_{i=1}^n \log(m_{s,t}(X_i)) > \sum_{i=1}^n \log(m_{s,j}(X_i))\right\} \bigcap \\ &\bigcap \{\Upsilon_2 = m_s\}. \end{aligned} \tag{2.48}$$

By a similar argument to the measurability of Υ_2 it follows that Υ_3 is also measurable. As such, given the recursive construction of the finite sequence $\Upsilon := (\Upsilon_k)_{k=1}^{\overline{J}}$ via our maximal packing set tree traversal, this pattern inductively repeats for each $k \in \{4, \ldots, \overline{J}\}$. Since $\nu^*(\mathbf{X}) := \Upsilon_{\overline{J}}$, this implies the measurability of $\nu^*(\mathbf{X})$, as required. $\square$

Remark 2.37. We note that the arguments in the proof above hold, even if the cardinality of the set of children densities at any iteration were at most countable (not *just* finite). That is, (2.46) would still return a measureable set even if $|P_{\Upsilon_1}| = \infty$, since Borel measurability is preserved over countable intersections and unions. The packing sets in our construction are necessarily at most countable, since all of the subsets of $\mathcal{F}$ we consider are separable in the L_2-metric (i.e., contain a countably dense subset). This follows from Lemma 2.34.

2.B.7 Proof of Theorem 2.19

We begin with a useful result, which will enable us to construct upper bounds for estimator $\nu^*(\mathbf{X})$.

Lemma 2.38. *The finite sequence $\Upsilon := (\Upsilon_k)_{k=1}^{\overline{J}}$, as defined in the construction of our estimator $\nu^*(\mathbf{X})$, satisfies*

$$\|\Upsilon_J - \Upsilon_{J'}\|_2 \leq \frac{d}{2^{J'-2}}, \tag{2.49}$$

for each pair of positive integers $J' < J$.

Proof of Lemma 2.38. Let $\Upsilon_{J'}, \Upsilon_J \in \Upsilon$, for any positive integers $J > J' \geq 1$. We then have

$$\|\Upsilon_J - \Upsilon_{J'}\|_2 \leq \sum_{i=J'}^{J-1} \|\Upsilon_{i+1} - \Upsilon_i\|_2 \leq \sum_{i=J'}^{J-1} \frac{d}{2^{i-1}} \leq \frac{d}{2^{J'-2}}. \tag{2.50}$$

As required. □

Lemma 2.39 (Telescoping sum of conditional probabilities)**.** *Let $n \geq 2$ be a fixed integer, and $\{A_1, A_2, \ldots, A_n\}$ denote events on a common probability space, with $\mathbb{P}(A_j^{\mathsf{c}}) > 0$ for each $j \geq 1$. We then have*

$$\mathbb{P}(A_n) \leq \sum_{j=n}^{2} \mathbb{P}(A_j \mid A_{j-1}^{\mathsf{c}}) + \mathbb{P}(A_1). \tag{2.51}$$

Proof of Lemma 2.39. We will prove this by induction on $n \geq 2$. We check the induction base case for $n = 2$. We first observe that

$$A_2 \subseteq A_1 \cup A_2 = (A_2 \cap A_1^{\mathsf{c}}) \sqcup A_1, \tag{2.52}$$

where the latter set is a *disjoint* union. It then follows that

$$\begin{aligned}
\mathbb{P}(A_2) &\leq \mathbb{P}\left(A_2 \cap A_1^{\mathsf{c}}\right) + \mathbb{P}(A_1) && \text{(by monotonicity of } \mathbb{P} \text{ applied to (2.52))} \\
&\leq \frac{\mathbb{P}\left(A_2 \cap A_1^{\mathsf{c}}\right)}{\mathbb{P}(A_1^{\mathsf{c}})} + \mathbb{P}(A_1) && \text{(since } \mathbb{P}(A_1^{\mathsf{c}}) \in (0,1]\text{, by assumption)} \\
&=: \mathbb{P}(A_2 \mid A_1^{\mathsf{c}}) + \mathbb{P}(A_1),
\end{aligned}$$

which proves the base case for $n = 2$. Now, by induction assume the result is true for each integer $n = k > 2$. We then have for $n = k + 1$ that:

$$\begin{aligned}
\mathbb{P}(A_{k+1}) &\leq \mathbb{P}(A_{k+1} \mid A_k^{\mathsf{c}}) + \mathbb{P}(A_k) && \text{(using induction base case)} \\
&\leq \mathbb{P}(A_{k+1} \mid A_k^{\mathsf{c}}) + \sum_{j=k}^{2} \mathbb{P}(A_j \mid A_{j-1}^{\mathsf{c}}) + \mathbb{P}(A_1) \\
& && \text{(using induction hypothesis)} \\
&= \sum_{j=k+1}^{2} \mathbb{P}(A_j \mid A_{j-1}^{\mathsf{c}}) + \mathbb{P}(A_1),
\end{aligned} \tag{2.53}$$

as required. So the result is true for $n = k + 1$, and thus by induction holds for each integer $n \geq 2$. □

Theorem 2.19 (Upper bound rate for the multistage sieve MLE $\nu^*(\mathbf{X})$)**.** *Let, $\nu^*(\mathbf{X}) = \Upsilon_{\overline{J}}$ be the output of the multistage sieve MLE which is run for*

$\overline{J} \in \mathbb{N}$ steps. Here $\overline{J}$ is defined as the maximal integer $J \in \mathbb{N}$, such that $\varepsilon_J := \frac{\sqrt{L(\alpha,\beta,c/2-1)}d}{2^{(J-2)}c}$ satisfies[9]

$$n\varepsilon_J^2 > 2\log M^{\mathsf{loc}}_{\mathcal{F}}\left(\varepsilon_J \frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}}, c\right) \vee \log 2, \tag{2.11}$$

or $\overline{J} = 1$ if no such J exists. Then

$$\mathbb{E}\|\nu^*(\mathbf{X}) - f\|_2^2 \leq \bar{C}\varepsilon^{*2},$$

for some universal constant $\bar{C}$, and where $\varepsilon^* := \varepsilon_{\overline{J}}$. We remind the reader that $c := 2(C+1)$ is the constant from the definition of local metric entropy, which is assumed to be sufficiently large. Here C is assumed to satisfy (2.9), and $L(\alpha,\beta,C)$ is defined as per (2.10).

Proof of Theorem 2.19. Combining the results of Lemma 2.14 (with $c := 2(C+1)$ where c is the constant from the definition of local packing entropy) and Lemma 2.15 we conclude that for each $j \in \{2,\ldots,J\}$ we have

$$\begin{aligned}
&\mathbb{P}\left(\|f - \Upsilon_j\|_2 > \frac{d}{2^{j-1}} \,\middle|\, \|f - \Upsilon_{j-1}\|_2 \leq \frac{d}{2^{j-2}}, \Upsilon_{j-1}\right) \\
&\leq |P_{\Upsilon_{j-1}}| \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(j-1)}(C+1)^2}\right) \qquad (2.54)\\
&\leq M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right) \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(j-1)}(C+1)^2}\right) \qquad (2.55)
\end{aligned}$$

where P_{Υ_j} are the maximal packing sets described in the construction of $\nu^*(\mathbf{X})$. Crucially, we observe that the RHS of (2.55) does not depend on the conditioned random variables, i.e., Υ_{j-1}, for each $j \in \{2,\ldots,J\}$ hence we can drop Υ_{j-1} from the conditioning. Now let denote $A_j := \{\|f - \Upsilon_j\|_2 > \frac{d}{2^{j-1}}\}$, for each integer $j \geq 1$. Then we can proceed by working with the unconditional events A_j in (2.55).

Moreover, we then have that $A^{\mathsf{c}}_{j-1} := \{\|f - \Upsilon_{j-1}\|_2 \leq \frac{d}{2^{j-2}}\}$ for each integer $j \geq 2$. In particular $\mathbb{P}(A^{\mathsf{c}}_1) = \{\|f - \Upsilon_1\|_2 \leq d\} = 1$, since $f, \Upsilon_1 \in \mathcal{F}$, so indeed $\|f - \Upsilon_1\|_2 \leq \operatorname{diam}_2(\mathcal{F}) =: d$ almost surely. By aligning our notation directly

[9]Observe that by the definition of $\varepsilon_{\overline{J}}$ and (2.11) we have that all packing sets used in the construction of the estimator must be finite, even though we are not assuming that the set $\mathcal{F}$ is totally bounded.

with Lemma 2.39, we can apply the telescoping bound to $\mathbb{P}(A_j)$ as follows

$$\begin{aligned}
\mathbb{P}(A_J) &\coloneqq \mathbb{P}\left(\|f - \Upsilon_J\|_2 > \frac{d}{2^{J-1}}\right) && \text{(by definition)} \\
&\leq M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right) \sum_{j=1}^{J-1} \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2j}(C+1)^2}\right) && \text{(per (2.55))} \\
&\leq M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right) a(1 + a^{4-1} + a^{16-1} + \ldots)\mathbb{1}(J > 1) && (2.56) \\
&\leq M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right) a(1 + a + a^2 + \ldots)\mathbb{1}(J > 1) \\
&\leq M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right) \frac{a}{1-a}\mathbb{1}(J > 1), && (2.57)
\end{aligned}$$

where for brevity in (2.56) we denote

$$a \coloneqq \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(J-1)}(C+1)^2}\right).$$

Since C is assumed to satisfy (2.9), and $L(\alpha,\beta,C)$ is defined as per (2.10), it follows that $a < 1$. Note here that the above bound (2.57) holds, provided that $\mathbb{P}(A_j^{\mathsf{c}}) > 0$ for $j < J$ as required by Lemma 2.39. Suppose that the RHS of (2.57) is strictly smaller than 1. In that case for all j, $\mathbb{P}(A_j^{\mathsf{c}}) > 0$ since bound (2.57) holds inductively for all $\mathbb{P}(A_j)$ for $j \leq J$. On the other hand, if the RHS of (2.57) is ≥ 1 then (2.57) trivially holds. In both cases we conclude that (2.57) holds.

If one sets $\varepsilon_J \coloneqq \frac{\sqrt{L(\alpha,\beta,C)}\,d}{2^{(J-1)}(C+1)}$, we have that if

$$n\varepsilon_J^2 > 2\log M^{\mathsf{loc}}_{\mathcal{F}}\left(\varepsilon_J \frac{2(C+1)}{\sqrt{L(\alpha,\beta,C)}}, c\right) = 2\log M^{\mathsf{loc}}_{\mathcal{F}}\left(\frac{d}{2^{J-2}}, c\right),$$

and $a \coloneqq \exp(-n\varepsilon_J^2) < 1/2 \iff n\varepsilon_J^2 > \log 2$, the above probability in (2.57) will be bounded from above by $2\exp(-n\varepsilon_J^2/2)$. This condition is implied when

$$n\varepsilon_J^2 > 2\log M^{\mathsf{loc}}_{\mathcal{F}}\left(\varepsilon_J \frac{2(C+1)}{\sqrt{L(\alpha,\beta,C)}}, c\right) \vee \log 2. \tag{2.58}$$

We now have

$$\|\nu^*_{\overline{J}} - f\|_2 \leq \|\Upsilon_{\overline{J}} - \Upsilon_J\|_2 + \|\Upsilon_J - f\|_2 \leq 3\varepsilon_J \frac{C+1}{\sqrt{L(\alpha,\beta,C)}}, \tag{2.59}$$

with probability at least $1-2\exp(-n\varepsilon_J^2/2)$ which holds for all J satisfying (2.58) (including $\overline{J}$). Here we want to clarify that the last inequality in (2.59) follows from the fact that $\|\Upsilon_{\overline{J}} - \Upsilon_J\|_2 \leq d/2^{J-2}$, as seen when we verified that Υ forms a Cauchy sequence in Lemma 2.38 (and since $\overline{J} \geq J$). Let J^* be selected as the maximum integer J such that (2.58) holds, or otherwise if such J does not exist $J^* = 1$, i.e. $J^* \equiv \overline{J}$. Let $\eta = 3\frac{C+1}{\sqrt{L(\alpha,\beta,C)}}$, $\underline{C} = 2$ and $C' = 1/2$. We have established that the following bound holds

$$\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > \eta\varepsilon_J) \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J > 1) \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J^* > 1),$$

for all $1 \leq J \leq J^*$, where this bound also holds in the case when $J^* = 1$ by exception. Observe that we can extend this bound to all $J \in \mathbb{Z}$ and $J \leq J^*$, since for $J < 1$ we have $\eta\varepsilon_J \geq 6d$ and so

$$\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > \eta\varepsilon_J) \leq 0 \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J^* > 1).$$

We conclude that

$$\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > \eta\varepsilon_J) \leq 0 \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J^* > 1),$$

for any $J \leq J^*$. Now for any $\varepsilon_{J-1} > x \geq \varepsilon_J$ for $J \leq J^*$ we have that

$$\begin{aligned}
\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > 2\eta x) &\leq \mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > \eta\varepsilon_{J-1}) \\
&\leq \underline{C}\exp(-C'n\varepsilon_{J-1}^2)\mathbb{1}(J^* > 1) \\
&\leq \underline{C}\exp(-C'nx^2)\mathbb{1}(J^* > 1),
\end{aligned}$$

where the last inequality follows due to the fact that the map $x \mapsto \underline{C}\exp(-C'nx^2)$ is monotonically decreasing for positive reals. We will now integrate the tail bound:

$$\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > 2\eta x) \leq \underline{C}\exp(-C'nx^2)\mathbb{1}(J^* > 1), \tag{2.60}$$

which holds true for $x \geq \varepsilon^*$, where $\varepsilon_J = \frac{\sqrt{L(\alpha,\beta,C)}\,d}{2^{(J-1)}(C+1)}$, always (since even if $J^* = 1$ by exception, this bound is still valid). We then have

$$\begin{aligned}
\mathbb{E}\|f - \nu^*_{\overline{J}}\|_2^2 &= \int_0^\infty 2x\mathbb{P}(\|f - \nu^*_{\overline{J}}\|_2 > x)\,\mathrm{d}x \\
&\leq C'''\varepsilon^{*2} + \int_{2\eta\varepsilon^*}^\infty 2x\underline{C}\exp(-C''nx^2)\mathbb{1}(J^* > 1)\,\mathrm{d}x \\
&= C'''\varepsilon^{*2} + C''''n^{-1}\exp(-C'''''n\varepsilon^{*2})\mathbb{1}(J^* > 1).
\end{aligned}$$

Now $n\varepsilon^{*2}$ is bigger than a constant (i.e., $\log 2$) otherwise $J^* = 1$. Hence, the above is smaller than $\bar{C}\varepsilon^{*2}$ for some absolute constant $\bar{C}$. $\square$

2.B.8 Proof of Theorem 2.20

Theorem 2.20 (Minimax rate)**.** *Define* $\varepsilon^* \coloneqq \sup\{\varepsilon : n\varepsilon^2 \le \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)\}$*, where* c *in the definition of local metric entropy is a sufficiently large absolute constant. Then the minimax rate is given by* $\varepsilon^{*2} \wedge d^2$ *up to absolute constant factors.*

Proof of Theorem 2.20. First suppose that ε^* satisfies $n\varepsilon^{*2} > 4\log 2$. Then for $\delta^* \coloneqq \varepsilon^*/\sqrt{4(1/\alpha \vee 1)}$ we have $\log M^{\mathsf{loc}}_{\mathcal{F}}(\delta^*, c) \ge \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon^*, c) \ge n\varepsilon^{*2}/2 + n\varepsilon^{*2}/2 > 2n\delta^{*2}/\alpha + 2\log 2$ and so this implies the sufficient condition for the lower bound per Lemma 2.11. Let $\eta \coloneqq \frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}} \wedge 1$. For a constant C such that $C\eta > 1$, we have

$$\begin{aligned} C^2 n\varepsilon^{*2} &\ge 1/\eta^2 \log M^{\mathsf{loc}}_{\mathcal{F}}(C\eta\varepsilon^*, c) \ge \log M^{\mathsf{loc}}_{\mathcal{F}}(C\eta\varepsilon^*, c) \\ &\ge \log M^{\mathsf{loc}}_{\mathcal{F}}\left(C\varepsilon^* \frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}}, c\right) \end{aligned}$$

Setting $\delta \coloneqq C\varepsilon^*$ we obtain that

$$n\delta^2 \ge \log M^{\mathsf{loc}}_{\mathcal{F}}\left(\delta \frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}}, c\right).$$

In addition since $C > 1$, δ satisfies (2.11) (taking into account that $n\varepsilon^{*2} > 4\log 2$, which implies $n\delta^2 \ge 4\log 2C^2 > \log 2$). We note that the map $0 < x \mapsto nx^2 - \log M^{\mathsf{loc}}_{\mathcal{F}}\left(x\frac{c}{\sqrt{L(\alpha,\beta,c/2-1)}}, c\right) \vee \log 2$ is non-decreasing by Lemma 2.15. Now, with ε_{J^*} defined as per Theorem 2.19, this implies that $\delta \ge \varepsilon_{J^*}/2$. This shows that the rate in this case is of the order ε^{*2}.

Next, suppose that ε^* defined by $\sup\{\varepsilon : n\varepsilon^2 \le \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)\}$ satisfies $n\varepsilon^{*2} \le 4\log 2$. For $2\varepsilon^*$, we have $16\log 2 \ge 4\varepsilon^{*2}n \ge \log M^{\mathsf{loc}}_{\mathcal{F}}(2\varepsilon^*, c)$. If c in the definition of local packing is large enough, we could put points in the diameter of the ball with radius $2\varepsilon^*$ such that the packing set has more than $\exp(16\log 2)$ many points. But that implies that the set $\mathcal{F}$ is entirely inside a ball of radius $\sqrt{16\log 2}\, n^{-1/2}$ (as $\varepsilon^{*2} \le (4\log 2)n^{-1}$). To see the latter, one can take the midpoint of the line segment connecting the endpoints of a diameter of $\mathcal{F}$ and position a ball of radius $2\varepsilon^*$ there. In such a case, for the lower bound, we could pick ε to be proportional to the diameter of the set (with a small proportionality constant). That will ensure that $\varepsilon\sqrt{n}$ is upper bounded by some constant (as $2\sqrt{(16\log 2)}\, n^{-1/2}$ is bigger than the diameter), and at the same time $\log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)$ can be made bigger than a constant (provided that c in the definition of a local packing is large enough) – by taking θ (where θ is the center of the localized set $B_2(\theta, \varepsilon) \cap \mathcal{F})$ to be the midpoint of a diameter of the set K and then placing equispaced points on the diameter. Hence, the diameter of the set is a lower bound (up to constant factors) in this case, which is of course always an upper bound too (up to constant factors). So we conclude that either for ε^* defined by $\sup\{\varepsilon : \varepsilon^2 n \leq \log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)\}$ satisfies $\varepsilon^{*2} n > 4\log 2$ or the lower and upper bounds are of the order of the diameter of the set. In summary the rate is given by the $\varepsilon^{*2} \wedge d^2$. This is true since in the second case, $4\varepsilon^*$ is bigger than the diameter of the set. ☐

2.B.9 Proof of Proposition 2.22

Proposition 2.22 (Extending results to $\mathcal{F}_B^{[0,\beta]}$)**.** *Let $\mathcal{F} \subset \mathcal{F}_B^{[0,\beta]}$ be a convex class of densities, with at least one $f_\alpha \in \mathcal{F}$ that is α-lower bounded, with $\alpha > 0$. Then the minimax rate in the squared L_2-metric is $\varepsilon^{*2} \wedge d^2$, where $\varepsilon^* := \sup\{\varepsilon : n\varepsilon^2 \leq \log M_{\mathcal{F}}^{\mathsf{loc}}(\varepsilon, c)\}$.*

Proof of Proposition 2.22. We argue this as follows. Let $f_\alpha \in \mathcal{F}$, which is lower bounded by some $\alpha > 0$. Now consider the following set of f_α-mixture densities, i.e., $\mathcal{F}' = \{(1/2)f_\alpha + (1/2)f : f \in \mathcal{F}\} \subset \mathcal{F}$. By construction, all densities in $\mathcal{F}'$ are thus lower bounded by $\alpha/2$, i.e. $\mathcal{F}' \subset \mathcal{F}_B^{[\alpha/2,\beta]}$. Moreover, $\mathcal{F}'$ forms a convex density class. Hence, the minimax rate would be given by $\varepsilon^2 \wedge \operatorname{diam}_2(\mathcal{F}')^2$ where $\varepsilon = \sup\{\varepsilon : n\varepsilon^2 \leq \log M_{\mathcal{F}'}^{\mathsf{loc}}(\varepsilon, c)\}$. We can artificially create variables from the class $\mathcal{F}'$ by randomizing X_i as follows

$$Z_i = \begin{cases} T_i \overset{\text{i.i.d.}}{\sim} f_\alpha & \text{with probability } 1/2, \\ X_i & \text{with probability } 1/2. \end{cases}$$

Then let $\widehat{f}$ be our estimator of $(1/2)f_\alpha + (1/2)f$. We know:

$$\mathbb{E}_Z \|\widehat{f} - ((1/2)f_\alpha + (1/2)f)\|_2^2 \lesssim \varepsilon^2 \wedge \operatorname{diam}(\mathcal{F}')^2,$$

so that

$$\mathbb{E}_{X,T,V} \|(2\widehat{f} - f_\alpha) - f\|_2^2 \lesssim 4\varepsilon^2 \wedge \operatorname{diam}(\mathcal{F}')^2,$$

where $T = (T_1, \ldots, T_n)$ and $V = (V_1, \ldots, V_n)$ are the values of the coin flips in the definition of Z_i. Hence, $\mathbb{E}_{T,V} 2\widehat{f} - f_\alpha$ achieves the same rate for f since by Jensen's inequality

$$\mathbb{E}_Y \|\mathbb{E}_{T,V}(2\widehat{f} - f_\alpha) - f\|_2^2 \leq \mathbb{E}_{Y,T,V} \|(2\widehat{f} - f_\alpha) - f\|_2^2 \lesssim 4\varepsilon^2 \wedge \operatorname{diam}(\mathcal{F}')^2.$$

Moreover, note that since $\widehat{f} \in \mathcal{F}'$ for each of value of T, V we have $\mathbb{E}_{T,V} 2\widehat{f} - f_\alpha \in \mathcal{F}$. Thus, the upper bound is the same for the two sets. On the other hand since $\mathcal{F}' \subset \mathcal{F}$ the lower bound is also of the same rate. Finally, observe that $\log M^{\mathsf{loc}}_{\mathcal{F}'}(\varepsilon, c) = \log M^{\mathsf{loc}}_{\mathcal{F}}(2\varepsilon, c)$ so that the order of $\varepsilon^* = \sup\{\varepsilon : n\varepsilon^2 \leq \log M^{\mathsf{loc}}_{\mathcal{F}'}(\varepsilon, c)\}$ is the same as that of the equation $\varepsilon^* = \sup\{\varepsilon : n\varepsilon^2 \leq \log M^{\mathsf{loc}}_{\mathcal{F}}(\varepsilon, c)\}$. In addition, it is also clear that $2\,\mathrm{diam}_2(\mathcal{F}') = \mathrm{diam}_2(\mathcal{F})$. $\square$

2.B.10 Proof of Theorem 2.24

We first prove the following simple lemma.

Lemma 2.40. *Suppose $\nu, \mu \in \mathcal{F}$ are two densities such that $\|\nu - \nu\|_2 \leq \delta$. If $\delta \leq \varepsilon$ then have $M(\nu, \varepsilon, c) \leq M(\mu, 2\varepsilon, 2c)$.*

Proof of Lemma 2.40. It suffices to show that $B(\nu, \varepsilon) \subseteq B(\mu, 2\varepsilon)$. For any $x \in B(\nu, \varepsilon)$ we have $\|x - \nu\|_2 \leq \varepsilon$, and hence by the triangle inequality we obtain

$$\|x - \mu\|_2 \leq \|x - \nu\|_2 + \|\nu - \mu\|_2 \leq \varepsilon + \delta \leq 2\varepsilon,$$

which completes the proof. $\square$

Theorem 2.24 (Adaptive upper bound rate for the multistage sieve MLE $\nu^*(\mathbf{X})$)**.** *Let, $\nu^*(\mathbf{X}) = \Upsilon_{\overline{J}}$ be the output of the multistage sieve MLE which is run for $\overline{J}$ iterations where $\overline{J}$ is defined as the maximal solution to*

$$n\varepsilon_J^2 > 2 \inf_{f \in \mathcal{F}} M^{\mathsf{adloc}}_{\mathcal{F}}\left(f, 2\varepsilon_J \frac{c}{\sqrt{L(\alpha, \beta, c/2 - 1)}}, 2c\right) \vee \log 2,$$

where $\varepsilon_J := \frac{\sqrt{L(\alpha,\beta,c/2-1)}\,d}{2^{(J-2)}c}$ and $\overline{J} = 1$ if no such J exists[10]*. Let J^* be defined as the maximal integer $J \in \mathbb{N}$, such that $\varepsilon_J := \frac{\sqrt{L(\alpha,\beta,c/2-1)}\,d}{2^{(J-2)}c}$ such that*[11]*.,*

$$n\varepsilon_J^2 > 2 M^{\mathsf{adloc}}_{\mathcal{F}}\left(f, 2\varepsilon_J \frac{c}{\sqrt{L(\alpha, \beta, c/2 - 1)}}, 2c\right) \vee \log 2, \tag{2.12}$$

and $J^* = 1$ if no such J exists. Then

$$\mathbb{E}\|\nu^*(\mathbf{X}) - f\|_2^2 \leq \bar{C}\varepsilon^{*2},$$

[10] Note that running the estimator with $\overline{J}$ many steps, may result into having non-finite packing sets — that is not an issue however.

[11] Observe that by the definition of $\varepsilon_{\overline{J}}$ and (2.11) we have that *some* packing sets used in the construction of the *adaptive* estimator may not be finite, but will be at most countable. This follows from the L_2-separability of $\mathcal{F}$ and is formalized in Lemma 2.34 in the appendix. We note that the measurability of the adaptive estimator still holds as per Proposition 2.18 in this case.

for some universal constant $\bar{C}$, and where $\varepsilon^* \coloneqq \varepsilon_{J^*}$. We remind the reader that $c \coloneqq 2(C+1)$ is the constant from the definition of local metric entropy, which is assumed to be sufficiently large. Here C is assumed to satisfy (2.9)*, and $L(\alpha, \beta, C)$ is defined as per* (2.10)*.*

Proof of Theorem 2.24. Combining the results of Lemma 2.14 (with $c \coloneqq 2(C+1)$ where c is the constant from the definition of local packing entropy) and Lemma 2.15 we conclude that for each $j \in \{2, \dots, J\}$ we have

$$\begin{aligned}
&\mathbb{P}\left(\|f - \Upsilon_j\|_2 > \frac{d}{2^{j-1}} \,\middle|\, \|f - \Upsilon_{j-1}\|_2 \le \frac{d}{2^{j-2}}, \Upsilon_{j-1}\right) \\
&\le |P_{\Upsilon_{j-1}}| \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(j-1)}(C+1)^2}\right) && (2.61)\\
&\le M(f, \frac{d}{2^{j-3}}, 2c) \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(j-1)}(C+1)^2}\right) && (2.62)\\
&\le M(f, \frac{d}{2^{J-3}}, 2c) \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(j-1)}(C+1)^2}\right) && (2.63)
\end{aligned}$$

where P_{Υ_j} are the maximal packing sets described in the construction of $\nu^*(\mathbf{X})$. Furthermore, inequality (2.62) follows from Lemma 2.40. Crucially, we observe that the RHS of (2.63) does not depend on the conditioned random variables, i.e., Υ_{j-1}, for each $j \in \{2, \dots, J\}$ hence we can drop Υ_{j-1} from the conditioning. Now let denote $A_j \coloneqq \{\|f - \Upsilon_j\|_2 > \frac{d}{2^{j-1}}\}$, for each integer $j \ge 1$. Then we can proceed by working with the unconditional events A_j in (2.55).

Moreover, we then have that $A_{j-1}^{\complement} \coloneqq \{\|f - \Upsilon_{j-1}\|_2 \le \frac{d}{2^{j-2}}\}$ for each integer $j \ge 2$. In particular $\mathbb{P}(A_1^{\complement}) = \{\|f - \Upsilon_1\|_2 \le d\} = 1$, since $f, \Upsilon_1 \in \mathcal{F}$, so indeed $\|f - \Upsilon_1\|_2 \le \operatorname{diam}_2(\mathcal{F}) =: d$ almost surely. By aligning our notation directly with Lemma 2.39, we can apply the telescoping bound to $\mathbb{P}(A_j)$ as follows

$$\begin{aligned}
\mathbb{P}(A_J) &\coloneqq \mathbb{P}\left(\|f - \Upsilon_J\|_2 > \frac{d}{2^{J-1}}\right) && \text{(by definition)}\\
&\le M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \frac{d}{2^{J-3}}, 2c\right) \sum_{j=1}^{J-1} \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2j}(C+1)^2}\right) && \text{(per (2.55))}\\
&\le M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \frac{d}{2^{J-3}}, 2c\right) a(1 + a^{4-1} + a^{16-1} + \dots)\mathbb{1}(J > 1) && (2.64)\\
&\le M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \frac{d}{2^{J-3}}, 2c\right) a(1 + a + a^2 + \dots)\mathbb{1}(J > 1)\\
&\le M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \frac{d}{2^{J-3}}, 2c\right) \frac{a}{1-a}\mathbb{1}(J > 1), && (2.65)
\end{aligned}$$

where for brevity in (2.64) we denote

$$a := \exp\left(-\frac{nL(\alpha,\beta,C)d^2}{2^{2(J-1)}(C+1)^2}\right).$$

Since C is assumed to satisfy (2.9), and $L(\alpha,\beta,C)$ is defined as per (2.10), it follows that $a < 1$. Note here that the above bound (2.65) holds, provided that $\mathbb{P}(A_j^{\mathsf{c}}) > 0$ for $j < J$ as required by Lemma 2.39. Suppose that the RHS of (2.65) is strictly smaller than 1. In that case for all j, $\mathbb{P}(A_j^{\mathsf{c}}) > 0$ since bound (2.65) holds inductively for all $\mathbb{P}(A_j)$ for $j \leq J$. On the other hand, if the RHS of (2.65) is ≥ 1 then (2.65) trivially holds. In both cases we conclude that (2.65) holds.

If one sets $\varepsilon_J := \frac{\sqrt{L(\alpha,\beta,C)}\,d}{2^{(J-1)}(C+1)}$, we have that if

$$n\varepsilon_J^2 > 2M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \varepsilon_J\frac{4(C+1)}{\sqrt{L(\alpha,\beta,C)}}, 2c\right)$$

and $a := \exp(-n\varepsilon_J^2) < 1/2 \iff n\varepsilon_J^2 > \log 2$, the above probability in (2.65) will be bounded from above by $2\exp(-n\varepsilon_J^2/2)$. This condition is implied when

$$n\varepsilon_J^2 > 2M_{\mathcal{F}}^{\mathsf{adloc}}\left(f, \varepsilon_J\frac{4(C+1)}{\sqrt{L(\alpha,\beta,C)}}, 2c\right) \vee \log 2. \tag{2.66}$$

We now have

$$\|\nu_{\overline{J}}^* - f\|_2 \leq \|\Upsilon_{\overline{J}} - \Upsilon_J\|_2 + \|\Upsilon_J - f\|_2 \leq 3\varepsilon_J\frac{C+1}{\sqrt{L(\alpha,\beta,C)}}, \tag{2.67}$$

with probability at least $1 - 2\exp(-n\varepsilon_J^2/2)$ which holds for all J satisfying (2.66). Here we want to clarify that the last inequality in (2.67) follows from the fact that $\|\Upsilon_{\overline{J}} - \Upsilon_J\|_2 \leq d/2^{J-2}$, as per Lemma 2.38 (and since $\overline{J} \geq J$). Let J^* be selected as the maximum integer J such that (2.66) holds, or otherwise if such J does not exist $J^* = 1$. Let $\eta = 3\frac{C+1}{\sqrt{L(\alpha,\beta,C)}}$, $\underline{C} = 2$ and $C' = 1/2$. Observe that by the definition of J^* it follows that all packing sets encountered prior J^*, will have been finite packing sets. We have established that the following bound holds

$$\mathbb{P}(\|f - \nu_{\overline{J}}^*\|_2 > \eta\varepsilon_J) \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J > 1) \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J^* > 1),$$

for all $1 \leq J \leq J^*$, where this bound also holds in the case when $J^* = 1$ by exception. Observe that we can extend this bound to all $J \in \mathbb{Z}$ and $J \leq J^*$, since for $J < 1$ we have $\eta\varepsilon_J \geq 6d$ and so

$$\mathbb{P}(\|f - \nu_{\overline{J}}^*\|_2 > \eta\varepsilon_J) \leq 0 \leq \underline{C}\exp(-C'n\varepsilon_J^2)\mathbb{1}(J^* > 1).$$

We conclude that

$$\mathbb{P}(\|f - \nu_{\bar{J}}^*\|_2 > \eta\varepsilon_J) \leq 0 \leq \underline{C}\exp(-C' n\varepsilon_J^2)\mathbb{1}(J^* > 1),$$

for any $J \leq J^*$. Now for any $\varepsilon_{J-1} > x \geq \varepsilon_J$ for $J \leq J^*$ we have that

$$\begin{aligned}\mathbb{P}(\|f - \nu_{\bar{J}}^*\|_2 > 2\eta x) &\leq \mathbb{P}(\|f - \nu_{\bar{J}}^*\|_2 > \eta\varepsilon_{J-1}) \\ &\leq \underline{C}\exp(-C' n\varepsilon_{J-1}^2)\mathbb{1}(J^* > 1) \\ &\leq \underline{C}\exp(-C' n x^2)\mathbb{1}(J^* > 1),\end{aligned}$$

where the last inequality follows due to the fact that the map $x \mapsto \underline{C}\exp(-C' nx^2)$ is monotonically decreasing for positive reals. We will now integrate the tail bound:

$$\mathbb{P}(\|f - \nu_{\bar{J}}^*\|_2 > 2\eta x) \leq \underline{C}\exp(-C' n x^2)\mathbb{1}(J^* > 1), \tag{2.68}$$

which holds true for $x \geq \varepsilon^*$, where $\varepsilon_J = \frac{\sqrt{L(\alpha,\beta,C)}\,d}{2^{(J-1)}(C+1)}$, always (since even if $J^* = 1$ by exception, this bound is still valid). We then have

$$\begin{aligned}\mathbb{E}\|f - \nu_{\bar{J}}^*\|_2^2 &= \int_0^\infty 2x\mathbb{P}(\|f - \nu_{\bar{J}}^*\|_2 > x)\,\mathrm{d}x \\ &\leq C'''\varepsilon^{*2} + \int_{2\eta\varepsilon^*}^\infty 2x\underline{C}\exp(-C'' nx^2)\mathbb{1}(J^* > 1)\,\mathrm{d}x \\ &= C'''\varepsilon^{*2} + C''''n^{-1}\exp(-C''''' n\varepsilon^{*2})\mathbb{1}(J^* > 1).\end{aligned}$$

Now $n\varepsilon^{*2}$ is bigger than a constant (i.e., $\log 2$) otherwise $J^* = 1$. Hence, the above is smaller than $\bar{C}\varepsilon^{*2}$ for some absolute constant $\bar{C}$. □

Remark 2.41 (Early stopping in adaptive estimation). Suppose that in our adaptive estimation, that we traverse the maximal packing set tree construction and encounter a density Υ_i, such that the cardinality of the set of its children densities is countably infinite, i.e., $|P_{\Upsilon_i}| = \infty$. Then we can simply return $\nu^*(\mathbf{X}) = \Upsilon_i$, in such a case. The reason for this is that the index i will be necessarily at least equal to J^* as defined in (2.12), which is what is required for (2.67) to hold.

2.C Proofs of Section 2.3

2.C.1 Formal justification for Example 2.25

Before proving Examples 2.25 to 2.27, we first prove a useful lemma. This lemma will provide a sufficient condition to ensure that L_2-*local* and L_2-*global* metric entropies are of the same order for various forms of the density class $\mathcal{F}$, as specified in our chosen examples.

Lemma 2.42 (Asymptotic order global metric entropy)**.** *Let $\mathcal{F} \subset \mathcal{F}_B^{[\alpha,\beta]}$, such that for any fixed $\eta > 0$, we have $0 < \varepsilon \mapsto \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp \varepsilon^{-1/\eta}$. Then there exists a $c > 0$, such that the following holds*

$$\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \tag{2.69}$$

Proof of Lemma 2.42. We firstly note that (2.69) has the following equivalence

$$\begin{aligned} &\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \\ \iff \exists 0 < k_1 < k_2 \text{ s.t. } &k_1 \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \leq \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \leq k_2 \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \end{aligned} \tag{2.70}$$

In general, for Equation (2.70) we observe that since $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) > 0$, it follows that $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \leq \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c)$. So taking $k_2 = 1$ will always suffice to ensure (2.70) holds. It remains to check that we can also find a $k_1 \in (0,1)$ such that (2.70) also holds. In our case, since $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp \varepsilon^{-1/\eta}$ by assumption, we have that $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) \geq C_1(\varepsilon/c)^{-1/\eta}$ and $\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \leq C_2\varepsilon^{-1/\eta}$ for some universal constants $C_1, C_2 > 0$. It then follows

$$\begin{aligned} \frac{\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c) - \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon)}{\log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon/c)} &\geq 1 - \left(\frac{C_2}{C_1}\right) c^{-\frac{1}{\eta}} \\ &\geq k_1 && \text{(as required, for } k_1 \in (0,1)) \\ &=: 1 - \delta && \text{(for some } \delta \in (0,1)\text{, since } k_1 \in (0,1)) \\ \text{if } c &\geq \left(\frac{C_2}{C_1 \delta}\right)^{\eta}. \end{aligned} \tag{2.71}$$

That is, there exists such a $k_1 \in (0,1)$, if we choose $c \geq \left(\frac{C_2}{C_1\delta}\right)^{\eta}$, for each $\eta > 0$. So indeed (2.69) holds, for the specified class $\mathcal{F}$, as required. □

Example 2.25 (Lipschitz density class $\mathcal{F}$). Let $1 < \Psi < \beta < \infty$, $\max\{1/q - 1/2, 0\} < \gamma \leq 1$, and $1 \leq q \leq \infty$ be fixed constants, and $B \coloneqq [0,1]$. Now, let

$\mathcal{F} \coloneqq \mathsf{Lip}_{\gamma,q}(\Psi)$ denote the space of (γ, q, Ψ)-Lipschitz densities with total variation at most β. That is,

$$\mathsf{Lip}_{\gamma,q}(\Psi) \coloneqq \left\{ f \colon B \to [0, \Psi] \,\middle|\, \|f(x+h) - f(x)\|_q \leq \Psi h^{\gamma}, \|f\|_q \leq \Psi, \int_B f \, \mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{2.15}$$

and $\|f\|_q \coloneqq \left(\int_B |f(x)|^q \, \mathrm{d}\mu\right)^{1/q}$. Note that in (2.15) we have that $x \in B$, and only consider $h > 0$ such that $x + h \in B$, so that the predicate of $\mathsf{Lip}_{\gamma,q}(\Psi)$ is well-defined. Then $\mathsf{Lip}_{\gamma,q}(\Psi)$ is a convex density class, there exists a density $f_\alpha \in \mathsf{Lip}_{\gamma,q}(\Psi)$ that is strictly positively bounded away from 0, and the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{Lip}_{\gamma,q}(\Psi)$ is of the order $n^{-\frac{2\gamma}{2\gamma+1}}$.

Proof of Example 2.25. In order to establish the minimax rate for $\mathsf{Lip}_{\gamma,q}(\Psi)$, we need to show that $\mathsf{Lip}_{\gamma,q}(\Psi)$ is a convex density class, and that there exists a density $f_\alpha \in \mathsf{Lip}_{\gamma,q}(\Psi)$ that is strictly positively bounded away from 0. We can then apply Proposition 2.22. We first verify that $\mathsf{Lip}_{\gamma,q}(\Psi)$ here is a convex density class. To that end, let $f, g \in \mathsf{Lip}_{\gamma,q}(\Psi)$, and let $\kappa \in [0, 1]$, be arbitrary. Then for each $x \in B \coloneqq [0, 1]$, we observe that

$$(\kappa f + (1-\kappa)g)(x) \coloneqq \kappa f(x) + (1-\kappa)g(x) \geq \kappa(0) + (1-\kappa)(0) = 0 \tag{2.72}$$

$$(\kappa f + (1-\kappa)g)(x) \coloneqq \kappa f(x) + (1-\kappa)g(x) \leq \kappa\Psi + (1-\kappa)\Psi = \Psi \tag{2.73}$$

From (2.72) and (2.73), it follows that

$$\kappa f + (1-\kappa)g \colon B \to [0, \Psi]. \tag{2.74}$$

Moreover, since $\int_B f \, \mathrm{d}\mu = \int_B g \, \mathrm{d}\mu = 1$, we have

$$\int_B (\kappa f + (1-\kappa)g) \, \mathrm{d}\mu = \kappa \int_B f \, \mathrm{d}\mu + (1-\kappa) \int_B g \, \mathrm{d}\mu = 1. \tag{2.75}$$

Since $f, g \in \mathsf{Lip}_{\gamma,q}(\Psi)$, we have both $\|f\|_q, \|g\|_q \leq \Psi$. Then by the triangle inequality it follows

$$\|\kappa f + (1-\kappa)g\|_q \leq \|\kappa f\|_q + \|(1-\kappa)g\|_q \leq \kappa\Psi + (1-\kappa)\Psi = \Psi. \tag{2.76}$$

Since f, g are measurable functions, then so is their convex combination, i.e.,

$\kappa f + (1-\kappa)g$. Now we observe

$$\begin{aligned}
&\|(\kappa f + (1-\kappa)g)(x+h) - (\kappa f + (1-\kappa)g)(x)\|_q \\
&= \|\kappa(f(x+h) - f(x)) + (1-\kappa)(g(x+h) - g(x))\|_q \\
&\leq \|\kappa(f(x+h) - f(x))\|_q + \|(1-\kappa)(g(x+h) - g(x))\|_q \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by the triangle inequality.)} \\
&\leq \kappa h^{\gamma} + (1-\kappa)h^{\gamma} \qquad\qquad\qquad\qquad (\text{since } f, g \in \mathsf{Lip}_{\gamma,q}(\Psi)) \\
&= h^{\gamma}, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2.77)
\end{aligned}$$

as required. Combining (2.74),(2.75),(2.76), and (2.77) we have shown that $\kappa f + (1-\kappa)g \in \mathsf{Lip}_{\gamma,q}(\Psi)$. This proves the convexity of $\mathsf{Lip}_{\gamma,q}(\Psi)$, as required.

Now let $f_{\alpha} \sim \mathsf{Unif}[B]$, i.e. $f_{\alpha}(x) \coloneqq \mathbb{I}_{[0,1]}(x)$. Therefore, $\|f_{\alpha}(x+h) - f(x)\|_q = 0 \leq \Psi h^{\gamma}$, for each $x \in B$, and $h > 0$ such that $f_{\alpha}(x+h)$ is defined. Moreover, $\|f_{\alpha}(x)\|_q = 1 < \Psi$, by assumption, for each $1 \leq q \leq \infty$. Now we have that $\int_B f_{\alpha} \,\mathrm{d}\mu = \int_B \mathbb{I}_{[0,1]}(x) \,\mathrm{d}\mu(x) = 1$, and f is measurable since it is a simple function. So indeed we have found $f_{\alpha} \in \mathsf{Lip}_{\gamma,q}(\Psi)$, such that it is α-lower bounded (with $\alpha = 1$).

We now proceed to check that L_2-*global* metric entropy is of the same order as the L_2-*local* metric entropy for $\mathsf{Lip}_{\gamma,q}(\Psi)$. That is, we want to check that (2.69) holds. Here $\mathcal{F} = \mathsf{Lip}_{\gamma,q}(\Psi)$, with $0 < \varepsilon \mapsto \log M_{\mathcal{F}}^{\mathsf{glo}}(\varepsilon) \asymp \varepsilon^{-1/\gamma}$. Thus, we can apply Lemma 2.42 with $\eta \coloneqq \gamma \in (0,1]$ to conclude that indeed (2.69) holds, as required.

Since we have checked all the sufficient conditions in order to apply Proposition 2.22 for $\mathsf{Lip}_{\gamma,q}(\Psi)$, we can obtain the minimax rate of density estimation by solving

$$n\varepsilon^2 \asymp \varepsilon^{-\frac{1}{\gamma}} \iff \varepsilon \asymp n^{-\frac{\gamma}{2\gamma+1}} \iff \varepsilon^2 \asymp n^{-\frac{2\gamma}{2\gamma+1}}. \tag{2.78}$$

So the minimax rate is (up to constants) the order of $n^{-\frac{2\gamma}{2\gamma+1}}$ as required. □

2.C.2 Formal justification for Example 2.26

Example 2.26 (Bounded total variation density class $\mathcal{F}$)**.** Let $1 < \zeta < \beta < \infty$ be a fixed constant, and $B \coloneqq [0,1]$. Now, let $\mathcal{F} \coloneqq \mathsf{BV}_{\zeta}$ denote the space of univariate densities with total variation at most β. That is,

$$\mathsf{BV}_{\zeta} \coloneqq \left\{ f \colon B \to [0,\zeta] \;\middle|\; \|f\|_{\infty} \leq \zeta, V(f) \leq \zeta, \int_B f \,\mathrm{d}\mu = 1, f \text{ measurable} \right\}, \tag{2.16}$$

where we define the total variation of f, i.e., $V(f)$ as

$$V(f) \coloneqq \sup_{\{x_1,\ldots,x_m \,|\, 0 \leq x_1 < \cdots < x_m \leq 1, m \in \mathbb{N}\}} \sum_{i=1}^{m-1} |f(x_{i+1}) - f(x_i)|, \tag{2.17}$$

and $\|f\|_\infty \coloneqq \sup_{x \in B} |f(x)|$. Then the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{BV}_\zeta$ is of the order $n^{-2/3}$.

Proof of Example 2.26. In order to establish the minimax rate for BV_ζ, we need to show that BV_ζ is a convex density class, and that there exists a density $f_\alpha \in \mathsf{BV}_\zeta$ that is strictly positively bounded away from 0. We can then apply Proposition 2.22. We first verify that BV_ζ here is a convex density class. To that end, let $f, g \in \mathsf{BV}_\zeta$, and let $\kappa \in [0, 1]$, be arbitrary. Then for each $x \in B \coloneqq [0, 1]$, it follows by an identical argument to (2.72) and (2.73) that

$$\kappa f + (1 - \kappa) g \colon B \to [0, \zeta]. \tag{2.79}$$

Moreover, since $\int_B f \,\mathrm{d}\mu = \int_B g \,\mathrm{d}\mu = 1$, we have

$$\int_B (\kappa f + (1 - \kappa) g) \,\mathrm{d}\mu = \kappa \int_B f \,\mathrm{d}\mu + (1 - \kappa) \int_B g \,\mathrm{d}\mu = 1. \tag{2.80}$$

Since $f, g \in \mathsf{BV}_\zeta$, we have both $\|f\|_\infty, \|g\|_\infty \le \zeta$. Then by the triangle inequality it follows

$$\|\kappa f + (1 - \kappa) g\|_\infty \le \|\kappa f\|_\infty + \|(1 - \kappa) g\|_\infty \le \kappa \zeta + (1 - \kappa) \zeta = \zeta. \tag{2.81}$$

Since f, g are measurable functions, then so is their convex combination, i.e., $\kappa f + (1 - \kappa) g$. Finally, fix any $m \in \mathbb{N}$, and let $a \le x_1 < \cdots < x_m \le b$ be any fixed partition of B. Now we observe

$$\begin{aligned}
&\sum_{i=1}^{m-1} |(\kappa f + (1 - \kappa) g)(x_{i+1}) - (\kappa f + (1 - \kappa) g)(x_i)| \\
&= \sum_{i=1}^{m-1} |\kappa (f(x_{i+1}) - f(x_i)) + (1 - \kappa)(g(x_{i+1}) - g(x_i))| \\
&\le \kappa \sum_{i=1}^{m-1} |f(x_{i+1}) - f(x_i)| + (1 - \kappa) \sum_{i=1}^{m-1} |g(x_{i+1}) - g(x_i)| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by the triangle inequality.)} \\
&\le \kappa V(f) + (1 - \kappa) V(g) \\
&\le \kappa(\zeta) + (1 - \kappa)(\zeta) \qquad \text{(since } V(f), V(g) \le \zeta \text{, by definition of } \mathsf{BV}_\zeta.) \\
&= \zeta.
\end{aligned}$$

Taking the supremum over all $m \in \mathbb{N}$ and all partitions of length m of B of the LHS sum we obtain:

$$V(\kappa f + (1 - \kappa) g) \le \zeta, \tag{2.82}$$

as required. Combining (2.79),(2.80),(2.81), and (2.82) we have shown that $\kappa f + (1-\kappa)g \in \mathsf{BV}_\zeta$. This proves the convexity of BV_ζ, as required.

Similar to the proof of Example 2.25, we let $f_\alpha \sim \mathsf{Unif}[B]$, i.e. $f_\alpha(x) := \mathbb{I}_{[0,1]}(x)$. Therefore, $\|f\|_\infty = 1 \leq \zeta$ by assumption. Also, $V(f) = 0 < \zeta$, by assumption. Now we have that $\int_B f_\alpha \,\mathrm{d}\mu = \int_B \mathbb{I}_{[0,1]}(x)\,\mathrm{d}\mu(x) = 1$, and f is measurable since it is a simple function. So indeed we have found $f_\alpha \in \mathsf{BV}_\zeta$, such that it is α-lower bounded (with $\alpha = 1$).

We now proceed to check that L_2-*global* metric entropy is of the same order as the L_2-*local* metric entropy for BV_ζ. That is, we want to check that (2.69) holds. Here $\mathcal{F} = \mathsf{BV}_\zeta$, with $0 < \varepsilon \mapsto \log M^{\mathsf{glo}}_{\mathcal{F}}(\varepsilon) \asymp \varepsilon^{-1}$. Thus, we can apply Lemma 2.42 with $\eta := 1$ to conclude that indeed (2.69) holds, as required.

Since we have checked all the sufficient conditions in order to apply Proposition 2.22 for BV_ζ, we can obtain the minimax rate of density estimation by solving

$$n\varepsilon^2 \asymp \varepsilon^{-1} \iff \varepsilon \asymp n^{-\frac{1}{3}} \iff \varepsilon^2 \asymp n^{-\frac{2}{3}}. \tag{2.83}$$

So the minimax rate is (up to constants) the order of $n^{-\frac{2}{3}}$ as required. □

2.C.3 Formal justification for Example 2.27

Example 2.27 (Quadratic functional density class $\mathcal{F}$)**.** Let $0 < \alpha < 1 < \beta < \infty$, and $\gamma > 1$ be fixed constants, with $B := [0,1]$. Now, let $\mathcal{F} := \mathsf{Quad}_\gamma$ denote the space of univariate quadratic functional densities. That is,

$$\mathsf{Quad}_\gamma := \left\{ f\colon B \to [\alpha,\beta] \,\middle|\, \|f''\|_\infty \leq \gamma, \int_B f\,\mathrm{d}\mu = 1, f \text{ measurable} \right\}. \tag{2.18}$$

Then Quad_γ is a convex density class, there exists a density $f_\alpha \in \mathsf{Quad}_\gamma$ that is strictly positively bounded away from 0, and the minimax rate (in the squared L_2-metric) for estimating $f \in \mathsf{Quad}_\gamma$ is of the order $n^{-4/5}$.

Proof of Example 2.27. In order to establish the minimax rate for Quad_γ, we need to show that Quad_γ is a convex density class. We can then apply Proposition 2.22. We first verify that Quad_γ here is a convex density class. To that end, let $f, g \in \mathsf{Quad}_\gamma$, and let $\kappa \in [0,1]$, be arbitrary. Then for each $x \in B := [0,1]$, it follows by an identical argument to (2.72) and (2.73) that

$$\kappa f + (1-\kappa)g\colon B \to [0,\beta]. \tag{2.84}$$

Moreover, since $\int_B f\,\mathrm{d}\mu = \int_B g\,\mathrm{d}\mu = 1$, we have

$$\int_B (\kappa f + (1-\kappa)g)\,\mathrm{d}\mu = \kappa \int_B f\,\mathrm{d}\mu + (1-\kappa)\int_B g\,\mathrm{d}\mu = 1. \tag{2.85}$$

Since f, g are measurable functions, then so is their convex combination, i.e., $\kappa f + (1-\kappa)g$. Now we observe

$$
\begin{aligned}
\|(\kappa f + (1-\kappa)g)''\|_\infty &= \|\kappa f'' + (1-\kappa)g''\|_\infty && \text{(by linearity of 2}^{\text{nd}}\text{ derivative.)} \\
&\leq \|\kappa f''\|_\infty + \|(1-\kappa)g''\|_\infty \\
& && \text{(by the triangle inequality.)} \\
&= \kappa\|f''\|_\infty + (1-\kappa)\|g''\|_\infty \\
&\leq \kappa\gamma + (1-\kappa)\gamma && \text{(since } f, g \in \mathsf{Quad}_\gamma\text{)} \\
&= \gamma, && (2.86)
\end{aligned}
$$

as required. Combining (2.84),(2.85), and (2.86) we have shown that $\kappa f + (1-\kappa)g \in \mathsf{Quad}_\gamma$. This proves the convexity of Quad_γ, as required.

Similar to the proof of Example 2.25, we let $f_\alpha \sim \mathsf{Unif}[B]$, i.e. $f_\alpha(x) \coloneqq \mathbb{I}_{[0,1]}(x)$. Since $\|f''\|_\infty = 0 \leq \gamma$. Here, for the boundary points of $B \coloneqq [0,1]$, we are careful to take all derivatives of $f_\alpha(x)$ at $x = 0$ from the right, and all derivatives from the left at $x = 1$. Now we have that $\int_B f_\alpha \, \mathrm{d}\mu = \int_B \mathbb{I}_{[0,1]}(x) \, \mathrm{d}\mu(x) = 1$, and f is measurable since it is a simple function. So indeed we have found $f_\alpha \in \mathsf{Quad}_\gamma$, such that it is α-lower bounded (with $\alpha = 1$).

We now proceed to check that L_2-*global* metric entropy is of the same order as the L_2-*local* metric entropy for Quad_γ. That is, we want to check that (2.69) holds. Here $\mathcal{F} = \mathsf{Quad}_\gamma$, with $0 < \varepsilon \mapsto \log M^{\mathsf{glo}}_{\mathcal{F}}(\varepsilon) \asymp \varepsilon^{-1/4}$. Thus, we can apply Lemma 2.42 with $\eta \coloneqq 4$ to conclude that indeed (2.69) holds, as required.

Since we have checked all the sufficient conditions in order to apply Proposition 2.22 for Quad_γ, we can obtain the minimax rate of density estimation by solving

$$
n\varepsilon^2 \asymp \varepsilon^{-\frac{1}{2}} \iff \varepsilon \asymp n^{-\frac{2}{5}} \iff \varepsilon^2 \asymp n^{-\frac{4}{5}}. \tag{2.87}
$$

So the minimax rate is (up to constants) the order of $n^{-\frac{4}{5}}$ as required. $\square$

2.C.4 Formal justification for Example 2.28

Example 2.28 (Convex mixture density class $\mathcal{F}$)**.** Let $\mathcal{F} \coloneqq \mathsf{Conv}_k$ where

$$
\mathsf{Conv}_k \coloneqq \left\{ \sum_{i=1}^{k} \alpha_i f_i \;\middle|\; \sum_{i=1}^{k} \alpha_i = 1, \alpha_i \geq 0, f_i \in \mathcal{F}_B^{[\alpha,\beta]} \right\}, \tag{2.19}
$$

for some <u>fixed</u> $k \in \mathbb{N}$ and $f_i \in \mathcal{F}_B^{[\alpha,\beta]}$ for each $i \in [k]$. Further, let $\mathbf{G} = (\mathbf{G}_{ij})_{i,j\in[k]}$ denote the Gram matrix with $\mathbf{G}_{ij} \coloneqq \int_B f_i f_j \mu(\,\mathrm{d}x)$, which we assume is positive definite, i.e., $\mathbf{G} \succ \mathbf{0}$. Then the minimax rate for estimating $f \in \mathsf{Conv}_k$ is bounded from above by $\sqrt{\frac{k}{n}}$ up to absolute constant factors.

Proof of Example 2.28. Let $\mathbf{G} = (\mathbf{G}_{ij})_{i,j\in[k]}$ denote the Gram matrix $\mathbf{G}_{ij} \coloneqq \int_B f_i f_j \mu(\,\mathrm{d}x)$. Then it is simple to see that for some point $\theta \in \mathcal{F}$ which can be represented as the convex combination $\theta = \sum_{i\in[k]} \alpha_i f_i$, the packing set should consist of functions $g_i = \sum_{j\in[k]} \beta_{ij} f_j$ satisfying both

$$(\alpha - \beta_i)^{\mathrm{T}} \mathbf{G} (\alpha - \beta_i) \le \varepsilon^2,$$
$$(\beta_i - \beta_j)^{\mathrm{T}} \mathbf{G} (\beta_i - \beta_j) > \varepsilon^2/c^2, \text{ for } i \ne j,$$

where β_i are vectors from the k-dimensional unit simplex, i.e., $\sum_{j\in[k]} \beta_{ij} = 1$, $\beta_{ij} \ge 0$. Now suppose that $\mathbf{G} \succ 0$. Then upon substituting $\alpha' = \sqrt{\mathbf{G}}\,\alpha$, $\beta_i' = \sqrt{\mathbf{G}}\,\beta_i$ and dropping the simplex requirements on the β we obtain the set

$$\|\alpha' - \beta_i'\| \le \varepsilon$$
$$\|\beta_i' - \beta_j'\| > \varepsilon/c,$$

which is like packing the unit sphere at a distance $1/c$. Hence, the log cardinality of such a packing is always $\lesssim k$ (Wainwright, 2019, see Chapter 5). If k is not allowed to scale with n, we conclude therefore that the minimax rate is upper bounded by $n^{-1/2}$ which is the parametric rate as we would expect. If k is allowed to scale with n the rate is smaller than $\sqrt{\frac{k}{n}}$. $\square$

Three

Adversarial Sign-Corrupted Isotonic Regression

Abstract: Classical univariate isotonic regression involves nonparametric estimation under a monotonicity constraint of the true signal. We consider a variation of this generating process, which we term adversarial sign-corrupted isotonic (ASCI) regression. Under this ASCI setting, the adversary has full access to the true isotonic responses, and is free to sign-corrupt them. Estimating the true monotonic signal given these sign-corrupted responses is a highly challenging task. Notably, the sign-corruptions are designed to violate monotonicity, and possibly induce heavy dependence between the corrupted response terms. In this sense, ASCI regression may be viewed as an adversarial stress test for isotonic regression. Our motivation is driven by understanding whether efficient robust estimation of the monotone signal is feasible under this adversarial setting. We develop `ASCIFIT`, a three-step estimation procedure under the ASCI setting. The `ASCIFIT` procedure is conceptually simple, easy to implement with existing software, and consists of applying the `PAVA` with crucial pre- and post-processing corrections. We formalize this procedure, and demonstrate its theoretical guarantees in the form of sharp high probability upper bounds and minimax lower bounds. We illustrate our findings with detailed simulations.

The work in this chapter was done jointly with Matey Neykov. It is based on a preprint with the title *"Adversarial Sign-Corrupted Isotonic Regression"*.

3.1 Introduction

Isotonic regression is a classically studied nonparametric regression problem in which the underlying signal satisfies a monotonicity constraint. In the univariate case, this regression setup provides a flexible nonparametric generalization to simple linear regression. That is, the underlying signal may be non-linear, but still satisfies monotonicity as in the simple linear model. The classically studied isotonic regression generating process is formally described in Definition 3.1:

Definition 3.1 (Classical isotonic regression)**.** We consider n observations, $\{Y_i \mid i \in [n]\}$, where each observation Y_i is generated from the following model:

$$
\begin{aligned}
Y_i &= \mu_i + \varepsilon_i && (3.1)\\
\text{s.t. } \mu_1 &\leq \mu_2 \leq \ldots \leq \mu_n && (3.2)\\
\text{and } \varepsilon_i &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (3.3)
\end{aligned}
$$

The statistical goal under this classical setup is to estimate the underlying signal vector $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^\top$, subject the monotonicity constraint in Equation (3.2), while σ is an unknown (nuisance) parameter. Throughout this paper we will adopt the convention, without loss of generality, that the signal vector is monotonically increasing (as per Equation (3.2)). Additionally we will assume that all estimation errors are computed under square loss (in Euclidean metric), in high probability.

3.1.1 Adversarial sign-corrupted isotonic (ASCI) regression

Our work in this paper is motivated by a variation of the classical isotonic regression estimation problem, per Definition 3.1. We refer to this newly proposed model as *adversarial sign-corrupted isotonic* (ASCI) *regression*. The generating process for this ASCI estimation problem is formalized in Definition 3.2.

Definition 3.2 (Adversarial sign-corrupted isotonic (ASCI) regression)**.** We consider n observations, $\{R_i \mid i \in [n]\}$, where each observation R_i is generated from the following model:

$$
\begin{aligned}
R_i &= \xi_i(\mu_i + \varepsilon_i) && (3.4)\\
\text{s.t. } 0 < \eta &\leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n && (3.5)\\
\text{and } \varepsilon_i &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (3.6)\\
\text{and } \xi_i &\in \{-1, 1\} && (3.7)
\end{aligned}
$$

Remark 3.3. We note that the constant $\eta > 0$ is known to the observer and adversary as part of the generating process. This means that the true signal is positive, since it is uniformly bounded away from η. It is an artefact of our method but as we will see later in Examples 3.5 and 3.7, highly non-trivial estimation tasks are still contained under this constraint.

Remark 3.4. Throughout this paper we will interchangeably use the terms ASCI *regression*, *setting*, *setup*, *model*, and *generating process* to refer to Definition 3.2.

By comparing Definitions 3.1 and 3.2, this ASCI regression generating process is a partial generalization of the classical isotonic regression. It can be briefly described as follows. Here the classical isotonic regression responses, $\mu_i + \varepsilon_i$ in Equation (3.1), are *sign-corrupted* in a manner chosen by an adversary, as captured by the multiplicative ξ_i terms. Here the $\xi_i \in \{-1, 1\}$ are *sign-corruptions* for the *true* data generating process, i.e., $Y_i \coloneqq \mu_i + \varepsilon_i$. It is important to note that the $\xi_i \in \{-1, 1\}$ for each $i \in [n]$, are chosen given that the adversary has full access to the true responses, i.e., $\{\mu_i + \varepsilon_i \,|\, i \in [n]\}$. As such, Equation (3.1) in the classical isotonic regression setup represents a special case of Equations (3.4) and (3.7) by taking $\xi_i \overset{a.s.}{=} 1$ for each $i \in [n]$. However, we note that in this ASCI setting, in Equation (3.5) the monotonically increasing signal vector $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^\top$ is bounded below by η, which is some fixed and known positive constant. As such this represents a restriction of the classical isotonic regression condition described in Equation (3.2). In summary, ASCI regression represents both a restriction and relaxation of the classical isotonic regression generating process. We will see why the restriction is necessary in this work, but we will later suggest possible ways in which it can be relaxed in future work.

3.1.2 Interesting special cases of ASCI regression

Interestingly, we note that even some special cases of this ASCI regression setup can result in highly non-trivial estimation tasks. Two particular ASCI regression special cases are formalized in Examples 3.5 and 3.7.

Example 3.5 (Two-component Gaussian mixture ASCI regression special case)**.** We consider n observations, $\{R_i \,|\, i \in [n]\}$, where each observation R_i is generated from the following model:

$$R_i = \xi_i \mu_i + \varepsilon_i \tag{3.8}$$

$$\text{s.t. } 0 < \eta \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n \tag{3.9}$$

$$\text{and } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{3.10}$$

$$\text{and } \xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{Rademacher}(p),\ p \in (0, 1),\ \text{and } \xi_i \perp\!\!\!\perp \varepsilon_i \tag{3.11}$$

Remark 3.6. Note that $\xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{Rademacher}(p)$ for each $i \in [n]$, means that $\xi_i = +1$ with probability p, and $\xi_i = -1$ with probability $1 - p$. We note that the model defined in Example 3.5 is a special case of Definition 3.2. This is formally proved in Section 3.B.3.

We note that in the univariate setting, Example 3.5 represents a generalization of the two-component Gaussian mixture model studied in detail in Balakrishnan et al. (2017, Section 3.2.1). Our model generalizes their setting in the sense that we allow a different mean, i.e., μ_i, for each of the n univariate observations. Interestingly, in this more general univariate mixture setting, our proposed `ASCIFIT` estimator (see Section 3.2) provides an efficient alternative to the EM algorithm (Dempster et al., 1977). Such models have extensive applications, e.g., community detection (Royer, 2017; Giraud and Verzelen, 2018).

Example 3.7 (Non-convex ASCI regression special case)**.** We consider n observations, $\{R_i \,|\, i \in [n]\}$, where each observation R_i is generated from the following model:

$$R_i = \gamma_i + \varepsilon_i \tag{3.12}$$

$$\text{s.t. } 0 < \eta \leq |\gamma_1| \leq |\gamma_2| \leq \ldots \leq |\gamma_n| \tag{3.13}$$

$$\text{and } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{3.14}$$

Remark 3.8. Verifying that Example 3.7 is a special case of Definition 3.2 is not a priori obvious, and is formally proved in Section 3.B.4.

In Example 3.7, one can think of this model being generated from the ASCI model as per Definition 3.2. In this special case, the adversary randomly chooses sign-corruptions *independently* of the error terms, i.e., $\xi_i \perp\!\!\!\perp \varepsilon_i$ for each $i \in [n]$. Under this setup, the resulting response term in Equation (3.12) is the same as the classical isotonic regression response, as seen by comparing to Equation (3.1). However, interestingly the adversarial sign-corruption is now absorbed into the revised monotonicity constraint $0 < \eta \leq |\gamma_1| \leq |\gamma_2| \leq \ldots \leq |\gamma_n|$, per Equation (3.13). As a result, this generating process is a highly non-convex constrained estimation problem. In this case `ASCIFIT` will allow one to recover $|\gamma_i|$, where as `PAVA` will not provide *any* information on the $|\gamma_i|$ (or γ_i), given the non-convex constraint.

3.1.3 Motivation and focus of our work

With the ASCI regression setup clearly defined, we turn our attention to describing the focus of our analysis in this work. This is summarized by the following core question of interest:

> ***Core question:*** *Given the adversarial sign-corrupted isotonic* (ASCI) *regression setup in Definition 3.2, can we find a computationally efficient estimator for $\boldsymbol{\mu}$, and demonstrate its precise (non-asymptotic) statistical optimality?*

To the best of our knowledge the ASCI model, and our core question of interest, have not been explicitly studied before in the literature. We note that this ASCI estimation problem is inherently challenging, and thus interesting, for three main reasons, i.e., **Challenge I** – **Challenge III**:

Challenge I (Dependent responses): in this estimation problem the adversary is free to choose the sign-corruption terms ξ_i, *after* observing all samples, possibly resulting in a strong dependence between the original isotonic responses. As such, any ASCI estimator must be able to handle arbitrary dependence structure between the sign-corrupted responses.

Challenge II (Violating signal monotonicity): qualitatively speaking, the sign-corruptions are in a sense 'extreme' in that by selectively changing the sign of the observations the adversary fundamentally 'attacks' the isotonic monotonicity constraint directly. It is this convex monotone constraint which classical isotonic estimators, i.e., `PAVA`, are *designed* to exploit.

Challenge III (Interesting special cases): The ASCI setting contains interesting non-trivial special cases as described in Examples 3.5 and 3.7. Naively applying typical least squares estimation techniques will be unable to provide any relevant information on the estimated quantity of interest.

Given these three formidable challenges posed by ASCI regression, any computationally and statistically efficient estimator here needs to utilize new techniques to exploit the potential non-convex structure in our setting. Our motivations here are thus driven by understanding the robustness of existing isotonic regression estimators under such adversarial settings. Moreover, for the ASCI setting to be worth studying, we wanted ensure practical algorithms for estimation under this adversarial setting, with sharp minimax (worst-case) statistical guarantees, both of which we were able to provide. We thus view the ASCI setting as *stimulating prototype* for more such research into adversarial robustness in isotonic regression.

3.1.4 Prior and related work

As noted, to the best of knowledge our core question of interest, i.e., isotonic regression under the proposed ASCI setup, has not been previously studied. Our work however builds on and utilizes known estimators from the classical isotonic regression literature. As such we limit our prior and related work summary on known risk bounds (and rates) for such isotonic regression estimators, and the efficient algorithms (i.e., the `PAVA`) and practical implementations thereof.

Isotonic regression (classical):

A lively historical overview of isotonic regression estimation from a computational lens is given in de Leeuw et al. (2009, Section 1). In brief, we note that the origins of isotonic regression can be traced back to a number of independently written papers in the 1950s. In particular it was studied by Brunk (1955); Ayer et al. (1955). Such estimators for "ordered parameters" were also analyzed in the series of papers van Eeden (1956, 1957a,b,c) which culminated into a PhD thesis in by the same author (van Eeden, 1958). Shortly thereafter the articles (Bartholomew, 1959a,b) also investigated the related idea of hypothesis testing under monotonicity constraints. We refer the interested reader to the classical comprehensive references Barlow et al. (1972); Robertson et al. (1988), for further reading.

The classical isotonic regression setup per Definition 3.1 under square loss is a convex optimization problem. As such, it has a unique solution, i.e., the Euclidean projection onto the closed convex monotone cone given by the constraint in Equation (3.2). In this case, the non-asymptotic risk bounds for the least squares estimator (LSE) of the monotone parameters μ_i are of the order $n^{-2/3}$ in sample complexity. This convergence rate has been established across a number of papers including van de Geer (1990, 1993); Donoho (1990); Birgé and Massart (1993b); Wang (1996); Meyer and Woodroofe (2000); Zhang (2002); Chatterjee et al. (2015). Broadly speaking, these results typically vary in the generality of their underlying assumptions on the normality or independence of the error terms in classical isotonic regression. As noted in the excellent recent survey Guntuboyina and Sen (2018), the same risk rate for this (and for more general) LSEs was established using an alternative approach in Chatterjee (2014). Moreover, in the case of minimax lower bounds, the matching risk rate (up to constant terms) for isotonic regression was established in Chatterjee et al. (2015) and also in Bellec and Tsybakov (2015), in both high probability and expectation terms.

Pool Adjacent Violators Algorithm (`PAVA`):

Rather remarkably, despite the nonparametric setup of classical isotonic regression, the LSE in this case has an explicit 'max-min' formulation (Barlow et al., 1972, Equation (1.9)). However, in practice it is efficiently computed using the pool adjacent violators algorithm (`PAVA`). As described in Tibshirani et al. (2011) the PAVA in effect estimates the ordered parameters by scanning through the (sorted) observations. For each adjacent pair of observations, the monotonicity constraint is checked. If the constraint is 'violated' by a given observation, the average of the observations is used as the estimate, with appropriate (minimal) backtracking to ensure that any restrospectively incurred violations are similarly corrected for. Efficient PAVA implementations, e.g., as described in Grotzinger and Witzgall (1984); Best and Chakravarti (1990), have a computational complexity of $\mathcal{O}(n)$, where n is the sample size. Since we will use the PAVA in just one step in our proposed three-step estimator for the ASCI regression parameter $\boldsymbol{\mu}$, we will not detail it further here. However, such open-source PAVA implementation details can be found in de Leeuw et al. (2009); Pedregosa et al. (2011).

3.1.5 Main contributions

Our contributions in this paper are twofold and are summarized as follows:

- **Computable estimators with non-asymptotic upper bounds:** We propose a computationally efficient three-step algorithm ASCIFIT, to estimate the required parameter $\boldsymbol{\mu}$, under the ASCI setting. Our ASCIFIT estimator converges at a $n^{-2/3}$ rate, with high probability. We illustrate our findings with extensive numerical simulations.
- **Sharp minimax lower bounds:** we provide matching high probability lower bounds (up to constant and log factors) under square loss, and thus demonstrate that our estimators are minimax optimal in this sense.

In particular, our upper bound proofs involve rather subtle theoretical details about the PAVA, and our use of method of moment techniques is quite unique in this literature. We believe these proof techniques will be of independent interest to researchers in isotonic regression. In particular, for similar adversarial estimation tasks, where traditional convex M-estimation techniques are infeasible.

3.1.6 Organization of the paper

The rest of this paper is organized as follows. In Section 3.2 we introduce ASCIFIT, our three-step estimation procedure for $\boldsymbol{\mu}$. In Section 3.3 we provide high probability upper bounds on estimation rates using ASCIFIT. In Section 3.4 we establish sharp minimax lower bounds for the parameter estimation in our ASCI setting. In Section 3.5 we provide extensive numerical ASCI simulations, to illustrate our findings. In Section 3.6 we summarize our results and describe exciting future research directions.

3.1.7 Notation

Throughout this paper, we typically use lowercase for scalars in $\mathbb{R}$, e.g., $(x, y, z, \ldots)$, bold lowercase for vectors, e.g., $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$, and bold uppercase for matrices, e.g., $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots)$. We use $\lesssim$ and $\gtrsim$ to mean $\leq$ and $\geq$, respectively, up to positive universal constants. We denote $a \vee b := \max\{a, b\}$ for each $a, b \in \mathbb{R}$. We say that a sequence $a_n := \mathcal{O}(1)$ if there exists $C > 0, N \in \mathbb{N}$ such that $|a_n| < C$ for each $n > N$. Similarly, $a_n = \mathcal{O}(b_n)$ iff $\frac{a_n}{b_n} = \mathcal{O}(1)$. We say that a sequence $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, $a_n = o(b_n)$ iff $\frac{a_n}{b_n} = o(1)$. We denote the finite set $\{1, \ldots, n\}$ by $[n]$. We define the *indicator function* $\mathbb{I}_\Omega(\mathbf{x})$ to take the value 1 when $\mathbf{x} \in \Omega \subseteq \mathbb{R}^d$, and 0 otherwise. We say that a function $f : \Omega \to \mathbb{R}$ is increasing, if for all $u, v \in \Omega \subset \mathbb{R}$ such that $u \leq v$, implies $f(u) \leq f(v)$. We use *strictly* increasing in the case where these inequalities are strict. Similarly we note that f is decreasing (or *strictly* decreasing) when these respective inequalities are reversed. We provide a useful notation summary table in Section 3.A.1.

3.2 `ASCIFIT`: A THREE-STEP ESTIMATION PROCEDURE FOR $\boldsymbol{\mu}$

As per our core question of interest, we now turn our attention to `ASCIFIT`, i.e. our proposed estimation procedure for $\boldsymbol{\mu}$, under the ASCI setup. The *Folded Normal* distribution, and in particular its mean and variance, will be fundamental to `ASCIFIT`. As such, we first formalize the key properties of the Folded Normal distribution in Definition 3.9.

Definition 3.9 (Folded Normal distribution)**.** Suppose $R \sim \mathcal{N}(\mu, \sigma^2)$, and let $T := |R|$. We then say that $T \sim \mathsf{FoldNorm}(\mu, \sigma)$, is a *Folded Normal* distribution. We denote the mean and variance of T, by $f(\mu, \sigma)$ and $g(\mu, \sigma)$, respectively. They are given as follows:

$$f(\mu, \sigma) := \mathbb{E}(T) = \sigma\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2)) - \mu(1 - 2\Phi(\mu/\sigma)). \tag{3.15}$$

$$g(\mu, \sigma) := \mathbb{V}(T) = \mu^2 + \sigma^2 - f(\mu, \sigma)^2. \tag{3.16}$$

Remark 3.10. We refer the reader to Tsagris et al. (2014); Elandt (1961) for more details. We only consider $\mu > \eta > 0$ per Equation (3.5), and we use the shorthand notation $f(\mu, \sigma)^2 := (f(\mu, \sigma))^2$.

We now describe `ASCIFIT`, our three-step procedure to estimate $\boldsymbol{\mu}$ under the ASCI setting, as follows:

`ASCIFIT`: Three-step procedure to estimate $\boldsymbol{\mu}$ under the ASCI setting

Step I (***Pre-processing and*** **`PAVA`**)**:**
Obtain an initial *naive* estimate of $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^\top$ by fitting isotonic regression (using the `PAVA`) on $T_i \coloneqq |R_i|$. Denote these estimates by $\widehat{\boldsymbol{\mu}}_{\text{naive}} \coloneqq (\widehat{T}_1, \ldots, \widehat{T}_n)^\top$.

Step II (***Second moment matching***)**:**
Estimate σ in the following way. Pick the σ solving the following equation, and denote the corresponding solution as $\widehat{\sigma}$:

$$G(\sigma) \coloneqq \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta, \sigma), \sigma))^2 = \frac{1}{n}\sum_{i=1}^{n} T_i^2. \tag{3.17}$$

Here $f^{-1}(\cdot, \sigma)$, denotes the inverse function of $f(\mu, \sigma)$ with respect to μ, when we hold σ fixed to the value σ.

Step III (***Post-processing via plug-in***)**:**
From $\widehat{\boldsymbol{\mu}}_{\text{naive}}$ in **Step I**, and $\widehat{\sigma}$ in **Step II**, compute $\widehat{\boldsymbol{\mu}}_{\text{ascifit}} \coloneqq (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^\top$ as follows:

$$\widehat{\mu}_i \coloneqq f^{-1}(\widehat{T}_i \vee f(\eta, \widehat{\sigma}), \widehat{\sigma}), \text{ for each } i \in [n]. \tag{3.18}$$

3.2.1 Intuition for the three `ASCIFIT` steps

We now provide more precise intuition for each of the three `ASCIFIT` steps, i.e., **Step I** – **Step III**.

Intuition for Step I:

Here, we begin with the pre-processing operation $T_i \coloneqq |R_i|$. This serves the critical dual purpose of removing the effect of the sign-corruptions ξ_i, and also induces *independence* of the resulting observations $(T_1, \ldots, T_n)^\top$. This helps directly address **Challenge I** and **Challenge II** under the ASCI setup. To better understand this dual effect, note that in the ASCI setup, the $\xi_i \in \{-1, 1\}$ may be arbitrarily chosen by the adversary (without a precise distributional assumption). However, the critical information in our model is given by pre-processing each observation, R_i, as $T_i \coloneqq |R_i|$. More specifically we have that $T_i = |\xi_i\,(\mu_i + \varepsilon_i)| = |\mu_i + \varepsilon_i|$. Since $\mu_i + \varepsilon_i \overset{\text{i.n.i.d.}}{\sim} \mathcal{N}\left(\mu_i, \sigma^2\right)$, per Definition 3.9 we have that $T_i \overset{\text{i.n.i.d.}}{\sim} \mathsf{FoldNorm}\left(\mu_i, \sigma\right)$, per Definition 3.9. We note that our preprocessed observations $\{T_1, \ldots T_n\}$ are all i.n.i.d.[1], since they have a common variance σ^2 but varying means μ_i for each observation $i \in [n]$. Moreover, fitting an isotonic regression to T_i intuitively estimates $f(\mu_i, \sigma)$ which are the mean of each T_i. This step is formally justified by the results of Zhang (2002).

Intuition for Step II:

This is motivated by second moment matching to estimate σ. Specifically, using the fact that the expected value of $\frac{1}{n}\sum_{i=1}^{n} T_i^2$, is $\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\mu_i^2$. The left hand side of Equation (3.17) directly estimates the term $\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\mu_i^2$. In **Step II** it is not clear a priori whether such an inverse function $f^{-1}(\cdot, \sigma)$ exists, or whether there exists a unique positive solution for σ in Equation (3.17). We will demonstrate that both assertions are true, and that the unique solution $\widehat{\sigma}$, to estimate σ, can be computed efficiently with a binary search approach. We would like to note here that estimating σ is not an easy problem (it is not by coincidence that in classical isotonic regression that σ is viewed as a nuisance parameter). This difficulty explains why we need to impose some additional assumptions on the vector μ and on σ later on. Next, we provide the intuition on why we use the factor $\widehat{T}_i \vee f(\eta, \sigma)$ in **Step II**, for each $i \in [n]$. This is summarized in Proposition 3.11.

Proposition 3.11 (Reason for the "$\vee f(\eta, \sigma)$"-correction in **Step II**). *The need for defining the $\vee f(\eta, \sigma)$ in Equation (3.17) in* **Step II** *in* `ASCIFIT`*, is that the solution to the problem*

$$\underset{\widetilde{T}_1, \ldots, \widetilde{T}_n}{\arg\min} \sum_{i=1}^{n} (T_i - \widetilde{T}_i)^2 \ \text{s.t.}\ f(\eta, \sigma) \leq \widetilde{T}_1 \leq \ldots \leq \widetilde{T}_n, \tag{3.19}$$

is related to the solution to

$$\underset{\widehat{T}_1, \ldots, \widehat{T}_n}{\arg\min} \sum_{i=1}^{n} (T_i - \widehat{T}_i)^2 \ \text{s.t.}\ \widehat{T}_1 \leq \ldots \leq \widehat{T}_n, \tag{3.20}$$

as $\widetilde{T}_i \coloneqq \widehat{T}_i \vee f(\eta, \sigma)$.

To understand the significance of Proposition 3.11, first note that we apply the `PAVA` to the T_i values in **Step I**. As such, the corresponding least squares `PAVA` estimates, $\widehat{T}_i$, actually project onto the *unconstrained* monotone cone, as per Equation (3.20). However, in our setup we actually want to solve the *constrained non-negative* monotone means, as per Equation (3.19). Fortunately, this is not an issue since we can simply post hoc correct each of the fitted

[1] i.e., independent but not identically distributed.

unconstrained PAVA solutions as $\widetilde{T}_i \coloneqq \widehat{T}_i \vee f(\eta, \sigma)$, for each $i \in [n]$. This follows by adapting Németh and Németh (2012, Corollary 1) to our ASCIFIT setup. From all of the above discussion, intuitively it follows that the term $\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2$ in (3.17) also estimates $\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\mu_i^2$, which explains why in **Step II** we equate that term to $\frac{1}{n}\sum_{i=1}^{n}T_i^2$.

Intuition for Step III:

To understand the need for this step, one needs to realize that the PAVA will estimate the means of T_i which are $f(\mu_i, \sigma)$. Hence in order for us to go back at the original μ_i scale, we need to invert the value of the PAVA estimates $\widehat{T}_i$. Ideally we would use the true value of σ in the inversion, but since it is unavailable to us, we use the plug-in estimate $\widehat{\sigma}$ as computed in **Step II**. In addition the term "$\vee f(\eta, \widehat{\sigma})$" in Equation (3.18) is present, since by assumption the value of each μ_i (or sufficiently just μ_1) must be at least η, after inverting.

3.3 Analysis of ASCIFIT: Upper bounds

We have now described the details and key intuition behind our three-step ASCIFIT estimator $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$, for $\boldsymbol{\mu}$. We now turn our attention to formalizing this intuition into least squares estimation risk bounds. More specifically, our end goal in this section is to describe our high probability non-asymptotic upper risk bound for $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$, and understand its dependence on the sample complexity, and other ASCI parameters. We also provide summary sketch behind the main proof techniques used and what insight they offer for estimation purposes. Before we state the results we will define the rate of convergence $r_{n,2}(\mu_n, \mu_1, \sigma)$, which plays an important role in all of the Theorems to follow. For an absolute constant $C_2 > 0$ define

$$r_{n,2}(\mu_n, \mu_1, \sigma) \coloneqq \min\left[2\sigma^2 C_2^2, \frac{27}{4}\left(\frac{\mu_n - \mu_1}{n}\right)^{\frac{2}{3}}(\sigma C_2)^{\frac{4}{3}} + \frac{2\sigma^2 C_2^2}{n}(1 + \log n)\right]. \tag{3.21}$$

Importantly, assuming that $\mu_n - \mu_1, \sigma$ are constants not scaling with the sample size n, we have that $r_{n,2}(\mu_n, \mu_1, \sigma) \lesssim \max\left\{\left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}}, \frac{\sigma^2 \log n}{n}\right\}$, where $V \coloneqq \mu_n - \mu_1$, is the total variation of the underlying monotone signal. With this essential background, we are ready to state our first result in Theorem 3.12.

Theorem 3.12 (Equation (3.17) has a unique root)**.** *Assume that there exist constants $\psi, \Psi, C > 0$ such that $\psi \le \sigma \le \Psi$ and $\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \le C$, for each $n \in \mathbb{N}$. In addition let $r_{n,2}(\mu_n, \mu_1, \sigma) = o(1)$, where the quantity $r_{n,2}(\mu_n, \mu_1, \sigma)$*

is defined in (3.21)*. Then for sufficiently large* n*,* $\delta = o\left((r_{n,2}(\mu_n, \mu_1, \sigma))^{-1}\right)$*, and* $\gamma = o\left(n^{1/2}\right)$*, under the ASCI setup, Equation* (3.17) *in* ***`ASCIFIT`*** *has a unique root* $\sigma^* \in \left[0, \sqrt{\frac{1}{n}\sum_{i=1}^n T_i^2}\right]$ *for* σ *with probability at least* $1 - \delta^{-1} - 2\gamma^{-2}$*.*

The key insight of Theorem 3.12 from a statistical perspective, is that our second moment matching approach in **Step II** will ensure that our proposed estimator $\widehat{\sigma}$, for σ, will be unique with high probability. The core idea behind the proof of Theorem 3.12 is that the map $\sigma \mapsto G(\sigma) := \sigma^2 + \frac{1}{n}\sum_{i=1}^n (f^{-1}(\widehat{T}_i \vee f(\eta, \sigma), \sigma))^2$ is monotone increasing over $\sigma \geq 0$. This enables the use of the intermediate value theorem to check that two endpoints of $G(\sigma) - \frac{1}{n}\sum_{i=1}^n T_i^2$, evaluated at $\sigma \in \{0, \sqrt{\frac{1}{n}\sum_{i=1}^n T_i^2}\}$ have opposite sign with high probability. This has important practical implications for estimation purposes. In effect, it means that we can efficiently *compute* $\widehat{\sigma}$, by solving $G(\sigma) = \frac{1}{n}\sum_{i=1}^n T_i^2$ (per Equation (3.17)) using a binary search approach between the two identified endpoints. We would like to mention that while using the intermediate value theorem sounds like an easy task, it turns out that it is extremely hard to verify that $G(0) \leq \frac{1}{n}\sum_{i=1}^n T_i^2$, for which the bulk of the proof of Theorem 3.12 is dedicated to.

Although Theorem 3.12 gives us a high probability bound on estimating $\widehat{\sigma}$ uniquely, it is important to next understand how efficiently $\widehat{\sigma}$ estimates σ. This is summarized in Theorem 3.13.

Theorem 3.13 ($\widehat{\sigma}$ is close to σ)**.** *Under the assumptions of Theorem 3.12, we have that* $|\sigma - \widehat{\sigma}| \lesssim (\delta r_{n,2}(\mu_n, \mu_1, \sigma))^{1/2} + \gamma n^{-1/2}$ *with probability at least* $1 - \delta^{-1} - 2\gamma^{-2}$*, where* $\delta^{-1}, \gamma^{-2} \in (0, 1)$*.*

From Theorem 3.13 we see that $\widehat{\sigma}$ converges to σ roughly at a $n^{-1/3}$ rate. In both Theorems 3.12 and 3.13, we require that there exist constants $\psi, \Psi, C > 0$ such that $\psi \leq \sigma \leq \Psi$ and $\frac{1}{n}\sum_{i=1}^n \mu_i^2 \leq C$, for each $n \in \mathbb{N}$. For transparency, we note that such assumptions are an artefact of our methodology and ensure that our risk bounds can be tightly controlled using the second moment matching approach. Given the highly adversarial corruptions and non-convex constraints that can arise under ASCI estimation, e.g., in Example 3.7, these are slightly stronger assumptions required for classical convex isotonic regression setup. They effectively represent a trade-off for the flexibility, and simplicity of using `ASCIFIT` under these adversarial settings, whilst still ensuring precise control in the parameter estimation risk bounds.

$\widehat{\sigma}$ in our post-processing correction for $\widehat{\boldsymbol{\mu}}_{\text{ascifit}} := (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^\top$. That is, our final estimate for each μ_i, is given by $\widehat{\mu}_i := f^{-1}(\widehat{T}_i \vee f(\eta, \widehat{\sigma}), \widehat{\sigma})$. With this explicit form and our tightly controlled bounds in Theorem 3.12 and Equation (3.17) we are finally able derive the required least squares risk rate of $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$. This is summarized in Theorem 3.14. We will shortly discuss this result further in Section 3.4 when we derive high probability minimax lower bounds.

Theorem 3.14 ($\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$ is close to $\boldsymbol{\mu}$)**.** *Under the assumptions of Theorem 3.12 and Theorem 3.13, we have that*

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma}) - \mu_i)^2 \lesssim \delta r_{n,2}(\mu_n,\mu_1,\sigma) + \gamma^2 n^{-1}, \tag{3.22}$$

with probability at least $1 - \delta^{-1} - 2\gamma^{-2}$.

Remark 3.15. We note that η is currently absorbed in our constants in Theorems 3.12 to 3.14. The exact form is complicated (but the smaller the η the bigger the constants). For more details, please refer to Section 3.D.

3.4 Lower bounds

We now derive high probability minimax lower bounds under the ASCI setting. We accordingly first introduce the relevant related notation and definitions here for classes of monotonic sequences. We denote $\mathcal{S}^{\uparrow} \coloneqq \left\{\boldsymbol{\mu} \coloneqq (\mu_1,\ldots,\mu_n)^{\top} \,\middle|\, \mu_1 \leq \ldots \leq \mu_n\right\}$ to be the set of all non-decreasing sequences. We define $k(\boldsymbol{\mu}) \geq 1$, for $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$, to be the integer such that $k(\boldsymbol{\mu}) - 1$ is the number of inequalities $\mu_i \leq \mu_{i+1}$ that are strict for $i \in [n-1]$ (i.e., the number of 'jumps' of $\boldsymbol{\mu}$). The class of bounded monotone functions are $\mathcal{S}^{\uparrow}(V^*) \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow} \,\middle|\, V(\boldsymbol{\mu}) \leq V^*\right\}$, for some fixed $V^* \geq 0$, and $V(\boldsymbol{\mu}) \coloneqq \mu_n - \mu_1$, is the total variation of any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$. We focus on the ASCI-restricted class of monotone sequences, i.e., $\mathcal{S}^{\uparrow}(V^*,\eta,C) \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}(V^*) \,\middle|\, \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leq C, \mu_1 > \eta > 0\right\}$.

We closely follow the approach of Bellec and Tsybakov (2015, Proposition 4) but non-trivially adapt it to our ASCI setting by ensuring the monotonicity constraint in Equation (3.5) is satisfied in the lower bound construction. The proof uses well established techniques including the Varshamov-Gilbert bound Tsybakov (2009, Lemma 2.9), and Fano's Lemma arguments using Tsybakov (2009, Theorem 2.7). This leads to our minimax lower bound result in Proposition 3.16.

Proposition 3.16 (Minimax lower bounds)**.** *Let* $n \geq 2, V^* > 0$ *and* $\sigma > 0$, *and define* $\widetilde{r}_{n,2}(V^*,\sigma) \coloneqq \max\left\{\left(\frac{\sigma^2 V^*}{n}\right)^{\frac{2}{3}}, \frac{\sigma^2}{n}\right\}$. *Then, there exist absolute constants*

$c, c' > 0$ *such that:*

$$\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathcal{S}^{\uparrow}(V^*, \eta, C)} \mathbb{P}_{\boldsymbol{\mu}} \left(\frac{1}{n} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq c\widetilde{r}_{n,2}(V^*, \sigma) \right) > c' \tag{3.23}$$

Crucially, Proposition 3.16 demonstrates that our high probability upper bounds for `ASCIFIT` in Theorem 3.14 are sharp in the minimax sense, up to constants and log factors. This is evident by directly comparing $\widetilde{r}_{n,2}(V^*, \sigma)$ to $r_{n,2}(\mu_n, \mu_1, \sigma)$ per Equation (3.21).

3.5 Simulations

We now demonstrate our `ASCIFIT` estimation algorithm in action through a variety of simulations[2]. Specifically, for simulation purposes we consider n observations, $\{R_i \,|\, i \in [n]\}$, where each observation R_i is generated from Example 3.5. For sufficiently large n, the ASCI model in Example 3.5 roughly translates to $(1-p)$-proportion of observations being independently sign-corrupted by the adversary. Moreover we assume per Equation (3.11) that the adversarial sign-corruptions, ξ_i, are chosen independently of all true errors, ε_i, for each $i \in [n]$. The true monotone signal is defined to be $\mu_i \coloneqq \eta + (1-\eta)\frac{i-1}{n}$, for each $i \in [n]$. We run this generating process over the following parameter grid: $\eta \coloneqq 0.2$, $p \coloneqq 0.5$, $\sigma \in \{0.5, 1, 1.5, 2\}$, $n \in \{100, 250, 500, 1000\}$. We perform 50 replications for each combination of simulation grid parameters. In each replication of this generating process we fit the `ASCIFIT` estimator $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$, for $\boldsymbol{\mu}$. The main summary result from running our simulation, is shown in Figure 3.5.1.

To clarify, given $\eta = 0.2, p = 0.5$, Figure 3.5.1 plots the sample mean-MSE, $\frac{1}{n}\|\widehat{\boldsymbol{\mu}}_{\text{ascifit}} - \boldsymbol{\mu}\|^2$, over 50 `ASCIFIT` replications for each value of $n \in \{100, 250, 500, 1000\}$. Here the sample mean-MSE is a useful simulation proxy for the least squares error, our core theoretical risk measure of interest. This sample mean-MSE is plotted separately for each of the four sigma values, $\sigma \in \{0.5, 1, 1.5, 2\}$. The mean-MSE value of each replication ($\pm$ 2 standard errors) are shown using error bars in an effort to quantify replication uncertainty. The plot in Figure 3.5.1 is as expected in that all of the sample mean-MSE values show a steady decreasing trend in n. Importantly the relative uncertainty in sample mean-MSE reduces in n, as seen by the smaller error bars to the right

[2]Reproducible code for all figures in this paper is found at: `https://github.com/shamindras/ascifit`. All of the simulation results in this section were run on a personal Macbook laptop with macOS, Intel Core i9 CPU, and 64GB RAM. The total runtime for a single run of all simulations is approximately 90 minutes.

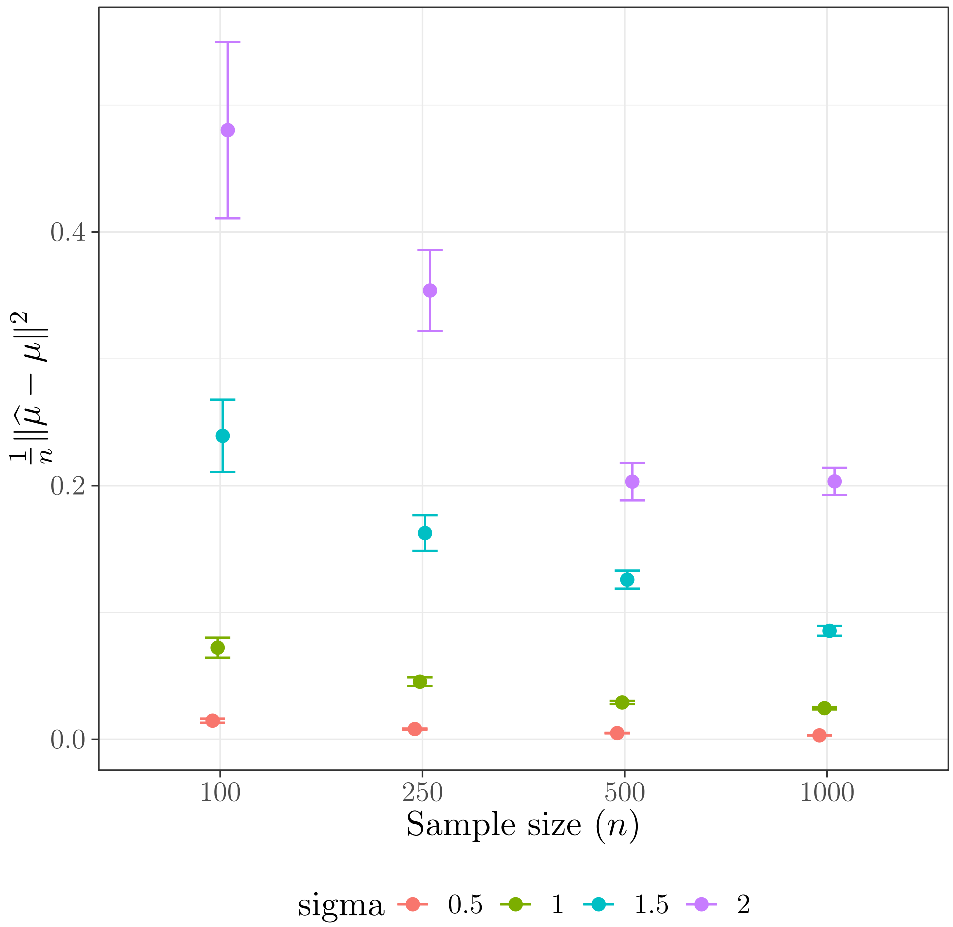

Figure 3.5.1: Mean (sample) MSE for estimating $\boldsymbol{\mu}$ via `ASCIFIT` as a function of n, σ.

of Figure 3.5.1. For smaller σ values, i.e., smaller variance in the underlying generating model, we see a much lower sample mean-MSE on average compared to higher σ-valued simulations. That is, our `ASCIFIT` estimator achieves better accuracy, with smaller underlying variability in the model, on average when other factors are held constant.

Finally, in order to precisely gauge how well the `ASCIFIT` estimator $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$, actually fits the true signal $\boldsymbol{\mu}$, it is instructive to plot both directly on the original generating sample data. This is seen for one instance over our parameter grid of simulations in Figure 3.5.2.

Specifically, for $\eta = 0.2, p = 0.5, n = 1000, \sigma = 1.5$, Figure 3.5.2 plots the simulated true generating process, $\boldsymbol{\mu}$, against the `ASCIFIT` estimator, $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$.

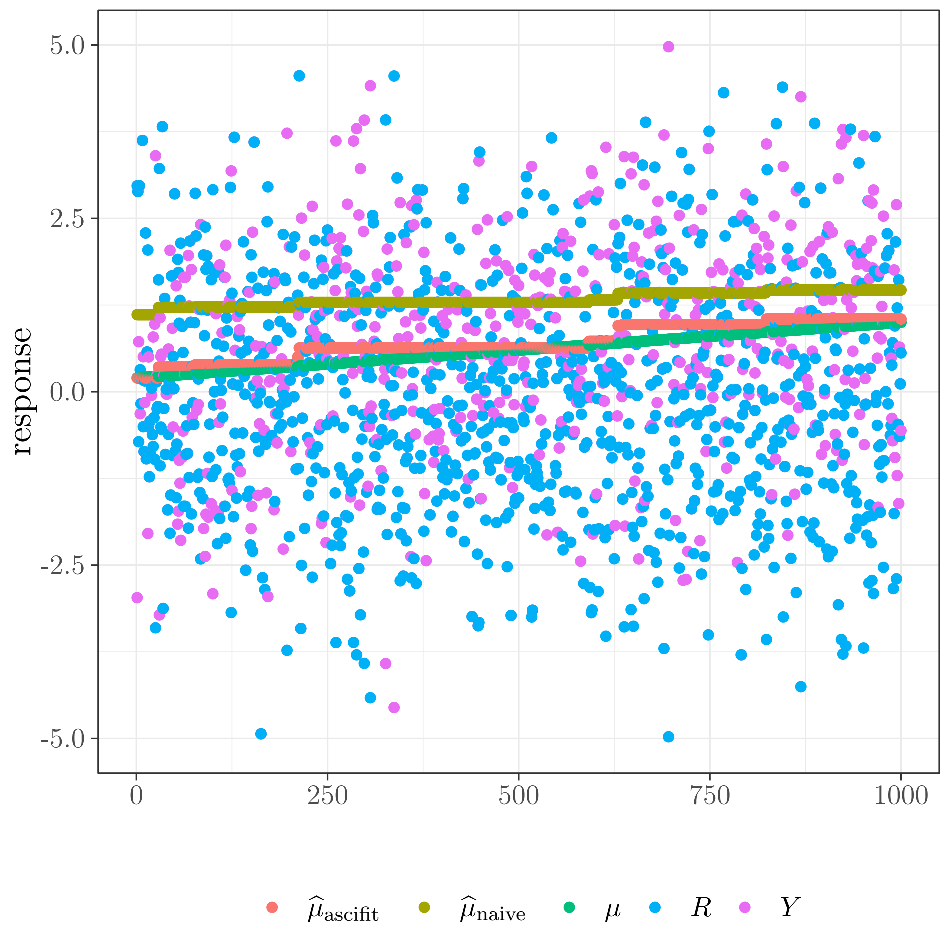

Figure 3.5.2: Example of $\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$ vs. $\boldsymbol{\mu}$ for $\eta = 0.2, p = 0.5, n = 1000, \sigma = 1.5$.

Additionally both the original and sign-corrupted individual observations are plotted to emphasize the difficulty of this estimation task. Moreover, for comparison purposes we also plot the naive estimator, i.e., $\widehat{\boldsymbol{\mu}}_{\text{naive}}$. Here $\widehat{\boldsymbol{\mu}}_{\text{naive}}$ represents the estimator by stopping at **Step I** in `ASCIFIT`. That is, estimating $\boldsymbol{\mu}$, by simply fitting isotonic regression (using the `PAVA`) on $T_i := |R_i|$. Furthermore, since $p = 0.5$, as expected, on average roughly half of the true responses are adversarially sign-corrupted. Despite this, one can see that `ASCIFIT` is relatively stable and reasonably recovers the true signal. This shows more directly (in such an instance), the robustness of `ASCIFIT` under such randomized adversarial sign-corruptions. Moreover since $n = 1000$, we can see that `ASCIFIT` indeed fits well with increasing sample complexity. In addition it highlights the importance of **Step II** and **Step III** in `ASCIFIT`.

3.6 Conclusion

In this paper we have considered a variation of the original isotonic regression problem in which the observations can be adversarially corrupted in their sign value. In this ASCI setting, adversarially refers to the fact that the sign-corruptions can be chosen to have strong dependence with the error terms in the original model. Our simple three-step estimation procedure, `ASCIFIT`, is easy to implement with existing software and has sharp non-asymptotic minimax guarantees on the estimation error, under square loss. For future directions we note that that true signal is required to be strictly positive for our guarantees to hold. We believe this restriction can be lifted if one uses *unimodal* regression instead of isotonic regression in **Step I**. However, sharp risk guarantees need to first be proven similar to Zhang (2002) under this unimodal setting. It would also be interesting to see if the moment matching technique could be extended subgaussian error terms. We leave these exciting directions for future work.

3.7 Acknowledgments

We would like to thank Arun Kumar Kuchibhotla, Alex Reinhart, Alessandro Rinaldo, Larry Wasserman from the Carnegie Mellon University (CMU) Department of Statistics & Data Science, and Yang Ning from the Cornell Department of Statistics & Data Science. Their encouragement, and extensive feedback throughout this work greatly shaped the final outcome. This paper extensively utilizes the `R` statistical software (R Core Team, 2021) for conducting simulations and plots. In particular, we relied primarily on the `tidyverse` (Wickham et al., 2019) collection of `R` packages.

Appendix - Chapter 3

3.A Preliminary

In this appendix we provide detailed proofs of all key statements from the main paper. Since our work relies a variety of core ideas from isotonic regression we first introduce some common definitions which will be referred to in subsequent proofs.

3.A.1 Notation Summary

To ensure that the Appendix is can be read in a standalone manner, we consolidate key notation used in the paper in Table 3.A.1. Unless stated otherwise $K \subseteq \mathbb{R}^d$ is a closed, non-empty convex set, and $\Omega \subseteq \mathbb{R}^d$.

3.A.2 Useful miscellaneous results

Here we prove some useful standard results that are used in several of the remaining proofs. For reader convenience, we provide short proofs to ensure that our work is self-contained.

We start with some elementary inequalities, which will be used repeatedly. First, in Lemma 3.17 we introduce a differencing inequality we use repeatedly to construct lower bounds.

Lemma 3.17 (Difference of squares lower bound)**.** *For each $a, b, l \in \mathbb{R}$, such that $b, l \geq 0$ and $a - b \geq l$, the following holds:*

$$a^2 - b^2 \geq la \geq l^2 \tag{3.24}$$

Table 3.A.1: Notation and conventions used in this chapter

	Variables and inequalities
$a \wedge b$	$\min\{a,b\}$ for each $a, b \in \mathbb{R}$
$a \vee b$	$\max\{a,b\}$ for each $a, b \in \mathbb{R}$
scalars	$x, y, z \in \mathbb{R}$
vectors	$\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$
matrices	$\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{d \times m}$
$\lesssim$	$\leq$ up to positive universal constants
$\gtrsim$	$\geq$ up to positive universal constants
$a_n = \mathcal{O}(1)$	$(\exists C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\lvert a_n \rvert < C)$
$a_n = \mathcal{O}(b_n)$	$\frac{a_n}{b_n} = \mathcal{O}(1)$
$a_n = o(1)$	$(\forall C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\lvert a_n \rvert < C)$
$a_n = o(b_n)$	$\frac{a_n}{b_n} = o(1)$
$X_n = o_P(1)$	$(\forall \varepsilon > 0)(\mathbb{P}(\lvert X_n \rvert \geq \varepsilon) \xrightarrow{n \to \infty} 0)$
$X_n = \mathcal{O}_P(1)$	$(\forall \varepsilon > 0)(\exists C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\mathbb{P}(\lvert X_n \rvert \geq C) \leq \varepsilon)$
	Functions and sets
$[n]$	$\{1, \ldots, n\}$, for $n \in \mathbb{N}$
Indicator function $\mathbb{I}_\Omega(\mathbf{x})$	Takes value 1 when $x \in \Omega$, and 0 otherwise
$\Pi_K : \mathbb{R}^d \to K$	ℓ_2-projection of any $\mathbf{x} \in \mathbb{R}^d$ onto K
$f : \Omega \to \mathbb{R}$ is increasing	If $\forall u, v \in \Omega$ such that $u \leq v \implies f(u) \leq f(v)$
$f : \Omega \to \mathbb{R}$ is decreasing	If $\forall u, v \in \Omega$ such that $u \leq v \implies f(u) \geq f(v)$
$\Phi : \mathbb{R} \to [0,1]$	Cumulative density function of $\mathcal{N}(0,1)$
$\phi : \mathbb{R} \to \mathbb{R}$	Probability density function of $\mathcal{N}(0,1)$
$\mathcal{S}^\uparrow$	$\left\{\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^\top \,\middle\vert\, \mu_1 \leq \ldots \leq \mu_n\right\}$
$\mathcal{S}^\uparrow_+$	$\left\{\boldsymbol{\mu} \in \mathcal{S}^\uparrow \,\middle\vert\, \mu_1 \geq 0\right\}$
$\mathcal{S}^\uparrow(V^*)$	$\left\{\boldsymbol{\mu} \in \mathcal{S}^\uparrow \,\middle\vert\, V(\boldsymbol{\mu}) \leq V^*\right\}$
$V(\boldsymbol{\mu})$	$\mu_n - \mu_1$ for $\boldsymbol{\mu} \in \mathcal{S}^\uparrow$
$\mathcal{S}^\uparrow_{k^*}$	$\left\{\boldsymbol{\mu} \in \mathcal{S}^\uparrow \,\middle\vert\, k(\boldsymbol{\mu}) \leq k^*\right\}$
$\mathcal{S}^\uparrow(V^*, \eta, C)$	$\left\{\boldsymbol{\mu} \in \mathcal{S}^\uparrow(V^*) \,\middle\vert\, \frac{1}{n}\sum_{i=1}^n \mu_i^2 \leq C, \mu_1 > \eta > 0\right\}$

Proof of Lemma 3.17. We proceed as follows. First note that since $b, l \geq 0$ by assumption, we have that $a - b \geq l \iff a \geq b + l \geq 0$. Now observe:

$$\begin{aligned}
a^2 - b^2 &= (a+b)(a-b) \\
&\geq l(a+b) && \text{(since } a - b \geq l \geq 0 \text{ by assumption.)} \\
&\geq la && \text{(since } b \geq 0 \text{ by assumption.)} \\
&\geq l^2 && \text{(since } a \geq l\text{)}
\end{aligned}$$

as required. □

Lemma 3.18 (Lower bound via difference of squares)**.** *For each $a, b, C, K \in \mathbb{R}$, such that $b \geq 0, a^2 - b^2 \geq C > 0, a \in [0, K]$, the following holds:*

$$a - b \geq \frac{C}{2K} \tag{3.25}$$

Proof of Lemma 3.18. We proceed as follows. First note that since $a^2 - b^2 \geq C > 0$ by assumption, we have that $a > 0$, and hence $a > b, K > 0$ since both a, b are non-negative. Now observe:

$$\begin{aligned}
& a^2 - b^2 \geq C && \text{(by assumption.)} \\
\implies \quad & a - b \geq \frac{C}{a+b} && \text{(since } a > 0, b \geq 0 \implies a + b > 0.\text{)} \\
& \qquad\quad \geq \frac{C}{2a} && \text{(since } a \geq b.\text{)} \\
& \qquad\quad \geq \frac{C}{2K} && \text{(since } a \leq K.\text{)}
\end{aligned}$$

as required. □

Lemma 3.19 (Maximum difference square inequality)**.** *For each $a, b, c \in \mathbb{R}$ such that $b \leq c$ the following inequality holds:*

$$((a \vee b) - c)^2 \leq (a - c)^2 \tag{3.26}$$

Proof of Lemma 3.19. Under the assumption that $a, b, c \in \mathbb{R}$ such that $b \leq c$, let $d := a \vee b$. We then observe:

$$\begin{aligned}
& (d - c)^2 \leq (a - c)^2 \\
\iff \quad & d^2 - a^2 \leq 2dc - 2ac && \text{(expanding and simplifying.)} \\
\iff \quad & (d + a)(d - a) \leq 2c(d - a) && (3.27)
\end{aligned}$$

So we need to equivalently prove that Equation (3.27). To that end we only need to consider 2 cases. Namely $a \geq b$, and $a < b$. Note that in the first case $a \geq b \implies d \coloneqq a \vee b = a$. In this case, both LHS/RHS of Equation (3.27) are 0, and the statement holds. Next consider the case $a < b$. Here we have $a < b \implies d \coloneqq a \vee b = b > a$. We then observe the following:

$$\begin{aligned} a + d &= a + b && \text{(since } d = b.\text{)} \\ &\leq 2b && \text{(since } a < b \text{ by assumption.)} \\ &\leq 2c && \text{(since } b \leq c \text{ by assumption.)} \end{aligned}$$

That is, we have that $a + d \leq 2c$. Substituting back to Equation (3.27) we have that $(d + a)(d - a) \leq 2c(d - a)$, which is what we wanted to show. Which completes the proof Equation (3.26), as required. □

Lemma 3.20 (Square sum inequality)**.** *For each $a, b \in \mathbb{R}$ the following holds:*

$$(a + b)^2 \leq 2(a^2 + b^2) \tag{3.28}$$

Proof of Lemma 3.20. We proceed as follows:

$$\begin{aligned} (a + b)^2 &= a^2 + 2ab + b^2 && (3.29) \\ &\leq a^2 + b^2 + 2\,|ab| && \text{(since } x \leq |x| \text{ for each } x \in \mathbb{R}\text{)} \\ &\leq a^2 + b^2 + 2(|a|^2 + |b|^2) && \text{(by AM-GM we have } 2\,|ab| \leq |a|^2 + |b|^2\text{)} \\ &= 2(a^2 + b^2) && (3.30) \end{aligned}$$

as required. □

As a result of Lemma 3.20 we obtain Corollary 3.21.

Corollary 3.21. *For random variables X_1, X_2 the following holds:*

$$\mathbb{V}\left(X_1 - X_2\right) \leq 2(\mathbb{V}\left(X_1\right) + \mathbb{V}\left(X_1\right)) \tag{3.31}$$

Proof of Corollary 3.21. First let the centered versions of the random variables be denoted as

$$\widetilde{X}_i \coloneqq X_i - \mathbb{E}\left(X_i\right), \text{ for each } i \in [2]. \tag{3.32}$$

It then follows that:

$$\begin{aligned}
\mathbb{V}(X_1 - X_2) &= \mathbb{V}(X_1 - X_2 + \mathbb{E}(X_2) - \mathbb{E}(X_1)) && (3.33)\\
&= \mathbb{V}\left(\widetilde{X}_1 - \widetilde{X}_2\right) && (3.34)\\
&= \mathbb{E}\left((\widetilde{X}_1 - \widetilde{X}_2)^2\right) && \text{(since } \widetilde{X}_1, \widetilde{X}_2 \text{ are both centered.)}\\
&= \mathbb{E}\left(2(\widetilde{X}_1^2 + \widetilde{X}_2^2)\right) && \text{(using Lemma 3.20)}\\
&= 2\left(\mathbb{E}\left(\widetilde{X}_1^2\right) + \mathbb{E}\left(\widetilde{X}_2^2\right)\right) && \text{(linearity of expectation.)}\\
&= 2(\mathbb{V}(X_1) + \mathbb{V}(X_2)) && \text{(since } \widetilde{X}_1, \widetilde{X}_2 \text{ are both centered.)}
\end{aligned}$$

as required. □

The following is a standard result from real analysis, which we use repeatedly.

Lemma 3.22 (B-Lipschitz characterization via bounded derivative)**.** *Let $f : I \to \mathbb{R}$ be continuous and once differentiable, where $I \subseteq \mathbb{R}$ is an interval (possibly unbounded).*

$$f \text{ is } B\text{-Lipschitz, with } B > 0 \iff (\exists B > 0)(\forall x \in \mathbb{R}) : (\left|f'(x)\right| \leq B) \quad (3.35)$$

Proof of Lemma 3.22. We prove both directions. In both parts we assume that $f : I \to \mathbb{R}$ be continuous and once differentiable, where $I \subseteq \mathbb{R}$ is an interval (possibly unbounded).
($\implies$). Suppose that f is B-Lipschitz, with $B > 0$. We then have that, for some fixed (but arbitrary) $c \in I$:

$$\begin{aligned}
&|f(x) - f(c)| \leq B\,|x - c| && \text{(by definition of } B\text{-Lipschitz property.)}\\
\implies &\left|\frac{f(x) - f(c)}{x - c}\right| \leq B && \text{(taking limits as } x \to c.)\\
\implies &\left|f'(c)\right| \leq B
\end{aligned}$$

Since $c \in I$ is arbitrary, indeed $|f'(x)| \leq B$, for each $x \in I$, as required.

($\impliedby$). Suppose that $|f'(x)| \leq B$, with $B > 0$. Further let $x, y \in I$, such that $x < y$. Since f is differentiable on I, we have:

$$\begin{aligned}
|f(x) - f(y)| &\leq \left|f'(c)\right|\,|x - y|\\
&\qquad \text{(by the mean value theorem, for some } c \in (x, y).)\\
&\leq B\,|x - y| && \text{(by assumption.)}
\end{aligned}$$

Which implies that f is B-Lipschitz, as required. □

Lemma 3.23 (Standard normal upper bound). *Let $\phi(x), \Phi(x)$ respectively denote the probability density function, and cumulative density function of a standard normal variable. Then the following inequality holds:*

$$\frac{x\phi(x)}{2\Phi(x)-1} \leq \frac{1}{2}, \text{ for each } x \geq 0 \tag{3.36}$$

With equality if and only if $x = 0$.

Proof of Lemma 3.23. We first note that at $x = 0$, that $\frac{x\phi(x)}{2\Phi(x)-1}$ is an indeterminate form of type $\frac{0}{0}$. As such we have:

$$\begin{aligned}
\lim_{x\to 0} \frac{x\phi(x)}{2\Phi(x)-1} &= \frac{\lim_{x\to 0} \frac{\partial}{\partial x} x\phi(x)}{\lim_{x\to 0} \frac{\partial}{\partial x} 2\Phi(x)-1} && \text{(using L'Hospital's rule.)} \\
&= \frac{\lim_{x\to 0} \phi(x) + x\phi'(x)}{\lim_{x\to 0} 2\phi(x)} \\
&= \frac{\lim_{x\to 0} \phi(x)}{\lim_{x\to 0} 2\phi(x)} \\
&= \frac{\phi(0)}{2\phi(0)} && \text{(by continuity of } \phi(x) \text{ at } x = 0.) \\
&= \frac{1}{2} && (3.37)
\end{aligned}$$

With our given function now defined to be $\frac{1}{2}$ at $x = 0$, we now proceed to prove our given inequality. Observe that we can equivalently reformulate it as:

$$\Phi(x) - x\phi(x) - \frac{1}{2} \geq 0 \tag{3.38}$$

Setting $h(x) \coloneqq \Phi(x) - x\phi(x) - \frac{1}{2}$, we observe that $h(0) = \Phi(0) - \frac{1}{2} = 0$. We need to show that $h(x) \geq 0$, for each $x \geq 0$, which will imply the result. We will show that $h(x)$ is increasing, i.e., or equivalently that $h'(x) \geq 0$, for each $x \geq 0$. We then have that:

$$\begin{aligned}
h'(x) &= \phi(x) - (\phi(x) + x\phi'(x)) \\
&= -x\phi'(x) \\
&= -x\left(-x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\right) && \text{(using } \phi(x) \coloneqq \tfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.) \\
&= x^2\phi(x) \\
&\geq 0 && (3.39)
\end{aligned}$$

Note that the inequality in Equation (3.39) is strict when $x > 0$ and equality holds if and only if $x = 0$. This means the function is strictly increasing and bounded away from 0 when for each $x > 0$, and equal to 0 only when $x = 0$, as required. $\square$

3.A.3 The Folded Normal Distribution

For convenience, we begin by quickly recalling the definition of the Folded Normal distribution.

Definition 3.9 (Folded Normal distribution)**.** Suppose $R \sim \mathcal{N}(\mu, \sigma^2)$, and let $T \coloneqq |R|$. We then say that $T \sim \mathsf{FoldNorm}(\mu, \sigma)$, is a *Folded Normal* distribution. We denote the mean and variance of T, by $f(\mu, \sigma)$ and $g(\mu, \sigma)$, respectively. They are given as follows:

$$f(\mu, \sigma) \coloneqq \mathbb{E}\left(T\right) = \sigma\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2)) - \mu(1 - 2\Phi(\mu/\sigma)). \tag{3.15}$$

$$g(\mu, \sigma) \coloneqq \mathbb{V}\left(T\right) = \mu^2 + \sigma^2 - f(\mu, \sigma)^2. \tag{3.16}$$

Remark 3.24. We note that Equation (3.15) can be equivalently written as follows:

$$\sigma\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2)) + \mu(1 - 2\Phi(-\mu/\sigma)) \tag{3.40}$$

Note that this equivalence follows from the symmetry of the standard normal CDF, i.e., $\Phi(x) = 1 - \Phi(-x)$ for each $x \in \mathbb{R}$. For our purposes we typically use the form of Equation (3.15).

3.A.4 Properties of the folded normal mean: $f(\mu, \sigma)$

Let's start setting up some notation. First we note as previously $T_i \coloneqq |R_i| = |\mu_i + \varepsilon_i|$. Where we then have $T_i \sim \mathsf{FoldNorm}\left(\mu_i, \sigma^2\right)$. Now denote $f(\mu_i, \sigma) \coloneqq \mathbb{E}\left(T_i\right)$, for each $i \in [n]$. Moreover the T_i random variables are all mutually independent, but not identically distributed (since their mean's, i.e., $f(\mu_i, \sigma)$ differ for each $i \in [n]$). Since we run PAVA on $(T_1, \ldots, T_n)$ we have the resulting estimators $(\widehat{T}_1, \ldots, \widehat{T}_n)$. We will also denote the population level error terms for this transformed (mean centered) response as $\delta_i \coloneqq T_i - f(\mu_i, \sigma)$. We note that the $(\delta_1, \ldots, \delta_n)$ are all mutually independent, but not identically distributed.

Lemma 3.25 (Properties of the Folded Normal mean)**.** *Suppose* $R \sim \mathcal{N}(\mu, \sigma^2)$. *Let* $T \overset{a.s.}{=} |R|$, *then* $T \sim \mathsf{FoldNorm}\left(\mu, \sigma^2\right)$ *per* Definition 3.9. *We denote the mean of the Folded Normal distribution by* $f(\mu, \sigma) \coloneqq \mathbb{E}\left(T\right)$. *Given this setup,* *and fixing $\sigma > 0$, we note the following important properties of $f(\mu, \sigma)$:*

$$f(\mu, \sigma) \geq 0 \textit{ for each } \mu \in \mathbb{R} \tag{3.41}$$

$$f(\mu, \sigma) \geq \mu \textit{ for each } \mu \in \mathbb{R} \tag{3.42}$$

$$f(\mu, \sigma) \textit{ is strictly increasing in } \mu \in \mathbb{R}_{>0} \tag{3.43}$$

$$\frac{\partial f(\mu, \sigma)}{\partial \mu} \in (0, 1) \textit{ is for each } \mu \in \mathbb{R}_{>0} \tag{3.44}$$

$$f(\mu, \sigma) \textit{ is 1-Lipschitz for each } \mu \in \mathbb{R}_{>0} \tag{3.45}$$

$$f(\mu, \sigma)^2 \leq \mu^2 + \sigma^2 \textit{ for each } \mu \in \mathbb{R}_{\geq 0} \tag{3.46}$$

Additionally for $\mu_1 \leq \ldots \leq \mu_n$ we have that the relationship holds for $V(f, \boldsymbol{\mu}, \sigma)$, i.e., the total variation of the mean of the Folded Normal distribution:

$$V(f, \boldsymbol{\mu}, \sigma) \coloneqq \sum_{i=1}^{n-1} |f(\mu_{i+1}, \sigma) - f(\mu_i, \sigma)| \leq \mu_n - \mu_1 \tag{3.47}$$

Proof of Lemma 3.25. We prove each property (Equations (3.41) to (3.47)) in turn. As per the assumption $\sigma > 0$ is fixed and that $R \sim \mathcal{N}\left(\mu, \sigma^2\right)$ for $\mu \in \mathbb{R}$. ■

(*Proof of Equation* (3.41).) We have that $T \coloneqq |R| \geq 0$ a.s. $\implies f(\mu, \sigma) \coloneqq \mathbb{E}(T) = \mathbb{E}(|R|) \geq 0$ by the monotonicity of expectation, as required. ■

(*Proof of Equation* (3.42).) We have that $R \leq |R|$ a.s. $\implies \mu \coloneqq \mathbb{E}(R) \leq \mathbb{E}(|R|) = \mathbb{E}(T) \eqqcolon f(\mu, \sigma)$ again by the monotonicity of expectation, as required. ■

(*Proof of Equation* (3.43).) For any $\mu > 0$ we have that:

$$\begin{aligned}
f(\mu, \sigma) &\coloneqq \sigma \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) - \mu\left(1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\right) && \text{(per Equation (3.15))} \\
\implies \frac{\partial f(\mu, \sigma)}{\partial \mu} &= -\frac{\mu}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) - 1 + 2\Phi\left(\frac{\mu}{\sigma}\right) + \frac{2\mu}{\sigma}\phi\left(\frac{\mu}{\sigma}\right) \\
&= 2\Phi\left(\frac{\mu}{\sigma}\right) - 1 && \left(\text{since } \tfrac{\mu}{\sigma}\sqrt{\tfrac{2}{\pi}} \exp\left(-\tfrac{\mu^2}{2\sigma^2}\right) = \tfrac{2\mu}{\sigma}\phi\left(\tfrac{\mu}{\sigma}\right)\right) \\
&> 0 && \left(\text{since } \mu, \sigma > 0 \text{ and } \Phi\left(\tfrac{\mu}{\sigma}\right) > \tfrac{1}{2}\right)
\end{aligned}$$

as required. ■

(*Proof of Equation* (3.44).) By the previous proof, we note that $\frac{\partial f(\mu,\sigma)}{\partial \mu} > 0$. Also using the previous proof and noting that $\Phi(x) > \frac{1}{2}$ for each $x > 0$, it follows that $\frac{\partial f(\mu,\sigma)}{\partial \mu} = -1 + 2\Phi\left(\frac{\mu}{\sigma}\right) < 1$. Combining both parts we have that $\frac{\partial f(\mu,\sigma)}{\partial \mu} \in (0, 1)$, as required. ■

(*Proof of Equation* (3.45).) By the previous proof, we note that $\frac{\partial f(\mu,\sigma)}{\partial \mu} \in (0, 1) \implies \left|\frac{\partial f(\mu,\sigma)}{\partial \mu}\right| \leq 1$ for each $\mu > 0$. It follows by the mean value theorem, that $f(\mu, \sigma)$ is 1-Lipschitz as required. ■

(*Proof of Equation* (3.46).) Observe that from Equation (3.16) we have that $g(\mu, \sigma) \coloneqq \mathbb{V}(T) = \mu^2 + \sigma^2 - f(\mu, \sigma)^2$. Since $\mathbb{V}(T) \geq 0$, it follows that $f(\mu, \sigma)^2 \leq \mu^2 + \sigma^2$ for each $\mu \in \mathbb{R}$, as required. ■

(*Proof of Equation* (3.47).) Let $i \in [n]$ be arbitrary. Now note that by the Equation (3.45) property it follows that , we then have that:

$$\begin{aligned}
V(f, \boldsymbol{\mu}, \sigma) &\coloneqq \sum_{i=1}^{n-1} |f(\mu_{i+1}, \sigma) - f(\mu_i, \sigma)| && \text{(by definition)} \\
&= \sum_{i=1}^{n-1} f(\mu_{i+1}, \sigma) - f(\mu_i, \sigma) \\
&\qquad \text{(using Equation (3.43) and monotonicity of } \mu_1 \leq \ldots \leq \mu_n) \\
&= f(\mu_n, \sigma) - f(\mu_1, \sigma) && \text{(by telescoping sum)} \\
&\leq |\mu_n - \mu_1| && \text{(using Equation (3.45))} \\
&= \mu_n - \mu_1 && \text{(by monotonicity of } \mu_1 \leq \ldots \leq \mu_n)
\end{aligned}$$

as required. ■

Thus all properties specified in Equations (3.41) to (3.47) are now proved. □

3.A.5 Properties of the folded normal variance: $g(\mu, \sigma)$

Lemma 3.26 (Properties of the Folded Normal variance)**.** *Let* $T \sim \mathsf{FoldNorm}\left(\mu, \sigma^2\right)$ *per Definition 3.9, and let* $g(\mu, \sigma) \coloneqq \mathbb{V}(T)$*. Given this setup, and fixing* $\sigma > 0$*, we note the following properties of* $g(\mu, \sigma)$*:*

$$g(\mu, \sigma) \leq \sigma^2, \text{ for each } \mu \in \mathbb{R} \tag{3.48}$$
$$g(\mu, \sigma) \geq g(0, \sigma), \text{ for each } \mu \in \mathbb{R}_{>0} \tag{3.49}$$
$$\mathbb{V}\left(T^2\right) = 4\mu^2\sigma^2 + 2\sigma^4, \text{ for each } \mu \in \mathbb{R} \tag{3.50}$$

Proof of Lemma 3.26. We prove each properties specified in Equations (3.48) to (3.50) in turn.

(*Proof of Equation* (3.48).) We have for each $\mu \geq 0$

$$
\begin{aligned}
g(\mu, \sigma) &:= \mu^2 + \sigma^2 - f(\mu, \sigma)^2 && \text{(per Equation (3.16))} \\
&\leq \sigma^2 && \text{(since } f(\mu,\sigma)^2 \geq \mu^2 \text{ using Equation (3.42))}
\end{aligned}
$$

as required. ■

(*Proof of Equation* (3.49).) First note that $g(0, \sigma) = \sigma^2 - f(0, \sigma)^2 = \sigma^2 - \left(\frac{2}{\pi}\right)\sigma^2$. It then follows that:

$$
\begin{aligned}
& g(\mu, \sigma) \geq g(0, \sigma) \\
\iff & \mu^2 + \sigma^2 - f(\mu, \sigma)^2 \geq \sigma^2 - \left(\frac{2}{\pi}\right)\sigma^2 \\
\iff & \mu^2 + \left(\frac{2}{\pi}\right)\sigma^2 \geq f(\mu, \sigma)^2 \qquad (3.51)
\end{aligned}
$$

We will then prove the equivalent statement Equation (3.51). Since $\mu, \sigma > 0$ in our case, let $\nu := \frac{\mu}{\sigma} > 0$ in what follows. Then dividing both sides of Equation (3.51) by ν we obtain:

$$
\sqrt{\nu^2 + \frac{2}{\pi}} \geq \nu(2\Phi(\nu) - 1) + \sqrt{\frac{2}{\pi}}\, e^{-\frac{\nu^2}{2}}
$$

Let us then define

$$
g(\nu) := \sqrt{\nu^2 + \frac{2}{\pi}} - \nu(2\Phi(\nu) - 1) - \sqrt{\frac{2}{\pi}}\, e^{-\frac{\nu^2}{2}} \qquad (3.52)
$$

Taking the derivative of $g(\nu)$ with respect to ν we obtain:

$$
\begin{aligned}
g'(\nu) &= \frac{\nu}{\sqrt{\nu^2 + \frac{2}{\pi}}} - 2\Phi(\nu) + 1 - \frac{2\nu}{\sqrt{2\pi}} e^{-\frac{\nu^2}{2}} + \nu\sqrt{\frac{2}{\pi}}\, e^{-\frac{\nu^2}{2}} && (3.53) \\
&= \frac{\nu}{\sqrt{\nu^2 + \frac{2}{\pi}}} - 2\Phi(\nu) + 1 && (3.54)
\end{aligned}
$$

As such all global and local extrema are obtained by setting $g'(\nu) = 0$, that is:

$$g'(\nu) = 0 \tag{3.55}$$
$$\iff \frac{\nu}{\sqrt{\nu^2 + \frac{2}{\pi}}} = 2\Phi(\nu) - 1 \tag{3.56}$$

Now, $\nu = 0$ is a clear solution, at which our function is exactly equal to 0. Also, we need to look at $\nu = \infty$, where we also have an identity. So we need to take care of other possible roots to the equation $\frac{\nu}{\sqrt{\nu^2+\frac{2}{\pi}}} = 2\Phi(\nu) - 1$. Now observe that since when $\nu \geq 0$ the function $\Phi(\nu)$ is concave and therefore $2\Phi(\nu) - 1 = 2(\Phi(\nu) - \Phi(0)) \leq 2\nu\phi(0) = \sqrt{\frac{2}{\pi}}\nu$. Thus for any non-zero solution $\bar{\nu}$ to the equation $\frac{\nu}{\sqrt{\nu^2+\frac{2}{\pi}}} = 2\Phi(\nu) - 1$, we must have $\frac{\bar{\nu}}{\sqrt{\bar{\nu}^2+\frac{2}{\pi}}} \leq \bar{\nu}\sqrt{\frac{2}{\pi}}$.
This implies that $\bar{\nu}^2 \geq \frac{\pi}{2} - \frac{2}{\pi}$. Now, going back to the original function we need to show

$$\sqrt{\bar{\nu}^2 + \frac{2}{\pi}} \geq \bar{\nu}(2\Phi(\bar{\nu}) - 1) + \sqrt{\frac{2}{\pi}}e^{-\bar{\nu}^2/2} = \frac{\bar{\nu}^2}{\sqrt{\bar{\nu}^2 + \frac{2}{\pi}}} + \sqrt{\frac{2}{\pi}}e^{-\bar{\nu}^2/2}.$$

The latter is equivalent to $\bar{\nu}^2 + \frac{2}{\pi} \geq \bar{\nu}^2 + \sqrt{\frac{2}{\pi}}e^{-\bar{\nu}^2/2}\sqrt{\bar{\nu}^2 + \frac{2}{\pi}}$ which is equivalent to $\frac{2}{\pi}e^{\bar{\nu}^2} \geq \bar{\nu}^2 + \frac{2}{\pi}$. since the function $\nu \mapsto \frac{2}{\pi}e^{\nu^2} - \nu^2$ is increasing for positive ν it suffices to check that $\frac{2}{\pi}e^{\bar{\nu}^2} \geq \bar{\nu}^2 + \frac{2}{\pi}$ for $\bar{\nu} = \frac{\pi}{2} - \frac{2}{\pi}$ (since as we know from before $\bar{\nu}$ is at least that value). This is true, and completes the proof, as required. ■

(*Proof of Equation* (3.50).) By direct calculation we have:

$$\begin{aligned}
\mathbb{V}\left(T^2\right) &= \mathbb{E}\left(T^4\right) - \left(\mathbb{E}\left(T^2\right)\right)^2 \\
&= \mathbb{E}\left(R^4\right) - \left(\mathbb{E}\left(R^2\right)\right)^2 && (\text{since } T \overset{a.s.}{=} |R|) \\
&= (\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4) - (\mu^2 + \sigma^2)^2 && (2^{\text{nd}}, 4^{\text{th}} \text{ moments of } \mathcal{N}(\mu, \sigma^2)) \\
&= 4\mu^2\sigma^2 + 2\sigma^4
\end{aligned}$$

as required. ■

Thus all properties specified in Equations (3.48) to (3.50) are now proved. □

3.A.6 Properties of the inverse folded normal mean: $f^{-1}(\mu, \sigma)$

Lemma 3.27 (Properties of the Folded Normal mean inverse)**.** *Suppose* $R \sim \mathcal{N}(\mu, \sigma^2)$*. Let* $T \overset{a.s.}{=} |R|$*, then* $T \sim \mathsf{FoldNorm}\left(\mu, \sigma^2\right)$ *per* Definition 3.9*. We denote the mean of the Folded Normal distribution by* $f(\mu, \sigma) \coloneqq \mathbb{E}\left(T\right)$*. Given this setup, and fixing* $\sigma > 0$*, we note the following important properties of* $f^{-1}(u, \sigma)$ *(which denotes the inverse with respect to* μ *function of* $f(\mu, \sigma)$ *when* σ *is held fixed):*

$$f^{-1}(u, \sigma) \text{ exists,} \tag{3.57}$$

$$\frac{\partial}{\partial \sigma} f^{-1}(u, \sigma) = -\frac{\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2))}{2\Phi(\mu/\sigma) - 1}, \tag{3.58}$$

$$\frac{\partial}{\partial u} f^{-1}(u, \sigma) = 1/(2\Phi(\mu/\sigma) - 1), \tag{3.59}$$

$$f^{-1}(u, \sigma) \text{ is a Lipschitz function for each } u > f(\eta, \sigma) > 0 \text{ for a fixed } \sigma, \tag{3.60}$$

where in the above $u = f(\mu, \sigma)$ *(or in other words* $\mu = f^{-1}(u, \sigma)$*).*

Proof of Lemma 3.27. We prove each properties specified in Equations (3.57) to (3.60) in turn.

(*Proof of Equation* (3.57).) Note that for a fixed $\sigma > 0$ the function $f(\mu, \sigma)$ is invertible (as it is increasing, per Lemma 3.25), as required. ■

(*Proof of Equation* (3.58).) In order to find the derivative of $\frac{\partial}{\partial \sigma} f^{-1}(\cdot, \sigma)$, we can parametrize as follows:

$$u = f(\mu, \sigma) \tag{3.61}$$

$$v = \sigma \tag{3.62}$$

We will use the inverse function theorem which says that under certain conditions $\mu = F(u, v) = F(u, \sigma)$ and $\sigma = G(u, v) = v$, for some functions F and G. Note that for a fixed $\sigma > 0$ the function $f(\mu, \sigma)$ is invertible (per Equation (3.57)). Thus

$$\frac{\partial}{\partial \sigma} f^{-1}(u, \sigma) = \frac{\partial \mu}{\partial \sigma} = \frac{\partial F(u, v)}{\partial v} = -\frac{\frac{\partial u}{\partial \sigma}}{J} = -\frac{\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2))}{J} \tag{3.63}$$

Where J is the Jacobian of the transformation

$$
\begin{aligned}
J &= \begin{vmatrix} \frac{\partial u}{\partial \mu} & \frac{\partial u}{\partial \sigma} \\ \frac{\partial v}{\partial \mu} & \frac{\partial v}{\partial \sigma} \end{vmatrix} \\
&= 2\Phi(\mu/\sigma) - 1 \\
&> 0 && \text{(since } \mu \geq \eta > 0\text{)}
\end{aligned}
$$

It follows that

$$\frac{\partial}{\partial \sigma} f^{-1}(u, \sigma) = -\frac{\sqrt{2/\pi}\exp(-\mu^2/(2\sigma^2))}{2\Phi(\mu/\sigma) - 1} \tag{3.64}$$

As required. ■

(*Proof of Equation* (3.59).) We similarly evaluate the derivative $\frac{\partial}{\partial u} f^{-1}(u, \sigma)$ as follows:

$$
\begin{aligned}
\frac{\partial}{\partial u} f^{-1}(u, \sigma) &= \frac{\partial}{\partial u}\mu && (3.65) \\
&= \frac{\frac{\partial v}{\partial \sigma}}{J} && (3.66) \\
&= \frac{1}{2\Phi(\mu/\sigma) - 1} && (3.67)
\end{aligned}
$$

As required. ■

(*Proof of Equation* (3.60).) We note that Equation (3.59) implies that

$$\frac{\partial}{\partial u} f^{-1}(u, \sigma) \leq \frac{1}{2\Phi(\eta/\sigma) - 1} \tag{3.68}$$

since $\mu \geq \eta > 0$ under our setting. In this case, this holds for each $u > f(\eta, \sigma) > 0$, for a fixed σ. Since this derivative is bounded by this constant, it follows that $f^{-1}(u, \sigma)$ is $\frac{1}{2\Phi(\eta/\sigma)-1}$-Lipschitz by applying Lemma 3.22. ■

Thus all properties specified in Equations (3.57) to (3.60) are now proved. □

3.A.7 Properties of: $J(\sigma)$

Definition 3.28 ($J(\sigma)$)**.** Let $\eta > 0$ be fixed, and $\sigma \geq 0$ per Equations (3.4) and (3.5), respectively. We define the function, $J : \mathbb{R}_{\geq 0} \to \mathbb{R}$, as:

$$J(\sigma) := \begin{cases} 0 & \text{if } \sigma = 0 \\ \sigma\left(\frac{1}{2} - \frac{\eta/\sigma\phi(\eta/\sigma)}{2\Phi(\eta/\sigma)-1}\right) & \text{otherwise} \end{cases} \tag{3.69}$$

In order to prove the key properties of $J(\sigma)$, we will first need to prove a useful result in Lemma 3.29.

Lemma 3.29. *We define the function, $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, as:*

$$M(x) \coloneqq \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{x\phi(x)}{2\Phi(x)-1} & \text{otherwise.} \end{cases} \tag{3.70}$$

Note that $M(0) = \frac{1}{2}$, by Equation (3.37). Then $M(x)$ is strictly decreasing for each $x > 0$.

Proof of Lemma 3.29. In order to show that $M(x)$ is decreasing for each $x > 0$, we will show that $M'(x) < 0$ for each $x > 0$. To see this, first observe that:

$$\begin{aligned} M'(x) &= \frac{(2\Phi(x)-1)\left(\phi(x)+x\phi'(x)\right)-2x\phi^2(x)}{(2\Phi(x)-1)^2} \\ &= \frac{(2\Phi(x)-1)\left(\phi(x)-x^2\phi(x)\right)-2x\phi^2(x)}{(2\Phi(x)-1)^2} \quad (\text{since } \phi'(x)+x\phi(x)=0) \\ &= \frac{\phi(x)}{(2\Phi(x)-1)^2}\left((2\Phi(x)-1)\left(1-x^2\right)-2x\phi(x)\right) \\ &= \frac{\phi(x)}{(2\Phi(x)-1)^2}\left(2\left(\Phi(x)-\frac{1}{2}\right)\left(1-x^2\right)-2x\phi(x)\right). \end{aligned} \tag{3.71}$$

Now we see that:

$$2\left(\Phi(x)-\frac{1}{2}\right) = 2\left(\Phi(x)-\Phi(0)\right) = 2\int_0^x \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,\mathrm{d}t = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}\,\mathrm{d}t \leq \frac{2x}{\sqrt{2\pi}}, \tag{3.72}$$

where the last inequality in Equation (3.72) followed from the fact that $e^{-\frac{t^2}{2}} \leq 1$ for each $t \geq 0$. It then follows that:

$$\begin{aligned} \left(2\left(\Phi(x)-\frac{1}{2}\right)\left(1-x^2\right)-2x\phi(x)\right) &= \left(\frac{2}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}\,\mathrm{d}t\right)\left(1-x^2\right)-\frac{2x}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \\ &\leq \frac{2x}{\sqrt{2\pi}}\left(1-x^2\right)-\frac{2x}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \\ &\qquad\qquad (\text{using Equation (3.72)}) \\ &< 0, \end{aligned} \tag{3.73}$$

Where Equation (3.73) followed by observing that since $1-x^2 < 1-\frac{x^2}{2} < e^{-\frac{x^2}{2}}$ for each $x > 0$. Now since $\frac{\phi(x)}{(2\Phi(x)-1)^2} > 0$ for each $x > 0$, we have by applying Equation (3.73) to Equation (3.72) that $M'(x) < 0$, for each $x > 0$, as required. ☐

Lemma 3.30 (Properties of $J(\sigma)$)**.** *Let $J(\sigma)$ be defined as per Equation (3.69). Then $J(\sigma)$ satisfies the following properties:*

$$J(\sigma) > 0 \textit{ for each } \sigma \in \mathbb{R}_{>0} \textit{ and 0 if and only if } \sigma = 0 \tag{3.74}$$

$$J(\sigma) \textit{ is continuous for each } \sigma \in \mathbb{R}_{>0} \tag{3.75}$$

$$\textit{For any } 0 < \sigma_1 < \sigma_2, \min_{\sigma \in [\sigma_1, \sigma_2]} J(\sigma) \geq \sigma_1 \left(\frac{1}{2} - \frac{\eta/\sigma_2 \phi(\eta/\sigma_2)}{2\Phi(\eta/\sigma_2) - 1} \right) > 0 \tag{3.76}$$

Proof of Lemma 3.30. We prove each property (Equations (3.74) to (3.76)) in turn. Throughout these proofs, we write:

$$J(\sigma) \coloneqq J_1(\sigma) J_2(\sigma), \text{ where } J_1(\sigma) \coloneqq \sigma, \text{ and } J_2(\sigma) \coloneqq \frac{1}{2} - \frac{\eta/\sigma \phi(\eta/\sigma)}{2\Phi(\eta/\sigma) - 1} \tag{3.77}$$

(*Proof of Equation* (3.74).) Observe that both $J_1(\sigma), J_2(\sigma)$ are zero if and only if $\sigma = 0$. In the case of $J_2(\sigma)$ this follows from Lemma 3.23. Now for $\sigma > 0$, $J_1(\sigma) \coloneqq \sigma > 0$, by assumption. And the fact that $J_2(\sigma) > 0$, for $\sigma > 0$ again follows directly from Lemma 3.23. As such, $J(\sigma) > 0$ for each $\sigma > 0$, since it is the product of two strictly positive functions over this support, as required. ■

(*Proof of Equation* (3.75).) $J_1(\sigma)$ is continuous for $\sigma > 0$. Moreover since $\phi(x), \Phi(x)$ for a standard normal are continuous over their support, $\mathbb{R}$, it follows that $J_2(\sigma)$ is also continuous for $\sigma > 0$. As such, $J(\sigma)$ is continuous for each $\sigma > 0$, since it is the product of two continuous functions, as required. ■

(*Proof of Equation* (3.76).) Note that for any two fixed σ_1, σ_2, such that $0 < \sigma_1 < \sigma_2$, the interval $[\sigma_1, \sigma_2]$ is compact. From Equation (3.75), we know that $J(\sigma)$ is continuous for $\sigma > 0$, and so it attains it's minimum (and maximum) on this interval. Moreover, from Equation (3.74), it follows that $\min_{\sigma \in [\sigma_1, \sigma_2]} J(\sigma) > 0$. Now we note that $\sigma \mapsto J_1(\sigma) \coloneqq \sigma$, is increasing in σ. Moreover for each $\sigma > 0$, we have that $J_2(\sigma) \coloneqq \frac{1}{2} - M(\frac{\eta}{\sigma})$, where the function M is as defined in Equation (3.70). Moreover it follows from Lemma 3.29 that $J_2(\sigma)$ is strictly decreasing for each $\sigma > 0$. By the non-negativity of $J(\sigma)$ over it's domain, we have that $J(\sigma) > J_1(\sigma_1) J_2(\sigma_2)$ for each $\sigma \in [\sigma_1, \sigma_2]$. From this we have that $\min_{\sigma \in [\sigma_1, \sigma_2]} J(\sigma) \geq J_1(\sigma_1) J_2(\sigma_2) = \sigma_1 \left(\frac{1}{2} - \frac{\eta/\sigma_2 \phi(\eta/\sigma_2)}{2\Phi(\eta/\sigma_2) - 1} \right) > 0$, as required. ■

Thus all properties specified in Equations (3.74) to (3.76) are now proved. □

3.A.8 Properties of: $G(\sigma)$

Definition 3.31 ($G(\sigma)$)**.** Under the setup of ASCI generating process per Definition 3.2, and per the `ASCIFIT` model we define the function, $G : \mathbb{R}_{\geq 0} \to \mathbb{R}$, as:

$$G(\sigma) \coloneqq \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta, \sigma), \sigma))^2 \tag{3.78}$$

Lemma 3.32 (Properties of $G(\sigma)$)**.** *Under the setup of ASCI generating process per Definition 3.2, and with $G(\sigma)$ defined as per Definition 3.31, we note the following important properties of $G(\sigma)$:*

$$\frac{\partial}{\partial\sigma}G(\sigma) = \frac{4}{n}\sum_{i=1}^{n}\sigma\left(\frac{1}{2} - \frac{f^{-1}(\widehat{T}_i, \sigma)/\sigma\phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma)\mathbb{1}(\widehat{T}_i \geq f(\eta, \sigma))}{2\Phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma) - 1}\right), \tag{3.79}$$

$$G(\sigma) \text{ is increasing for } \sigma \geq 0, \text{ and strictly increasing for } \sigma > 0. \tag{3.80}$$

Proof of Lemma 3.32. We prove each property (Equations (3.79) and (3.80)) in turn. Throughout these proofs, $J(\sigma)$ is as defined in Definition 3.28, and $G(\sigma)$ is as defined in Definition 3.31.

(*Proof of Equation* (3.79).). Using the definition, we have:

$$\begin{aligned}
&\frac{\partial}{\partial\sigma}G(\sigma) \\
&= 2\sigma - \frac{2}{n}\sum_{i=1}^{n}\frac{\sqrt{2/\pi}\,f^{-1}(\widehat{T}_i, \sigma)\exp(-f^{-1}(\widehat{T}_i, \sigma)^2/(2\sigma^2))\mathbb{1}(\widehat{T}_i \geq f(\eta, \sigma))}{2\Phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma) - 1} \\
&\hspace{20em}\text{(using Equation (3.58))} \\
&= 2\sigma - \frac{4\sigma}{n}\sum_{i=1}^{n}\frac{f^{-1}(\widehat{T}_i, \sigma)/\sigma\phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma)\mathbb{1}(\widehat{T}_i \geq f(\eta, \sigma))}{2\Phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma) - 1} \qquad (3.81) \\
&= \frac{4}{n}\sum_{i=1}^{n}\sigma\left(\frac{1}{2} - \frac{f^{-1}(\widehat{T}_i, \sigma)/\sigma\phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma)\mathbb{1}(\widehat{T}_i \geq f(\eta, \sigma))}{2\Phi(f^{-1}(\widehat{T}_i, \sigma)/\sigma) - 1}\right)
\end{aligned}$$

as required. ■

(*Proof of Equation* (3.80).). Now per Lemma 3.23, we have that $x \mapsto x\phi(x)/(2\Phi(x)-1) \leq 1/2$ for all $x \geq 0$, and moreover it is decreasing for $x > 0$. Therefore the derivative, $\frac{\partial}{\partial\sigma}G(\sigma)$, is bounded from below by 0. As such $G(\sigma)$ is increasing in σ, for $\sigma \geq 0$. In fact, since $\eta, \sigma > 0$, it follows that $\frac{\eta}{\sigma} > 0$. In turn, we have that $\frac{\partial}{\partial\sigma}G(\sigma)$ is bounded from below by $J(\sigma) \coloneqq \sigma\left(\frac{1}{2} - \frac{\eta/\sigma\phi(\eta/\sigma)}{2\Phi(\eta/\sigma)-1}\right) > 0$, for each $\sigma > 0$, using Lemma 3.30. It follows that $G(\sigma)$ is *strictly* increasing in σ, for $\sigma > 0$, as required. ■

Thus all properties specified in Equations (3.79) and (3.80) are now proved. □

3.B Proofs of Section 3.1

3.B.1 Mathematical Preliminaries

Lemma 3.33 (Symmetrization with Rademacher random variables)**.** *Suppose that ε is a symmetric distribution i.e. $\varepsilon \overset{d}{=} -\varepsilon$, $\xi \sim \mathsf{Rademacher}(\alpha)$, with $\alpha \in [0, 1]$. If $\xi \perp\!\!\!\perp \varepsilon$ then $\xi\varepsilon \overset{d}{=} \varepsilon$.*

Proof of Lemma 3.33. Let us define $Q \coloneqq \xi\varepsilon$. We then have the following:

$$\begin{aligned}
\mathbb{P}(Q \geq q) &\coloneqq \mathbb{P}(\xi\varepsilon \geq q) && \text{(since } Q \coloneqq \xi\varepsilon.\text{)} \\
&= \mathbb{P}(\xi\varepsilon \geq q \,|\, \xi = -1)\,\mathbb{P}(\xi = -1) + \mathbb{P}(\xi\varepsilon \geq q \,|\, \xi = 1)\,\mathbb{P}(\xi = 1) \\
& && \text{(since } \xi \sim \mathsf{Rademacher}(\alpha).\text{)} \\
&= \mathbb{P}(-\varepsilon \geq q)(1 - \alpha) + \mathbb{P}(\varepsilon \geq q)(\alpha) && \text{(since } \xi \perp\!\!\!\perp \varepsilon.\text{)} \\
&= \mathbb{P}(\varepsilon \geq q)(1 - \alpha) + \mathbb{P}(\varepsilon \geq q)(\alpha) && \text{(since } \varepsilon \overset{d}{=} -\varepsilon.\text{)} \\
&= \mathbb{P}(\varepsilon \geq q)(1 - \alpha + \alpha) \\
&= \mathbb{P}(\varepsilon \geq q)
\end{aligned}$$

So we have that $Q \coloneqq \xi\varepsilon \overset{d}{=} \varepsilon$, as required. □

The setting can be simplified if the adversary chooses the sign-corruptions independent of the error terms. To see this, first note that ε_i are centered (i.e. symmetric) Gaussian random variables. Now, if the $(\xi_1, \ldots, \xi_n)$ are picked independently from $(\varepsilon_1, \ldots, \varepsilon_n)$, the ASCI generating process response reduces to $R_i = \xi_i\mu_i + \varepsilon_i$. That is our setting encompasses this more simplified setting, and is shown formally in Corollary 3.34. Further, we note that in the case where $\xi_i \overset{a.s.}{=} 1$ then and $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$, then this is equivalent to the standard univariate isotonic regression setup.

Corollary 3.34. *In the case where $\xi_i \perp\!\!\!\perp \varepsilon_i$, for each $i \in [n]$ we have that the ASCI generating process simplifies to $R_i = \xi_i\mu_i + \varepsilon_i$.*

Proof of Corollary 3.34. We note that the underlying adversarial generating process is given by $R_i = \xi_i(\mu_i + \varepsilon_i) = \xi_i\mu_i + \xi_i\varepsilon_i$, for each $i \in [n]$. Now since $\xi_i \perp\!\!\!\perp \varepsilon_i$ we have by applying Lemma 3.33 for each $i \in [n]$ that $\xi_i\varepsilon_i \overset{\text{i.i.d.}}{\sim} \varepsilon_i$. And so the required adversarial model can be written as $R_i = \xi_i\mu_i + \varepsilon_i$, as required. □

3.B.2 Important Model Definitions

First, we formally (redefine) the generating model described in Example 3.5.

Definition 3.35 (Two-component Gaussian mixture ASCI special case from Example 3.5)**.** We consider n observations, $\{R_i \mid i \in [n]\}$, where each observation R_i is generated from the following model:

$$\begin{aligned} R_i &= \xi_i \mu_i + \varepsilon_i && (3.82)\\ \text{s.t. } 0 &< \eta \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n && (3.83)\\ \text{and } \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (3.84)\\ \text{and } \xi_i &\overset{\text{i.i.d.}}{\sim} \mathsf{Rademacher}(p),\ p \in (0,1),\ \text{ and } \xi_i \perp\!\!\!\perp \varepsilon_i && (3.85) \end{aligned}$$

Second, we formally define the generating model described in Example 3.7.

Definition 3.36 (Non-convex generating model from Example 3.7)**.** We consider n observations, $\{R_i \mid i \in [n]\}$, where each observation R_i is generated from the following model:

$$\begin{aligned} R_i &= \gamma_i + \varepsilon_i && (3.86)\\ \text{s.t. } 0 &< \eta \leq |\gamma_1| \leq |\gamma_2| \leq \ldots \leq |\gamma_n| && (3.87)\\ \text{and } \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (3.88) \end{aligned}$$

Remark 3.37. From a simulation perspective, each γ_i is generated first subject to Equation (3.87), then ξ_i is sampled independently, and both are added to give each response R_i.

Third, we introduce an alternative model as per Definition 3.38.

Definition 3.38 (Alternative non-convex model)**.**

$$\begin{aligned} R_i &= \xi_i a_i + \varepsilon_i && (3.89)\\ \text{s.t. } 0 &< \eta \leq a_1 \leq a_2 \leq \ldots \leq a_n && (3.90)\\ \text{and } \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) && (3.91)\\ \text{and } \xi_i &= \operatorname{sgn}\left(\gamma_i\right) && (3.92)\\ \text{and } \xi_i &\perp\!\!\!\perp \varepsilon_i && (3.93)\\ \text{and } a_i &= |\gamma_i| && (3.94) \end{aligned}$$

Finally, for convenience we recall Definition 3.2 as follows.

Definition 3.2 (Adversarial sign-corrupted isotonic (ASCI) regression)**.** We consider n observations, $\{R_i \mid i \in [n]\}$, where each observation R_i is generated from the following model:

$$R_i = \xi_i(\mu_i + \varepsilon_i) \tag{3.4}$$
$$\text{s.t. } 0 < \eta \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n \tag{3.5}$$
$$\text{and } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{3.6}$$
$$\text{and } \xi_i \in \{-1, 1\} \tag{3.7}$$

3.B.3 Formal justification for Example 3.5

Proposition 3.39 (Justification for Example 3.5)**.** *Under the model generating process described in Example 3.5 (i.e., per Definition 3.35), the following model definition inclusion holds.*

$$\text{Definition 3.35} \subseteq \text{Definition 3.2} \tag{3.95}$$

Remark 3.40. Here, each definitional inclusion is to be read as the former generating model definition being a special case of the latter generating model definition.

Proof of Example 3.5. Our basic strategy is to show each model inclusion in turn.

(Definition 3.35 $\subseteq$ Definition 3.2). Observe that Equations (3.83) and (3.84) are definitionally equivalent to Equations (3.5) and (3.6), respectively. Moreover we have that Equation (3.85) is a special case of Equation (3.7). Finally, from Equation (3.85) we have that $\xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{Rademacher}(p)$, $p \in (0, 1)$, and $\xi_i \perp\!\!\!\perp \varepsilon_i$. Thus from Corollary 3.34, it follows that Equation (3.82) is a special case of Equation (3.82).

In summary we have shown that Equation (3.95) holds, from which it follows that Example 3.5 (or equivalently Definition 3.35) is a special case of Definition 3.2, as required. □

3.B.4 Formal justification for Example 3.7

We now provide a formal Formal justification that Example 3.7 is a special case of the generating process described in Definition 3.2.

Lemma 3.41 (Justification for Example 3.7)**.** *Under the model generating process described in Example 3.7, the following model definition inclusion holds.*

$$\text{Definition 3.36} = \text{Definition 3.38} \subseteq \text{Definition 3.2} \tag{3.96}$$

Remark 3.42. As with Proposition 3.39, each definitional inclusion is to be read as the former generating model definition being a special case of the latter generating model definition. In the case of equivalence, we note that that both inclusions hold between the model definitions.

Proof of Example 3.7. Our basic strategy is to show each model inclusion in turn.

(Definition 3.36 = Definition 3.38) This follows by construction. Observe that Equations (3.92) and (3.94) imply that $\xi_i a_i = \operatorname{sgn}(\gamma_i)\,|\gamma_i| = \gamma_i$, so that Equations (3.86) and (3.89) are equivalent. In addition from Equation (3.94), we have that $a_i = |\gamma_i|$ and thus Equations (3.87) and (3.90) are equivalent, as are Equations (3.88) and (3.91). As such the equality is established between the two generating model definitions.

(Definition 3.38 $\subseteq$ Definition 3.2). Observe that by Equations (3.6) and (3.91) are definitionally equivalent. Observe from Equation (3.92) that $\xi_i = \operatorname{sgn}(\gamma_i) \in \{-1, 1\}$ which is a special case of Equation (3.7). For each observation $i \in [n]$ using Equation (3.94) that by setting $a_i \coloneqq \mu_i$ that Equations (3.5) and (3.90) are equivalent. Finally since $\xi_i \perp\!\!\!\perp \varepsilon_i$ from Equation (3.93), we note that Equation (3.89) is a special case of Equation (3.4) by applying Corollary 3.34 to the observation R_i, for each $i \in [n]$.

In summary we have shown that Equation (3.96) holds, from which it follows that Definition 3.36 is a special case of Definition 3.2, as required. $\square$

3.C Proofs of Section 3.2

3.C.1 Mathematical Preliminaries

Theorem 3.43 (Projection onto the nonnegative monotone cone)**.** *Suppose that $\mathcal{S}^{\uparrow} \subseteq \mathbb{R}^n$ is the monotone cone, that is,*

$$\mathcal{S}^{\uparrow} \coloneqq \left\{ \boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^{\top} \in \mathbb{R}^n \,\middle|\, \mu_1 \leq \ldots \leq \mu_n \right\}.$$

and $\mathcal{S}_{+}^{\uparrow} \subseteq \mathbb{R}^n$ is the nonnegative monotone cone, that is,

$$\mathcal{S}_{+}^{\uparrow} \coloneqq \left\{ \boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^{\top} \in \mathcal{S}^{\uparrow} \,\middle|\, \mu_1 \geq 0 \right\}.$$

Then for an arbitrary $\mathbf{v} \in \mathbb{R}^n$ it holds that

$$\Pi_{\mathcal{S}_{+}^{\uparrow}}(\mathbf{v}) = \left(\Pi_{\mathcal{S}^{\uparrow}}(\mathbf{v})\right)^{+},$$

where for any $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z}^{+} \in \mathbb{R}^n$ stands for the lattice operation defined by the order induced by the nonnegative orthant in $\mathbb{R}^n$. That is, we define the operation componentwise as $\left(\mathbf{z}^{+}\right)_i \coloneqq (\mathbf{z})_i \vee 0$ for each component index $i \in [n]$.

Proof of Theorem 3.43. See Németh and Németh (2012, Corollary 1) for details. □

Remark 3.44. In effect, Theorem 3.43 basically states that in order to project onto the nonegative monotone cone, K, one can instead first project onto the monotone cone, W, first, and then take the non-negative part along each component. This is useful, since one can leverage algorithms like the PAVA which already efficiently handle projection onto the unrestricted monotone cone, W.

3.C.2 Proof of Proposition 3.11

Proposition 3.11 (Reason for the "$\vee f(\eta, \sigma)$"-correction in **Step II**)**.** *The need for defining the $\vee f(\eta, \sigma)$ in Equation (3.17) in* **Step II** *in* ***ASCIFIT****, is that the solution to the problem*

$$\underset{\widetilde{T}_1, \ldots, \widetilde{T}_n}{\arg\min} \sum_{i=1}^{n} (T_i - \widetilde{T}_i)^2 \text{ s.t. } f(\eta, \sigma) \leq \widetilde{T}_1 \leq \ldots \leq \widetilde{T}_n, \tag{3.19}$$

is related to the solution to

$$\underset{\widehat{T}_1, \ldots, \widehat{T}_n}{\arg\min} \sum_{i=1}^{n} (T_i - \widehat{T}_i)^2 \text{ s.t. } \widehat{T}_1 \leq \ldots \leq \widehat{T}_n, \tag{3.20}$$

as $\widetilde{T}_i \coloneqq \widehat{T}_i \vee f(\eta, \sigma)$.

Proof of Proposition 3.11. This follows along the following lines. First subtract $f(\eta,\sigma)$ from all $\widetilde{T}_i$ to bring the first problem to

$$\arg\min_{\bar{T}_i} \sum_{i=1}^{n} ((T_i - f(\eta,\sigma)) - \bar{T}_i)^2 \text{ s.t. } 0 \leq \bar{T}_1 \leq \ldots \leq \bar{T}_n, \tag{3.97}$$

where $\bar{T}_i = \widetilde{T}_i - f(\eta,\sigma)$. Now the solution to the unrestricted problem

$$\arg\min_{T_i^*} \sum_{i=1}^{n} ((T_i - f(\eta,\sigma)) - T_i^*)^2 \text{ s.t. } T_1^* \leq \ldots \leq T_n^*,$$

is $T_i^* = \widehat{T}_i - f(\eta,\sigma)$. Next we apply Theorem 3.43, we see that $\bar{T}_i = T_i^* \vee 0$, so that $\widetilde{T}_i = \bar{T}_i + f(\eta,\sigma) = T_i^* \vee 0 + f(\eta,\sigma) = (T_i^* + f(\eta,\sigma)) \vee f(\eta,\sigma) = \widehat{T}_i \vee f(\eta,\sigma)$ which is what we wanted to show. $\square$

3.D Proofs of Section 3.3

3.D.1 Mathematical Preliminaries

The key idea to prove this theorem here is to apply (Zhang, 2002, Theorem 2.2(ii)) to our specific setting. To ensure our work is self-contained, we translate this result into the notation of our paper:

Theorem 3.45 (Theorem 2.2 (ii) (Zhang, 2002))**.** *Let* $R_{n,p}(f, \boldsymbol{\mu}, \sigma, \sigma_p) \coloneqq \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\left|\widehat{T}_i - f(\mu_i, \sigma)\right|^p\right)\right)^{\frac{1}{p}}$*. Let* $\delta_i \coloneqq T_i - f(\mu_i, \sigma)$ *be independent, with* $\mathbb{E}(\delta_i) = 0$ *and* $\mathbb{E}\left(|\delta_i|^{p\vee 2}\right) \leq \sigma_p^{p\vee 2}$*,* $p \geq 1$ *then:*

$$R_{n,p}(f, \boldsymbol{\mu}, \sigma, \sigma_p) \leq 2^{\frac{1}{p}} \sigma_p C_p \min\left[1, \frac{3}{2}\left\{\frac{3}{(3-p)_+}\left(\frac{V(f, \boldsymbol{\mu}, \sigma)}{n\sigma_p C_p}\right)^{\frac{p}{3}} + \frac{1}{n}\int_0^n \frac{dx}{(x \vee 1)^{\frac{p}{2}}}\right\}^{\frac{1}{p}}\right] \tag{3.98}$$

where C_p *are constants depending on* p *only in general.*

Proof of Theorem 3.45. See Zhang (2002, Theorem 2.2(ii)) for details. Note that to translate between our notation and theirs respectively, we have $T_i \equiv y_i, \widehat{T}_i \equiv \widehat{f}_n(t_i), f(\mu_i, \sigma) \equiv f(t_i), \delta_i \equiv \varepsilon_i$ for each $i \in [n]$. ☐

Corollary 3.46 (Upper bound for $R_{n,2}^2(f, \boldsymbol{\mu}, \sigma, \sigma_2)$)**.** *In our setting, define* $X \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{T}_i - f(\mu_i, \sigma)\right)^2$*. We then have:*

$$\begin{aligned}\mathbb{E}(X) &\leq \min\left[2\sigma^2 C_2^2, \frac{27}{4}\left(\frac{\mu_n - \mu_1}{n}\right)^{\frac{2}{3}}(\sigma C_2)^{\frac{4}{3}} + \frac{2\sigma^2 C_2^2}{n}(1 + \log n)\right] \\ &\eqqcolon r_{n,2}(\mu_n, \mu_1, \sigma)\end{aligned} \tag{3.99}$$

where C_2 *is a constant.*

Proof of Corollary 3.46. Since $X \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{T}_i - f(\mu_i, \sigma)\right)^2$, then $X = R_{n,2}^2(f, \boldsymbol{\mu}, \sigma, \sigma_2)^2$, by definition in the setting of Theorem 3.45, assuming the relevant sufficient conditions are met. We now need to check the sufficient condition for Theorem 3.45. Here we have, for each $i \in [n]$, that $\delta_i \coloneqq T_i - f(\mu_i, \sigma)$. Note that by definition $\mathbb{E}(\delta_i) = \mathbb{E}(T_i) - f(\mu_i, \sigma) = 0$. We observe that $(\delta_1, \ldots, \delta_n)$ are independent since the original responses, i.e. $(R_1, \ldots, R_n)$ are independent by assumption. And taking absolute values and centering are measurable transformations which preserve their independence. We note that as per Zhang (2002, Theorem 2.2(ii)), we are required to check the sufficient condition $\mathbb{E}\left(|\delta_i|^{p\vee 2}\right) \leq \sigma_p^{p\vee 2}$. In our case, with $p=2$, this is equivalent to showing that $\mathbb{E}\left(\delta_i^2\right) \leq \sigma_2^2$. Then for each $i \in [n]$ we have:

$$
\begin{aligned}
\mathbb{E}\left(|\delta_i|^{p\vee 2}\right) &= \mathbb{E}\left(\delta_i^2\right) && \text{(since } p=2.\text{)} \\
&= \mathbb{V}\left(T_i\right) && \text{(since } \delta_i \text{ are mean centered } T_i \text{ values.)} \\
&=: g(\mu_i, \sigma) && \text{(by definition.)} \\
&\leq \sigma^2 && \text{(using Equation (3.48))} \\
&=: \sigma_2^2 && (3.100)
\end{aligned}
$$

As required, by defining $\sigma_2 \coloneqq \sigma$. So we meet this sufficient condition. Additionally observe that

$$
\begin{aligned}
\int_0^n \frac{dx}{(x\vee 1)} &= \int_0^1 \frac{dx}{(x\vee 1)} + \int_1^n \frac{dx}{(x\vee 1)} && \text{(by truncation)} \\
&= \int_0^1 dx + \int_1^n \frac{dx}{x} \\
&= 1 + \log n && (3.101)
\end{aligned}
$$

Now, in our setting note that $V(f, \boldsymbol{\mu}, \sigma) \leq \mu_n - \mu_1$ using Equation (3.47), it follows that:

$$
\begin{aligned}
\mathbb{E}\left(X\right) &\coloneqq R_{n,2}^2(f, \boldsymbol{\mu}, \sigma, \sigma_2) && \text{(by definition.)} \\
&\leq \min\left\{2\sigma_2^2 C_2^2, \frac{27}{4}\left(\frac{\mu_n - \mu_1}{n}\right)^{\frac{2}{3}} (\sigma_2 C_2)^{\frac{4}{3}} + \frac{2\sigma_2^2 C_2^2}{n} \int_0^n \frac{dx}{(x\vee 1)}\right\} \\
& && \text{(setting } p=2 \text{ in Theorem 3.45.)} \\
&= \min\left\{2\sigma_2^2 C_2^2, \frac{27}{4}\left(\frac{\mu_n - \mu_1}{n}\right)^{\frac{2}{3}} (\sigma_2 C_2)^{\frac{4}{3}} + \frac{2\sigma_2^2 C_2^2}{n} (1 + \log n)\right\} \\
& && \text{(using Equation (3.101))} \\
&= \min\left\{2\sigma^2 C_2^2, \frac{27}{4}\left(\frac{\mu_n - \mu_1}{n}\right)^{\frac{2}{3}} (\sigma C_2)^{\frac{4}{3}} + \frac{2\sigma^2 C_2^2}{n} (1 + \log n)\right\} \\
& && \text{(since } \sigma_2 \coloneqq \sigma \text{ per Equation (3.100))} \\
&=: r_{n,2}(\mu_n, \mu_1, \sigma) && (3.102)
\end{aligned}
$$

as required. □

Lemma 3.47 (Concentration for mean Folded Normal)**.** *In our setting we assume that* $\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leq C$*, for each* $n \in \mathbb{N}$*. Define* $X \coloneqq \frac{1}{n}\sum_{i=1}^{n}(T_i - f(\mu_i, \sigma))^2$*. We then have:*

$$|X - \mathbb{E}(X)| \leq 2\gamma\sigma\sqrt{\frac{5\sigma^2 + 4C}{n}} \tag{3.103}$$

with probability at least $1-\gamma^{-2}$*, where* $\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} g(\mu_i, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma^2 - f(\mu_i, \sigma)^2\right)$*.*

Proof of Lemma 3.47. First we determine $\mathbb{E}(X)$ as follows:

$$\begin{aligned}
\mathbb{E}(X) &\coloneqq \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(T_i - f(\mu_i, \sigma))^2\right) && \text{(by definition of } X\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((T_i - f(\mu_i, \sigma))^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}(T_i) && \text{(by since } f(\mu_i, \sigma) \coloneqq \mathbb{E}(T_i)\text{.)} \\
&= \frac{1}{n}\sum_{i=1}^{n} g(\mu_i, \sigma) && \text{(by since } g(\mu_i, \sigma) \coloneqq \mathbb{V}(T_i)\text{.)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma^2 - f(\mu_i, \sigma)^2\right) && \text{(using Equation (3.16))}
\end{aligned}$$

as required. Next we determine $\mathbb{V}(X)$ as follows:

$$\begin{aligned}
\mathbb{V}(X) &\coloneqq \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}(T_i - f(\mu_i, \sigma))^2\right) && \text{(by definition of } X\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left((T_i - f(\mu_i, \sigma))^2\right) && \text{(by the independence of } T_i\text{)}
\end{aligned}$$

Now note that for each $i \in [n]$ we have:

$$
\begin{aligned}
\mathbb{V}\left((T_i - f(\mu_i,\sigma))^2\right) &= \mathbb{V}\left(T_i^2 + f(\mu_i,\sigma)^2 - 2f(\mu_i,\sigma)T_i\right) \\
&= \mathbb{V}\left(T_i^2 - 2f(\mu_i,\sigma)T_i\right) && \text{(by translation invariance.)} \\
&\leq 2\left(\mathbb{V}\left(T_i^2\right) + \mathbb{V}\left(2f(\mu_i,\sigma)T_i\right)\right) && \text{(using Corollary 3.21)} \\
&= 2\mathbb{V}\left(T_i^2\right) + 8f(\mu_i,\sigma)^2\mathbb{V}\left(T_i\right) \\
&= 2\mathbb{V}\left(T_i^2\right) + 8f(\mu_i,\sigma)^2 g(\mu_i,\sigma) && \text{(since } g(\mu_i,\sigma) \coloneqq \mathbb{V}\left(T_i\right)) \\
&\leq 2(4\mu_i^2\sigma^2 + 2\sigma^4) + 8f(\mu_i,\sigma)^2\sigma^2 \\
& && \text{(using Equations (3.48) and (3.50))} \\
&= 8\mu_i^2\sigma^2 + 8f(\mu_i,\sigma)^2\sigma^2 + 4\sigma^4 \\
&\leq 16f(\mu_i,\sigma)^2\sigma^2 + 4\sigma^4 && \text{(using Equation (3.42))} \\
&\leq 16(\mu_i^2 + \sigma^2)\sigma^2 + 4\sigma^4 && \text{(using Equation (3.46))} \\
&= 16\mu_i^2\sigma^2 + 20\sigma^4 && (3.104)
\end{aligned}
$$

Therefore we have that

$$
\begin{aligned}
\mathbb{V}\left(X\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left((T_i - f(\mu_i,\sigma))^2\right) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\left(16f(\mu_i,\sigma)^2\sigma^2 + 4\sigma^4\right) && \text{(using Equation (3.42))} \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\left(16\mu_i^2\sigma^2 + 20\sigma^4\right) && \text{(using Equation (3.104))} \\
&\leq \frac{16C\sigma^2 + 20\sigma^4}{n} && \text{(assuming } \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mu_i^2 \leq C \text{, for each } n \in \mathbb{N}.)
\end{aligned}
$$

From this it follows that:

$$
\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| \geq t\right) &\leq \frac{\mathbb{V}\left(X\right)}{t^2}, \quad \text{for each } t > 0 \\
& \qquad\qquad \text{(using Chebychev's inequality)} \\
&\leq \frac{16C\sigma^2 + 20\sigma^4}{nt^2}, \quad \text{for each } t > 0 \qquad (3.105)
\end{aligned}
$$

It then follows that by setting the upper bound (RHS) to $\gamma^{-2} \in (0,1)$, that

$$
\frac{16C\sigma^2 + 20\sigma^4}{nt^2} = \frac{1}{\gamma^2} \implies t = \gamma\sigma\sqrt{\frac{5\sigma^2 + 4C}{n}}
$$

We then have that $|X - \mathbb{E}(X)| \leq 2\gamma\sigma\sqrt{\frac{5\sigma^2+4C}{n}}$, with probability at least $1-\gamma^{-2}$, as required. $\square$

Our end goal is to to show a the following high probability result described in Theorem 3.48.

Theorem 3.48 (Concentration of fitted Folded Normal)**.**

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{T}_i - f(\mu_i,\sigma)\right)^2 \leq \delta r_{n,2}(\mu_n,\mu_1,\sigma) \tag{3.106}$$

with probability at least $1-\delta^{-1}$*.*

Proof of Theorem 3.48. First to simplify notation we let $X \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{T}_i - f(\mu_i,\sigma)\right)^2$ represent the quantity of interest. Observe that $X \geq 0$ a.s. by definition, so that $|X| \overset{a.s.}{=} X$. Then for any $t > 0$ we have:

$$\begin{aligned}
\mathbb{P}(X \geq t) &\leq \frac{\mathbb{E}(X)}{t} && \text{(by Markov's inequality)}\\
&\leq \frac{R_{n,2}^2(f(\mu_i,\sigma))}{t} && \text{(by definition, per Corollary 3.46)}\\
&\leq \frac{r_{n,2}(\mu_n,\mu_1,\sigma)}{t} && \text{(using Corollary 3.46)}
\end{aligned}$$

It then follows that by setting the upper bound (RHS) to $\delta^{-1} \in (0,1)$, that

$$\frac{r_{n,2}(\mu_n,\mu_1,\sigma)}{t} = \frac{1}{\delta} \implies t = \delta r_{n,2}(\mu_n,\mu_1,\sigma)$$

We then have that $|X| \overset{a.s.}{=} X \leq \delta r_{n,2}(\mu_n,\mu_1,\sigma)$, with probability at least $1-\delta^{-1}$, as required. $\square$

Lemma 3.49 (Concentration of $\frac{1}{n}\sum_{i=1}^{n} T_i^2$)**.** *In our setting, define* $X \coloneqq \frac{1}{n}\sum_{i=1}^{n} T_i^2$*. We then have:*

$$|X - \mathbb{E}(X)| \leq 2\gamma\sigma\sqrt{\frac{2\sigma^2+4C}{n}} \tag{3.107}$$

with probability at least $1-\gamma^{-2}$*, where* $\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2+\sigma^2\right)$*.*

Proof of Lemma 3.49. Let $X \coloneqq \frac{1}{n}\sum_{i=1}^{n} T_i^2$. First we determine $\mathbb{E}(X)$ as follows:

$$
\begin{aligned}
\mathbb{E}(X) &\coloneqq \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} T_i^2\right) && \text{(by definition of } X\text{)}\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(T_i^2\right) && \text{(by linearity of expectation.)}\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{V}(T_i) + (\mathbb{E}(T_i))^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\mu_i^2 + \sigma^2 - f(\mu_i,\sigma)^2 + f(\mu_i,\sigma)^2\right)\\
& && \text{(using Equations (3.15) and (3.16))}\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\mu_i^2 + \sigma^2\right)
\end{aligned}
$$

as required. Next we determine $\mathbb{V}(X)$ as follows:

$$
\begin{aligned}
\mathbb{V}(X) &\coloneqq \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} T_i^2\right) && \text{(by definition of } X.\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\left(T_i^2\right) && \text{(by the independence of } T_i.\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left(4\mu_i^2\sigma^2 + 2\sigma^4\right) && \text{(using Equation (3.50))}\\
&\leq \frac{4C\sigma^2 + 2\sigma^4}{n} && \text{(assuming } \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mu_i^2 \leq C\text{, for each } n \in \mathbb{N}.\text{)}
\end{aligned}
$$

From this it follows that:

$$
\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\right) &\leq \frac{\mathbb{V}(X)}{t^2}, \quad \text{for each } t > 0\\
& \qquad\qquad \text{(using Chebychev's inequality)}\\
&\leq \frac{4C\sigma^2 + 2\sigma^4}{nt^2}, \quad \text{for each } t > 0
\end{aligned}
\tag{3.108}
$$

It then follows that by setting the upper bound (RHS) to $\gamma^{-2} \in (0,1)$, that

$$
\frac{4C\sigma^2 + 2\sigma^4}{nt^2} = \frac{1}{\gamma^2} \implies t = \gamma\sigma\sqrt{\frac{2\sigma^2 + 4C}{n}}
$$

We then have that $|X - \mathbb{E}(X)| \leq 2\gamma\sigma\sqrt{\frac{2\sigma^2+4C}{n}}$, with probability at least $1-\gamma^{-2}$, as required. ☐

3.D.2 Proof of Theorem 3.12

Theorem 3.12 (Equation (3.17) has a unique root)**.** *Assume that there exist constants $\psi, \Psi, C > 0$ such that $\psi \leq \sigma \leq \Psi$ and $\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leq C$, for each $n \in \mathbb{N}$. In addition let $r_{n,2}(\mu_n, \mu_1, \sigma) = o(1)$, where the quantity $r_{n,2}(\mu_n, \mu_1, \sigma)$ is defined in (3.21). Then for sufficiently large n, $\delta = o\left((r_{n,2}(\mu_n, \mu_1, \sigma))^{-1}\right)$, and $\gamma = o\left(n^{1/2}\right)$, under the ASCI setup, Equation (3.17) in `ASCIFIT` has a unique root $\sigma^* \in \left[0, \sqrt{\frac{1}{n}\sum_{i=1}^{n}T_i^2}\right]$ for σ with probability at least $1-\delta^{-1}-2\gamma^{-2}$.*

Proof of Theorem 3.12. First, under the `ASCIFIT` setup, we can rewrite Equation (3.17) as $H(\sigma) = 0$, where:

$$H(\sigma) \coloneqq G(\sigma) - \frac{1}{n}\sum_{i=1}^{n}T_i^2. \tag{3.109}$$

$$G(\sigma) \coloneqq \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta, \sigma), \sigma))^2 \tag{3.110}$$

Our goal in this proof is to show that $H(\sigma) = 0$ has a solution $\sigma^* \in \left[0, \sqrt{\frac{1}{n}\sum_{i=1}^{n}T_i^2}\right]$, which occurs with high probability. We note that per Lemma 3.32 that $G(\sigma)$ is increasing for $\sigma \geq 0$ and *strictly* increasing for $\sigma > 0$ (per Equation (3.80)). Moreover to see that the equation $H(\sigma) = 0$ has a unique root we appeal to the Intermediate Value Theorem. Specifically we are required to find two values for σ, i.e. $\{\sigma_1, \sigma_2\}$, such that the following conditions hold:

$$G(\sigma_2) \geq \frac{1}{n}\sum_{i=1}^{n}T_i^2 \tag{3.111}$$

$$G(\sigma_1) \leq \frac{1}{n}\sum_{i=1}^{n}T_i^2 \tag{3.112}$$

By taking $\sigma_2 \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^n T_i^2}$, we observe that a.s.:

$$G(\sigma_2) = \frac{1}{n}\sum_{i=1}^n T_i^2 + \underbrace{\frac{1}{n}\sum_{i=1}^n \left(f^{-1}\left(\widehat{T}_i \vee f\left(\eta, \frac{1}{n}\sum_{i=1}^n T_i^2\right), \frac{1}{n}\sum_{i=1}^n T_i^2\right)\right)^2}_{\geq 0 \text{ a.s.}} \tag{3.113}$$

$$\geq \sum_{i=1}^n T_i^2 \tag{3.114}$$

So indeed $\sigma_2 \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^n T_i^2}$ satisfies the required condition in Equation (3.111). We now claim that $\sigma_1 \coloneqq 0$ will satisfy Equation (3.112). First observe that:

$$G(0) = \frac{1}{n}\sum_{i=1}^n f^{-1}(\widehat{T}_i \vee f(\eta, 0), 0)^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{T}_i \vee \eta)^2, \tag{3.115}$$

we then want to show that

$$\frac{1}{n}\sum_{i=1}^n (\widehat{T}_i \vee \eta)^2 \leq \frac{1}{n}\sum_{i=1}^n T_i^2, \tag{3.116}$$

holds with high probability, to be specified later.

Furthermore, since $\widehat{T}_i \vee \eta$ is the solution to an optimization problem we have that a.s.:

$$\sum_{i=1}^n (\widehat{T}_i \vee \eta - \eta)(T_i - \eta) = \sum_{i=1}^n (\widehat{T}_i \vee \eta - \eta)^2. \tag{3.117}$$

We see that Equation (3.117) holds since when you project any vector $\mathbf{v} \in \mathbb{R}^n$ on a monotone cone $K \subseteq \mathbb{R}^n$, then $\Pi_K(\mathbf{v})^\top \mathbf{v} = \|\Pi_K(\mathbf{v})\|_2^2$ per Bellec (2018, Equation 1.16). Specifically, in our case we have that $K = \mathcal{S}_+^{\uparrow} \coloneqq \{\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n : 0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n\}$, and $\mathbf{v} \coloneqq (T_1 - \eta, \ldots, T_n - \eta)^\top$.

We further observe that Equation (3.117) can be rewritten as follows a.s.:

$$
\begin{aligned}
& \sum_{i=1}^{n}(\widehat{T}_i \vee \eta - \eta)^2 = \sum_{i=1}^{n}(\widehat{T}_i \vee \eta - \eta)(T_i - \eta) && \text{(Equation (3.117))} \\
\iff & \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 + 2\eta \sum_{i=1}^{n}(\widehat{T}_i \vee \eta - \eta) + \eta^2 = \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i - \eta \sum_{i=1}^{n}(\widehat{T}_i \vee \eta - \eta) \\
& \qquad - \eta \sum_{i=1}^{n} \widehat{T}_i + \eta^2 && \text{(expanding LHS/RHS.)} \\
\iff & \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 = \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i - \eta \left(\sum_{i=1}^{n} T_i - \sum_{i=1}^{n} \widehat{T}_i \vee \eta \right) && (3.118) \\
\iff & \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 = \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i - \eta \left(\sum_{i=1}^{n} \widehat{T}_i - \sum_{i=1}^{n} \widehat{T}_i \vee \eta \right), && (3.119)
\end{aligned}
$$

where to go from Equation (3.118) to Equation (3.119) we used the fact that $\sum_{i=1}^{n} T_i = \sum_{i=1}^{n} \widehat{T}_i$. This holds since we know that $\widehat{T}_i$ are the `PAVA` solutions. Now we derive the following upper bound a.s.:

$$
\begin{aligned}
& \frac{1}{n}\left(\sum_{i=1}^{n} \widehat{T}_i - \sum_{i=1}^{n} \widehat{T}_i \vee \eta \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i - \frac{1}{n}\sum_{i=1}^{n} f(\mu_i, \sigma) + \frac{1}{n}\sum_{i=1}^{n} f(\mu_i, \sigma) - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i \vee \eta \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{T}_i - f(\mu_i, \sigma) \right) + \frac{1}{n}\sum_{i=1}^{n} \left(f(\mu_i, \sigma) - \widehat{T}_i \vee \eta \right) && (3.120) \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\widehat{T}_i - f(\mu_i, \sigma) \right)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\widehat{T}_i \vee \eta - f(\mu_i, \sigma) \right)^2}, && (3.121)
\end{aligned}
$$

where the transition between Equations (3.120) and (3.121) was by applying the Cauchy-Schwartz inequality to each summand. Note that for each $i \in [n]$, we have that $f(\mu_i, \sigma) \ge \mu_i \ge \eta$ per Equations (3.5) and (3.42). Then using Lemma 3.19 we have a.s.:

$$\frac{1}{n}\sum_{i=1}^{n}\Big((\widehat{T}_i \vee \eta) - f(\mu_i,\sigma)\Big)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i - f(\mu_i,\sigma))^2. \tag{3.122}$$

Hence by first using the inequality in Equation (3.122) to upper bound Equation (3.121), we can in turn upper bound the LHS of Equation (3.119) as follows a.s.:

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i + 2\eta\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i - f(\mu_i,\sigma))^2}\,. \tag{3.123}$$

On the other hand we have by Theorem 3.48 that:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{T}_i - f(\mu_i,\sigma)\Big)^2 \leq \delta r_{n,2}(\mu_n,\mu_1,\sigma), \tag{3.124}$$

with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$. Thus from Equation (3.123), we have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 &\leq \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i + 2\eta\left(\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}} && \text{(using Equation (3.124))}\\ &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)\widehat{T}_i + 2\eta\left(\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}}. && (3.125)\end{aligned}$$

Note that the final equality in Equation (3.125) holds, since $\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)T_i = \sum_{i=1}^{n}(\widehat{T}_i \vee \eta)\widehat{T}_i$, by again since we know that $\widehat{T}_i$ are the `PAVA` solutions. We then apply Cauchy-Schwartz to this summand of Equation (3.125) to obtain the following upper bound with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$.

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 \leq \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2}\sqrt{\sum_{i=1}^{n}\widehat{T}_i^2} + 2\eta\left(\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}}, \tag{3.126}$$

Now observe that since $\eta > 0$, the following holds a.s.:

$$\eta = |\eta| = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\eta^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2} \tag{3.127}$$

Then using Equation (3.127) we have the following:

$$\begin{aligned}
&\eta\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{T}_i^2}\right) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{T}_i^2}\right) \\
&\hfill \text{(using Equation (3.127))} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2 - \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2}\sqrt{\sum_{i=1}^{n}\widehat{T}_i^2}. \qquad (3.128)
\end{aligned}$$

By applying the upper bound derived in Equation (3.126) to Equation (3.128) we obtain the following:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{T}_i^2} \leq 2\left(\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}}, \qquad (3.129)$$

with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$.

Now we will show that $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{T}_i^2}$ is a constant distance away from $\sqrt{\frac{1}{n}\sum_{i=1}^{n}T_i^2}$, which will imply that for large n the value at 0 is smaller than the target value, i.e., $G(\sigma_1) \coloneqq G(0) \leq \frac{1}{n}\sum_{i=1}^{n}T_i^2$ as required per Equation (3.112).

On the other hand using Lemma 3.47, we have:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(T_i - f(\mu_i,\sigma))^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma^2 - f(\mu_i,\sigma)^2\right)\right| \leq l(\gamma, C, \sigma), \qquad (3.130)$$

with probability at least $1-\gamma^{-2}$, where $l(\gamma, C, \sigma) \coloneqq \gamma\sigma\sqrt{\frac{5\sigma^2+4C}{n}}$. Subtracting the inequalities in Equations (3.124) and (3.130) we then obtain:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 + \frac{2}{n}\sum_{i=1}^{n}(\widehat{T}_i - T_i) f(\mu_i, \sigma) \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma^2 - f(\mu_i,\sigma)^2\right) - l(\gamma, C, \sigma) - \delta r_{n,2}(\mu_n, \mu_1, \sigma) \qquad (3.131) \\
\iff &\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma^2 - f(\mu_i,\sigma)^2\right) \\
&- \left(l(\gamma, C, \sigma) + \delta r_{n,2}(\mu_n, \mu_1, \sigma) + \frac{2}{n}\sum_{i=1}^{n}(\widehat{T}_i - T_i) f(\mu_i, \sigma)\right). \qquad (3.132)
\end{aligned}
$$

Now, in order sharpen the lower bound in Equation (3.132), we upper bound the term $\frac{2}{n}\sum_{i=1}^{n}(\widehat{T}_i - T_i)f(\mu_i, \sigma)$ as follows:

$$
\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}(\widehat{T}_i - T_i) f(\mu_i, \sigma) \\
&= \frac{2}{n}\sum_{i=1}^{n}(\widehat{T}_i - T_i)(f(\mu_i, \sigma) - \widehat{T}_i) \qquad \text{(since } \widehat{T}_i \text{ are the } \texttt{PAVA} \text{ solutions.)} \\
&\leq \frac{2}{n}\sqrt{\sum_{i=1}^{n}(\widehat{T}_i - T_i)^2}\sqrt{\sum_{i=1}^{n}(f(\mu_i, \sigma) - \widehat{T}_i)^2} \qquad \text{(by Cauchy-Schwartz.)} \\
&\leq \frac{2}{n}\sqrt{\sum_{i=1}^{n}(f(\mu_i, \sigma) - T_i)^2}\sqrt{\sum_{i=1}^{n}(f(\mu_i, \sigma) - \widehat{T}_i)^2} \\
&\qquad \text{(since } \widehat{T}_i \text{ are } \texttt{PAVA}\text{, i.e., LSE solutions.)} \\
&= 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mu_i, \sigma) - T_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(\mu_i, \sigma) - \widehat{T}_i\right)^2} \qquad (3.133) \\
&\leq 2\left(l(\gamma, C, \sigma)\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}}, \qquad (3.134)
\end{aligned}
$$

with probability at least with probability at least $1 - \delta^{-1} - \gamma^{-2}$, by the union bound. Note that to obtain Equation (3.134) we applied the bounds in Equations (3.146) and (3.130) to Equation (3.133). Now using the bound in

Equation (3.134) in Equation (3.132) we conclude that:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 \\
&\geq \frac{1}{n}\sum_{i=1}^{n} g(\mu_i, \sigma) \\
&- \left(l(\gamma, C, \sigma) + \delta r_{n,2}(\mu_n, \mu_1, \sigma) + 2\left(l(\gamma, C, \sigma)\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}} \right)
\end{aligned} \tag{3.135}$$

Now from Equation (3.49) we have that $g(\mu, \sigma) \geq g(0, \sigma) = \sigma^2 \left(1 - \frac{2}{\pi}\right)$, for each $\mu > 0$. Hence if $\sigma \geq \psi > 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 \geq \psi^2 \left(1 - \frac{2}{\pi}\right) - \left(l(\gamma, C, \sigma) + \delta r_{n,2}(\mu_n, \mu_1, \sigma) + 2\left(l(\gamma, C, \sigma)\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}} \right) > \tag{3.136}$$

$\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 \geq \psi^2 \left(1 - \frac{2}{\pi}\right) > 0$, with probability at least $1 - \delta^{-1} - \gamma^{-2}$. Hence under the assumption that $r_{n,2}(f, \mu_n, \mu_1, \sigma) = o(1)$, for sufficiently large n the above will be bigger than a constant. Now by Lemma 3.49 we have $\frac{1}{n}\sum_{i=1}^{n} T_i^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\mu_i^2 + \sigma^2) + 2\gamma\sigma\sqrt{\frac{2\sigma^2 + 4C}{n}}$ which is upper bounded by some constant for sufficiently large n given our assumption that $\frac{1}{n}\sum_{i=1}^{n} \mu_i^2 \leq C$, for each $n \in \mathbb{N}$, and $\sigma \leq \Psi$ for some constants $C, \Psi > 0$. It follows that by applying Lemma 3.18 to Equation (3.136) we have:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} T_i^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2} \geq \kappa > 0, \tag{3.137}$$

for sufficiently large n, where κ is some positive constant, with probability at least $1 - \delta^{-1} - 2\gamma^{-2}$.

Going back to equation (3.129), it follows that required equation will have a solution between $\left[0, \sqrt{\frac{1}{n}\sum T_i^2}\right]$, with probability at least with probability at least $1 - \delta^{-1} - 2\gamma^{-2}$, as required. $\square$

3.D.3 Proof of Theorem 3.13

Lemma 3.50 (Upper and lower bounds for $\widehat{\sigma}$)**.** *Assume that there exist constants $\psi, \Psi, C > 0$ such that $\psi \leq \sigma \leq \Psi$ and $\frac{1}{n}\sum_{i=1}^{n} \mu_i^2 \leq C$, for each $n \in \mathbb{N}$.*

Then under the ASCI setting per Definition 3.2, the following hold:

$$\widehat{\sigma} \geq K_1 \text{ with probability at least } 1 - \delta^{-1} - 2\gamma^{-2}, \text{ for sufficiently large } n. \tag{3.138}$$

$$\widehat{\sigma} \leq K_2 \text{ with probability at least } 1 - \delta^{-1} - 2\gamma^{-2}, \text{ for sufficiently large } n, \tag{3.139}$$

where $K_1, K_2 > 0$ *are fixed constants and* $\gamma^{-2}, \delta^{-1} \in (0, 1)$ *are as in the proof of Theorem 3.12.*

Proof of Lemma 3.50. We prove each property (Equations (3.138) and (3.139)) in turn.

(*Proof of Equation* (3.138).) We note that by assumption we have $0 < \psi \leq \sigma \leq \Psi$. We now want to show that $\widehat{\sigma}$ is positively bounded away from 0, with high probability. First, observe that per Theorem 3.12 that $\widehat{\sigma}$ uniquely solves $G(\widehat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} T_i^2$, with high probability. Per Equation (3.115), we then have that:

$$G(\widehat{\sigma}) - G(0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{T}_i^2 - \frac{1}{n}\sum_{i=1}^{n} (\widehat{T}_i \vee \eta)^2. \tag{3.140}$$

We then have by the Mean Value Theorem, and the fact that $G(\sigma)$ is increasing for each $\sigma \geq 0$ (per Lemma 3.32), that there exists a $\widetilde{\sigma} \in [0, \widehat{\sigma}]$ such that

$$G(\widehat{\sigma}) - G(0) = G'(\widetilde{\sigma})\widehat{\sigma}. \tag{3.141}$$

Now we have $\widetilde{\sigma} \leq \widehat{\sigma}$, or equivalently that $2\widetilde{\sigma} \leq 2\widehat{\sigma}$. Since $G'(\sigma) \leq 2\sigma$ using Equation (3.81), it follows that $G'(\sigma) \leq 2\widetilde{\sigma} \leq 2\widehat{\sigma}$. Using this and Equation (3.141), we see that:

$$G(\widehat{\sigma}) - G(0) = G'(\widetilde{\sigma})\widehat{\sigma} \leq (2\widehat{\sigma})\,\widehat{\sigma} \leq 2\widehat{\sigma}^2, \tag{3.142}$$

Now using Equation (3.142) and the proof of Theorem 3.12 we have that $G(\widehat{\sigma}) - G(0) = \frac{1}{n}\sum_{i=1}^{n} T_i^2 - \frac{1}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee \eta)^2$ is positively bounded away from 0 with high probability. So it follows that $\widehat{\sigma} \geq \sqrt{\frac{G(\widehat{\sigma}) - G(0)}{2}} > 0$, with high probability, as required. ■

(*Proof of Equation* (3.139).) First, observe that per Theorem 3.12 that $\widehat{\sigma}$ uniquely solves $G(\widehat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} T_i^2$, with high probability. By Definition 3.31 this implies that $\widehat{\sigma} \leq \frac{1}{n}\sum_{i=1}^{n} T_i^2$, with high probability. Moreover by Lemma 3.49 we have $\frac{1}{n}\sum_{i=1}^{n} T_i^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\mu_i^2 + \sigma^2) + 2\gamma\sigma\sqrt{\frac{2\sigma^2 + 4C}{n}}$ with probability at least γ^{-1}, where $1-\gamma^{-2}$ for $\gamma \in (0,1)$. This in turn is bounded, in high probability, by some constant, $K_2 > 0$ for sufficiently large n given our assumptions $\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leq C$, for each $n \in \mathbb{N}$, and $\sigma \leq \Psi$, as required. ■

Thus all properties specified in Equations (3.138) and (3.139) are now proved. □

Theorem 3.13 ($\widehat{\sigma}$ is close to σ)**.** *Under the assumptions of Theorem 3.12, we have that* $|\sigma - \widehat{\sigma}| \lesssim (\delta r_{n,2}(\mu_n, \mu_1, \sigma))^{1/2} + \gamma n^{-1/2}$ *with probability at least* $1-\delta^{-1}-2\gamma^{-2}$*, where* $\delta^{-1}, \gamma^{-2} \in (0,1)$.

Proof of Theorem 3.13. Recall our map $G(\sigma) \coloneqq \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2$ as originally defined in Equation (3.110). We will first try to show that $G(\sigma)$ is close to $\frac{1}{n}\sum_{i=1}^{n}T_i^2$. First note that $f^{-1}(\cdot \vee f(\eta,\sigma),\sigma)$ is a $L \coloneqq \frac{1}{2\Phi(\eta/\sigma)-1}$-Lipschitz function per Lemma 3.27 and the fact that σ is a (both upper and lower) bounded quantity by assumption. Thus it follows that

$$\left|f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma) - f^{-1}(f(\mu_i,\sigma),\sigma)\right| \leq L\left|\widehat{T}_i \vee f(\eta,\sigma) - f(\mu_i,\sigma)\right|, \tag{3.143}$$

and therefore

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma) - \mu_i)^2 && (3.144)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left|f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma) - f^{-1}(f(\mu_i,\sigma),\sigma)\right|^2 \\
&\leq \frac{L^2}{n}\sum_{i=1}^{n}(\widehat{T}_i \vee f(\eta,\sigma) - f(\mu_i,\sigma))^2 && \text{(using Equation (3.143))}\\
&\leq \frac{L^2}{n}\sum_{i=1}^{n}(\widehat{T}_i - f(\mu_i,\sigma))^2 && \text{(using Equation (3.43) and Lemma 3.19.)}
\end{aligned}$$

In sum, we have established:

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma) - \mu_i)^2 \leq \frac{L^2}{n}\sum_{i=1}^{n}(\widehat{T}_i - f(\mu_i,\sigma))^2, \tag{3.145}$$

We saw earlier by Theorem 3.48 we have that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{T}_i - f(\mu_i,\sigma)\right)^2 \leq \delta r_{n,2}(\mu_n,\mu_1,\sigma), \tag{3.146}$$

with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$. Combining Equations (3.145) and (3.146) we have that

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)-\mu_i)^2 \leq L^2\delta r_{n,2}(\mu_n,\mu_1,\sigma) \tag{3.147}$$

with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$. Thus by the triangle inequality, and reverse triangle inequality we have

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2 \in \left[\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 - h_n, \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 + L^2\delta r_{n,2}(\mu_n,\mu_1,\sigma)+h_n\right] \tag{3.148}$$

where $h_n \coloneqq 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mu_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)-\mu_i)^2}$. Given our assumption that $\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leq C$, for each $n \in \mathbb{N}$, we have that:

$$h_n \leq 2L\left(C\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}} \tag{3.149}$$

with probability at least $1-\delta^{-1}$, for $\delta^{-1} \in (0,1)$. Then combining Equations (3.148) and (3.149), we have that there exists some $l_1 \in [-2,2]$ for sufficiently large n such that

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2 = \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 + l_1L\left(2C\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}}. \tag{3.150}$$

with probability at least $1-2\delta^{-1}$, for $\delta^{-1} \in (0,1)$, using the union bound.

Similarly, using Lemma 3.49 we have that there exists some $l_2(\sigma,C,\gamma) \in \mathbb{R}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}T_i^2 = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 + l_2(\sigma,C,\gamma)n^{-1/2}. \tag{3.151}$$

with probability at least $1-\gamma^{-2}$, for $\gamma^{-2} \in (0,1)$. Moreover per Theorem 3.12 we have that $\widehat{\sigma}$ uniquely solves $G(\widehat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n}T_i^2$, with high probability. That is:

$$G(\widehat{\sigma}) = \widehat{\sigma}^2 + \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma}))^2 = \frac{1}{n}\sum_{i=1}^{n}T_i^2. \tag{3.152}$$

Then combining Equations (3.151) and (3.152), we conclude that

$$|G(\sigma)-G(\widehat{\sigma})| \leq l_1L\left(2C\delta r_{n,2}(\mu_n,\mu_1,\sigma)\right)^{\frac{1}{2}} + l_2(\sigma,C,\gamma)n^{-1/2}. \tag{3.153}$$

We now consider two cases, namely $\widehat{\sigma} > \sigma$ and $\sigma > \widehat{\sigma}$. In the first case, with $\widehat{\sigma} > \sigma$, we seek to show that $G'(\xi) \geq K_1 > 0$, in high probability, for each $\xi \in (\sigma, \widehat{\sigma})$. Here K_1 represents a positive constant. By Lemma 3.50 both σ and $\widehat{\sigma}$ are upper and lower bounded by some constants which implies that ξ is also upper and lower bounded by some constants call them C_1 and C_2, i.e., $C_1 \leq \xi \leq C_2$. Since $G'(\xi) \geq J(\xi) = \xi\left(\frac{1}{2} - \frac{\eta/\xi\phi(\eta/\xi)}{2\Phi(\eta/\xi)-1}\right)$. As we argued earlier $J(\xi)$ is positive and since it is a continuous function and the set $[C_1, C_2]$ is compact it achieves its minimum, which is strictly positive. Hence $G'(\xi) \geq K_1 > 0$.

Similarly, in the second case, with $\sigma > \widehat{\sigma}$, we can also show that $G'(\xi) \geq K_2 > 0$, in high probability, for each $\xi \in (\sigma, \widehat{\sigma})$. Where again, K_2 represents a positive constant.

Then by using the Mean Value Theorem we have that there exists some $\xi \in (\sigma, \widehat{\sigma})$ such that $|G(\sigma) - G(\widehat{\sigma})| = G'(\xi)\,|\sigma - \widehat{\sigma}| > \min(K_1, K_2)\,|\sigma - \widehat{\sigma}|$. Thus from equation (3.153) we have

$$l_1 L\left(2C\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}} + l_2(\sigma, C, \gamma)n^{-1/2} = |G(\sigma) - G(\widehat{\sigma})| \geq \min(K_1, K_2)\,|\sigma - \widehat{\sigma}|, \tag{3.154}$$

and hence $|\sigma - \widehat{\sigma}| \lesssim l_1 L\left(2C\delta r_{n,2}(\mu_n, \mu_1, \sigma)\right)^{\frac{1}{2}} + l_2(\sigma, C, \gamma)n^{-1/2}$. □

3.D.4 Proof of Theorem 3.14

Theorem 3.14 ($\widehat{\boldsymbol{\mu}}_{\text{ascifit}}$ is close to $\boldsymbol{\mu}$)**.** *Under the assumptions of Theorem 3.12 and Theorem 3.13, we have that*

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta, \widehat{\sigma}), \widehat{\sigma}) - \mu_i)^2 \lesssim \delta r_{n,2}(\mu_n, \mu_1, \sigma) + \gamma^2 n^{-1}, \tag{3.22}$$

with probability at least $1 - \delta^{-1} - 2\gamma^{-2}$.

Proof of Theorem 3.14. We will now consider $\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta, \widehat{\sigma}), \widehat{\sigma}) - \mu_i)^2$.

We observe that a.s.:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma})-\mu_i)^2 && (3.155)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma})-f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2\\
&+\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)-\mu_i)^2\\
&+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)-\mu_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma})-f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2}, && (3.156)
\end{aligned}
$$

where the transition between Equations (3.155) and (3.156) was by applying adding and subtracting $f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)$, then applying the triangle inequality, and finally applying the Cauchy-Schwartz inequality to the cross product summand.

We now set to upper bound the Equation (3.156) further. First, we saw in Equation (3.147) that $\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)-\mu_i)^2 \leq L^2\delta r_{n,2}(\mu_n,\mu_1,\sigma)$, with probability at least $1-\delta^{-1}$, for $\delta^{-1}\in(0,1)$. Next, we will tackle the term

$$
\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma})-f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2. \tag{3.157}
$$

Note that map $\sigma \mapsto f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma)$ is a $L := \frac{\sqrt{2/\pi}\exp(-\mu^2/\sigma^2/2)}{2\Phi(\mu/\sigma)-1} \leq \frac{\sqrt{2/\pi}}{2\Phi(\eta/\sigma)-1}$-Lipschitz per Lemmas 3.22 and 3.27, and in addition both $\sigma,\widehat{\sigma}$ are upper and lower bounded by constants. It follows that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma})-f^{-1}(\widehat{T}_i \vee f(\eta,\sigma),\sigma))^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|L(\sigma-\widehat{\sigma})|^2 && \text{(using L-Lipschitz property.)}\\
&= L^2(\sigma-\widehat{\sigma})^2\\
&\lesssim 2C\delta r_{n,2}(\mu_n,\mu_1,\sigma)+\gamma^2 n^{-1}, && (3.158)
\end{aligned}
$$

where Equation (3.158) follows from Theorem 3.13 with probability at least $1-\delta^{-1}-2\gamma^{-2}$.

Then applying the upper bounds in Equations (3.157) and (3.158) appropriately to each corresponding summand of Equation (3.156), we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\widehat{T}_i \vee f(\eta,\widehat{\sigma}),\widehat{\sigma}) - \mu_i)^2 \lesssim \delta r_{n,2}(\mu_n,\mu_1,\sigma) + \gamma^2 n^{-1}, \tag{3.159}$$

with probability at least $1-\delta^{-1}-2\gamma^{-2}$. $\square$

3.E Proofs of Section 3.4

3.E.1 Mathematical Preliminaries

Since we adapt the lower bound construction from Bellec and Tsybakov (2015) for our ASCI setting, we first introduce the relevant related notation and definitions here first for classes of monotonic sequences. We denote $\mathcal{S}^{\uparrow} \coloneqq \left\{\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)^{\top} \,\middle|\, \mu_1 \leq \ldots \leq \mu_n\right\}$ to be the set of all non-decreasing sequences. We define $k(\boldsymbol{\mu}) \geq 1$, for $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$, to be the integer such that $k(\boldsymbol{u}) - 1$ is the number of inequalities $\mu_i \leq \mu_{i+1}$ that are strict for $i \in [n-1]$ (i.e., number of jumps of $\boldsymbol{\mu}$). The class of monotone functions we will consider are $\mathcal{S}^{\uparrow}(V^*) \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow} \,\middle|\, V(\boldsymbol{\mu}) \leq V^*\right\}$, for some fixed $V^* \in \mathbb{R}$, and $V(\boldsymbol{\mu}) = \mu_n - \mu_1$, is the total variation of any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$. We also consider the restricted class of monotone sequences, $\mathcal{S}^{\uparrow}_{k^*} \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow} \,\middle|\, k(\boldsymbol{\mu}) \leq k^*\right\}$, and $\mathcal{S}^{\uparrow}(V^*, \eta, C) \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}(V^*) \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} \mu_i^2 \leq C, \mu_1 > \eta > 0\right\}$.

3.E.2 Proof of Proposition 3.51

We follow directly the proof technique and construction from Bellec and Tsybakov (2015, Proposition 4), but make suitable adaptations for our ASCI setup. We largely follow their notation to help readers align the commonalities and differences in the underlying constructions used. Our first lower bound result is stated in Proposition 3.51.

Proposition 3.51 (Minimax lower bounds)**.** *Let $n \geq 2, V^* > 0$ and $\sigma > 0$. There exist absolute constants $c, c' > 0$ such that for any positive integer $k^* \leq n$ satisfying $(k^*)^3 \leq \frac{16n(V^*)^2}{\sigma^2}$ we have*

$$\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathcal{S}^{\uparrow}_{k^*} \cap \mathcal{S}^{\uparrow}(V^*, \eta, C)} \mathbb{P}_{\boldsymbol{\mu}}\left(\frac{1}{n}\left\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right\|^2 \geq \frac{c\sigma^2 k}{n}\right) > c' \tag{3.160}$$

where $\eta > 0$ is a fixed positive constant per Equation (3.5), $C \geq (V^*)^2 + 4\gamma^2 + 2\eta^2$, $\gamma \coloneqq \frac{1}{8}\sqrt{\frac{\sigma^2 k^*}{n}}$, $\mathbf{P}_{\boldsymbol{\mu}}$ denotes the distribution of $(R_1, \ldots, R_n)^{\top}$ satisfying Equation (3.4), and $\inf_{\widehat{\boldsymbol{\mu}}}$ is the infimum over all estimators.

Proof of Proposition 3.51. Let n be a multiple of $k^* \in \mathbb{N}$. Then for any $\boldsymbol{\omega}, \boldsymbol{\omega'} \in \{0,1\}^{k^*}$, using the Varshamov-Gilbert bound (Tsybakov, 2009, Lemma 2.9), there exists a set $\Omega \in \{0,1\}^{k^*}$ such that:

$$\mathbf{0} = (0, \ldots, 0)^{\top} \in \Omega, \quad \log(|\Omega| - 1) \geq \frac{k^*}{8}, \quad \text{and} \quad \mathrm{d}_{\mathsf{HAMM}}\left(\boldsymbol{\omega}, \boldsymbol{\omega}'\right) > \frac{k^*}{8} \tag{3.161}$$

for any two distinct $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. For each $\boldsymbol{\omega} \in \Omega$, define a vector $\mathbf{u}^{\boldsymbol{\omega}} \in \mathbb{R}^n$ componentwise, for each component index $i \in [n]$ as follows:

$$\mathbf{u}_i^{\boldsymbol{\omega}} := \frac{\left\lfloor (i-1)\frac{k^*}{n} \right\rfloor V^*}{2k^*} + \gamma \boldsymbol{\omega}_{\left\lfloor (i-1)\frac{k^*}{n} \right\rfloor + 1}. \tag{3.162}$$

$$\bar{\mathbf{u}}_i^{\boldsymbol{\omega}} := \mathbf{u}_i^{\boldsymbol{\omega}} + \eta. \tag{3.163}$$

where $\gamma := \frac{1}{8}\sqrt{\frac{\sigma^2 k^*}{n}}$ and $\eta > 0$ is a fixed positive constant per Equation (3.5). Importantly we note that $\mathbf{u}_i^{\boldsymbol{\omega}}$ per Equation (3.162) is precisely as constructed in Bellec and Tsybakov (2015, Proposition 4). However, critically the construction in Equation (3.163) is adapted to our ASCI setting, by componentwise translation by $\eta > 0$. More compactly, it is also convenient to represent this construction as $\bar{\mathbf{u}}^{\boldsymbol{\omega}} := \mathbf{u}^{\boldsymbol{\omega}} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} := (\eta, \ldots, \eta)^\top \in \mathbb{R}^n$.

As per Bellec and Tsybakov (2015, Proposition 4) we first note the following properties for $\mathbf{u}_i^{\boldsymbol{\omega}}$, for each $i \in [n]$. For any $\boldsymbol{\omega} \in \Omega, \mathbf{u}^{\boldsymbol{\omega}}$ is a piecewise constant sequence with $k\left(\mathbf{u}^{\boldsymbol{\omega}}\right) \leq k^*, \mathbf{u}^{\boldsymbol{\omega}}$ is a non-decreasing sequence because $\gamma \leq \frac{V^*}{2k^*}$, and by construction $V\left(\mathbf{u}^{\boldsymbol{\omega}}\right) \leq V^*$. Thus, $\mathbf{u}^{\boldsymbol{\omega}} \in \mathcal{S}_{k^*}^{\uparrow} \cap \mathcal{S}^{\uparrow}(V)$ for all $\boldsymbol{\omega} \in \Omega$.

Now we observe the following corresponding properties of the $\boldsymbol{\eta}$-translated sequence $\bar{\mathbf{u}}^{\boldsymbol{\omega}}$. First note that since for any $\boldsymbol{\omega} \in \Omega, \mathbf{u}^{\boldsymbol{\omega}}$ is a piecewise constant non-decreasing sequence, so is $\bar{\mathbf{u}}_j^{\boldsymbol{\omega}}$, by translation invariance. Next, consider any arbitrary index $j \in [n]$ relating to a 'jump' in $\mathbf{u}^{\boldsymbol{\omega}}$, i.e., $\mathbf{u}_j^{\boldsymbol{\omega}} < \mathbf{u}_{j+1}^{\boldsymbol{\omega}}$ (note the *strict* inequality). We then have that:

$$\begin{aligned}
& \mathbf{u}_j^{\boldsymbol{\omega}} < \mathbf{u}_{j+1}^{\boldsymbol{\omega}} && \text{(by assumption.)} \\
\iff\ & \mathbf{u}_j^{\boldsymbol{\omega}} + \eta < \mathbf{u}_{j+1}^{\boldsymbol{\omega}} + \eta \\
\iff\ & \bar{\mathbf{u}}_j^{\boldsymbol{\omega}} < \bar{\mathbf{u}}_{j+1}^{\boldsymbol{\omega}} && \text{(using Equation (3.163))}
\end{aligned}$$

So any 'jump' in the original sequence $\mathbf{u}^{\boldsymbol{\omega}}$ corresponds to a jump in the $\boldsymbol{\eta}$-translated sequence $\bar{\mathbf{u}}^{\boldsymbol{\omega}}$. That is, we have $k\left(\bar{\mathbf{u}}^{\boldsymbol{\omega}}\right) = k\left(\mathbf{u}^{\boldsymbol{\omega}}\right) \leq k^*$. In addition, we note that

$$\begin{aligned}
V\left(\bar{\mathbf{u}}^{\boldsymbol{\omega}}\right) &= \bar{\mathbf{u}}_n^{\boldsymbol{\omega}} - \bar{\mathbf{u}}_1^{\boldsymbol{\omega}} \\
&= \left(\mathbf{u}_n^{\boldsymbol{\omega}} + \eta\right) - \left(\mathbf{u}_1^{\boldsymbol{\omega}} + \eta\right) \\
&= \mathbf{u}_n^{\boldsymbol{\omega}} - \mathbf{u}_1^{\boldsymbol{\omega}} \\
&= V\left(\mathbf{u}^{\boldsymbol{\omega}}\right) \\
&\leq V^* && \text{(by construction of } \mathbf{u}^{\boldsymbol{\omega}}.)
\end{aligned}$$

By construction we also have that $\bar{\mathbf{u}}_1^{\boldsymbol{\omega}} := \mathbf{u}_1^{\boldsymbol{\omega}} + \eta \geq \eta > 0$, since $\mathbf{u}_1^{\boldsymbol{\omega}} \geq 0$ by construction (in fact each component is non-negative). Finally, per our ASCI setting, we want to check if there exists a $C > 0$, such that $\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{u}}_i^{\boldsymbol{\omega}})^2 \leq C$, for each $n \in \mathbb{N}$. Given $\bar{\mathbf{u}}^{\boldsymbol{\omega}}$ we observe the following for each component index $i \in [n]$:

$$\begin{aligned}
(\bar{\mathbf{u}}_i^{\boldsymbol{\omega}})^2 &:= (\mathbf{u}_i^{\boldsymbol{\omega}} + \eta)^2 && \text{(using Equation (3.162))}\\
&\leq 2\left((\mathbf{u}_i^{\boldsymbol{\omega}})^2 + \eta^2\right) && \text{(using Lemma 3.20)}\\
&= 2\left(\left(\frac{\left\lfloor (i-1)\frac{k}{n}\right\rfloor V^*}{2k} + \gamma\boldsymbol{\omega}_{\left\lfloor (i-1)\frac{k}{n}\right\rfloor+1}\right)^2 + \eta^2\right) && \text{(using Equation (3.162))}\\
&\leq 2\left(2\left[\left(\frac{\left\lfloor (i-1)\frac{k}{n}\right\rfloor V^*}{2k}\right)^2 + \left(\gamma\boldsymbol{\omega}_{\left\lfloor (i-1)\frac{k}{n}\right\rfloor+1}\right)^2\right] + \eta^2\right) && \text{(using Lemma 3.20)}\\
&\leq 2\left(2\left[\left(\frac{V^*k}{2k}\right)^2 + \gamma^2\right] + \eta^2\right) && \text{(since } \tfrac{i-1}{n} \leq 1 \text{ for each } i \in [n].)\\
&\leq (V^*)^2 + 4\gamma^2 + 2\eta^2 && (3.164)
\end{aligned}$$

So indeed it follows from Equation (3.164) that:

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{u}}_i^{\boldsymbol{\omega}})^2 \leq (V^*)^2 + 4\gamma^2 + 2\eta^2 =: C \tag{3.165}$$

So that we have $\bar{\mathbf{u}}^{\boldsymbol{\omega}} \in \mathcal{S}_{k^*}^{\uparrow} \cap \mathcal{S}^{\uparrow}(V^*, \eta, C)$. Moreover, for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$, we observe that:

$$\begin{aligned}
\left|\bar{\mathbf{u}}^{\boldsymbol{\omega}} - \bar{\mathbf{u}}^{\boldsymbol{\omega}'}\right|^2 &:= \left|(\mathbf{u}^{\boldsymbol{\omega}} + \boldsymbol{\eta}) - (\mathbf{u}^{\boldsymbol{\omega}'} + \boldsymbol{\eta})\right|^2 && \text{(by construction } \bar{\mathbf{u}}^{\boldsymbol{\omega}} := \mathbf{u}^{\boldsymbol{\omega}} + \boldsymbol{\eta}.)\\
&= \left\|\mathbf{u}^{\boldsymbol{\omega}} - \mathbf{u}^{\boldsymbol{\omega}'}\right\|^2\\
&= \frac{\gamma^2}{k^*}\mathrm{d}_{\mathsf{HAMM}}\left(\boldsymbol{\omega}, \boldsymbol{\omega}'\right)\\
&\geq \frac{\gamma^2}{8}\\
&= \frac{\sigma^2 k^*}{512n}
\end{aligned}$$

Set for brevity $\mathbf{P}_{\boldsymbol{\omega}} = \mathbf{P}_{\bar{\mathbf{u}}^{\boldsymbol{\omega}}}$. The Kullback-Leibler divergence $d_{\mathsf{KL}}(\mathbf{P}_{\boldsymbol{\omega}} \,||\, \mathbf{P}_{\boldsymbol{\omega}'})$, between $\mathbf{P}_{\boldsymbol{\omega}}$ and $\mathbf{P}_{\boldsymbol{\omega}'}$, is equal to $\frac{n}{2\sigma^2} \left\| \mathbf{u}^{\boldsymbol{\omega}} - \mathbf{u}^{\boldsymbol{\omega}'} \right\|^2$ for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. Thus,

$$d_{\mathsf{KL}}\Big(\mathbf{P}_{\boldsymbol{\omega}} \,||\, \mathbf{P}_{\mathbf{0}}\Big) = \frac{\gamma^2 n \mathrm{d}_{\mathsf{HAMM}}\left(\mathbf{0}, \boldsymbol{\omega}\right)}{2k^* \sigma^2} \leq \frac{k^*}{128} \leq \frac{\log(|\Omega| - 1)}{16} \tag{3.166}$$

Applying Tsybakov (2009, Theorem 2.7) with $\alpha = 1/16$ completes the proof. □

3.E.3 Proof of Proposition 3.16

From Proposition 3.51, in line with Bellec and Tsybakov (2015, Corollary 5), we immediately obtain the following result in Proposition 3.16. Once again, we utilize the technique of Bellec and Tsybakov (2015, Corollary 5) to obtain the following corollary. The important changes to ensure that we adapt to our ASCI setting are captured in Proposition 3.51 and our proof thereof.

Proposition 3.16 (Minimax lower bounds). *Let $n \geq 2, V^* > 0$ and $\sigma > 0$, and define $\widetilde{r}_{n,2}(V^*, \sigma) := \max\left\{ \left(\frac{\sigma^2 V^*}{n}\right)^{\frac{2}{3}}, \frac{\sigma^2}{n} \right\}$. Then, there exist absolute constants $c, c' > 0$ such that:*

$$\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathcal{S}^{\uparrow}(V^*, \eta, C)} \mathbb{P}_{\boldsymbol{\mu}} \left(\frac{1}{n} \left\| \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \right\|^2 \geq c \widetilde{r}_{n,2}(V^*, \sigma) \right) > c' \tag{3.23}$$

Proof of Proposition 3.51. As per Bellec and Tsybakov (2015, Corollary 5), to prove this corollary it is enough to note that if $\frac{16n(V^*)^2}{\sigma^2} \geq 1$, by choosing k^* in Proposition 3.51 as the integer part of $\left(\frac{16n(V^*)^2}{\sigma^2}\right)^{\frac{1}{3}}$, we obtain the lower bound corresponding to $\left(\frac{\sigma^2 V^*}{n}\right)^{\frac{2}{3}}$ under the maximum in Equation (3.23). On the other hand, if $\frac{16n(V^*)^2}{\sigma^2} < 1$ the term $\frac{\sigma^2}{n}$ is dominant, so that we need to have the lower bound of the order $\frac{\sigma^2}{n}$, which is trivial (it follows from a reduction to the bound for the class composed of two constant functions). □

Part II

Location-Scale Estimation

Four

Uniform Location Estimation on Convex Bodies

Abstract: In this paper we generalize classical univariate uniform location-scale estimation over an interval, to multivariate uniform location-scale estimation over general convex bodies. Unlike the univariate setting, the sample observations are no longer totally ordered and previous estimation techniques prove insufficient to account for the more refined geometry of the generating process. Our focus is location estimation, though we consider both known and unknown scale parameter regimes. Under both scaling regimes, considering the dimension d as fixed, our proposed location estimators converge at an n^{-1} rate. In fact, our high probability upper bound convergence guarantees hold for any location estimator lying in a region known as the "critical set". We provide minimax lower bounds to justify the optimality of our estimators in terms of the sample complexity. To ensure practicality of our estimators, we provide algorithms with provable convergence rates for our estimators, over a wide class of convex bodies. We illustrate our findings with extensive simulations.

The work in this chapter was done jointly with Matey Neykov. It is based on a (forthcoming) preprint with the title *"Uniform Location Estimation on Convex Bodies"*.

4.1 Introduction

Many problems in statistical theory can be reformulated as *location* and *scale* parameter estimation problems. In particular, an elementary classical example is univariate uniform location estimation (Lehmann and Casella, 1998, Example 3.19). This underlying generating process for estimation purposes is described as follows. Suppose we independently sample n points uniformly over a fixed compact interval, e.g., $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\frac{1}{2}, \frac{1}{2}]$ for each $i \in [n]$. Now assume that we instead *observe* the translated sample, i.e., $Y_i \overset{a.s.}{=} X_i + \theta$. We then have that $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, for each $i \in [n]$. In this setting, the translation quantity $\theta \in \mathbb{R}$ is a fixed but unknown *location parameter*, to be estimated. Thus, the *location estimation* problem here involves estimating θ in a statistically and computationally efficient manner.

There are various approaches location estimation in this univariate uniform setting. For example, the MLE of θ can be shown to lie in an interval, i.e., $\widehat{\theta}_{\text{MLE}} \in [Y_{(n)} - \frac{1}{2}, Y_{(1)} + \frac{1}{2}]$. The MLE in this specific case is a particularly popular pedagogical example (Whittinghill and Hogg, 2001). This is largely because the estimated location parameter of interest, θ, appears in the support of the uniform density function. As a result this requires a more careful analysis of the resulting likelihood function. Under square loss, the risk optimal equivariant estimator for θ in this case has closed form and is known as the Pitman location estimator. It is given by $\widehat{\theta}_{\text{pit}} = \frac{Y_{(1)} + Y_{(n)}}{2}$ (Pitman, 1939a, Section 2). The simple closed form of the Pitman estimator and the MLE in this univariate uniform setting are both efficient to compute, and also easy to update in online settings.

Notably, while $\widehat{\theta}_{\text{MLE}}$ is *any* convex combination of the interval endpoints $\left\{Y_{(1)} + \frac{1}{2}, Y_{(n)} - \frac{1}{2}\right\}$, $\widehat{\theta}_{\text{pit}}$ is *uniquely* given by their midpoint (centroid). Furthermore, both the MLE and Pitman estimators here highlight the interesting geometry of the location estimation process in that they both depend on the extremal order statistics (i.e. $\left\{Y_{(1)}, Y_{(n)}\right\}$) of the observed samples. These order statistics arise naturally here since both estimators specifically exploit the total ordering of our real-valued univariate observations. We further note that in this univariate case the Pitman estimator $\widehat{\theta}_{\text{pit}}$ is well studied in that its asymptotic distribution is known and importantly that it converges to the true parameter θ in at an n^{-1} rate (Robbins and Zhang, 1986a, Equation 1.15).

It is also simple to generalize the aforementioned univariate location estimation problem to a problem with unknown scale. In this case, suppose one instead observes samples $Y_i \overset{a.s.}{=} \sigma X_i + \theta$, for some fixed but unknown *scale parameter* $\sigma > 0$. We then have that $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[\theta - \frac{\sigma}{2}, \theta + \frac{\sigma}{2}]$. Treating the scaling σ, as a nuisance parameter, a natural estimator of the location parameter θ in this case would be the MLE which takes the form[1] $\widehat{\theta}_{\text{MLE}} = \frac{Y_{(1)}+Y_{(n)}}{2}$. Once again this estimator is simple to compute and one can show that it will converge to its target parameter at a n^{-1} rate. This leads us to three natural follow-up questions:

> ***Three core questions:*** *First, under both known and unknown scaling regimes, how can one generalize this univariate uniform location estimation problem to a multivariate setting? Second, how can one derive statistically optimal location estimators and understand their geometry in this multivariate setting? Third, what are practical algorithms to compute such estimators under a wide variety of use-cases with convergence rate guarantees?*

Investigating these three core questions of interest motivates our work in this paper.

4.1.1 Multivariate uniform location estimation on convex bodies

In order to investigate our three proposed questions of interest, we first formally define the multivariate uniform distribution on a convex body $K \in \mathcal{K}^d$. Throughout our paper, a convex body refers to compact, convex set, with a non-empty interior in $\mathbb{R}^d$.

Definition 4.1 (Multivariate uniform distribution on a convex body)**.** Let $d \in \mathbb{N}$, and $K \in \mathcal{K}^d$ be fixed. We say that $\mathbf{X} \sim \mathsf{Unif}[K]$ if and only if its probability density function, $f_{\mathbf{X}} \colon \mathbb{R}^d \to \mathbb{R}_{>0}$, is defined as follows:

$$f_{\mathbf{X}}(\mathbf{x}) \coloneqq \frac{\mathbb{I}_K(\mathbf{x})}{\operatorname{vol}_d(K)} \tag{4.1}$$

Remark 4.2. Importantly, per Definition 4.1, the uniform density of $\mathbf{X}$ is completely characterized by its support, i.e., the underlying (fixed) convex body K. Since K has non-empty interior, it follows that $\operatorname{vol}_d(K) > 0$, for each $d \in \mathbb{N}$. So $f_{\mathbf{X}}(\mathbf{x})$ is well-defined and non-degenerate.

Remark 4.3. Given that the support of $K \in \mathcal{K}^d$, of $\mathbf{X}$, is a convex body, the *convex geometric* properties of K will play a *crucial* role in our work. We therefore introduce the required convex analytic details as needed, to ensure our work is self-contained.

[1] A formal justification for this claim is provided in Section 4.B.1.

We then consider the following generalized multivariate uniform *location-scale* generating process on a convex body.

Definition 4.4 (Multivariate uniform location-scale generating process)**.** Let $d \geq 1$ be a fixed positive integer. Further, let $K \in \mathcal{K}^d$ also be a fixed, and assume that d, K, and centroid (K) are all known to the observer. Let $\mathbf{v} \in \mathbb{R}^d$ be a fixed but unknown location parameter. We then consider n observations, $(\mathbf{Y}_i)_{i=1}^n$, where each observation $\mathbf{Y}_i$ is generated from the following model:

$$\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \sigma \mathbf{X}_i \tag{4.2}$$

$$\text{s.t. } \mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[K] \tag{4.3}$$

$$\text{and } \sigma > 0 \tag{4.4}$$

In (4.4), we consider both *scaling regimes* in which the fixed *scale parameter* σ is either known or unknown to the observer, and we assume that σ does not scale with n. In the latter case σ is treated as a nuisance parameter.

Remark 4.5. We want to emphasize that for $d = 1$, the *univariate* uniform location generating process case described earlier is indeed a very special case of Definition 4.4 for two main reasons. First it is the only case where observations are totally ordered. Second, compact intervals are the *only* convex bodies when $d = 1$. Both of these facts which simplify the univariate estimation process no longer directly apply when $d \geq 2$. Hence our described setting above is a minimal *strict* generalization of the classical univariate uniform location estimation setting.

For simplicity, we illustrate the geometry of this underlying *location-scale* data generating process in Figure 4.1.1, for the particular case where $K \subset \mathbb{R}^2$ is a regular convex nonagon[2]. Our primary goal here is to estimate the location parameter $\mathbf{v}$ (treating σ as a nuisance parameter) ensuring statistical and computational guarantees. In our multivariate setting we hold the dimension $d \in \mathbb{N}$, of our location parameter space to be fixed and allow the number of observed samples n, to increase asymptotically. Additionally we will assume that all estimation errors are computed under square loss, in high probability.

We note that our proposed multivariate generating process described now answers our first core question of interest. Namely, per Remark 4.5 we have a minimal assumption multivariate generalization of the original univariate uniform location estimation problem in both known and unknown scaling regimes.

[2]For convenience, this will be used as a running illustrative example throughout the paper.

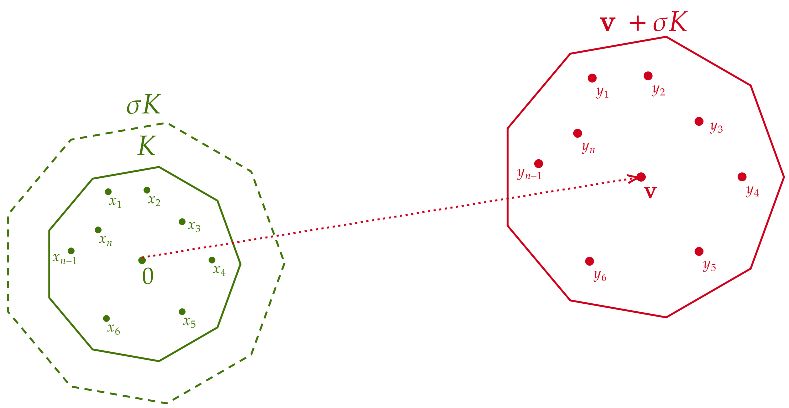

Figure 4.1.1: We observe samples $\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \sigma \mathbf{X}_i$, with $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[K]$, goal is to estimate $\mathbf{v}$.

Remark 4.6. Given that the centroid (K) is assumed to be known to the observer, we will assume WLOG that centroid $(K) = \mathbf{0} \in \mathbb{R}^d$. This is so, since the entire convex body K can always be translated by $-$ centroid (K) to shift its centroid to $\mathbf{0}$. Recall that $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[K]$, for each $i \in [n]$. In our multivariate uniform setting, we then have that each $\mathbf{X}_i$ is then also mean centered at $\mathbf{0}$ i.e. $\mathbb{E}(\mathbf{X}_i) = \mathbf{0}$ for each $i \in [n]$. The observer having this knowledge of the centroid of the original supporting distribution is a standard assumption in location estimation.

Despite the univariate uniform location-scale estimation being a well studied estimation problem, to the best of our knowledge our proposed multivariate generalization of uniform location-scale estimation has not been explicitly studied previously. A particular emphasis of this work is to consider a wide variety of location estimators under this setting and understand the statistical and computational and trade-offs that arise in the estimation process.

4.1.2 Prior and related work

This paper is fundamentally motivated by better understanding multivariate uniform location parameter estimation over convex bodies under general scaling regimes. As previously noted, to the best of our knowledge our proposed multivariate generalization to the classical univariate uniform location-scale estimation has not been formally studied in the literature. However, we observe that our multivariate uniform location-scale estimation setting lies at the intersection of two main research areas. These include location-equivariant estimation, and more indirectly Chebyshev or ℓ_∞-estimation under uniform noise. As such, we review the most relevant literature from these two broad fields as relevant for our multivariate uniform location estimation purposes.

Location Equivariant Estimation

The general location-scale estimation problem was first formally studied by Pitman in Pitman (1939a) for the estimation of *univariate* location and scale parameters under square risk loss. Notably, Pitman applied a general location estimation technique to the univariate gaussian distribution, the shifted exponential distribution, and also derived the univariate uniform location estimator, $\widehat{\theta}_{\text{pit}}$, described previously. Pitman's location estimator is remarkably flexible in that it provides a closed-form integral to estimate the location parameter in *any* univariate location family. Pitman's work in the estimation of parameters of general univariate location-scale families under squared error loss was quickly extended to inference, i.e., hypothesis testing of the location-scale parameters in Pitman (1939b). Since this work, the minimaxity and additional decision-theoretic properties of univariate Pitman location estimators on the real line are well documented in the statistics literature (Lehmann and Casella, 1998, Chapter 3) and (Strawderman, 2000, Example 3). These minimaxity results for univariate Pitman estimators were extended to the more general case where observations are independent, but not identically distributed in (van Eeden, 2006, Chapter 4).

Notably, this seminal work of Pitman (1939a) led to the development of the more general theory of *equivariant estimation*. Broadly speaking, equivariant estimation studies the statistical decision-theoretic properties of estimators that satisfy certain 'symmetry restrictions'. Since our emphasis in this paper is location parameter estimation, we will focus only on location-equivariant estimators. The 'symmetry restriction' that such location-equivariant estimators satisfy is an invariance under the common translation of the data. The sample mean, sample median, and the Pitman estimator are popular examples of such location-equivariant estimators. A key property of the Pitman location estimator is that under square loss, of all location-equivariant estimators of θ the Pitman estimator is the one with minimum risk. Commonly, the Pitman location estimator is referred to as the *minimum risk equivariant* (MRE) estimator of the location parameter (Lehmann and Casella, 1998, Theorem 3.20). For more details we refer the interested reader to (Lehmann and Casella, 1998, Chapter 3) and (Keener, 2010, Chapter 10) as the standard modern references.

Moreover the univariate Pitman location estimator can be analagously extended to the *multivariate* setting per (Maruyama and Strawderman, 2021, Equation (3.8)). Under square loss, the multivariate Pitman location estimator is also known to be minimax optimal for the location parameter (see Maruyama and Strawderman, 2021, Remark 3.1). Furthermore, Pitman estimators were shown to be locally minimax in (Strasser, 1982), and globally minimax under even more general equivariant settings in (Milbrodt, 1987). Despite the versatility of Pitman's technique, we note that its rate of convergence to the true location parameter is not known in general. Such rates need to be derived with respect to the underlying structure of the specific generative location family being studied. In particular, in our specific multivariate uniform location estimation setting we are not aware of convergence rates of Pitman's estimator to $\mathbf{v}$, the location vector. This presents another challenge of deriving convergence rates for the Pitman estimator in our setting.

Chebyshev estimation under uniform noise

Furthermore, we re-emphasize here that our work in this paper is focused on the multivariate *uniform* location estimation over convex bodies. As such we note that some of the techniques we use in this paper are broadly related to those used in Chebyshev or ℓ_∞ estimation under *uniform* noise. Recall, in linear regression the Chebyshev estimator is an alternative to the ordinary least squares estimator which minimizes estimation error with respect to the ℓ_∞ risk loss. Although our paper is concerned with multivariate uniform location estimation, not regression, we note that (Robbins and Zhang, 1986b; Schechtman and Schechtman, 1986) studied the convergence rates of the Chebyshev estimator for the simple linear regression setting with *uniformly* distributed noise. More recently Yi and Neykov (2021) studied the favorable optimality properties of the Chebyshev estimator for multivariate regression under uniform noise. The specific connections of Chebyshev estimation to our work will become clear in our multivariate uniform location estimation in the unknown scaling regime.

4.1.3 Main contributions

Our contributions in this paper are threefold and are summarized as follows:

- **Computable estimators with non-asymptotic upper bounds:** we consider both known and unknown σ regimes, and propose estimators of $\mathbf{v}$. In both scaling regimes, considering the dimension d as fixed, our proposed location estimators converge at a n^{-1} rate. In fact, our high probability upper bound convergence guarantees hold for *any* location estimator lying in a region known as the *critical set*, which we define later. We demonstrate our proof techniques for convex body polytopes in $\mathbb{R}^d$ and then extend them to general convex bodies $K \in \mathcal{K}^d$. We also show that our scale estimator also converges at a n^{-1} rate over general convex bodies.

- **Minimax lower bounds:** we provide matching high probability lower bounds (up to constant) under the Euclidean norm and thus prove that our location estimators are minimax optimal in this sense. This once again considers the dimension d as fixed (and the proportionality constant depends on d). Our proof techniques hold for arbitrary convex bodies $K \in \mathcal{K}^d$.

- **Provably efficient algorithms:** We provide provably efficient algorithms to demonstrate the computability of our location and scale estimators in a variety of practical cases for our general problem setting. We illustrate our key findings with detailed simulations.

Throughout our paper we emphasize the *(convex) geometric intuition* behind our the various location estimators, and try use illustrations where possible to highlight the convex geometry of the estimation process, or proof techniques thereof.

4.1.4 Organization of the paper

The rest of this paper is organized as follows. In Section 4.1.5 we define the main required notation used throughout the paper. In Section 4.2 we first investigate a variety of location estimators in the (easier) known scaling regime. In Section 4.3 we propose a location estimator in the (more challenging) unknown scaling regime. In Section 4.4 we derive the upper and lower bounds for a class of location estimators. For upper bounds, we begin with a "warm-up" case where $K \in \mathcal{K}^d$ is a convex polytope. We then show how to extend the proof techniques to the case where K is a general convex body. We then derive minimax lower bounds for location estimation in both scaling regimes. In Section 4.5 we proceed to give practical algorithmic implementations of our proposed estimators with provable convergence guarantees in relatively general settings. In Section 4.6 we then illustrate our findings with extensive simulations. In Section 4.7 we conclude by summarizing our results and some future research directions.

4.1.5 Notation

Throughout this paper we use the following notational conventions. Any additional section-specific notation will be introduced as needed.

Variables and inequalities

Unless specified otherwise, we typically use lowercase for scalars in $\mathbb{R}$, e.g. $(x, y, z, \ldots)$, boldface lowercase for vectors, e.g. $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$, and boldface uppercase for matrices[3], e.g. $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots)$. We use $\lesssim$ and $\gtrsim$ to mean $\leq$ and $\geq$ up to positive universal constants. We say that a sequence $a_n = \mathcal{O}(1)$ if there exists $C > 0, N \in \mathbb{N}$ such that $|a_n| < C$ for each $n > N$. Similarly, $a_n = \mathcal{O}(b_n)$ if $\frac{a_n}{b_n} = \mathcal{O}(1)$. We say that a sequence $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, $a_n = o(b_n)$ if $\frac{a_n}{b_n} = o(1)$. We say that $X_n = o_P(1)$ if for every $\varepsilon > 0$, $\mathbb{P}(|X_n| \geq \varepsilon) \xrightarrow{n \to \infty} 0$. Similarly $X_n = \mathcal{O}_P(1)$ if for every $\varepsilon > 0$ there is a finite $C_\varepsilon > 0$ such that, for all n large enough $\mathbb{P}(|X_n| \geq C_\varepsilon) \leq \varepsilon$.

Sets and related operations

We denote the finite set $\{1, \ldots, n\}$ by $[n]$. We define $\mathcal{K}^d$ to be the space of convex bodies in $\mathbb{R}^d$. That is, $\mathcal{K}^d$ is the collection of all compact, convex sets, with non-empty interior in $\mathbb{R}^d$. The *Minkowski sum* of two non-empty sets $A, B \subset \mathbb{R}^d$ is defined as $A + B \coloneqq \{\mathbf{a} + \mathbf{b} \,|\, \mathbf{a} \in A,\ \mathbf{b} \in B\}$. Similarly for a fixed vector $\mathbf{v} \in \mathbb{R}^d$ and a non-empty set $K \subset \mathbb{R}^d$ we define the *translation* of K by $\mathbf{v}$ as $\mathbf{v} + K \coloneqq \{\mathbf{v}\} + K = \{\mathbf{v} + \mathbf{k} \,|\, \mathbf{k} \in K\}$. For a given scalar $\mu \in \mathbb{R} \setminus \{0\}$ we define $\mu A \coloneqq \{\mu \mathbf{a} \,|\, \mathbf{a} \in A\}$ to be the *scaling* (or *dilation*) of the non-empty set A by μ. When we write $A - B$, we *define* it to be $A - B \coloneqq A + (-B) = \{\mathbf{a} - \mathbf{b} \,|\, \mathbf{a} \in A,\ \mathbf{b} \in B\}$. For a closed, non-empty convex set K we define $\Pi_K(\mathbf{x})$ to be the Euclidean, i.e., ℓ_2-projection of $\mathbf{x}$ onto K. We further denote $\overline{B}_2^d(\mathbf{x}, r) \coloneqq \left\{\mathbf{y} \in \mathbb{R}^d \,\middle|\, \|\mathbf{y} - \mathbf{x}\|_2 \leq r\right\}$ to be the *closed* ℓ_2-ball in $\mathbb{R}^d$ centered at $\mathbf{x} \in \mathbb{R}^d$ with radius r, and similarly $B_2^d(\mathbf{x}, r)$ denotes the *open* ℓ_2-ball in $\mathbb{R}^d$. Finally $\mathbb{S}^{d-1} \coloneqq \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \|\mathbf{x}\|_2 = 1\right\}$ denotes the unit ℓ_2-sphere in $\mathbb{R}^d$.

Functions

For a given set $K \subset \mathbb{R}^d$, we define the *indicator function* $\mathbb{I}_K(x)$ to take the value 1 when $x \in K$, and 0 otherwise. We have that λ_d denotes the d-dimensional Lebesgue measure, and $\mathrm{vol}_d(K)$ to be the volume of the set K with respect to λ_d. Additionally, for such a set K, we denote its interior by $\mathrm{int}(K)$, and its centroid by $\mathrm{centroid}(K)$. Unless stated otherwise we will always work in the metric space $\left(\mathbb{R}^d, \|\cdot\|_2\right)$.

4.2 Location estimation (known σ regime)

Given our observed samples $(\mathbf{Y}_i)_{i=1}^n$ generated according to Definition 4.4, we begin by investigating location estimation in the (easier) known scaling regime. We sequentially introduce three natural location estimators for $\mathbf{v}$. These include the multivariate Pitman location estimator $(\widehat{\mathbf{v}}_{\mathrm{pit}})$, the sample

[3] We will also use boldface uppercase for *random* vectors. Their dimension will be made clear in context.

mean ($\overline{\mathbf{v}}$), and what we refer to as the marginal projection location estimator ($\widehat{\mathbf{v}}_{\text{marg}}$). We review the basic statistical decision-theoretic properties of each estimator in general settings. Furthermore, for each estimator we comment on the statistical vs. computational optimality trade-off which describes the practicability of each estimator. Ultimately we will see that there is a need to develop a new location estimator that better balances this statistical vs. computational trade-off in more general settings.

4.2.1 Parameter identifiability

Before discussing our first three location estimators, a fundamental question is whether the underlying data generating process is identifiable for statistical estimation purposes. This is indeed true and summarized in the following proposition.

Proposition 4.7 (Parameter identifiability)**.** *The data generating process per Definition 4.4, satisfies parameter identifiability for location parameter* $\mathbf{v}$*, and scale parameter* σ*.*

With Proposition 4.7, we are ready to now consider multivariate location estimation in both known and unknown scaling regimes.

4.2.2 The critical set and its geometric properties

This intersection set arises repeatedly throughout this work, and we refer to it henceforth as the *critical set*. We briefly note the key geometric properties of the critical set in Proposition 4.8.

Proposition 4.8 (The critical set and its geometric properties)**.** *Let* $(\mathbf{Y}_i)_{i=1}^n$ *be generated according to Definition 4.4. We define critical set to be* $\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$*, for each* $n \in \mathbb{N}$*. Moreover, the critical set contains the true location vector* $\mathbf{v}$ *almost surely for each* $n \in \mathbb{N}$*, and is thus non-empty. Furthermore it is a compact convex set, and is thus closed and bounded.*

We note that although $K \in \mathcal{K}^d$ is known to the observer, the critical set $\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$, in general will have a much more complicated form (and representation) to K, due to its random intersection construction. We will shortly see this more complex geometry, even for our simple nonagon example given in Figure 4.1.1, as simulated in Figure 4.2.1.

4.2.3 Multivariate Pitman location estimator: $\widehat{\mathbf{v}}_{\text{pit}}$

Location and scale estimation as pioneered by Pitman (Pitman, 1939b), was developed in the case of univariate estimation. However the Pitman location estimator, which is the equivariant rate optimal estimator under the squared loss, can be extended naturally to the multivariate estimation setting, which is of direct interest to us. We first note the form of the general multivariate Pitman location estimator in Theorem 4.9. The proof can be found in Ibragimov and Has'minskiĭ (1981, Lemma 2.1) and Bickel and Doksum (2016, Theorem 8.3.1). In Section 4.C we also provide an alternative proof which is effectively a step-by-step multivariate extension of the relevant univariate results of (Lehmann and Casella, 1998, Chapter 3).

Theorem 4.9 (Multivariate Pitman location estimator)**.** *Consider the more general location estimation problem, under the known scaling regime. That is, let $d \geq 1$ be a fixed positive integer, and denote $\mathbf{v} \in \mathbb{R}^d$ to be the fixed but unknown location parameter. We then consider n observations, $(\mathbf{Y}_i)_{i=1}^n$, where each observation $\mathbf{Y}_i \in \mathbb{R}^d$ is generated from the following model:*

$$\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \mathbf{X}_i \tag{4.5}$$

$$s.t. \;\; (\mathbf{X}_1, \ldots, \mathbf{X}_n) \sim f, \tag{4.6}$$

where f is a valid joint probability density of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then under square loss risk, for this generating process, the multivariate mimimum risk equivariant location estimator is the Pitman estimator, $\widehat{\mathbf{v}}_{\mathrm{pit}}$, which is defined as follows:

$$\widehat{\mathbf{v}}_{\mathrm{pit}} = \frac{\int_{\mathbb{R}^d} \mathbf{u} f\left(\mathbf{y}_1 - \mathbf{u}, \ldots, \mathbf{y}_{n-1} - \mathbf{u}, \mathbf{y}_n - \mathbf{u}\right) \mathrm{d}\mathbf{u}}{\int_{\mathbb{R}^d} f\left(\mathbf{y}_1 - \mathbf{u}, \ldots, \mathbf{y}_{n-1} - \mathbf{u}, \mathbf{y}_n - \mathbf{u}\right) \mathrm{d}\mathbf{u}}. \tag{4.7}$$

Remark 4.10 (Location equivariant estimators). Consider the general location estimation model described in Theorem 4.9. Then a location estimator $\delta\colon \bigotimes_{i=1}^n \mathbb{R}^d \to \mathbb{R}^d$, for $\mathbf{v}$, is said to be *location equivariant* (or invariant) if $\delta(\mathbf{X}_1 + \mathbf{v}, \ldots, \mathbf{X}_n + \mathbf{v}) = \delta(\mathbf{X}_1, \ldots, \mathbf{X}_n) + \mathbf{v}$, for each $\mathbf{v} \in \mathbb{R}^d$.

Under square loss, the multivariate Pitman estimator is known to be minimax optimal for location (Maruyama and Strawderman, 2021, Remark 3.1). Moreover of all location-equivariant estimators, the Pitman location estimator has minimum mean squared error. It thus is referred to as the minimum risk equivariant estimator (MRE) for the location parameter. Given that the Pitman estimator is known to satisfy many favorable statistical decision-theoretic properties a natural first step is to then directly apply Theorem 4.9 to our multivariate uniform location estimation. The resulting form multivariate uniform location Pitman estimator, $\widehat{\mathbf{v}}_{\mathrm{pit}}$, is shown in Corollary 4.11.

Corollary 4.11 (Multivariate uniform Pitman location estimator)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Then* *the Pitman estimator $\widehat{\mathbf{v}}_{\text{pit}}$, of the location parameter $\mathbf{v}$ is the centroid of the critical set, that is:*

$$\widehat{\mathbf{v}}_{\text{pit}} = \text{centroid}\left(\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)\right). \tag{4.8}$$

Remark 4.12. One can readily check that in the case where $d = 1$, the form of $\widehat{\mathbf{v}}_{\text{pit}}$ given in (4.8), reduces to the univariate Pitman estimator, $\widehat{\theta}_{\text{pit}}$, as described in Section 4.1. A formal proof can be found in Section 4.C.4. We note that we used $\widehat{\mathbf{v}}_{\text{pit}}$ to denote the general multivariate Pitman location estimator in (4.7), and also for the special case for our multivariate uniform Pitman location estimation setting as per (4.8). Henceforth $\widehat{\mathbf{v}}_{\text{pit}}$ will *only* refer to (4.8).

We see that Corollary 4.11 already reveals the interesting geometry of location estimation even in the (easier) known scaling regime. Specifically, $\widehat{\mathbf{v}}_{\text{pit}}$ in this setting amounts to computing the the centroid of the critical set, as defined in Proposition 4.8.

As such it is not clear whether its centroid, $\widehat{\mathbf{v}}_{\text{pit}}$, is computable apriori over general classes of convex bodies $K \in \mathcal{K}^d$. However, there are at least two restricted settings in which $\widehat{\mathbf{v}}_{\text{pit}}$ *is* easily computable. First, suppose $K \in \mathcal{K}^d$ is an axis-aligned hyperrectangle with centroid at the origin (as usual). The resulting critical set in this case is always another axis-aligned hyperrectangle. The computation of its centroid (by symmetry) is just the vertex centroid of this axis-aligned hyperrectangle. Second, if $K \in \mathcal{K}^2$ is a convex polygon then the resulting critical set is another convex polygon in $\mathbb{R}^2$. In this special case, the Pitman location estimator can be efficiently computed using using closed form polygon centroid algorithms from computational geometry (Heckbert, 2013, Chapter I.1).

Unfortunately, computing the Pitman estimator over a broader class of convex bodies $\mathcal{K}^d$ is challenging. This is so since even *approximating* the centroid of convex bodies in higher dimensions, for even some special classes of convex bodies, is a NP-hard problem as shown in Rademacher (2007). This suggests that we need to consider alternative practically computable estimators for the location in our setting that still enjoy statistical optimality similar to the Pitman estimator, under square loss, but are computable in these more general practical settings.

4.2.4 Sample mean as a location estimator: $\overline{\mathbf{v}}$

As noted previously, the mulitvariate Pitman estimator enjoys many favorable statistical decision-theoretic properties, but it is not clear how to compute it

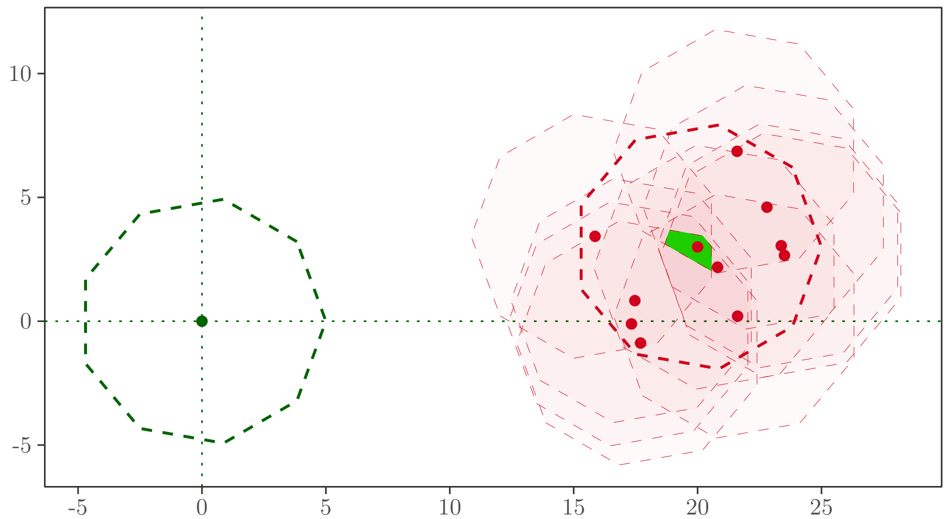

Figure 4.2.1: Simulated **critical set (right)**, $K \subset \mathbb{R}^2$ (nonagon), $\mathbf{v} = (20, 3)^\top$, and $n = 10$.

in more general settings for our purposes. A simple computable alternative to consider for multivariate uniform location estimation is the sample mean, i.e., $\overline{\mathbf{v}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i$. It is clear that $\overline{\mathbf{v}}$ is a consistent and unbiased estimator for $\mathbf{v}$. This is formally summarized in Proposition 4.13.

Proposition 4.13 (Sample mean is consistent and unbiased for location)**.** *Let* $(\mathbf{Y}_i)_{i=1}^{n}$ *be generated according to Definition 4.4, with* σ *known to the observer. The sample mean* $\overline{\mathbf{v}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i$ *is a consistent and unbiased estimator of the location parameter* $\mathbf{v}$*, regardless of the value of the true scale parameter* σ*.*

Since the sample mean is another location-equivariant estimator, we know from the previous section that the Pitman estimator achieves a smaller mean squared error (MSE). As such, with respect to the MSE, $\overline{\mathbf{v}}$ is not statistically optimal in a relative sense to $\widehat{\mathbf{v}}_{\text{pit}}$. However, since the specific rate of convergence of $\widehat{\mathbf{v}}_{\text{pit}}$ is not known, it is hard to gauge precisely how suboptimal $\overline{\mathbf{v}}$ is relative to $\widehat{\mathbf{v}}_{\text{pit}}$. To better understand this relative suboptimality of $\overline{\mathbf{v}}$, we first show in our setting $\overline{\mathbf{v}}$ converges to $\mathbf{v}$ at a rate proportional to $\sqrt{\frac{d}{n}}$, with high probability. This is detailed in Proposition 4.14.

Proposition 4.14 (Sample mean is $\sqrt{\frac{d}{n}}$-consistent)**.** *Let* $(\mathbf{Y}_i)_{i=1}^{n}$ *be generated according to Definition 4.4, with* σ *known to the observer. Then the sample* *mean estimator, i.e.* $\overline{\mathbf{v}}$*, satisfies* $\|\overline{\mathbf{v}} - \mathbf{v}\|_2 \leq \sigma \operatorname{diam}(K) \sqrt{\frac{d}{\gamma n}}$ *with probability at least* $1 - \gamma$*.*

Remark 4.15 (Sample mean is lower bounded at $n^{-\frac{1}{2}}$ risk rate)*.* It can be shown in fact that the risk rate of the sample mean location estimator has a lower bound at rate $\frac{1}{\sqrt{n}}$, suggesting that the upper bound risk rate in Proposition 4.14 can't be improved further in sample complexity. This result can be established by applying the Paley-Zygmund inequality coordinate-wise on the sample mean location estimator. A more formal proof can be found in Section 4.C.8.

We will later show that the rate of convergence of the Pitman estimator (and in fact *any* estimator lying in the critical set) to the true location parameter $\mathbf{v}$, is of the form $\frac{C(d,K)}{n}$, with high probability. Here $C(d, K)$ is some constant that depends on the dimension d and on the known convex body K. This is typically better than the $\frac{C(d,K)}{\sqrt{n}}$ rate achieved by the sample mean with the dimension held fixed, per our setting. Compared to the multivariate Pitman location estimator, it is clear that the sample mean is much easier to compute across more general settings, even in an online manner. However the relative statistical suboptimality compared to the Pitman estimator suggests that there are likely location estimators that better balance the trade-off. We investigate a another location estimation candidate in the next section.

4.2.5 Naive strategy - the marginal uniform projection estimator: $\widehat{\mathbf{v}}_{\text{marg}}$

We have now considered both the multivariate Pitman estimator ($\widehat{\mathbf{v}}_{\text{pit}}$) and the sample mean ($\overline{\mathbf{v}}$) for location parameter estimation $\mathbf{v}$. As discussed, they each highlight various statistical vs. computability trade-offs. Such trade-offs makes them impractical for location estimation purposes across a large class of convex bodies K, and for arbitrary dimension $d \geq 1$. As noted in Section 4.2.3 the multivariate Pitman estimator is easy to compute where $K \in \mathcal{K}^d$ is an axis-aligned hyperrectangle. This motivates an alternative location estimation strategy which may better balance our underlying optimality trade-offs, over more general convex bodies $K \in \mathcal{K}^d$. Under our multivariate uniform generating process, we formally describe such an estimator as follows:

Definition 4.16 (Marginal uniform projectionestimator)**.** Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Suppose we now fix any orthonormal basis $(\mathbf{a}_j)_{j=1}^d$ of $\mathbb{R}^d$. Then the marginal projection estimator for $\mathbf{v}$, i.e., $\widehat{\mathbf{v}}_{\text{marg}}$, is constructed coordinate-wise as: $[\widehat{\mathbf{v}}_{\text{marg}}]_j :=$

$f_j\left(\mathbf{a}_j^\top \mathbf{Y}_1, \ldots, \mathbf{a}_j^\top \mathbf{Y}_n\right)$, for each coordinate index $j \in [d]$. Here $f_j \colon \mathbb{R}^n \to \mathbb{R}$ is a measurable function of its inputs.

Remark 4.17. In Definition 4.16, it should be noted that each coordinate-wise value of the estimator $\widehat{\mathbf{v}}_{\text{marg}}$, *only* uses information from the the projection of the original data values along the given coordinate direction.

The key idea behind this estimation strategy is that given our multivariate uniform observations, one can potentially better utilize marginal univariate density information along pre-specified individual directions in $\mathbb{R}^d$. The end goal here is to try and better balance the previously mentioned statistical vs. computational trade-off observed in the multivariate Pitman estimator and the sample mean.

To understand why this estimation strategy may seem reasonable, let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Further we assume that $\mathcal{K}^d$ is an axis-aligned hyperrectangle[4] with centroid $(K) = \mathbf{0}$, per our usual setting. Given an observation $\mathbf{Y}_i$, we denote its coordinate components as $\mathbf{Y}_i \coloneqq (Y_{i1}, \ldots, Y_{id})^\top$, for each $i \in [n]$. We can then proceed to exploit marginal information for location estimation purposes as follows. Let us consider the orthonormal basis vectors $\mathbf{a}_j \coloneqq \mathbf{e}_j$, where $\mathbf{e}_j$ is the j^{th} standard basis vector in $\mathbb{R}^d$, for each $j \in [d]$. Here we can simply apply the uniform univariate Pitman location estimator *along each individual coordinate*. We can then combine all d coordinate-wise Pitman estimators into our single marginal projection estimator, for the hyperrectangle K, i.e., $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$. More formally we define $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ as follows:

Example 4.18 (Marginal uniform hyperrectangle projection estimator)**.** Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Furthermore, let $K \in \mathcal{K}^d$ be an axis-aligned hyperrectangle so that $K \coloneqq \prod_{j=1}^d [-\sigma\lambda_j, \sigma\lambda_j]$, where $\lambda_j > 0$, for each $j \in [d]$. Then the marginal uniform projection hyperrectangle estimator, $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$, is defined as follows:

$$\begin{aligned} \widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}} &\coloneqq (\widehat{\pi}_1, \ldots, \widehat{\pi}_d)^\top \qquad (4.9) \\ \text{where } \widehat{\pi}_j &\coloneqq \frac{\min\{Y_{1j}, \ldots, Y_{nj}\} + \max\{Y_{1j}, \ldots, Y_{nj}\}}{2}, \text{ for each } j \in [d] \end{aligned}$$

[4]The hyperrectangle is assumed WLOG to be axis-aligned with respect to the standard basis. If the hyperrectangle is not axis-aligned, one can rotate it (via change in basis) to be axis-aligned, perform location estimation, and then rotate the estimator back to the original basis. In short, this estimation strategy generalizes to *all* hyperrectangles in $\mathbb{R}^d$.

Remark 4.19. We note that $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ satisfies Definition 4.16, since π_j is a measurable function of the projected data for each $j \in [d]$. Moreover, in the case $d = 1$, this reduces to the univariate Pitman location estimator (by construction) as described in Section 4.1. However, we do not claim that in the case $d \geq 2$ where $K \in \mathcal{K}^d$ is an axis-aligned hyperrectangle, that $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ and $\widehat{\mathbf{v}}_{\text{pit}}$ are equivalent. We are simply trying to investigate whether the location estimator $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ *also* has sharp convergence to the true location parameter $\mathbf{v}$, i.e., at a n^{-1} rate in the multivariate setting. An simple illustrative example of $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ for $K \subset \mathbb{R}^2$, is shown in Figure 4.2.2.

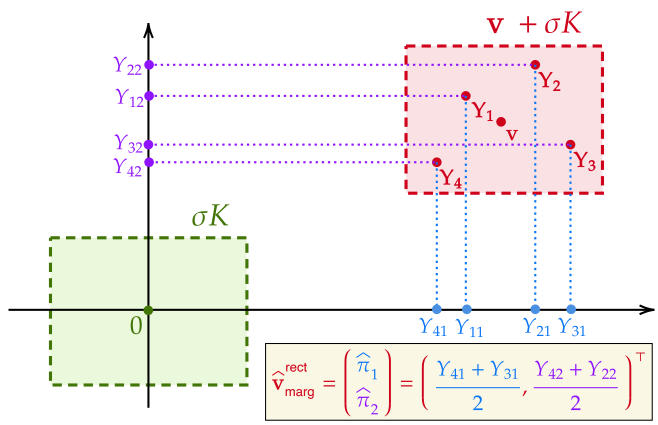

Figure 4.2.2: Example of $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ estimation, $K \subset \mathbb{R}^2$ (rectangle), and $n = 4$ samples.

Clearly, since the $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ applies the univariate uniform Pitman location estimator along each coordinate (i.e., $\widehat{\pi}_j$), it is efficient to compute even in online settings. But what about the rate of estimation in this specific setting? This is summarized in Proposition 4.20.

Proposition 4.20 (Projection Estimator for hyperrectangles)**.** *Under the setting of Example 4.18, let the marginal projection estimator,* $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$*, be defined as per* (4.9)*. Then* $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ *satisfies* $\left\| \widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}} - \mathbf{v} \right\|_2 \leq \frac{2\sigma \|\boldsymbol{\lambda}\|_2 \log\left(\frac{2d}{\gamma}\right)}{n}$ *with probability at least* $1 - \gamma$*, where* $\boldsymbol{\lambda} \coloneqq (\lambda_1, \ldots, \lambda_d)^\top$.

We see from Proposition 4.20 that in the case where $K \in \mathcal{K}^d$ is an axis-aligned hyperrectangle, $\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}$ allows us to obtain such sharp n^{-1} risk rates in sample complexity, and also dimension dependence. The immediate follow up question is whether we leverage this marginal uniform projection estimator strategy for other convex bodies $K \in \mathcal{K}^d$ and get similar risk rates? Unfortunately, we show that this is not the case in general. We do this by showing that we can in fact *lower bound* the risk rate for this marginalized projection estimation approach in the special case $K = \overline{B}_2^3(\mathbf{0}, R)$, i.e., K is the closed Euclidean ball of radius R centered at the origin in three dimensions. This particular counterexample is formalized in Theorem 4.21.

Theorem 4.21 (Lower bound of $K = \overline{B}_2^3(\mathbf{0}, R)$)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. Furthermore, let $d = 3$, $\sigma = 1$, and $K = \overline{B}_2^3(\mathbf{0}, R) \in \mathcal{K}^3$. Suppose WLOG we are estimating the first coordinate of the location parameter, v_1. Let W denote the class of all such marginal projection estimators for v_1, as per Definition 4.16. Then there exists some $C' \in (0, 1)$ such that the following holds:*

$$\inf_{\widetilde{w} \in W} \sup_{v_1 \in \mathbb{R}} \mathbb{P}\left(|\widetilde{w} - v_1| \geq (1 - C')n^{-\frac{3}{4}}\right) \geq C' \tag{4.10}$$

Theorem 4.21 demonstrates that in the case where $K = \overline{B}_2^3(\mathbf{0}, R) \in \mathcal{K}^3$, if the marginal projection information is used for a single coordinate, then the location parameter coordinate along that projection can't be estimated at better than an $n^{-\frac{3}{4}}$ rate. In turn, this implies that we certainly have a worse than n^{-1} rate across all coordinates in Euclidean norm. That is, the true location parameter $\mathbf{v}$ cannot be estimated at a n^{-1} in this specific setting for *any* marginal projection estimator satisfying Definition 4.16. However we do note that there *may* exist location estimators of $\mathbf{v}$ which better utilize marginal projection information to get faster convergence rates than the estimators described in Definition 4.16. Rather than pursue finding such estimators, we propose an entirely different projection location estimator altogether which better balances the statistical vs. computational optimality trade-offs discussed.

4.2.6 Our projection location estimator: $\widehat{\mathbf{v}}_{\text{prj}}$

As discussed, our marginal uniform projection location estimator strategy seems to not be promising in more general settings. Returning to the multivariate uniform Pitman location estimator, we observe that it requires the computation of the centroid of the critical set. Given the favorable statistical optimality of the Pitman estimator, we observe that it relies on using all d coordinate-wise information simultaneously (rather than marginally). This motivates an alternative projection strategy to estimate the location parameter $\mathbf{v}$. More specifically we propose the following *projection estimator*, denoted by $\widehat{\mathbf{v}}_{\text{prj}}$:

$$\widehat{\mathbf{v}}_{\text{prj}} \coloneqq \Pi_{\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)}(\overline{\mathbf{v}}) \coloneqq \operatorname*{arg\,min}_{\mathbf{w} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)} \|\overline{\mathbf{v}} - \mathbf{w}\|_2 \tag{4.11}$$

Geometrically, our proposed projection estimator $\widehat{\mathbf{v}}_{\text{prj}}$ simply takes the sample mean $\overline{\mathbf{v}}$ (per Section 4.2.4) and projects it onto the critical set $\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$. The motivation for projecting onto the critical set for $\widehat{\mathbf{v}}_{\text{prj}}$, is justified by its core geometric properties as noted in Proposition 4.8. More specifically, since the critical set is a non-empty closed convex set which contains $\mathbf{v}$ almost surely, it means that our projection estimator $\widehat{\mathbf{v}}_{\text{prj}}$ exists and is unique, i.e., returns a singleton. In this sense $\widehat{\mathbf{v}}_{\text{prj}}$ is well-defined. We then have that $\widehat{\mathbf{v}}_{\text{prj}}$ is always at least as accurate as the sample mean ($\overline{\mathbf{v}}$) for location estimation. This means that the projection operation moves the sample mean closer to all points in the critical set, and hence closer to $\mathbf{v}$. Of course, if $\overline{\mathbf{v}}$ already lies in the critical set, then $\overline{\mathbf{v}}$ will be returned. This all follows from basic properties of the Euclidean projection on (non-empty) closed convex sets and is formalized in Proposition 4.22.

Proposition 4.22 (Projection Estimator Motivation)**.** *Let $(\mathbf{Y}_i)_{i=1}^{n}$ be generated according to Definition 4.4, with σ known to the observer. Let $\overline{\mathbf{v}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i$ denote the sample mean, and $\widehat{\mathbf{v}}_{\text{prj}} = \Pi_{\bigcap_{i=1}^{n}(\mathbf{Y}_i-\sigma K)}(\overline{\mathbf{v}})$ denote the projection location estimator. Then for any $\mathbf{z} \in \bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$ we have:*

$$\|\overline{\mathbf{v}} - \mathbf{z}\|_2 \geq \|\widehat{\mathbf{v}}_{\text{prj}} - \mathbf{z}\|_2 \text{ a.s.} \tag{4.12}$$

Remark 4.23. Importantly, since $\mathbf{v} \in \bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$ a.s., we set $\mathbf{z} = \mathbf{v}$ in (4.12). This shows that $\widehat{\mathbf{v}}_{\text{prj}}$ will have the same or smaller risk than $\overline{\mathbf{v}}$ under square loss a.s.

Recall from Section 4.2.3 that in this known σ regime the Pitman estimator $\widehat{\mathbf{v}}_{\text{pit}}$ requires the computation of the centroid of the critical set $\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$, which is a hard task in general. On the other hand the projection estimator, $\widehat{\mathbf{v}}_{\text{prj}}$, simply requires us to project onto the critical set. We will see that relaxing from centroid estimation to projection onto the critical set is much easier computationally, for a larger class of convex bodies $K \in \mathcal{K}^d$. For instance it reduces to a quadratic program when K is a convex polytope. Moreover, from a statistical perspective, we will later show that *any* estimator lying in the critical set (including $\widehat{\mathbf{v}}_{\text{pit}}$ and $\widehat{\mathbf{v}}_{\text{prj}}$) converges in probability to $\mathbf{v}$ at a n^{-1} rate. Before discussing computational guarantees and statistical optimality of any estimators lying in the critical set, we first discuss our proposed estimators in the unknown σ regime.

4.3 Location-scale estimation (unknown σ regime)

We now turn our attention to the more challenging multivariate uniform location-scale estimation setting. We again observe $(\mathbf{Y}_i)_{i=1}^{n}$ to be generated according to Definition 4.4, however this time in the unknown σ regime. Here *both* parameters $\mathbf{v} \in \mathbb{R}^d$, and $\sigma \in \mathbb{R}_{>0}$ are unknown to the observer and need to be estimated. We propose to estimate them both simultaneously using their respective MLEs. In our case the likelihood function, for n observed samples $(\mathbf{Y}_i)_{i=1}^n$, is given by:

$$\begin{aligned} L(\mathbf{v}, \sigma \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n) &= \frac{\prod_{i=1}^n \mathbb{I}(\mathbf{Y}_i - \mathbf{v} \in \sigma K)}{(\mathrm{vol}_d(\sigma K))^n} \\ &= \frac{\prod_{i=1}^n \mathbb{I}(\mathbf{Y}_i - \mathbf{v} \in \sigma K)}{(\sigma^d \mathrm{vol}_d(K))^n} \qquad (4.13) \\ &= \frac{\mathbb{I}_{\cap_{i=1}^n \mathbf{Y}_i - \sigma K}(\mathbf{v})}{\sigma^{nd} (\mathrm{vol}_d(K))^n} \qquad (4.14) \end{aligned}$$

Unfortunately in dimensions two and higher we note the n observed samples $(\mathbf{Y}_i)_{i=1}^n$ are no longer totally ordered, unlike the univariate case. As such the typical univariate approach of computing the MLE using order statistics of the observations via indicator functions is not applicable here. This is further complicated by the fact that σ is now unknown. Before describing our MLEs for $\mathbf{v}$ and σ in this general setting, we first introduce and briefly review the Minkowski gauge functional. The gauge functional is a central object in formulating our MLEs. We first formalize it in Definition 4.24 and then provide a brief summary of its essential properties as relevant for our estimation purposes.

Definition 4.24 (Minkowski gauge functional on convex bodies)**.** Let $K \in \mathcal{K}^d$, then the Minkowski gauge functional of the set K, $\rho_K : \mathbb{R}^d \to [0, \infty)$, is defined as follows:

$$\rho_K(\mathbf{x}) \coloneqq \inf \{t > 0 \mid \mathbf{x} \in tK\}$$

Since $K \in \mathcal{K}^d$ is a convex body, then by definition $K \subset \mathbb{R}^d$ is a convex, compact set with a non-empty interior. Specifically we have that centroid $(K) = \mathbf{0} \in \mathrm{int}(K)$, which means that K is necessarily absorbing i.e. $\bigcup_{t>0} tK = \mathbb{R}^d$. It follows that the Minkowski functional is a positive homogeneous and sub-additive functional, and hence also a convex functional. For a proof of these facts and more details on the Minkowski gauge functional, see (Lindahl, 2016, Section 6.10). We also note that Minkowski gauge functionals are Lipschitz (Mordukhovich and Nam, 2014, Proposition 3.32).

Armed with this definition of the Minkowski gauge functional and its key properties we now make the following observation of the critical set.

$$
\begin{aligned}
\mathbf{x} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) &\iff \mathbf{x} \in Y_i - \sigma K,\ \forall\, i \in [n] \\
&\iff \mathbf{Y}_i - \mathbf{x} \in \sigma K,\ \forall\, i \in [n] \\
&\iff \rho_K(\mathbf{Y}_i - \mathbf{x}) \leq \sigma,\ \forall\, i \in [n] \\
&\iff \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}) \leq \sigma. \qquad (4.15)
\end{aligned}
$$

From (4.15) we observe that precise equivalence of containment of a vector, $\mathbf{x}$ in the critical set and the maximum of n, $\mathbf{Y}_i$-*translated* Minkowski gauge functional, which as noted is a convex function. Comparing this equivalence to the form of our likelihood in (4.14) we observe that the location-scale parameter MLEs, $\widehat{\mathbf{v}}_{\text{MLE}}$ and $\widehat{\sigma}_{\text{MLE}}$, are the solution to the following unconstrained convex optimization problem:

$$
\inf_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}). \qquad (4.16)
$$

We will in fact show that the infimum in the convex problem in (4.16) is in fact a minimum. Moreover our proposed MLEs for $\mathbf{v}$, and σ are then given by $\widehat{\mathbf{v}}_{\text{MLE}} \coloneqq \arg\min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$, and $\widehat{\sigma}_{\text{MLE}} \coloneqq \min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$ respectively.

Theorem 4.25 (Minimum is attained in the MLEs)**.** *The* MLEs *for* $\mathbf{v}$ *and* σ *as given by* $\widehat{\mathbf{v}}_{\text{MLE}} \in \arg\min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$*, and* $\widehat{\sigma}_{\text{MLE}} \coloneqq \min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$*, respectively, both exist.*

Remark 4.26. It is interesting to note that the gauge functional reformulation of our multivariate uniform location-scale MLEs has a very similar form to the Chebyshev or ℓ_∞ estimator used in regression. More specifically, the 'min-max' form of the estimator we propose here for the location-scale MLE is comparable to (Yi and Neykov, 2021, Equation 2.1) which is the Chebyshev estimator applied to multivariate linear regression under univariate uniform noise.

In fact since the minimum is attained for the MLE, we have the following key result which guarantees that the location parameter MLE ($\widehat{\mathbf{v}}_{\text{MLE}}$), in the unknown scaling regime lies in the critical set.

Proposition 4.27 (Location MLE is contained in the critical set)**.** *Since* $0 < \widehat{\sigma}_{\mathrm{MLE}} \leq \sigma$ *and* $\mathbf{0} \in \operatorname{int}(K)$ *(per Remark 4.6), we have that*

$$\bigcap_{i=1}^{n} (\mathbf{Y}_i - \widehat{\sigma}_{\mathrm{MLE}} K) \subseteq \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$$

The significance of this result is that $\widehat{\mathbf{v}}_{\mathrm{MLE}}$, much like $\widehat{\mathbf{v}}_{\mathrm{pit}}$, and $\widehat{\mathbf{v}}_{\mathrm{prj}}$ all lie in the critical set. We will shortly provide upper bounds for risk rates hold for *any* estimator that lies in the critical set $\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$. We will see that *all* of these estimators then are consistent for $\mathbf{v}$ at a rate of $\frac{C(d,K)}{n}$.

From a computational perspective we will also shortly see, that we propose a subgradient descent algorithm to estimate the above MLEs for general convex bodies, provided that we have a supporting hyperplane oracle for the underlying convex body $K \in \mathcal{K}^d$ and we can evaluate its Minkowski gauge functional efficiently. Moreover, these proposed algorithms will come with convergence guarantees based on the subgradient method.

4.4 Upper and lower bounds

4.4.1 Upper bounds warm-up: convex polytopes in $\mathbb{R}^d$

We seek a high probability upper bound for the mean squared risk error for our projection estimator, $\widehat{\mathbf{v}}_{\mathrm{prj}}$. We will in fact derive such bounds for *any* estimator lying in the critical set, i.e. $\widehat{\mathbf{v}}_{\mathrm{cri}} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$ a.s.. Before describing the result and proof sketch thereof where $K \in \mathcal{K}^d$ is a general convex body, we first demonstrate it in the special case where $K \in \mathcal{K}^d$ is a *convex polytope*, with m facets[5]. Per our setting in Definition 4.4, we have that $\operatorname{centroid}(K) = \mathbf{0} \in \operatorname{int}(K)$. This convex polytope proof construction will more clearly illustrate the geometric intuition behind the proof technique used in the general case when $K \subset \mathcal{K}^d$ is a *convex body.* The main result for the high probability upper bound for the rate of convergence of our estimator is provided in Theorem 4.28.

Theorem 4.28 (Consistency of location estimators in the critical set, $K \in \mathcal{K}^d$ polytope.)**.** *Let* $(\mathbf{Y}_i)_{i=1}^n$ *be generated according to Definition 4.4. Further assume that* $K \in \mathcal{K}^d$ *is a convex polytope with* m *facets. Let* $\widehat{\mathbf{v}}_{\mathrm{cri}}$ *denote any location estimator lying in the critical set, i.e.* $\widehat{\mathbf{v}}_{\mathrm{cri}} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$ a.s. *We then have that* $\widehat{\mathbf{v}}_{\mathrm{cri}}$ *satisfies* $\|\widehat{\mathbf{v}}_{\mathrm{cri}} - \mathbf{v}\|_2 \leq \sigma \alpha_n (\operatorname{diam}(K))$*, with probability at least* $1 - \gamma \in (0, 1)$*, if* $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$*. Here* $c_{\min} \leq \frac{1}{m}$*, is a constant that depends on the convex body* K*.*

[5]The formal definition of a convex polytope facet can be found in Section 4.E.1.

We now give a brief sketch of the proof technique in the case where $K \in \mathcal{K}^d$ is a convex polytope. The basic approach can be broken down intuitively into three key steps as follows:

Step I: Bound the estimation error by the diameter of the critical set.

First we observe by definition of $\widehat{\mathbf{v}}_{\text{cri}}$ and using Proposition 4.8, both $\widehat{\mathbf{v}}_{\text{cri}}, \mathbf{v} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$ a.s.. We immediately obtain:

$$\|\widehat{\mathbf{v}}_{\text{cri}} - \mathbf{v}\|_2 \leq \sup \left\{ \|\mathbf{w} - \mathbf{z}\|_2 \,\middle|\, \mathbf{w}, \mathbf{z} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \right\} =: \operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \right) \text{ a.s.} \tag{4.17}$$

The RHS in (4.17) is finite since the critical polytope is compact per Proposition 4.8, again from Proposition 4.8. Thus we can meaningfully try and bound this diameter with high probability.

Step II: Construct an enveloping polytope of the critical set.

In order to upper bound $\operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right)$, we in fact show that the critical polytope, $\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$, can be nested in another *enveloping polytope*, i.e., $-\alpha_n(\sigma K)+\mathbf{v}$, with high probability. Showing that this enveloping polytope can be constructed from n samples with high probability is the most challenging part of the proof. In brief, we first consider a sequence of dilations of the origin centered polytope, i.e., $\mathcal{P} := ((1 - \alpha_n)\sigma K)_{n=1}^{\infty}$. Here $(\alpha_n)_{n=1}^{\infty} \to 0$ is a non-negative sequence, where $\alpha_n \in (0, 1)$, for each $n \in \mathbb{N}$, to be determined. Each such dilation induces a nested polytope $(1 - \alpha_n)\sigma K \subseteq \sigma K$, which in turn induces a boundary polytope shell, i.e. $S := \sigma K \setminus (1 - \alpha_n)\sigma K$. This boundary shell can be decomposed into a union of m sets, one for each facet. We can in turn show that such an enveloping enveloping polytope can be constructed, with high probability, if we appropriately select $(\alpha_n)_{n=1}^{\infty}$ so that we can sample *at least* one point from each of the m boundary facet shells. In Section 4.E, we provide the details of how to select $(\alpha_n)_{n=1}^{\infty}$, to guarantee this high probability construction.

Step III: Bound the estimation error by the diameter of the enveloping polytope.

Our bounds depend on the diameters of convex polytopes. As such, using the monotonicity, translation invariance, and scaling properties of the ℓ_2-diameter of compact sets in $\mathbb{R}^d$ we obtain:

$$
\begin{aligned}
\|\widehat{\mathbf{v}}_{\text{cri}} - \mathbf{v}\|_2 &\leq \operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right) && \text{(using } \underline{\textbf{Step I}}.\text{)} \\
&\leq \operatorname{diam}\left(-\alpha_n(\sigma K) + \mathbf{v}\right) && \text{(using } \underline{\textbf{Step II}}.\text{)} \\
&= \alpha_n \sigma \operatorname{diam}(K),
\end{aligned}
$$

with probability at least $1 - \gamma \in (0, 1)$, if $\alpha_n = \frac{1}{c_{\min} n}\left(\log\left[\frac{m}{\gamma}\right]\right)$, as required. Note that $c_{\min} \leq \frac{1}{m}$, is a constant that depends on the convex body K. The full details of the proof are found in Section 4.E.

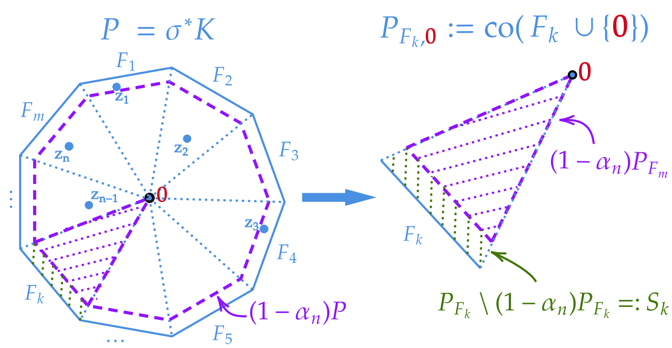

Figure 4.4.1: (Left) Triangulation of $P = \sigma K$. (Right) Decomposition of pyramid $P_{F_k,\mathbf{0}}$

4.4.2 Upper bounds: general convex bodies in $\mathbb{R}^d$

We now generalize our proof from convex polytopes to the case where the set $K \in \mathcal{K}^d$ is a general convex body (per Definition 4.41) with centroid $(K) = \mathbf{0}$, as usual. The main result is summarized in Theorem 4.29.

Theorem 4.29 (Consistency of location estimators in the critical set, $K \in \mathcal{K}^d$.)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. Let $\widehat{\mathbf{v}}_{\text{cri}}$ denote any location estimator lying in the critical set, i.e. $\widehat{\mathbf{v}}_{\text{cri}} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$ a.s.. We then have that $\widehat{\mathbf{v}}_{\text{cri}}$ satisfies $\|\widehat{\mathbf{v}}_{\text{cri}} - \mathbf{v}\|_2 \leq \frac{\sigma C_1 \kappa_n}{n}$, with probability at least $1 - 2\exp(-C_2\kappa_n / \operatorname{polylog}_d(\kappa_n))$, where $C_1 := C_1(d, K)$ and $C_2 := C_2(d, K)$ are constants which depend on the dimension d and the convex body K, κ_n is any slowly diverging sequence with n, and $\operatorname{polylog}_d(\kappa_n)$ is a poly-logarithmic factor of κ_n which also depends on the dimension d.*

Remark 4.30. From Proposition 4.27 we have that $\widehat{\mathbf{v}}_{\mathrm{MLE}} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$. As such Theorem 4.29 applies to $\widehat{\mathbf{v}}_{\mathrm{MLE}}$, and ensures the same rate of estimation under this unknown scaling location estimation setting.

We now give a brief sketch of the proof technique in the case where $K \in \mathcal{K}^d$ is a general convex body. The basic approach can be broken down intuitively into four key steps as follows:

Step I: Bound the estimation error by the diameter of the critical set.

Similar to Step I of the polytope proof, we note that per (4.17) that once again we have the basic inequality $\|\widehat{\mathbf{v}}_{\mathrm{cri}} - \mathbf{v}\|_2 \leq \operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right)$ a.s.. However the critical set here is no longer a polytope, but likely a more complicated, i.e., *non-polytopal* convex set depending on the underlying convex body K. Our goal still remains to upper bound the diameter of the critical set in this more general convex body setting.

Step II: Construct an enveloping convex polytope via supporting hyperplanes.

Our proof differs here from that used in the case where K is a convex polytope, since our critical set may not be a polytope, i.e. may not have explicit facets. However, we retain the spirit of our previous proof by first enveloping the entire convex body with a polytope formed by some of its specific supporting hyperplanes. To ensure that such a supporting hyperplane polytope can be constructed with high probability, we again take an α_n shell of the convex body, and show that we can sufficiently sample points along the boundary shell. Moreover we need to ensure that for each such point by taking the ray from the origin passing through the point, to the boundary of the convex body K. Then one can clearly envelope the convex body K with finitely many supporting hyperplanes resulting in a convex polyhedron. However we require this to be bounded, i.e. a polytope. Indeed by choosing α_n to be of the order of $\frac{\sigma C_1 \kappa_n}{n}$, where κ_n is a slowly diverging sequence, such a construction is guaranteed with high probability.

Step III: Construct an enveloping polytope of the critical set.

Now that we have enveloped our convex body σK entirely, we can indeed apply our reflection trick so that $\operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right) \subseteq -\alpha_n P + \mathbf{v}$, as with our convex polytope. Note that here, $\sigma K \subseteq P$ is our enveloping polytope for the convex body. In essence we have enveloped our (possibly) non-polytopal critical set, with a convex polytope, with high probability.

Step IV: Bound the estimation error by the diameter of the envelop-

ing polytope.

Finally, we can apply the identical approach from Step III in the convex polytope proof to bound our estimation error, over general convex bodies, as required. The full details of the proof are found in Section 4.E.

Additionally, although σ is treated as a nuisance parameter, we can show that estimation error using $\widehat{\sigma}_{\mathrm{MLE}}$, can also be done at an n^{-1} rate. This is formally described in Proposition 4.31.

Proposition 4.31 (Consistency of the scale parameter MLE, $K \in \mathcal{K}^d$.)**.** *Assume that the same conditions as Theorem 4.29 hold, and let $G > 0$ denote the Lipschitz constant of the Minkowski gauge functional $\rho_K(\mathbf{x})$. We then have that $|\widehat{\sigma}_{\mathrm{MLE}} - \sigma| \leq \frac{\sigma\kappa_n}{n}(GC_1 + 1)$, with probability at least $1 - 2\exp(-C_2\kappa_n/\operatorname{polylog}_d(\kappa_n))$, where C_1, C_2 are as defined in Theorem 4.29.*

4.4.3 Lower bounds

We now turn our attention to proving minimax lower bounds for our location estimator. The main result is summarized in Theorem 4.32.

Theorem 4.32 (Minimax lower bound for location estimation)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Let $\widehat{\mathbf{v}}$, be any estimator (measurable function) for the location parameter $\mathbf{v}$. We then have that the following holds:*

$$\inf_{\widehat{\mathbf{v}}} \sup_{\mathbf{v}\in\mathbb{R}^d} \mathbb{P}\left(\|\widehat{\mathbf{v}} - \mathbf{v}\|_2 \geq \sup_{\mathbf{z}\in\mathbb{S}^{d-1}} \frac{\sigma \operatorname{vol}_d(K)}{n \operatorname{vol}_{d-1}(K \mid \mathbf{z}^\perp)}\right) \geq \frac{1}{2}. \tag{4.18}$$

Here $K \mid \mathbf{z}^\perp$ is the image of the orthogonal projection of K onto the orthogonal complement of $\mathbf{z}$. Note that $\mathbf{z}^\perp := \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \langle \mathbf{x}, \mathbf{z}\rangle = 0\right\}$, i.e., the hyperplane through $\mathbf{0} \in \mathbb{R}^d$, with $\mathbf{z} \in \mathbb{R}^d$ as a normal vector.

From Theorem 4.32 we note that our lower bounds are of the form $\frac{c(d,K)}{n}$, and thus match the upper bounds in sample complexity, in our fixed dimension setting. The bounds also capture similar dependence on σ and K, which are both fixed constants in our setting. It is worthwhile here giving a sketch of the proof techniques used in constructing our lower bounds. We note from the outset that since we are working over a compact support the KL-divergence between to perturbed uniform densities over compact support is infinite. Thus many standard techniques, e.g., Fano's method (Yu, 1997) become challenging to naively apply in this setting. In short, we need to better exploit the underlying convex support structure of the multivariate uniform distributions. This proof sketch can be described as follows:

Step I: Set up the perturbed multivariate uniform distributions.

First we note that $K \in \mathcal{K}^d$ is fixed with centroid $(K) = \mathbf{0}$. Since we are in the known scaling regime, we will assume WLOG that $\sigma = 1$. Let $K_{\mathbf{z}} \coloneqq K + \mathbf{z} \in \mathcal{K}^d$ denote a translation of it by some perturbation $\mathbf{z} \in \mathbb{R}^d$. Further let $X_1 \sim \mathsf{Unif}[K]$ and $X_2 \sim \mathsf{Unif}[K_{\mathbf{z}}]$. As we will see, the minimax rate will be captured by the magnitude of this perturbation, i.e., $\|\mathbf{z}\|_2$. The perturbation $\mathbf{z}$, will be determined accordingly later.

Step II: Compute exact TV distance between the two perturbed distributions.

Since that we are working with multivariate uniform distributions, the total variation between X_1, X_2 can in fact be computed in *closed form*. This circumvents the previously noted issues with the KL-divergence related approaches. More specifically we have

$$\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right) = \frac{1}{2}\left(\frac{\mathrm{vol}_d\left(K \triangle K_{\mathbf{z}}\right)}{\mathrm{vol}_d\left(K\right)}\right) = \frac{\mathrm{vol}_d\left(K \setminus K_{\mathbf{z}}\right)}{\mathrm{vol}_d\left(K\right)} \tag{4.19}$$

Geometrically, it is a scaling of the symmetric set difference of $K, K + \mathbf{z}$. This is not so surprising, since the multivariate uniform distributions are in this case entirely defined by their compact convex supporting sets.

Step III: Upper bound the TV distance further using the 'sweep set'.

This set difference of (4.19) can be bounded further by considering the *sweep set* of the convex body K, by perturbation $\mathbf{z}$, i.e., $K + [0,1]\mathbf{z} \coloneqq \{\mathbf{k} + \lambda \mathbf{z} \mid \lambda \in [0,1], \mathbf{k} \in K\}$. This is shown visually in Section 4.4.3. One can show using geometric properties of the sweep set (Schymura, 2014; Gardner, 2006) that the required total variation distance can be upper bounded as:

$$\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right) = \frac{\mathrm{vol}_d\left(K \setminus K_{\mathbf{z}}\right)}{\mathrm{vol}_d\left(K\right)} \leq \frac{\mathrm{vol}_d\left(K \triangle K + [0,1]\mathbf{z}\right)}{2\,\mathrm{vol}_d\left(K\right)} = \frac{\|\mathbf{z}\|_2\,\mathrm{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)}{2\,\mathrm{vol}_d\left(K\right)}, \tag{4.20}$$

The RHS now ensures that $\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right)$ can be bounded by controlling the norm of the perturbation $\mathbf{z}$, i.e., $\|\mathbf{z}\|_2$.

Step IV: Apply Le Cam's Lemma to derive the lower bound.

The minimax lower bound can then be constructed with an application of Le Cam's lemma Yu (1997, Lemma 1) with n vectors each sampled uniformly from K, and $K + \mathbf{z}$. Using the subadditivity property of the total variation distance applied to the derived upper bound, Then by choosing $\mathbf{z}$ such that $\|\mathbf{z}\|_2 = \frac{\sigma\,\mathrm{vol}_d\left(K\right)}{n\,\mathrm{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)}$ implies that we lower bound our minimax risk away from 0. Finally, taking the supremum over all possible directions for our chosen perturbation vector, completes the proof. Full details are provided in Section 4.E.4.

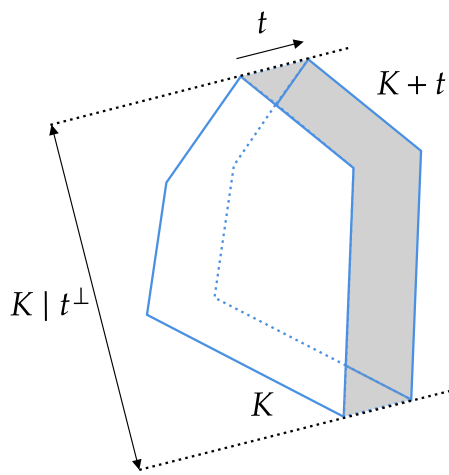

Figure 4.4.2: An illustration of the sweep set (shaded grey) of the convex body K by the vector **t**

Remark 4.33. When we have more knowledge of $K \in \mathcal{K}^d$, we can further refine the lower bound in Theorem 4.32. Observe the following formulas as per Koldobsky et al. (2016, Section 2):

$$\mathrm{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right) = \frac{1}{2}\int_{\mathbb{S}^{d-1}} |\langle x, \mathbf{z}\rangle| \,\mathrm{d}S_K, \tag{4.21}$$

$$\mathrm{vol}_d\,(K) = \frac{1}{d}\int_{\mathbb{S}^{d-1}} h_K \,\mathrm{d}S_K, \tag{4.22}$$

where $\mathbf{z}$ is a unit vector, $h_K \colon \mathbb{R}^d \to (-\infty, \infty]$ with $h_K(\mathbf{x}) := \sup_{\mathbf{y}\in K}\langle \mathbf{x}, \mathbf{y}\rangle$ is the support function of K, and S_K is its surface measure. When $K \in \mathcal{K}^d$ is symmetric, i.e., $K = -K$, one can pick $\mathbf{z}'$ to be the point on the boundary farthest from $\mathbf{0} \in \mathbb{R}^d$, so that $\|\mathbf{z}'\|_2 = \frac{\mathrm{diam}(K)}{2}$. Taking $\mathbf{z} := \frac{\mathbf{z}'}{\|\mathbf{z}'\|_2}$ gives $|\langle \mathbf{x}, \mathbf{z}\rangle| \leq \frac{h_K(\mathbf{x})}{\|\mathbf{z}'\|_2}$, and hence

$$\frac{\mathrm{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)}{\mathrm{vol}_d\,(K)} \leq \frac{d}{\mathrm{diam}\,(K)} \tag{4.23}$$

On the other hand for a general set K by the paper above, we know there exists a direction $\mathbf{z}$ such that $c\,\mathrm{vol}_d\,(K)^{(d-1)/d}\sqrt{d} \geq \mathrm{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)$.

4.5 Algorithmic Implementation

In this section, we review efficient algorithms with provable guarantees to compute our proposed estimators in practical settings. We consider both location-scale regimes i.e. where $\mathbf{v}$ is unknown with known scale parameter σ. And the more general regime where both $\mathbf{v}$ and σ are unknown. Unlike previous sections of our paper, we begin in Section 4.5.3 by first showing algorithmic approach to the unknown scaling regime using the subgradient method. We then demonstrate in Section 4.5.2 how we can utilize the constrained subgradient method to compute the estimator in the known σ setting.

4.5.1 Subgradients of the Minkowski gauge functional

As noted, in developing practical algorithms for estimating $\mathbf{v}$ and σ in we will primarily use variations of the subgradient method from convex optimization. In particular, we need a means of computing subgradients of the Minkowski gauge functional, $\rho_K(\mathbf{x})$, as defined in Definition 4.24. Here, we collect some useful results on the Minkowski gauge functional, which will ensure that our proposed algorithms will have precise convergence guarantees. As noted in Section 4.3, $\rho_K(\mathbf{x})$ is a convex and Lipschitz functional on $\mathbb{R}^d$. This gives

Lemma 4.34 (Bounded subgradient of $\rho_K(\mathbf{x})$)**.** *Let $G > 0$ be the Lipschitz constant for the Minkowski gauge functional, $\rho_K(\mathbf{x})$, as defined in Definition 4.24. Then for any $\mathbf{w} \in \mathbb{R}^d$, and for any $\mathbf{z} \in \partial(\rho_K(\mathbf{w}))$, we have $\|\mathbf{z}\|_2 \leq G$.*

We also note the well known fact that since the pointwise maximum of a finite set of convex functions is convex (see Mordukhovich and Nam (2014, Proposition 1.38(ii)) for a proof). Moreover, the pointwise maximum of a finite set of Lipschitz functions is also Lipschitz (see Lemma 4.61 in Section 4.A.3 for a proof). It follows that $\max_{i\in[n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$ is convex and Lipschitz. As such a natural algorithm would be to use subgradient method. We first show how to construct subgradients of $\rho_K(\mathbf{x})$, in the case where we have knowledge of supporting hyperplanes to the convex body $K \in \mathcal{K}^d$ at any boundary point of K.

Proposition 4.35 (Subgradient of $\rho_K(\mathbf{x})$)**.** *Let $K \in \mathcal{K}^d$ with $\mathbf{0} \in \operatorname{int}(K)$. Let $\rho_K(\mathbf{x})$ be it is Minkowski gauge functional, for all $\mathbf{x} \in \mathbb{R}^d$. Further, let $\mathbf{x}'$ be the vector that is parallel to $\mathbf{x}$ and lies on the boundary of K, and let $\mathbf{x}'^*$ be any supporting hyperplane through $\mathbf{x}'$. Then $\frac{\mathbf{x}'^*}{\mathbf{x}'^{*\top}\mathbf{x}'}$ is a subgradient at $\mathbf{x}'$.*

Next one needs to realize that a subgradient of a maximum of a finite set of convex functions is a subgradient of the convex function which achieves the maximum. This shows that if one can evaluate the Minkowski gauge functional efficiently one can find the max, and then one needs to find the supporting hyperplane at the max as described above.

4.5.2 Algorithmic implementation (v unknown, σ known)

Recall from (4.11), that in the case where σ is known our estimator for the location parameter $\mathbf{v} \in \mathbb{R}^d$ is given by:

$$\widehat{\mathbf{v}}_{\text{prj}} \coloneqq \Pi_{\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)}(\overline{\mathbf{v}}) \tag{4.24}$$

We then observe that this problem is finding the projection onto the (non-empty) intersection of a finite number of closed convex sets from $\overline{\mathbf{v}}$. Algorithms which are known to provably converge in such cases are given by the alternating projection based methods as developed in von Neumann (1950); Boyle and Dykstra (1986). However since we are working with an intersection of arbitrary convex bodies K, the convergence rates to the optimal solution are not directly available for this method. For example, the such convergence rates are typically known only for special *non-compact* closed convex sets such as subspaces as per Deutsch (1985). Moreover, such algorithms require knowledge of how to (Euclidean) project onto the specific convex body K, which can be a non-trivial exercise in itself, depending on its geometric structure. As such for practical purposes, we will utilize subgradient methods to find an efficient algorithm with not only convergence guarantees on the optimal value, but also on the *rate* of convergence.

In this known σ regime, instead of projecting onto the intersection of the convex bodies as per (4.24) it may be easier instead to solve the following *constrained* convex optimization problem for $\boldsymbol{\tau}$

$$\min \ \|\overline{\mathbf{v}} - \boldsymbol{\tau}\|_2 \tag{4.25}$$

$$\text{s.t.} \ \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \leq \sigma. \tag{4.26}$$

This constrained convex optimization problem is *equivalent* to the original problem i.e. an ℓ_2-projection onto the intersection of the convex bodies. To see this equivalence formally, we first observe that the ℓ_2-projection in (4.24) is a unique vector in the critical set, which minimizes the distance to the sample mean. This requirement is equivalently captured in (4.25). We further observe that per (4.15) we have the equivalence:

$$\boldsymbol{\tau} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \iff \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \leq \sigma. \tag{4.27}$$

Then using (4.27) shows the equivalence of the original projection onto the critical set problem in (4.24) to the constrained optimization problem as defined in (4.25) and (4.26). In summary, we have shifted our focus from the original estimation problem in (4.24), to solving the equivalent constrained convex optimization problem as defined by both (4.25) and (4.26).

For reasons which will be made clear later, we instead recommend solving the slightly *augmented* constrained convex optimization problem

$$\min \|\overline{\mathbf{v}} - \boldsymbol{\tau}\|_2 \tag{4.28}$$

$$\text{s.t.} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \leq \sigma\,(1 + 1/n)\,. \tag{4.29}$$

By directly comparing (4.26) to (4.29), we see that the solution to (4.28) and (4.29) is the same as that obtained by projecting $\overline{\mathbf{v}}$ onto the larger (dilated) set $\sigma\,(1+1/n)\,K$, rather than projecting onto σK. In other words, this new problem is an *augmented* version of Equation (4.24) in this precise sense. Since $\sigma\,(1+1/n)\,K$ is again a convex body, the projection has a unique solution. Let $\boldsymbol{\tau}^{\star}$ denote the *unique* optimal solution to the augmented convex program specified in (4.28) and (4.29). The main difference is Since $\sigma < \sigma\,(1+1/n)$, we then have using Proposition 4.27 the following nested relationship:

$$\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \subset \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma\,(1+1/n)\,K)\,, \tag{4.30}$$

and so $\mathbf{v} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma\,(1+1/n)\,K)$. Then letting $\mathbf{y}_1 = \mathbf{Y}_1$, it follows that:

$$\|\mathbf{y}_1 - \boldsymbol{\tau}^{\star}\|_2 \leq \sigma\,(1+1/n)\,\mathrm{diam}\,(K) \tag{4.31}$$

$$\|\mathbf{y}_1 - \mathbf{v}\|_2 \leq \sigma\,(1+1/n)\,\mathrm{diam}\,(K)\,. \tag{4.32}$$

Then let $R \coloneqq \sigma\,(1+1/n)\,\mathrm{diam}\,(K) > 0$, from which it follows that $\|\mathbf{y}_1 - \mathbf{v}\|_2 \leq \sigma\,\mathrm{diam}\,(K) < R$. Since $\max_{i\in[n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$ is convex and Lipschitz, we then can appeal to the *constrained* subgradient method per Boyd and Park (2014, Section 7) to solve this convex optimization problem. This is again applicable when one has a supporting hyperplane oracle of K and one can evaluate its Minkowski gauge functional efficiently. Moreover, in order for the constrained subgradient method to converge, we require that Slater's condition holds, i.e., the problem is strictly feasible. In our case, we this means that there exists some point $\mathbf{x}^{\mathrm{sf}} \in \mathbb{R}^d$, with $\max_{i\in[n]} \rho_K(\mathbf{Y}_i - \mathbf{x}^{\mathrm{sf}}) < \sigma\,(1+1/n)$. We also require diminishing nonsummable step sizes per Boyd and Park (2014, Section 2.2), i.e., step-sizes t_k chosen to satisfy

$$t_k > 0, \quad \lim_{k\to\infty} t_k = 0, \quad \sum_{k=1}^{\infty} t_k = \infty, \tag{4.33}$$

for each iteration index $k \in \mathbb{N}$. The formal constrained subgradient method as required for our setup is formalized in Algorithm 4.1, by setting $\varepsilon > 0$, $\mathbf{x}^{(0)}$ to $\mathbf{Y}_1$, and with t_k chosen per (4.33).

Algorithm 4.1 Subgradient Descent ($\mathbf{v}$ unknown, σ known)

procedure ConstrainedSubgradientDescentGauge$\left(\left(\mathbf{x}^{(0)}, (t_k)_{k=1}^{\infty}\right)\right)$

$\quad \mathbf{g}^{(0)} \leftarrow \frac{\mathbf{x}_0'^{*}}{\mathbf{x}_0'^{*\top}\mathbf{x}_0'}$, $\mathbf{x}_{\text{best}}^{(0)} \leftarrow \mathbf{x}^{(0)}$

$\quad$ **for** $k = 1, \ldots, L$ **do**

$$\mathbf{g}^{(k-1)} \in \begin{cases} \partial\left(\|\overline{\mathbf{v}} - \mathbf{x}^{(k-1)}\|_2\right), & \text{if } \max_i \rho_K(\mathbf{Y}_i - \mathbf{x}^{(k-1)}) \leq \sigma, \\ \partial\left(\rho_K(\mathbf{Y}_{i^*} - \mathbf{x}^{(k-1)})\right) & \text{for } i^* = \arg\max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}^{(k-1)}), \\ & , \text{ if } \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}^{(k-1)}) > \sigma \end{cases}$$

$\quad\quad \mathbf{x}^{(k)} \leftarrow \mathbf{x}^{(k-1)} - t_k \cdot \mathbf{g}^{(k-1)}$

$\quad\quad \mathbf{x}_{\text{best}}^{(k)} \leftarrow \arg\min\{\|\overline{\mathbf{v}} - \mathbf{x}_{\text{best}}^{(k-1)}\|_2, \|\overline{\mathbf{v}} - \mathbf{x}^{(k)}\|_2\}$

$\quad$ **return** $\mathbf{x}_{\text{best}}^{(L)}$

Proposition 4.36 (Convergence of subgradient method)**.** *Suppose that there exist constants $R, G > 0$ with $\left\|\mathbf{x}^{(0)} - \mathbf{x}^{\star}\right\|_2 \leq R$, $\left\|\mathbf{x}^{(0)} - \mathbf{v}\right\|_2 \leq R$, and $\left\|\mathbf{g}^{(k)}\right\|_2 \leq G$ for all k. Where $\mathbf{x}^{\star}$ is the unique optimal solution, then running Algorithm 4.1 ensures that*

$$\|\overline{\mathbf{v}} - \mathbf{x}_{\text{best}}^{(L)}\|_2 - \|\overline{\mathbf{v}} - \mathbf{x}^{\star}\|_2 \leq \frac{R^2 + G^2 \sum_{k=1}^{L} t_k^2}{2\sum_{k=1}^{L} t_k} \tag{4.34}$$

Remark 4.37 (Application to Polytopes). We note that the subgradient method described here readily applies to the case where K is a convex polytope. In particular if $[\mathbf{A} \mid \mathbf{b}]$ is the halfspace representation of K, then the supporting hyperplane at any given boundary point can computed directly from the supporting hyperplane on which it lies. The resulting subgradient can be readily computed as described using this supporting hyperplane information, as previously described. If the point lies on an edge or vertex of a polytope then, by definition it lies in the intersection of two or more halfspaces. This is not an issue, since *any* one of these halfspaces can be chosen at random in such cases to derive the subgradient as described. However, since polytopes have more geometric structure, we provide a more simplified algorithmic approach in this case (under both scaling regimes) using linear and quadratic programming. These details are provided in Section 4.5.5.

4.5.3 Algorithmic Implementation (v unknown, σ unknown)

In this case where σ is unknown our estimator for the location parameter $\mathbf{v} \in \mathbb{R}^d$ is given by the MLE per Theorem 4.25 as the solution $\boldsymbol{\tau}$ to the following unconstrained optimization

$$\arg\min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}). \tag{4.35}$$

Since we have established that $\max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$ is convex and Lipschitz, (4.35) is again a convex optimization problem. However, since σ is unknown, we no longer have it as a constraint as in Section 4.5.2. This time, we can directly appeal to the regular (i.e., *unconstrained*) subgradient method to compute $\widehat{\mathbf{v}}_{\text{MLE}}$ in our setting. This is formalized in Algorithm 4.2, for $\varepsilon > 0$, and setting $\mathbf{x}^{(0)}$ to $\mathbf{Y}_1$.

Algorithm 4.2 Subgradient Method Gauge Unconstrained (**v** unknown, σ unknown)

---

procedure CONSTRAINEDSUBGRADIENTDESCENTGAUGE$\left(\left(\varepsilon, \mathbf{x}^{(0)}, (t_k)_{k=1}^{\infty}\right)\right)$
    $\mathbf{g}^{(0)} \leftarrow \partial\left(\rho_K(\mathbf{Y}_i - \mathbf{x}^{(0)})\right)$, $\mathbf{x}_{\text{best}}^{(0)} \leftarrow \mathbf{x}^{(0)}$
    for $k = 1, \ldots, L$ **do**
        $\mathbf{g}^{(k-1)} \in \partial\left(\rho_K(\mathbf{Y}_i - \mathbf{x}^{(k-1)})\right)$
        $\mathbf{x}^{(k)} \leftarrow \mathbf{x}^{(k-1)} - t_k \cdot \mathbf{g}^{(k-1)}$
        $\mathbf{x}_{\text{best}}^{(k)} \leftarrow \arg\min\{\max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}_{\text{best}}^{(k-1)}), \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}^{(k)})\}$
    return $\mathbf{x}_{\text{best}}^{(L)}$

Importantly, a similar convergence guarantee as per Proposition 4.36, holds in this setting and summarized in Proposition 4.38.

Proposition 4.38 (Convergence of subgradient method)**.** *Suppose that there exist constants $R, G > 0$ with $\left\|\mathbf{x}^{(0)} - \mathbf{x}^{\star}\right\|_2 \leq R$, $\left\|\mathbf{g}^{(k)}\right\|_2 \leq G$ for all k. Where $\mathbf{x}^{\star}$ is the unique optimal solution, then running Algorithm 4.2 ensures that*

$$\max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}_{\text{best}}^{(L)}) - \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}^{\star}) \leq \frac{R^2 + G^2 \sum_{k=1}^{L} t_k^2}{2 \sum_{k=1}^{L} t_k} \tag{4.36}$$

4.5.4 Accounting for ε-suboptimality of the subgradient method

With the subgradient methods, as discussed in the previous sections, you can only get to within ε of the true minimum value. Here ε is the RHS of (4.36).

This in turn means that in theory our estimated $\widetilde{\sigma} \coloneqq \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \mathbf{x}_{\text{best}}^{(L)})$ can either be $0 < \widehat{\sigma}_{\text{MLE}} \leq \widetilde{\sigma} \leq \sigma$ or be bigger than the true scale parameter σ i.e. $0 < \widehat{\sigma}_{\text{MLE}} \leq \sigma \leq \widetilde{\sigma}$. However, the latter is not a concern[6]. This is formalized in Theorem 4.39.

Theorem 4.39. *If you run the subgradient descent in Algorithm 4.2, such that the* RHS *in* (4.36) *is at most* $\frac{C}{n}$ *for some sufficiently small* $C > 0$*, then*

$$\|\mathbf{x}_{\text{best}}^{(L)} - \mathbf{v}\|_2 \lesssim \frac{1}{n}, \tag{4.37}$$

with high probability (say .99).

Remark 4.40. An analagous theorem to Theorem 4.39 applies to Algorithm 4.1, where σ is known.

4.5.5 Estimating location-scale parameters of convex polytopes

As noted in Remark 4.37, if $K \in \mathcal{K}^d$ is a convex *polytope*, then location estimation subgradient method algorithms described in Sections 4.5.3 and 4.5.2 still apply, under both known and unknown scaling regimes. However, when $K \in \mathcal{K}^d$ is a convex polytope, we can more efficiently estimate the location parameter $\mathbf{v}$, under both scaling regimes. We can reduce such computations to linear and quadratic programs, making it particularly useful in practical situations.

To that end, let $K \coloneqq [\mathbf{A} \mid \mathbf{b}]$ be the $\mathcal{H}$-representation of the given polytope K. That is K is uniquely defined (up to ordering) by the set of m *coordinate-wise* inequalities (i.e., halfspaces) $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, for each $\mathbf{x} \in \mathbb{R}^d$. Here $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$ are both known. We then have that $\sigma K = [\mathbf{A} \mid \sigma \mathbf{b}]$. Since our observations are generated as $\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \sigma \mathbf{X}_i$, for each $i \in [n]$, we have the equivalent m coordinate-wise inequalities

$$\mathbf{A}(\mathbf{v} + \sigma \mathbf{X}_i) \leq \mathbf{A}\mathbf{v} + \sigma \mathbf{b} \iff \mathbf{A}\mathbf{Y}_i - \sigma \mathbf{b} \leq \mathbf{A}\mathbf{v}. \tag{4.38}$$

To simplify notation in what follows, we define the matrices $\mathbf{M}, \mathbf{B} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{M} \coloneqq \begin{bmatrix} \uparrow & & \uparrow & & \uparrow \\ \mathbf{A}\mathbf{Y}_1 & \cdots & \mathbf{A}\mathbf{Y}_i & \cdots & \mathbf{A}\mathbf{Y}_n \\ \downarrow & & \downarrow & & \downarrow \end{bmatrix}, \quad \text{and} \quad \mathbf{B} \coloneqq \begin{bmatrix} \uparrow & & \uparrow & & \uparrow \\ \mathbf{b} & \cdots & \mathbf{b} & \cdots & \mathbf{b} \\ \downarrow & & \downarrow & & \downarrow \end{bmatrix}. \tag{4.39}$$

[6]We note that σ is always *fixed* and *positive*, as per our assumed multivariate uniform location-scale generating process in Definition 4.4

Then for each coordinate $k \in [m]$, we then have the equivalent constraints:

$$\max_{i \in [n]} [\mathbf{M} - \sigma \mathbf{B}]_{ki} \leq [\mathbf{A}\mathbf{v}]_k \iff \max_{i \in [n]} [\mathbf{M}]_{ki} \leq [\mathbf{A}\mathbf{v} + \sigma \mathbf{b}]_k \tag{4.40}$$

where $[\mathbf{M} - \sigma \mathbf{B}]_{ki}$ denotes the entry in row k, and column i, of $\mathbf{M} - \sigma \mathbf{B} \in \mathbb{R}^{m \times n}$. Note that the maximum in (4.40) are always taken in a row-wise manner. Furthermore, the critical polytope here can then be represented as:

$$\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) = \left\{ \boldsymbol{\tau} \in \mathbb{R}^d \,\middle|\, \max_{i \in [n]} [\mathbf{M}]_{ki}, \text{for each } k \in [m] \leq [\mathbf{A}\boldsymbol{\tau} + \sigma \mathbf{b}]_k \right\} \tag{4.41}$$

With this set up, we can now address the estimation of location-scale parameters under this polytope setting under both scaling regimes.

Known σ regime (via Quadratic Programming):

In this case we can solve the following quadratic program (QP) for $\boldsymbol{\tau}$, to estimate $\widehat{\mathbf{v}}_{\text{prj}}$:

$$\min \; \|\overline{\mathbf{v}} - \boldsymbol{\tau}\|_2^2 \tag{4.42}$$

$$\text{s.t.} \;\; \max_{i \in [n]} [\mathbf{M}]_{ki} \leq [\mathbf{A}\boldsymbol{\tau} + \sigma \mathbf{b}]_k \,. \tag{4.43}$$

Note that the feasible points from (4.43) forms critical set as per (4.41).

Unknown σ regime (via Linear Programming):

In this case we can solve the following linear program (LP) for $\boldsymbol{\tau}$, and γ, to compute the estimates $\widehat{\mathbf{v}}_{\text{MLE}}$, and $\widehat{\sigma}_{\text{MLE}}$, respectively.

$$\min \gamma \tag{4.44}$$

$$\text{s.t.} \;\; \max_{i \in [n]} [\mathbf{M}]_{ki} \leq [\mathbf{A}\boldsymbol{\tau} + \gamma \mathbf{b}]_k \tag{4.45}$$

$$\text{and } \gamma \geq 0. \tag{4.46}$$

4.6 Simulations

We provide simulations[7] showing the efficiency of our estimator in artificial settings on various convex bodies, and also in an applied setting motivated by measurement error. Our simulations are done in the case where $d = 2$ in the case of a convex nonagon (9-sided regular polygon). The reason for this is twofold. First in $d = 2$ we can easily simulate our our sampling process for

[7] Reproducible code available at https://github.com/shamindras/sce.

$\mathbf{Y}_i$ and more importantly visualize it to confirm our intuition. Second, in this $2D$ setting (for $\sigma := 1$ fixed regime), we can calculate the Pitman estimator directly, since the intersection sets is a polygon, whose centroid has closed form. For example the plot Figure 4.6.1 shows the progressive simulation of our projection intersection set (in green) as the sample size n increases. The Pitman estimator is the blue centroid of this intersection set. As expected with an increase in the sample size n, the red dashed-line nonagon representing $\widehat{\mathbf{v}}_{\text{pit}} + K$, quickly overlaps the true shifted $\mathbf{v} + K$ black dashed-line nonagon. In other words $\widehat{\mathbf{v}}_{\text{pit}}$ quickly converges to the true parameter $\mathbf{v}$, as measured by square loss error. Such 2D centroid computations are readily implemented in open source software packages (Gillies et al., 2007; Pebesma, 2018).

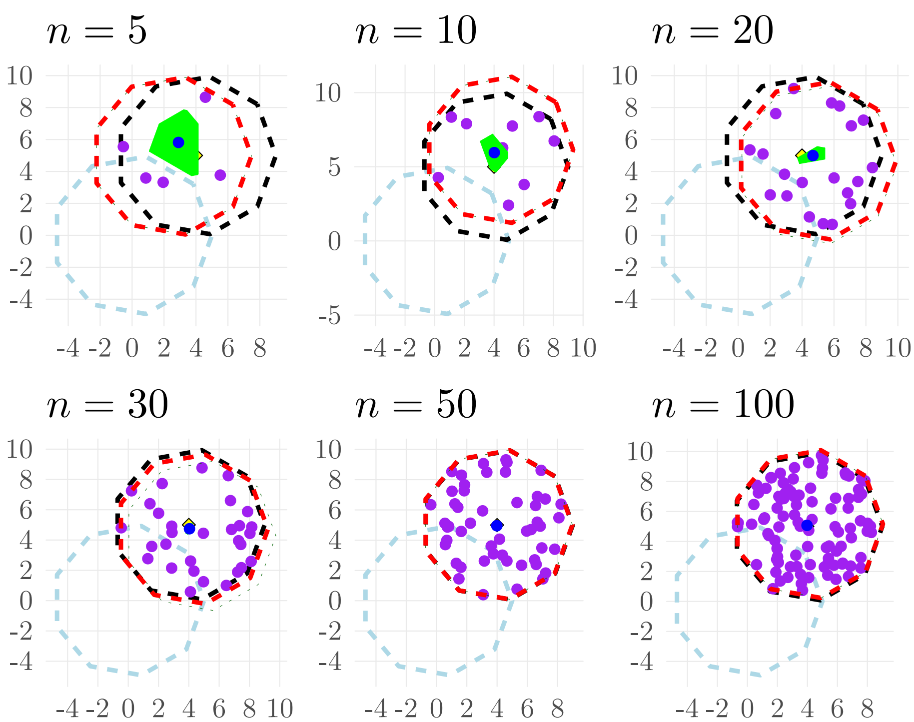

Figure 4.6.1: The intersection set for different sample sizes for a 2D nonagon

We can also view the squared loss errors as a function of our sample size n, as seen in Figure 4.6.2. Here we calculate the Pitman estimator explicitly as the centroid of the intersection polygon. We estimate the location parameter $\mathbf{v}$ using the sample mean ($\overline{\mathbf{v}}$), the Pitman location estimator ($\widehat{\mathbf{v}}_{\text{pit}}$), and our projection location estimator ($\widehat{\mathbf{v}}_{\text{prj}}$). We then measure the estimation error using the Euclidean norm with 100 replications for each value of $n \in \{5, 10, 20, 30, 50, 100\}$. We plot the mean and the 25^{th} and 75^{th} quantiles for each estimator, for each value of n. As expected the Pitman estimator is statistically optimal (uniformly) across all replications, but our projection location estimator begins to converge closer (in estimation error) to it, and both at a faster rate than the sample mean.

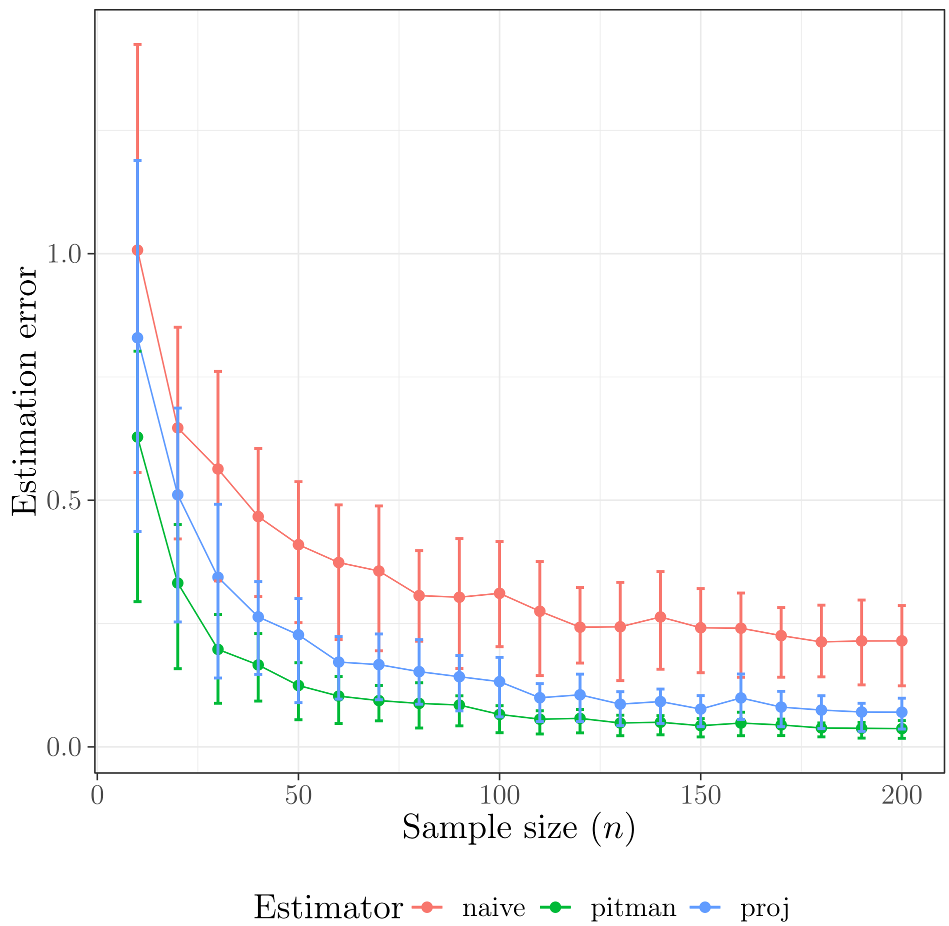

Figure 4.6.2: Estimation error in L_2-loss for estimators: $\overline{\mathbf{v}}$ (naive), $\widehat{\mathbf{v}}_{\mathrm{pit}}, \widehat{\mathbf{v}}_{\mathrm{prj}}$.

4.7 Discussion

In this paper we have proposed several estimators for location-scale parameters in our multivariate uniform setting over convex bodies $K \in \mathcal{K}^d$. In the known scaling regime we demonstrated a fundamental trade-off arising between the statistical optimality and the computational feasibility of estimation in general settings. In particular we showed that although the multivariate uniform Pitman estimator satisfies many favorable statistical decision-theoretic properties, it amounts to computing the centroid of the critical set, which is not practical apart from some very restricted settings. Motivated by this trade-off we proposed our projection location estimator in the known scaling regime, and demonstrate how to obtain location-scale MLEs in the unknown scaling regimes. Most importantly we show that these estimators lie in the critical set, and thus all converge at the rate of $\frac{C(d,K)}{n}$ with high probability. We support these rates with matching minimax lower bounds in sample complexity. Additionally we provide feasible algorithms with provable guarantees for our proposed estimators over more general settings compared to known estimators.

However, this opens up many exciting directions for further exploration. For example, our upper bounds hold for *any* estimator in the critical set, and as such suboptimal in the dimension dependent constant $C(d, K)$ compared to the minimax lower bounds. These could be tightened further by exploiting the convex-geometric structure of each individual estimator. If one seeks to perform inference on these location-scale parameters of interest, there are some approaches one could potentially adapt from techniques from Wasserman et al. (2020) which apply over non-regular models. However as the authors note that there are some known issues applying their techniques to uniform distributions. Working on extending techniques to our multivariate uniform setting poses an interesting new challenge. Another important direction is to understand whether our techniques yield minimax optimal estimators where the underlying distribution is non-uniform over $K \in \mathcal{K}^d$. We believe our upper bounds techniques can achieve similar risk convergence rates, provided that the underlying density is bounded away from zero in probability over the boundary of the convex supporting set K. We defer this and the above open problems to future work.

4.8 Acknowledgments

We would like to thank Arun Kumar Kuchibhotla, Alex Reinhart, Alessandro Rinaldo, Ryan J. Tibshirani, Larry Wasserman from the Carnegie Mellon University (CMU) Department of Statistics & Data Science, and Yang Ning from the Cornell Department of Statistics & Data Science. Their encouragement, and extensive feedback throughout this work greatly shaped the final outcome. This paper extensively utilizes the `R` statistical software (R Core Team, 2021) for conducting simulations and plots. In particular, we relied primarily on the

following `R` software packages (listed alphabetically): `cowplot` (Wilke, 2020), `patchwork` (Pedersen, 2020), `sf` (Pebesma, 2018), `tidyverse` (Wickham et al., 2019), and `uniformly` (Laurent, 2018). All other figures were drawn using the `Mathcha`[8] editor.

[8] https://www.mathcha.io/editor

Appendix - Chapter 4

4.A Preliminary

In this appendix we provide detailed proofs of all key statements from the main paper. We first review the notation, and then key facts from mathematical analysis, e.g. Lipschitz functions, convex analysis etc. They will be used repeatedly throughout the appendix.

4.A.1 Notation Summary

To ensure that the appendix is can be read in a standalone manner, we begin by consolidating the key notation used in the paper in Table 4.A.1[9].

4.A.2 Required convex analysis and convex geometry results

Since our work relies a variety of core ideas from convex analysis and convex geometry we first introduce some common definitions which will be referred to in subsequent proofs. Although many definitions and facts related to convex bodies are well known and found in convex analysis textbooks (e.g. (Deutsch, 2001; Niculescu and Persson, 2018)) we note them here to ensure that our work is largely self-contained[10].

Definition 4.41 (Convex Body)**.** A convex body in $K \subset \mathbb{R}^d$ is a compact convex set with a non-empty interior. Furthermore we denote the space of all convex bodies in $\mathbb{R}^d$ by $\mathcal{K}^d$, i.e., $\mathcal{K}^d \coloneqq \left\{ K \subset \mathbb{R}^d \,\middle|\, K \text{ is a convex body} \right\}$.

Lemma 4.42. *Let $d \geq 1$ be a fixed integer. Then, any convex body $K \in \mathcal{K}^d$ (as per Definition 4.41) has a strictly positive volume (Lebesgue measure) i.e.* $\mathrm{vol}_d\left(K\right) > 0$.

[9] Unless stated otherwise $K \subseteq \mathbb{R}^d$ is a closed, non-empty convex set, and $\Omega \subseteq \mathbb{R}^d$. We further assume that $\mathbf{v} \in \mathbb{R}^d$, and that $A, B \subseteq \mathbb{R}^d$ are non-empty sets.

[10] For reader convenience, where possible, we try to provide short proofs to keep the text self-contained or provide suitable detailed references thereof. On first reading we recommend skimming this section and thereafter referring to it as needed in subsequent proofs.

Table 4.A.1: Notation and conventions used in this chapter

	Variables and inequalities
$a \wedge b$	$\min\{a, b\}$ for each $a, b \in \mathbb{R}$
$a \vee b$	$\max\{a, b\}$ for each $a, b \in \mathbb{R}$
scalars	$x, y, z \in \mathbb{R}$
vectors	$\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$
matrices	$\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{d \times m}$
$\lesssim$	$\leq$ up to positive universal constants
$\gtrsim$	$\geq$ up to positive universal constants
$a_n = \mathcal{O}(1)$	$(\exists C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\lvert a_n \rvert < C)$
$a_n = \mathcal{O}(b_n)$	$\frac{a_n}{b_n} = \mathcal{O}(1)$
$a_n = o(1)$	$(\forall C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\lvert a_n \rvert < C)$
$a_n = o(b_n)$	$\frac{a_n}{b_n} = o(1)$
$X_n = o_P(1)$	$(\forall \varepsilon > 0)(\mathbb{P}(\lvert X_n \rvert \geq \varepsilon) \xrightarrow{n \to \infty} 0)$
$X_n = \mathcal{O}_P(1)$	$(\forall \varepsilon > 0)(\exists C > 0)(\exists N \in \mathbb{N})(\forall n \geq N)(\mathbb{P}(\lvert X_n \rvert \geq C) \leq \varepsilon)$
	Functions and sets
Indicator function $\mathbb{I}_\Omega(\mathbf{x})$	Takes value 1 when $x \in \Omega$, and 0 otherwise.
$\Pi_K : \mathbb{R}^d \to K$	ℓ_2-projection of any $\mathbf{x} \in \mathbb{R}^d$ onto K
$[n]$	$\{1, \ldots, n\}$, for $n \in \mathbb{N}$
$A + B$	$\{\mathbf{a} + \mathbf{b} \mid \mathbf{a} \in A,\ \mathbf{b} \in B\}$, i.e., *Minkowski Sum* of A, B.
$A - B$	$\{\mathbf{a} - \mathbf{b} \mid \mathbf{a} \in A,\ \mathbf{b} \in B\}$.
$\mathbf{v} + K$	$\{\mathbf{v} + \mathbf{k} \mid \mathbf{k} \in K\}$, i.e., *translation of* of A by $\mathbf{v}$.
μK, for $\mu \in \mathbb{R} \setminus \{0\}$	$\{\mu \mathbf{k} \mid \mathbf{k} \in K\}$
$\overline{B}_2^d(\mathbf{x}, r)$	$\left\{\mathbf{y} \in \mathbb{R}^d \,\middle\vert\, \lVert \mathbf{y} - \mathbf{x} \rVert_2 \leq r\right\}$.
$B_2^d(\mathbf{x}, r)$	$\left\{\mathbf{y} \in \mathbb{R}^d \,\middle\vert\, \lVert \mathbf{y} - \mathbf{x} \rVert_2 < r\right\}$.
$\mathbb{S}^{d-1}$	$\left\{\mathbf{x} \in \mathbb{R}^d \,\middle\vert\, \lVert \mathbf{x} \rVert_2 = 1\right\}$
λ_d	d-dimensional Lebesgue measure, $d \in \mathbb{N}$
$\mathrm{vol}_d(K)$	volume of K with respect to λ_d
$\mathrm{int}(K)$	interior of K.
$\mathrm{centroid}(K)$	centroid of K.

Proof of Lemma 4.42. Note that K is compact, and hence closed, thus it is Lebesgue measurable with respect to λ_d. Moreover, since $K \in \mathcal{K}^d$, it has a non-empty interior. So suppose WLOG (by translation invariance of λ_d), that $\mathbf{0} \in \text{int}(K)$. Then there exists an $\varepsilon > 0$ such that $B(\mathbf{0}, \varepsilon) \subseteq K$. Now, we have that $\text{vol}_d(B(\mathbf{0}, \varepsilon)) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}\varepsilon^d$, e.g., see Jones (1993, Chapter 9, Section C). As such $\text{vol}_d(B(\mathbf{0}, \varepsilon)) > 0$ for each $d \in \mathbb{N}$. We then have by monotonicity of the Lebesgue measure that

$$0 < \text{vol}_d(B(\mathbf{0}, \varepsilon)) \leq \text{vol}_d(K), \quad \text{for each } d \in \mathbb{N}. \tag{4.47}$$

As required. □

Definition 4.43 (Centroid of a Convex Body)**.** The centroid of a convex body $K \subset \mathbb{R}^d$, with strictly positive Lebesgue measure, is defined as follows:

$$\text{centroid}(K) := \frac{\int \mathbf{x}\mathbb{I}_K(\mathbf{x})\, \mathrm{d}\lambda_d(\mathbf{x})}{\int \mathbb{I}_K(\mathbf{x})\, \mathrm{d}\lambda_d(\mathbf{x})} = \frac{\int \mathbf{x}\mathbb{I}_K(\mathbf{x})\, \mathrm{d}\lambda_d(\mathbf{x})}{\text{vol}_d(K)}$$

Remark 4.44 (Centroid is Unique)*.* Since $\text{vol}_d(K) > 0$ per Lemma 4.42 we have that $\text{centroid}(K)$ is well defined for any convex body $K \subset \mathbb{R}^d$. Furthermore since $\text{centroid}(K)$, is defined in terms of the d-dimensional Lebesgue measure λ_d, it is unique in $\mathbb{R}^d$.

Definition 4.45 (Metric Projection onto a Convex Set)**.** Let K be a non-empty subset of $\mathbb{R}^d$, and let $\mathbf{x} \in \mathbb{R}^d$. The (possibly empty) set of Euclidean best approximations from $\mathbf{x}$ to K is denoted by $P_K(\mathbf{x})$, where

$$P_K(\mathbf{x}) := \left\{ \mathbf{y} \in K \;\middle|\; \|\mathbf{x} - \mathbf{y}\|_2 = \inf_{\mathbf{w} \in K} \|\mathbf{w} - \mathbf{y}\|_2 \right\}$$

This defines a mapping P_K from $\mathbb{R}^d$ into the subsets of K called the *metric projection* onto K.

Theorem 4.46 (Uniqueness of Metric Projections in Finite Dimensions)**.** *Let $K \subset \mathbb{R}^d$ be a non-empty, closed, and convex set. Then for each $\mathbf{x} \in \mathbb{R}^d$ we have that $P_K(\mathbf{x})$ is a unique mapping i.e. a singleton. In such cases we denote the singleton mapping $P_K(\mathbf{x})$ by $\Pi_K(\mathbf{x}) := \arg\min_{\mathbf{w} \in K} \|\mathbf{x} - \mathbf{w}\|_2$.*

Proof of Theorem 4.46. See (Deutsch, 2001, Theorem 3.5) and (Deutsch, 2001, Theorem 3.6) for details. □

Theorem 4.47 (Characterization of Best Approximations from Convex Sets)**.** *Let K be a convex subset of the inner product space $X, \mathbf{x} \in X$, and $\mathbf{y}_0 \in K$. Then $\{\mathbf{y}_0\} = P_K(\mathbf{x})$ if and only if*

$$\langle \mathbf{x} - \mathbf{y}_0, \mathbf{y} - \mathbf{y}_0 \rangle \leq 0$$

for all $\mathbf{y} \in K$

Proof of Theorem 4.47. See (Deutsch, 2001, Theorem 4.1) for details. □

Theorem 4.48 (Pythagorean Theorem of Convex Set Projections)**.** *Let $K \subset \mathbb{R}^d$ be a closed convex set, $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{y} = \Pi_K(\mathbf{x})$ Then for any $\mathbf{z} \in K$ we have:*

$$\|\mathbf{y} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{z}\|$$

Proof of Theorem 4.47. Since $K \subset \mathbb{R}^d$ be a closed convex set, from Theorem 4.46 $\mathbf{y} = \Pi_K(\mathbf{x})$ is well defined, and a singleton. We then observe that for any $\mathbf{z} \in K$:

$$\begin{aligned}
\|\mathbf{x} - \mathbf{z}\|^2 &= \|\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}\|^2 \\
&= \|\mathbf{x} - \mathbf{y}\|^2 + 2\langle \mathbf{x} - \mathbf{y}, \mathbf{y} - \mathbf{z} \rangle + \|\mathbf{y} - \mathbf{z}\|^2 \\
&\geq \|\mathbf{y} - \mathbf{z}\|^2 \quad (\text{From Theorem 4.47}, \langle \mathbf{x} - \mathbf{y}, \mathbf{y} - \mathbf{z} \rangle + \|\mathbf{y} - \mathbf{z}\| \geq 0)
\end{aligned}$$

This implies $\|\mathbf{y} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{z}\|$ as required. □

Theorem 4.49 (Translation and Scale invariance of Projection on Closed Convex Sets)**.** *Let $K \subset \mathbb{R}^d$ be a closed (non-empty) convex set, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and alpha $\in \mathbb{R} \setminus \{0\}$. We then have:*

$$\Pi_{K+\mathbf{y}}(\mathbf{x} + \mathbf{y}) = \Pi_K(\mathbf{x}) + \mathbf{y} \tag{4.48}$$

$$\Pi_{\alpha K}(\alpha \mathbf{x}) = \alpha \Pi_K(\mathbf{x}) \tag{4.49}$$

Proof of Theorem 4.49. See (Deutsch, 2001, Theorem 2.7) for details. □

Lemma 4.50. *For a given convex body $K \subset \mathbb{R}^d$ as per Definition 4.41, if $\mathbf{X} \sim \mathsf{Unif}[K]$ then $\text{centroid}(K) = \mathbb{E}(\mathbf{X})$. From this it follows that $\text{centroid}(K) = \mathbf{0} \iff \mathbb{E}(\mathbf{X}) = \mathbf{0}$.*

Proof of Lemma 4.50. Firstly, per Definition 4.1 $\mathbf{X} \sim \mathsf{Unif}[K] \iff f_{\mathbf{X}}(\mathbf{x}) \coloneqq \frac{\mathbb{I}_K(\mathbf{x})}{\mathrm{vol}_d(K)}$, where $f_{\mathbf{X}}(\mathbf{x})$ is the probability density function of $\mathbf{X}$. Now we have from Definition 4.43:

$$\begin{aligned}
\mathrm{centroid}\,(K) &\coloneqq \frac{\int_{\mathbb{R}^d} \mathbf{x}\mathbb{I}_K(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x})}{\int_{\mathbb{R}^d} \mathbb{I}_K(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x})} && \text{(by definition.)} \\
&= \frac{\int_{\mathbb{R}^d} \mathbf{x}\mathbb{I}_K(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x})}{\mathrm{vol}_d(K)} && \text{(since } \textstyle\int_{\mathbb{R}^d} \mathbb{I}_K(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x}) = \mathrm{vol}_d(K).\text{)} \\
&= \int_{\mathbb{R}^d} \mathbf{x}\frac{\mathbb{I}_K(\mathbf{x})}{\mathrm{vol}_d(K)}\,\mathrm{d}\lambda_d(\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \mathbf{x} f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x}) && \text{(since } f_{\mathbf{X}}(\mathbf{x}) \coloneqq \tfrac{\mathbb{I}_K(\mathbf{x})}{\mathrm{vol}_d(K)}.\text{)} \\
&=: \mathbb{E}(\mathbf{X}) && \text{(since } \mathbb{E}(\mathbf{X}) \coloneqq \textstyle\int_{\mathbb{R}^d} \mathbf{x} f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\lambda_d(\mathbf{x}).\text{)}
\end{aligned}$$

We have shown that $\mathrm{centroid}\,(K) = \mathbb{E}(\mathbf{X})$. From this it does indeed follow that $\mathrm{centroid}\,(K) = \mathbf{0} \iff \mathbb{E}(\mathbf{X}) = \mathbf{0}$ as required. □

Lemma 4.51. *Under the conditions of Lemma 4.50, if* $\mathbf{X} \sim \mathsf{Unif}[K]$ *and* $\mathbf{Y} = \mathbf{v} + \sigma\mathbf{X}$ *then* $\mathbf{Y} \sim \mathsf{Unif}[\mathbf{v} + \sigma K]$. *Further we have that* $\mathrm{centroid}\,(\mathbf{v} + \sigma K) = \mathbf{v}$.

Proof of Lemma 4.51. Firstly, since $\mathbf{X} \sim \mathsf{Unif}[K] \iff f_{\mathbf{X}}(\mathbf{x}) = \frac{\mathbb{I}_K(\mathbf{x})}{\mathrm{vol}_d(K)}$, where $f_{\mathbf{X}}(\mathbf{x})$ is the probability density function of $\mathbf{X}$. We observe that $\mathbf{Y}$ can be represented as an affine transformation $\mathbf{Y} = \mathbf{v} + (\sigma\mathbf{I}_p)\mathbf{X}$, where $\mathbf{I}_p \in \mathbb{R}^{p\times p}$ is the identity matrix. We then have:

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}) &= \frac{f_{\mathbf{X}}\left((\sigma\mathbf{I}_p)^{-1}(\mathbf{y} - \mathbf{v})\right)}{|\det(\sigma\mathbf{I}_p)|} && \text{(Affine density transformation.)} \\
&= \frac{f_{\mathbf{X}}\left(\frac{1}{\sigma}(\mathbf{y} - \mathbf{v})\right)}{\sigma^p} && \text{(Since } (\sigma\mathbf{I}_p)^{-1} = \tfrac{1}{\sigma}\mathbf{I}_p\text{, and } \det(\sigma\mathbf{I}_p) = \sigma^p\text{)} \\
&= \frac{\mathbb{I}_K\left(\frac{1}{\sigma}(\mathbf{y} - \mathbf{v})\right)}{\sigma^p\,\mathrm{vol}_d(K)} && \text{(Since } f_{\mathbf{X}}(\mathbf{x}) = \tfrac{\mathbb{I}_K(\mathbf{x})}{\mathrm{vol}_d(K)}\text{)} \\
&= \frac{\mathbb{I}_{\mathbf{v}+\sigma K}(\mathbf{y})}{\mathrm{vol}_d(\sigma K)} && \text{(since } \tfrac{1}{\sigma}(\mathbf{y} - \mathbf{v}) \in K \iff \mathbf{y} \in \mathbf{v} + \sigma K\text{)} \\
&= \frac{\mathbb{I}_{\mathbf{v}+\sigma K}(\mathbf{y})}{\mathrm{vol}_d(\mathbf{v} + \sigma K)} && \text{(translation invariance of Lebesgue measure)}
\end{aligned}$$

Since $f_{\mathbf{Y}}(\mathbf{y}) = \frac{\mathbb{I}_{\mathbf{v}+\sigma K}(\mathbf{y})}{\mathrm{vol}_d(\mathbf{v}+\sigma K)}$, we have that $\mathbf{Y} \sim \mathsf{Unif}[\mathbf{v} + \sigma K]$, as required. Additionally from Lemma 4.50 we have that: $\mathrm{centroid}\,(\mathbf{v} + \sigma K) = \mathbb{E}(\mathbf{Y})$.

But by linearity of expectation $\mathbb{E}(\mathbf{Y}) = \mathbb{E}(\mathbf{v} + \sigma\mathbf{X}) = \mathbf{v} + \sigma\mathbb{E}(\mathbf{X}) = \mathbf{v}$. So centroid $(\mathbf{v} + \sigma K) = \mathbf{v}$, as expected. $\square$

Lemma 4.52. *If $A, B \subset \mathbb{R}^d$ are both non-empty convex sets, then for $\delta, \mu \in \mathbb{R}$ $\delta A + \mu B \coloneqq \{\delta\mathbf{a} + \mu\mathbf{b} \mid \mathbf{a} \in A, \mathbf{b} \in B\}$, is a non-empty convex set.*

Proof of Lemma 4.52. Let $\mathbf{x}_1, \mathbf{x}_2 \in \delta A + \mu B$, we then have $\mathbf{x}_1 = \delta\mathbf{a}_1 + \mu\mathbf{b}_1$ and $\mathbf{x}_2 = \delta\mathbf{a}_2 + \mu\mathbf{b}_2$ for some $\mathbf{a}_1, \mathbf{a}_2 \in A$, $\mathbf{b}_1, \mathbf{b}_2 \in B$ (since A, B are both non-empty). Then for $\beta \in [0, 1]$, we have:

$$\begin{aligned}
x &\coloneqq \beta\mathbf{x}_1 + (1-\beta)\mathbf{x}_2 \\
&= \beta(\delta\mathbf{a}_1 + \mu\mathbf{b}_1) + (1-\beta)(\delta\mathbf{a}_2 + \mu\mathbf{b}_2) \\
&= \delta(\underbrace{\beta\mathbf{a}_1 + (1-\beta)\mathbf{a}_2}_{\in A, \text{ by convexity}}) + \mu(\underbrace{\beta\mathbf{b}_1 + (1-\beta)\mathbf{b}_2}_{\in B, \text{ by convexity}}) \\
&\in \delta A + \mu B
\end{aligned}$$ $\square$

Lemma 4.53 (Minkowski Sum of Convex Bodies is a Convex Body)**.** *If $A, B \subset \mathbb{R}^d$ are both convex bodies, then $A + B \subset \mathbb{R}^d$ is also a convex body.*

Proof of Lemma 4.53. Firstly, since $A, B \subset \mathbb{R}^d$ both have a non-empty interior per Definition 4.41, then $A + B \subset \mathbb{R}^d$ is also non-empty. In order to show that $A + B$ has a non-empty interior, we note firstly that since A is a convex body, let $\mathbf{a} \in \operatorname{int}(A)$. Then there exists $\varepsilon > 0$ such that $\mathbf{a} \in B(\mathbf{a}, \varepsilon) \subset A$. For a fixed $\mathbf{b} \in B$ we then have $\mathbf{a} + \mathbf{b} \in B(\mathbf{a} + \mathbf{b}, \varepsilon) = B(\mathbf{a}, \varepsilon) + \{\mathbf{b}\} \subset \bigcup_{\mathbf{d} \in B} (A + \mathbf{d}) = A + B$. As such, $A + B$ has a non-empty interior.

Now it suffices to show that $A + B$ is a compact set in $\mathbb{R}^d$. We observe that the addition map $f : A \times B \to A + B$ is continuous in $\mathbb{R}^d$, a normed space. As such we have that the image of a compact set $A \times B$, under the continuous addition map i.e. $f(A \times B) = A + B$, is compact. From Lemma 4.52 $A + B$ is a non-empty convex set, and we have shown that it has a non-empty interior and compact. Thus $A + B$ is a convex body in $\mathbb{R}^d$. $\square$

Lemma 4.54. *If $A \subset \mathbb{R}^d$ is a convex body with $\mathbf{0} \in \operatorname{int}(A)$, then for $0 < \mu_1 \leq \mu_2$ we have $\mu_1 A \subseteq \mu_2 A$.*

Proof of Lemma 4.54. Since $\mu_2 > 0$ by assumption, we can consider WLOG the scaling of A by μ where $\mu \coloneqq \frac{\mu_1}{\mu_2}$, and hence $0 < \mu \leq 1$. We now need to show for each $\mathbf{a} \in A \implies \mu\mathbf{a} \in A$. By assumption we have that $\mathbf{0} \in \operatorname{int}(A)$, it follows by the convexity of A that

$$A \supset [\mathbf{0}, \mathbf{a}] \coloneqq \{\mathbf{x} \in A \,|\, \mathbf{x} = (1-\gamma)\mathbf{0} + \gamma\mathbf{a}, \gamma \in [0,1]\} = \{\mathbf{x} \in A \,|\, \mathbf{x} = \gamma\mathbf{a}, \gamma \in [0,1]\}$$

Thus $\mu\mathbf{a} \in [\mathbf{0}, \mathbf{a}] \subset A$ as required. □

Remark 4.55. We note that it is essential to assume that $\mathbf{0} \in \operatorname{int}(A)$ in Lemma 4.54. Consider the case $\mu_1 = 1, \mu_2 = 5$ with $p = 2$, and $A = [2,3] \times [2,3] \subset \mathbb{R}^2$. Here $\mathbf{0} \notin \operatorname{int}(A)$. Then $\mu_1 A = A = [2,3] \times [2,3]$ and $\mu_2 A = [10,15] \times [10,15]$ and indeed $\mu_1 A \cap \mu_2 A = \emptyset$, in this case.

Theorem 4.56 (Centroid of a convex body lies in its interior)**.** *The centroid of a convex body $K \subset \mathbb{R}^n$ lies in* $\operatorname{int}(K)$*.*

Proof of Theorem 4.56. The proof is found in (Niculescu and Persson, 2018, Proposition 3.9.2), for barycentre of a non-empty convex set. In our setting the conditions are satisfied since we are working with convex bodies (which have a non-empty interior) and in the case of uniform distribution over convex bodies the barycentre and centroid coincide per Lemma 4.50. □

Theorem 4.57 (WLLN for Random Vectors)**.** *Let $(\mathbf{X}_i)_{i=1}^{\infty}$ be random vectors in $\mathbb{R}^d$ ($d \in \mathbb{N}$, fixed), and let $\overline{\mathbf{X}}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. If $\mathbb{E}(\|\mathbf{X}_1\|_2) < \infty$ then $\overline{\mathbf{X}}_n \xrightarrow{p} \mathbb{E}(\mathbf{X}_1)$*

Proof of Theorem 4.57. The proof is found in (Ferguson, 1996, Theorem 4(a)) □

Remark 4.58. We note that the proof in (Ferguson, 1996, Theorem 4(a)) relies on the use of characteristic functions of random vectors $\mathbf{X}_i \in \mathbb{R}^d$. As such the sufficient conditions for the WLLN to hold per Theorem 4.57 only require $\mathbb{E}(\|\mathbf{X}_1\|_2) < \infty$. This is a weaker condition than requiring bounded second moments in univariate WLLN theorems whose proofs rely on Chebychev's inequality.

We will use the following general lemma to later establish the compactness and convexity of the critical set, i.e., $\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$.

Lemma 4.59 (Preservation of compactness and convexity under affine maps)**.** *Let $K \in \mathcal{K}^d$ be a convex body. Further, for any fixed $\alpha \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^d$, let $f\colon \mathbb{R}^d \to \mathbb{R}^d$, such that $f(\mathbf{x}) \coloneqq \alpha\mathbf{x} + \mathbf{c}$ be an affine map. Then $\alpha K + \mathbf{c}$ is compact and convex.*

Proof of Lemma 4.59. First we show that the affine map f is continuous. Fix any $\varepsilon > 0$, and choose $\delta := \frac{\varepsilon}{|\alpha|+1} > 0$. Then for for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ such that $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 < \delta$. We then have that:

$$\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|_2 = \|\alpha\mathbf{x}_1 - \alpha\mathbf{x}_2\|_2 = |\alpha| \|\mathbf{x}_1 - \mathbf{x}_2\|_2 < |\alpha| \delta < \varepsilon \tag{4.50}$$

So indeed the affine map is continuous. Since $K \in \mathcal{K}^d$, it is indeed compact and convex. Moreover the since $\mathbf{b} + \alpha K := f[K]$, i.e., the image of K under f. Since f is affine, convexity of $\mathbf{b} + \alpha K$ is preserved. Similarly since f is continuous, the compactness of $\mathbf{b} + \alpha K$ is also preserved. □

4.A.3 Useful miscellaneous results

Here we prove some useful standard results that are used in several of the remaining proofs.

Definition 4.60 (Lipschitz function)**.** Let $X \subseteq \mathbb{R}^d$ be non-empty, with the metric $d\colon X \times X \to [0, \infty)$. Then for a given constant $B > 0$, we have that $f\colon X \to \mathbb{R}$ is a B-Lipschitz function if and only if:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq Bd(\mathbf{x}, \mathbf{y}), \text{ for each } \mathbf{x}, \mathbf{y} \in X. \tag{4.51}$$

Typically we take $d(\mathbf{x}, \mathbf{y})$ to be the Euclidean metric in $\mathbb{R}^d$ throughout this paper, i.e., $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_2$, for each $\mathbf{x}, \mathbf{y} \in X$.

Lemma 4.61 (Maximum of a finite collection of Lipschitz functions is Lipschitz)**.** *Let $X \subseteq \mathbb{R}^d$ be non-empty, with the Euclidean metric, and $m \in \mathbb{N}$ be fixed. Furthermore, let $f_j\colon X \to \mathbb{R}$ be a B_j-Lipschitz function for each $j \in [m]$, as per Definition 4.60. Then $g\colon X \to \mathbb{R}^d$, defined as $g(\mathbf{x}) := \max_{j\in[m]} f_j(\mathbf{x})$, is a B^*-Lipschitz function, for some $B^* > 0$. Here we can take $B^* := \max_{j\in[m]} B_j$.*

Proof of Lemma 4.61. We will prove this by induction on $m \in \mathbb{N}$. First for $m = 1$, we have that f_1 is B_j-Lipschitz by assumption, so the statement is true. Next for $m = 2$ we want to show that $g(\mathbf{x}) := \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ is B^*-Lipschitz, for some $B^* > 0$. To see this, first observe that:

$$\begin{aligned} f_1(\mathbf{x}) - f_1(\mathbf{y}) &\leq |f_1(\mathbf{x}) - f_1(\mathbf{y})| \leq B_1 \|\mathbf{x} - \mathbf{y}\|_2 && (4.52)\\ \implies f_1(\mathbf{x}) &\leq f_1(\mathbf{y}) + B_1 \|\mathbf{x} - \mathbf{y}\|_2 \\ &\leq g(\mathbf{y}) + B_1 \|\mathbf{x} - \mathbf{y}\|_2 && (4.53) \end{aligned}$$

Similarly we obtain:

$$f_2(\mathbf{x}) \leq g(\mathbf{y}) + B_2 \|\mathbf{x} - \mathbf{y}\|_2 \tag{4.54}$$

Combining Equations (4.53) and (4.54) we have that:

$$g(\mathbf{x}) \coloneqq \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\} \leq g(\mathbf{y}) + B^* \left\|\mathbf{x} - \mathbf{y}\right\|_2 \tag{4.55}$$
$$\implies g(\mathbf{x}) - g(\mathbf{y}) \leq B^* \left\|\mathbf{x} - \mathbf{y}\right\|_2 \tag{4.56}$$

Where $B^* \coloneqq \max\{B_1, B_2\}$. Similarly by interchanging the roles of $\mathbf{x}, \mathbf{y}$ in Equation (4.56) we obtain:

$$g(\mathbf{y}) - g(\mathbf{x}) \leq B^* \left\|\mathbf{x} - \mathbf{y}\right\|_2 \tag{4.57}$$

Combining Equations (4.56) and (4.57), we see that:

$$|g(\mathbf{x}) - g(\mathbf{y})| \leq B^* \left\|\mathbf{x} - \mathbf{y}\right\|_2 \tag{4.58}$$

Which shows that $g(\mathbf{x})$ is a B^*-Lipschitz function, as required. So the statement is true for $m = 2$. Now we assume the inductive hypothesis, i.e., $h(\mathbf{x}) \coloneqq \max_{j\in[k]} f_j(\mathbf{x})$, is a B^*-Lipschitz function, for some $B^* > 0$. Here we can take $B^* \coloneqq \max_{j\in[k]} B_j$. We now show that this is true for some $m \coloneqq k \in \mathbb{N}$. We now show that this is true for $m = k+1$. Let f_{k+1} be a B_{k+1}-Lipschitz function. Let $g(\mathbf{x}) \coloneqq \max_{j\in[k+1]} f_j(\mathbf{x})$. We then have that:

$$\begin{aligned}
g(\mathbf{x}) &\coloneqq \max_{j\in[k+1]} f_j(\mathbf{x}) \\
&= \max\{\max_{j\in[k]} f_j(\mathbf{x}), f_{k+1}(\mathbf{x})\} \\
&\text{(since } \max\{a, b, c\} = \max\{\max\{a, b\}, c\},\ \forall\, a, b, c \in \mathbb{R}.\text{)} \\
&= \max\{h(\mathbf{x}), f_{k+1}(\mathbf{x})\} \qquad (4.59)
\end{aligned}$$

Applying the $m = 2$ base case to Equation (4.59), this implies that $g(\mathbf{x})$ is a B'-Lipschitz function, where $B' \coloneqq \max_{j\in[k+1]} B_j$. As required. □

Lemma 4.62 (B-Lipschitz characterization via bounded derivative)**.** *Let $f : I \to \mathbb{R}$ be continuous and once differentiable, where $I \subseteq \mathbb{R}$ is an interval (possibly unbounded).*

$$f \text{ is } B\text{-Lipschitz, with } B > 0 \iff (\exists B > 0)(\forall x \in \mathbb{R}) : \left(\left|f'(x)\right| \leq B\right) \tag{4.60}$$

Proof of Lemma 4.62. We prove both directions. In both parts we assume that $f : I \to \mathbb{R}$ be continuous and once differentiable, where $I \subseteq \mathbb{R}$ is an interval (possibly unbounded).

($\implies$). Suppose that f is B-Lipschitz, with $B > 0$. We then have that, for some fixed (but arbitrary) $c \in I$:

$$
\begin{aligned}
|f(x) - f(c)| &\leq B\,|x - c| && \text{(by definition of } B\text{-Lipschitz property.)}\\
\implies \left|\frac{f(x) - f(c)}{x - c}\right| &\leq B && \text{(taking limits as } x \to c.)\\
\implies \left|f'(c)\right| &\leq B
\end{aligned}
$$

Since $c \in I$ is arbitrary, indeed $|f'(x)| \leq B$, for each $x \in I$, as required.

($\impliedby$). Suppose that $|f'(x)| \leq B$, with $B > 0$. Further let $x, y \in I$, such that $x < y$. Since f is differentiable on I, we have:

$$
\begin{aligned}
|f(x) - f(y)| &\leq \left|f'(c)\right| |x - y| \\
&\qquad \text{(by the mean value theorem, for some } c \in (x, y).)\\
&\leq B\,|x - y| && \text{(by assumption.)}
\end{aligned}
$$

Which implies that f is B-Lipschitz, as required. $\square$

Lemma 4.63 (Properties of ℓ_2-diameters of compact sets)**.** *Let $A \subseteq B \subseteq \mathbb{R}^d$ be two non-empty, compact sets with $\mathbf{0} \in \operatorname{int}(A)$. Moreover let $\alpha \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^d$ be fixed, but arbitrary. We then have the following definition and key facts.*

$$\operatorname{diam}(A) \coloneqq \sup\left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in A\right\} \tag{4.61}$$

$$\operatorname{diam}(A) \leq \operatorname{diam}(B) < \infty \tag{4.62}$$

$$\operatorname{diam}(\alpha A) = |\alpha| \operatorname{diam}(A) \tag{4.63}$$

$$\operatorname{diam}(A + \mathbf{c}) = \operatorname{diam}(A) \tag{4.64}$$

Proof of Lemma 4.63. We prove each of properties specified in Equations (4.62) and (4.64) in turn.

(*Proof of Equation* (4.62)). Now per Since the $\|\cdot\|_2 : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is a continuous function. It then follows that We have $\operatorname{diam}(A) \coloneqq \sup\left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in A\right\} < \infty$, since it is a supremum of a continuous function over a compact set, i.e., a maximum. Similarly it follows that $\operatorname{diam}(B) < \infty$. We then have:

$$
\begin{aligned}
\left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in A\right\} &\subseteq \left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in B\right\} && \text{(since } A \subseteq B.)\\
\implies \sup\left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in A\right\} &\leq \sup\left\{\|\mathbf{x} - \mathbf{y}\|_2 \,|\, \mathbf{x}, \mathbf{y} \in B\right\}\\
&\qquad \text{(by monotonicity of sup.)}\\
\iff \operatorname{diam}(A) &\leq \operatorname{diam}(B) && \text{(using Equation (4.61).)}
\end{aligned}
$$

as required. ■

(*Proof of Equation* (4.63)). First if $\alpha = 0$, we then have that $\alpha A = \{\mathbf{0}\}$. It then follows that:

$$\operatorname{diam}(\alpha A) = \operatorname{diam}(\{\mathbf{0}\}) = 0 = \alpha \operatorname{diam}(A) \qquad \text{(using Equation (4.62))}$$

as required. Next suppose $\alpha \in \mathbb{R} \setminus \{0\}$. We then observe that:

$$\begin{aligned}
\operatorname{diam}(\alpha A) &\coloneqq \sup\{\|\alpha\mathbf{x} - \alpha\mathbf{y}\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} && \text{(using Equation (4.61))} \\
&= \sup\{\|\alpha(\mathbf{x} - \mathbf{y})\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} \\
&= \sup\{|\alpha|\, \|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} \\
&= |\alpha| \sup\{\|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} \\
&\eqqcolon |\alpha| \operatorname{diam}(A) && \text{(using Equation (4.61))}
\end{aligned}$$

as required. ■

(*Proof of Equation* (4.64)). Given $\mathbf{c} \in \mathbb{R}^d$, we observe that:

$$\begin{aligned}
\operatorname{diam}(A + \mathbf{c}) &\coloneqq \sup\{\|(\mathbf{x} + \mathbf{c}) - (\mathbf{y} + \mathbf{c})\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} \\
& && \text{(using Equation (4.61))} \\
&= \sup\{\|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{x}, \mathbf{y} \in A\} \\
&\eqqcolon \operatorname{diam}(A) && \text{(using Equation (4.61))}
\end{aligned}$$

as required. ■

Thus all properties specified in Equations (4.62) and (4.64) are now proved. □

4.B Proofs of Section 4.1

4.B.1 Formal justification for $\widehat{\theta}_{\mathrm{MLE}}$ with unknown scale parameter

We first provide a Formal justification for the claim that $\widehat{\theta}_{\mathrm{MLE}} = \frac{Y_{(1)}+Y_{(n)}}{2}$, in Section 4.1.

Proposition 4.64 (MLE for univariate uniform with unknown scale)**.** *Let $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\frac{1}{2}, \frac{1}{2}]$, and let $Y_i = \sigma X_i + \theta$ for each $i \in [n]$, so that $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[\theta - \frac{\sigma}{2}, \theta + \frac{\sigma}{2}]$. Then the MLE, $\widehat{\theta}_{\mathrm{MLE}}$, for θ, is given by $\widehat{\theta}_{\mathrm{MLE}} = \frac{Y_{(1)}+Y_{(n)}}{2}$.*

Proof of Proposition 4.64. To ensure our paper is self-contained we quickly prove the result here. To do this, we observe that: $f_{Y_i}(y_i) = \frac{\mathbb{I}_{[\theta-\frac{\sigma}{2},\theta+\frac{\sigma}{2}]}(y_i)}{\sigma}$, for each $i \in [n]$. We then have that the likelihood is given by:

$$
\begin{aligned}
L(\theta, \sigma \mid Y_1, \ldots, Y_n) &= \frac{\prod_{i=1}^{n} \mathbb{I}_{[\theta-\frac{\sigma}{2},\theta+\frac{\sigma}{2}]}(y_i)}{\sigma^n} \\
&= \frac{\prod_{i=1}^{n} \mathbb{I}\big(y_i \geq \theta - \frac{\sigma}{2}, y_i \leq \theta + \frac{\sigma}{2}\big)}{\sigma^n} \qquad (4.65)
\end{aligned}
$$

The likelihood, as per Equation (4.65), is maximized when the numerator indicator function evaluates to 1, and the denominator is minimized over $\sigma > 0$. First, we note that the indicator evaluates to 1, when *both* the following conditions satisfied:

$$
y_{(n)} - \theta \leq \frac{\sigma}{2} \text{ and } y_{(1)} - \theta \geq -\frac{\sigma}{2}. \qquad (4.66)
$$

$$
\iff \sigma \geq 2(y_{(n)} - \theta) \text{ and } \sigma \geq 2(\theta - y_{(1)}) \qquad (4.67)
$$

Or equivalently, we have that:

$$
\sigma \geq \max\left\{2(\theta - y_{(1)}), 2(y_{(n)} - \theta)\right\} \qquad (4.68)
$$

But Equation (4.68) is minimized over $\sigma > 0$, when the RHS expression is minimized. This occurs when we have:

$$
2(\theta - y_{(1)}) = 2(y_{(n)} - \theta) \iff \theta = \frac{y_{(1)} + y_{(n)}}{2}. \qquad (4.69)
$$

Which implies that:

$$
\widehat{\theta}_{\mathrm{MLE}} = \frac{Y_{(1)} + Y_{(n)}}{2}. \qquad (4.70)
$$

As required. ☐

Remark 4.65. We note that the proof of Proposition 4.64 readily generalizes to the case where $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\lambda, \lambda]$, for any fixed $\lambda > 0$.

4.B.2 Proof of Proposition 4.7

Proposition 4.7 (Parameter identifiability). *The data generating process per Definition 4.4, satisfies parameter identifiability for location parameter* $\mathbf{v}$*, and scale parameter* σ*.*

Proof of Proposition 4.7. More formally consider pair of parametric family of densities, $P_{\theta_1}, P_{\theta_2}$, where $\theta_1, \theta_2 \in \Theta$. In our case let us denote $\theta_i = \begin{bmatrix} \mathbf{v}_i \\ \sigma_i^* \end{bmatrix}$ for $i \in \{1, 2\}$. We then have that the densities $P_{\theta_1}, P_{\theta_2}$ are identifiable if and only if $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$, for all $\theta_1, \theta_2 \in \Theta$. To demonstrate identifiability we observe that it is equivalent to showing the contrapositive statement $\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$. Suppose that $\theta_1 \neq \theta_2$ and by way of contradiction assume that $P_{\theta_1} = P_{\theta_2}$. Now WLOG we can assume that $\mathbf{v}_1 = \mathbf{0}$, and we now consider 2 cases. First, consider the case where $\mathbf{v}_2 = \mathbf{0}$. But this implies $\sigma_1^* \neq \sigma_2^*$. But since $\mathbf{v}_2 = \mathbf{0} \implies \sigma_1^* K = \sigma_2^* K \implies \sigma_1^* = \sigma_2^*$, a contradiction. Second, consider the case where $\mathbf{v}_2 \neq \mathbf{0}$. But we then have:

$$\begin{aligned} &\mathbf{v}_2 \neq \mathbf{0} \\ \implies &\sigma_1^* K = \mathbf{v}_2 + \sigma_2^* K \\ \implies &\operatorname{centroid}(\sigma_1^* K) = \operatorname{centroid}(\mathbf{v}_2 + \sigma_2^* K) \\ \implies &\mathbf{0} = \mathbf{v}_2 \end{aligned}$$

Which is a contradiction. Combining the above we do indeed observe that $\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$, or equivalently that $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$. As such we have shown that we have model parameter identifiability for estimation purposes. $\square$

4.C Proofs of Section 4.2

With the main mathematical preliminaries set up, we provide proofs of all mathematical statements in the main text in turn. For reader convenience we restate the corresponding statements from the paper in the appendix before providing the proof thereof.

4.C.1 Proof of Theorem 4.9

We note that the proof of Theorem 4.9 can be found in Ibragimov and Has'minskiĭ (1981, Lemma 2.1) and Bickel and Doksum (2016, Theorem 8.3.1). Here we provide an alternative proof which is effectively a step-by-step multivariate extension of the relevant univariate results of (Lehmann and Casella, 1998, Chapter 3).

Setup and key notation

We first setup the following notation for our multivariate setting. Firstly we define $\mathbf{x}^{(n,d)} := (\mathbf{x}_1^{(n,d)}, \ldots, \mathbf{x}_n^{(n,d)})^\top := (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$. Here $\mathbf{x}^{(n,d)} \in \mathbb{R}^{nd}$ denotes our *stacked sample* vector of the n i.i.d. d-dimensional samples i.e. $\mathbf{x}_i^\top := (x_{i,1}, \ldots, x_{i,d})^\top \in \mathbb{R}^d$, $\forall\, i \in [n]$. Additionally for convenience, we also denote this i^{th} component more explicitly as $\mathbf{x}_i^{(n,d)} := \mathbf{x}_i^\top$. For a fixed vector $\mathbf{v} \in \mathbb{R}^d$ we define the *stacked bar* vector $\overline{\mathbf{v}}^{(n,d)} := \underbrace{(\mathbf{v}^\top, \ldots, \mathbf{v}^\top)^\top}_{n \text{ times}}$. In this case $\overline{\mathbf{v}}^{(n,d)} \in \mathbb{R}^{nd}$ denotes n copies of the same d-dimensional fixed location vector i.e. $\overline{\mathbf{v}}_i^{(n,d)} = \mathbf{v}^\top = (v_1, \ldots, v_p)^\top \in \mathbb{R}^d$, $\forall\, i \in [n]$. Using this multivariate notation allows us to define translation of each component of $\mathbf{x}^{(n,d)} \in \mathbb{R}^{nd}$ by a common *location vector* $\mathbf{a} \in \mathbb{R}^d$ concisely as follows:

$$\begin{aligned}
\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)} &:= (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top + \underbrace{(\mathbf{a}^\top, \ldots, \mathbf{a}^\top)^\top}_{n \text{ times}} \\
&= (\mathbf{x}_1^\top + \mathbf{a}^\top, \ldots, \mathbf{x}_n^\top + \mathbf{a}^\top)^\top
\end{aligned}$$

Here $\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)} \in \mathbb{R}^{nd}$. This is a convenient multivariate extension of notation used in Lehmann and Casella (1998, Chapter 3), to ensure the proofs are easier to follow.

Definition 4.66 (Location equivariance and invariance)**.** An estimator $\delta : \mathbb{R}^{nd} \to \mathbb{R}^d$ is called *location equivariant* if:

$$\begin{aligned}
\delta(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) &= \delta(\mathbf{x}^{(n,d)}) + \overline{\mathbf{a}}_n^{(n,d)}, \ \forall\, \mathbf{x}^{(n,d)},\ \overline{\mathbf{a}}^{(n,d)} \in \mathbb{R}^{nd} && (4.71) \\
&= \delta(\mathbf{x}^{(n,d)}) + \mathbf{a} && (\text{since } \overline{\mathbf{a}}_i^{(n,d)} = \mathbf{a},\ \forall\, i \in [n])
\end{aligned}$$

An estimator $u : \mathbb{R}^{nd} \to \mathbb{R}^{d}$ is called *location invariant* if:

$$u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) = u(\mathbf{x}^{(n,d)}),\ \forall\, \mathbf{x}^{(n,d)},\ \overline{\mathbf{a}}^{(n,d)} \in \mathbb{R}^{nd} \tag{4.72}$$

Multivariate Pitman location estimation

Lemma 4.67 ((Lehmann and Casella, 1998), Lemma 3.6, Mulitvariate setting)**.** *If δ_0 is any equivariant estimator, then a necessary and sufficient condition for δ to be equivariant is that*

$$\delta(\mathbf{x}^{(n,d)}) = \delta_0\left(\mathbf{x}^{(n,d)}\right) + u(\mathbf{x}^{(n,d)}),\ \forall\, \mathbf{x}^{(n,d)} \in \mathbb{R}^{nd} \tag{4.73}$$

and

$$u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) = u(\mathbf{x}^{(n,d)}),\ \forall\, \mathbf{x}^{(n,d)},\ \overline{\mathbf{a}}^{(n,d)} \in \mathbb{R}^{nd} \tag{4.74}$$

Proof of Lemma 4.67. We can now prove both directions following the same approach as the univariate proof, with our multivariate equivariance definition and notation.
($\Longrightarrow$) Firstly assume that δ_0 and δ are both equivariant estimators. So we then define

$$u(\mathbf{x}^{(n,d)}) \coloneqq \delta(\mathbf{x}^{(n,d)}) - \delta_0\left(\mathbf{x}^{(n,d)}\right),\ \forall\, \mathbf{x}^{(n,d)} \in \mathbb{R}^{nd} \tag{4.75}$$

We then have that for all $\mathbf{x}^{(n,d)}, \overline{\mathbf{a}}^{(n,d)}$:

$$\begin{aligned}
u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) &= \delta(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) - \delta_0(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) \\
&\hspace{12em} \text{(using Equation (4.75))} \\
&= \delta(\mathbf{x}^{(n,d)}) + \overline{\mathbf{a}}_n^{(n,d)} - (\delta_0\left(\mathbf{x}^{(n,d)}\right) + \overline{\mathbf{a}}_n^{(n,d)}) \\
&\hspace{12em} \text{(by equivariance of } \delta, \delta_0\text{)} \\
&= \delta(\mathbf{x}^{(n,d)}) - \delta_0\left(\mathbf{x}^{(n,d)}\right) \\
&= u(\mathbf{x}^{(n,d)}) \qquad \text{(per Equation (4.75))}
\end{aligned}$$

($\Longleftarrow$) Now we conversely assume that δ_0 is an equivariant estimator and $u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) = u(\mathbf{x}^{(n,d)}),\ \forall \mathbf{x}^{(n,d)}$. We then define

$$\delta(\mathbf{x}^{(n,d)}) \coloneqq u(\mathbf{x}^{(n,d)}) + \delta_0\left(\mathbf{x}^{(n,d)}\right) \tag{4.76}$$

To show that δ is indeed equivariant, we proceed as follows:

$$\begin{aligned}
\delta(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) &= u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) + \delta_0(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) \\
&\qquad \text{(using Equation (4.76))} \\
&= u(\mathbf{x}^{(n,d)}) + \delta_0\left(\mathbf{x}^{(n,d)}\right) + \overline{\mathbf{a}}_n^{(n,d)} \\
&\qquad \text{(since } u \text{ is invariant, } \delta_0 \text{ is equivariant)} \\
&= \delta(\mathbf{x}^{(n,d)}) + \overline{\mathbf{a}}_n^{(n,d)} \qquad \text{(by definition of } \delta(\mathbf{x}^{(n,d)})\text{)}
\end{aligned}$$

Which shows that δ is indeed equivariant. ☐

Lemma 4.68 ((Lehmann and Casella, 1998), Lemma 3.7, Mulitvariate setting)**.** *A function u is invariant i.e. satisfies $u(\mathbf{x}^{(n,d)}+\overline{\mathbf{a}}^{(n,d)}) = u(\mathbf{x}^{(n,d)}), \forall \mathbf{x}^{(n,d)}, \overline{\mathbf{a}}^{(n,d)} \in \mathbb{R}^{nd}$, if and only if it is a function of the differences $\mathbf{y}_i^{(n,d)} = \mathbf{x}_i^{(n,d)} - \mathbf{x}_n^{(n,d)}$, $\forall i \in [n-1]$, $n \geq 2$ and for $n = 1$ if and only if is constant valued.*

Proof of Lemma 4.68. As suggested by the theorem statement, we will split into cases $n = 1$ and $n \geq 2$ and prove both directions. Let us consider the case $n = 1$.
($\Longrightarrow$) Suppose u is invariant. We then have $u(\mathbf{x}^{(1,d)}+\overline{\mathbf{a}}^{(1,d)}) = u(\mathbf{x}^{(1,d)}), \forall \mathbf{x}^{(1,d)}, \overline{\mathbf{a}}^{(1,d)}$. Here we can set $\overline{\mathbf{a}}^{(1,d)} = -\mathbf{x}^{(1,d)}$, which gives us $u(\mathbf{x}^{(1,d)}) = u(\mathbf{x}^{(1,d)} - \mathbf{x}^{(1,d)}) = u(\mathbf{0})$, which is indeed constant valued.
($\Longleftarrow$) Now suppose u is a constant valued function i.e. $u(\mathbf{x}^{(1,d)}) = \mathbf{c}$, $\forall\, \mathbf{x}^{(1,d)}$, for some constant $\mathbf{c} \in \mathbb{R}^d$. Then we have:

$$\begin{aligned}
u(\mathbf{x}^{(1,d)} + \overline{\mathbf{a}}^{(1,d)}) &= \mathbf{c} \\
&= u(\mathbf{x}^{(1,d)})
\end{aligned}$$

Indeed u is invariant to translations.
Now we consider the case $n \geq 2$. We introduce the difference invariant as $\underline{\mathbf{y}}^{(n,d)} := (\mathbf{x}_1^{(n,d)} - \mathbf{x}_n^{(n,d)}, \ldots, \mathbf{x}_{n-1}^{(n,d)} - \mathbf{x}_n^{(n,d)})^\top$.
($\Longrightarrow$) Suppose u is invariant. We then have:

$$\begin{aligned}
u\left((\mathbf{x}^{(n,d)} - \overline{\mathbf{x}}_n^{(n,d)})^\top\right) &= u\left((\mathbf{x}_1^{(n,d)} - \mathbf{x}_n^{(n,d)}, \ldots, \mathbf{x}_{n-1}^{(n,d)} - \mathbf{x}_n^{(n,d)}, \mathbf{x}_n^{(n,d)} - \mathbf{x}_n^{(n,d)})^\top\right) \\
&= u\left((\mathbf{x}_1^{(n,d)} - \mathbf{x}_n^{(n,d)}, \ldots, \mathbf{x}_{n-1}^{(n,d)} - \mathbf{x}_n^{(n,d)}, \mathbf{0})^\top\right) \\
&= u(\underline{\mathbf{y}}^{(n,d)}, \mathbf{0})) \qquad \text{(by definition of } \underline{\mathbf{y}}^{(n,d)}\text{)} \\
&= u'(\underline{\mathbf{y}}^{(n,d)}) \qquad \text{(by definition of } u'\text{)}
\end{aligned}$$

So u is indeed a function of the differences $\underline{\mathbf{y}}^{(n,d)}$, as required.
($\Longleftarrow$) Now suppose that u is a function of the difference invariant $\underline{\mathbf{y}}^{(n,d)}$ i.e. $u(\mathbf{x}^{(n,d)}) = v(\underline{\mathbf{y}}^{(n,d)})$, $\forall\, \mathbf{x}^{(n,d)}$, for some function v. We then have:

$$\begin{aligned}
u(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) &= v(\mathbf{x}_1^{(n,d)} + \mathbf{a} - (\mathbf{x}_n^{(n,d)} + \mathbf{a}), \dots, \mathbf{x}_{n-1}^{(n,d)} + \mathbf{a} - (\mathbf{x}_n^{(n,d)} + \mathbf{a})) \\
&= v(\mathbf{x}_1^{(n,d)} - \mathbf{x}_n^{(n,d)}, \dots, \mathbf{x}_{n-1}^{(n,d)} - \mathbf{x}_n^{(n,d)}) \\
&= v\left(\underline{\mathbf{y}}^{(n,d)}\right) && \text{(by definition)} \\
&= u(\mathbf{x}^{(n,d)}) && \text{(by assumption)}
\end{aligned}$$

So indeed u is an invariant function. $\square$

Theorem 4.69 ((Lehmann and Casella, 1998), Theorem 3.8, Mulitvariate setting)**.** *If δ_0 is any equivariant estimator, then a necessary and sufficient condition for δ to be equivariant is that there exists a function v of $n-1$ arguments for which*

$$\delta(\mathbf{x}^{(n,d)}) = \delta_0\left(\mathbf{x}^{(n,d)}\right) - v(\underline{\mathbf{y}}^{(n,d)}),\ \forall\, \mathbf{x}^{(n,d)} \in \mathbb{R}^{nd}$$

Proof of Theorem 4.69. This follows directly in the multivariate setting (as with the univariate case) by a combination of the results of Lemma 4.67 and Lemma 4.68. $\square$

Definition 4.70 ((Lehmann and Casella, 1998), Definition 3.2)**.** A family of densities $f(x \mid \xi)$, with parameter ξ, and a loss function $L(\xi, d)$ are location invariant if, respectively, $f\left(x' \mid \xi'\right) = f(x \mid \xi)$ and $L(\xi, d) = L\left(\xi', d'\right)$ whenever $\xi' = \xi + a$ and $d' = d + a$. If both the densities and the loss function are location invariant, the problem of estimating ξ is said to be location invariant under the transformations:

$$X_i' = X_i + a \tag{4.77}$$
$$\xi_i' = \xi_i + a \tag{4.78}$$
$$d' = d + a \tag{4.79}$$

Theorem 4.71 ((Lehmann and Casella, 1998), Theorem 3.10, Mulitvariate setting)**.** *Let $\mathbf{x}^{(n,d)}$ be distributed from a location family. Let $\underline{\mathbf{y}}^{(n,d)} \coloneqq (\mathbf{x}_1^{(n,d)} - \mathbf{x}_n^{(n,d)}, \dots, \mathbf{x}_{n-1}^{(n,d)} - \mathbf{x}_n^{(n,d)}, \mathbf{0})^\top$ be given. Suppose that the loss function ρ is location invariant and that there exists an equivariant estimator δ_0 of ξ with* *finite risk. Assume that for each $\underline{\mathbf{y}}^{(n,d)}$ there exists a number $v\left(\underline{\mathbf{y}}^{(n,d)}\right) = v^*(\underline{\mathbf{y}}^{(n,d)})$ which minimizes*

$$\mathbb{E}_0\left(\rho\left[\delta_0\left(\mathbf{x}^{(n,d)}\right) - v\left(\underline{\mathbf{y}}^{(n,d)}\right)\right] \mid \underline{\mathbf{y}}^{(n,d)}\right)$$

Then, a location equivariant estimator δ of ξ with minimum risk exists and is given by

$$\delta^*(\mathbf{x}^{(n,d)}) = \delta_0\left(\mathbf{x}^{(n,d)}\right) - v^*(\underline{\mathbf{y}}^{(n,d)})$$

Proof of Theorem 4.71. We follow the univariate proof approach taken in (Keener, 2010, Theorem 10.4), which extends naturally to our multivariate setting. First we define the risk of a location equivariant estimator to be $\mathbf{R}(\theta, \delta) \coloneqq \mathbb{E}_0\left(\rho\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) - v\left(\underline{\mathbf{y}}^{(n,d)}\right)\right)\right)$.

$$\begin{aligned}
\mathbf{R}(\theta, \delta) &\coloneqq \mathbb{E}_0\left(\rho\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) - v\left(\underline{\mathbf{y}}^{(n,d)}\right)\right)\right) \\
&= \mathbb{E}_0\left(\mathbb{E}_0\left(\rho\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) - v\left(\underline{\mathbf{y}}^{(n,d)}\right)\right) \mid \underline{\mathbf{y}}^{(n,d)}\right)\right) \\
&\geq \mathbb{E}_0\left(\mathbb{E}_0\left(\rho\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) - v^*(\underline{\mathbf{y}}^{(n,d)})\right) \mid \underline{\mathbf{y}}^{(n,d)}\right)\right) \\
&= \mathbb{E}_0\left(\rho\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) - v^*(\underline{\mathbf{y}}^{(n,d)})\right)\right) \\
&= \mathbb{E}_0\left(\rho\left(\delta^*(\mathbf{x}^{(n,d)})\right)\right) \\
&= \mathbf{R}\left(\theta, \delta^*\right)
\end{aligned}$$

□

4.C.2 Final proof of Theorem 4.9

We now come to the main proof of this section, namely that of Theorem 4.9.

Theorem 4.9 (Multivariate Pitman location estimator)**.** *Consider the more general location estimation problem, under the known scaling regime. That is, let $d \geq 1$ be a fixed positive integer, and denote $\mathbf{v} \in \mathbb{R}^d$ to be the fixed but unknown location parameter. We then consider n observations, $(\mathbf{Y}_i)_{i=1}^n$, where each observation $\mathbf{Y}_i \in \mathbb{R}^d$ is generated from the following model:*

$$\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \mathbf{X}_i \tag{4.5}$$

$$s.t. \ \ (\mathbf{X}_1, \ldots, \mathbf{X}_n) \sim f, \tag{4.6}$$

where f is a valid joint probability density of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then under square loss risk, for this generating process, the multivariate mimimum risk equivariant location estimator is the Pitman estimator, $\widehat{\mathbf{v}}_{\text{pit}}$, which is defined as follows:

$$\widehat{\mathbf{v}}_{\text{pit}} = \frac{\int_{\mathbb{R}^d} \mathbf{u} f\left(\mathbf{y}_1 - \mathbf{u}, \ldots, \mathbf{y}_{n-1} - \mathbf{u}, \mathbf{y}_n - \mathbf{u}\right) \mathrm{d}\mathbf{u}}{\int_{\mathbb{R}^d} f\left(\mathbf{y}_1 - \mathbf{u}, \ldots, \mathbf{y}_{n-1} - \mathbf{u}, \mathbf{y}_n - \mathbf{u}\right) \mathrm{d}\mathbf{u}}. \tag{4.7}$$

Proof of Theorem 4.9. From Theorem 4.71 we know that for any equivariant estimator, $\delta_0\left(\mathbf{x}^{(n,d)}\right)$, of $\mathbf{v}$, under the square error loss, we have that a minimum risk equivariant estimator (MRE) is given by:

$$\delta^*(\mathbf{x}^{(n,d)}) = \delta_0\left(\mathbf{x}^{(n,d)}\right) - \mathbb{E}\left(\delta_0\left(\mathbf{x}^{(n,d)}\right) \mid \underline{\mathbf{y}}^{(n,d)}\right)$$

Where $\delta_0\left(\mathbf{x}^{(n,d)}\right)$ is *any* equivariant estimator. We can proceed using a similar derivation to the univariate case discussed in (Lehmann and Casella, 1998). Let $\delta_0\left(\mathbf{x}^{(n,d)}\right) = \mathbf{x}_n^{(n,d)}$. This indeed is location equivariant since we have:

$$\begin{aligned}\delta_0(\mathbf{x}^{(n,d)} + \overline{\mathbf{a}}^{(n,d)}) &= \mathbf{x}_n^{(n,d)} + \overline{\mathbf{a}}_n^{(n,d)} && \text{(by definition)}\\ &= \delta_0\left(\mathbf{x}^{(n,d)}\right) + \overline{\mathbf{a}}_n^{(n,d)}\end{aligned}$$

As required. Further we know that $\mathbf{Y}_i = \mathbf{x}_i - \mathbf{x}_n,\ \forall\, i \in [n-1]$, and $\mathbf{y}_n = \mathbf{x}_n$.

Expanding, we have that $(\mathbf{x}_{1,1},\dots,\mathbf{x}_{1,d},\mathbf{x}_{2,1},\dots,\mathbf{x}_{2,d},\dots,\mathbf{x}_{n-1,1},\mathbf{x}_{n-1,d},\mathbf{x}_{n,1},\dots,\mathbf{x}_{n,d})^\top = (\mathbf{y}_{1,1}+\mathbf{x}_{n,1},\dots,\mathbf{y}_{1,d}+\mathbf{x}_{n,d},\mathbf{y}_{2,1}+\mathbf{x}_{n,1},\dots,\mathbf{y}_{2,d}+\mathbf{x}_{n,d},\dots,\mathbf{y}_{n-1,1}+\mathbf{x}_{n,1},\mathbf{y}_{n-1,d}+\mathbf{x}_{n,d},\mathbf{y}_{n,1},\dots,\mathbf{y}_{n,d})^\top$. We then have that Jacobian of this transformation is equal to:

$$\begin{aligned}\mathbf{J} &= \left(\frac{\partial \mathbf{x}_{i,j}}{\partial \mathbf{y}_{k,l}}\right)_{(i,j,k,l)\in[nd]} \in \mathbb{R}^{nd\times nd}\\ &= \begin{cases}1 & \text{, if } i=k, j=l \text{ or } k=n, j=l\\ 0 & \text{, otherwise}\end{cases}\end{aligned}$$

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x_{1,1}}{\partial y_{1,1}} & \cdots & \frac{\partial x_{1,1}}{\partial y_{n,d}}\\ \vdots & \ddots & \vdots\\ \frac{\partial x_{n,d}}{\partial y_{1,1}} & \cdots & \frac{\partial x_{n,d}}{\partial y_{n,d}}\end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ 0 & 0 & \ddots & 1\\ 0 & 0 & \cdots & 1\end{pmatrix}$$

Qualitatively, each row $j \in [nd]$ in the Jacobian matrix $\mathbf{J} \in \mathbb{R}^{nd\times nd}$, takes values 1 only at the diagonal and n^{th} entries respectively i.e. $\mathbf{J}_{j,j} = \mathbf{J}_{j,n} = 1$ and all other entries take value 0. Since $\mathbf{J}$ is lower triangular, we have that it is determinant is the product of its diagonal entries which all are equal to 1 i.e. $\det(\mathbf{J}) = \prod_{i\in[nd]} \mathbf{J}_{ii} = 1 \implies |\det(\mathbf{J})| = 1$.

To calculate the minimum risk equivariant estimator in this theorem explicitly, let us assume that the equivariant estimator

$$\delta_0(\mathbf{X}) = \mathbf{X}_n$$

has finite risk. To evaluate the conditional expectation in the theorem we need the conditional distribution of X_n given Y (under P_0), which we can obtain from the joint density. Using a change of variables $\mathbf{y}_i = \mathbf{x}_i - \mathbf{x}_n, i = 1, \ldots, n-1$, in the integrals against $d\mathbf{x}_i$,

$$\begin{aligned}
P_0\left[\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_n \end{pmatrix} \in B\right] &= E_0 \mathbb{I}_B(\mathbf{Y}_1, \ldots, \mathbf{Y}_{n-1}, \mathbf{X}_n) \\
&= E_0 \mathbb{I}_B(\mathbf{X}_1 - X_n, \ldots, \mathbf{X}_{n-1} - \mathbf{X}_n, \mathbf{X}_n) \\
&= \int \cdots \int \mathbb{I}_B(\mathbf{x}_1 - \mathbf{x}_n, \ldots, \mathbf{x}_n) f\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) d\mathbf{x}_1 \cdots d\mathbf{x}_n \\
&= \int \cdots \int f\left(\mathbf{y}_1 + \mathbf{x}_n, \ldots, \mathbf{y}_{n-1} + \mathbf{x}_n, \mathbf{x}_n\right) d\mathbf{y}_1 \cdots d\mathbf{y}_{n-1} d\mathbf{x}_n.
\end{aligned}$$

Therefore the joint density of $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is given by:

$$p_{\mathbf{Y}}\left(\mathbf{y}_1, \ldots, \mathbf{y}_n\right) = f\left(\mathbf{y}_1 + \mathbf{y}_n, \ldots, \mathbf{y}_{n-1} + \mathbf{y}_n, \mathbf{y}_n\right)$$

We thus have the conditional density $p_{\mathbf{Y}_n | (\mathbf{y}_1, \ldots, \mathbf{y}_{n-1})}$ is

$$\frac{f\left(\mathbf{y}_1 + \mathbf{y}_n, \ldots, \mathbf{y}_{n-1} + \mathbf{y}_n, \mathbf{y}_n\right)}{\int f\left(\mathbf{y}_1 + \mathbf{t}, \ldots, \mathbf{y}_{n-1} + \mathbf{t}, \mathbf{t}\right) d\mathbf{t}} \tag{4.80}$$

Now since $\mathbf{Y}_n = \mathbf{X}_n$, we can now directly compute $\mathbb{E}\left(\delta_0(\mathbf{X}) \mid \mathbf{Y}\right)$ as follows:

$$\begin{aligned}
\mathbb{E}\left(\delta_0(\mathbf{X}) \mid \mathbf{Y}\right) &= \mathbb{E}\left(\mathbf{X}_n \mid \mathbf{Y}\right) && (\text{since } \delta_0(\mathbf{X}) = \mathbf{X}_n) \\
&= \mathbb{E}\left(\mathbf{Y}_n \mid \mathbf{Y}\right) && (\text{since } \mathbf{Y}_n = \mathbf{X}_n) \\
&= \frac{\int \mathbf{t} f\left(\mathbf{y}_1 + \mathbf{t}, \ldots, \mathbf{y}_{n-1} + \mathbf{t}, \mathbf{t}\right) d\mathbf{t}}{\int f\left(\mathbf{y}_1 + \mathbf{t}, \ldots, \mathbf{y}_{n-1} + \mathbf{t}, \mathbf{t}\right) d\mathbf{t}} && (\text{using Equation } (4.80)) \\
&= \frac{\int \mathbf{t} f\left(\mathbf{x}_1 - \mathbf{x}_n + \mathbf{t}, \ldots, \mathbf{x}_{n-1} - \mathbf{x}_n + \mathbf{t}, \mathbf{t}\right) d\mathbf{t}}{\int f\left(\mathbf{x}_1 - \mathbf{x}_n, \ldots, \mathbf{x}_{n-1} - \mathbf{x}_n + \mathbf{t}, \mathbf{t}\right) d\mathbf{t}} \\
& && (\text{In terms of } \mathbf{x}_i\text{'s}) \\
&= \mathbf{x}_n - \frac{\int \mathbf{u} f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_{n-1} - \mathbf{u}, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}}{\int f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_{n-1} - \mathbf{u}, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}} \\
& && (\text{Using } \mathbf{u} = \mathbf{x}_n - \mathbf{t})
\end{aligned}$$

From this we derive the Pitman estimator of $\mathbf{v}$ as:

$$\begin{aligned}
\widehat{\mathbf{v}}_{\text{pit}} &= \delta^*(\mathbf{X}) \\
&= \delta_0(\mathbf{X}) - \mathbb{E}\left(\delta_0(\mathbf{X}) \mid \mathbf{Y}\right) \\
&= \frac{\int \mathbf{u} f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_{n-1} - \mathbf{u}, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}}{\int f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_{n-1} - \mathbf{u}, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}}
\end{aligned}$$

As required. $\square$

4.C.3 Proof of Corollary 4.11

Corollary 4.11 (Multivariate uniform Pitman location estimator)**.** *Let* $(\mathbf{Y}_i)_{i=1}^n$ *be generated according to Definition 4.4, with* σ *known to the observer. Then the Pitman estimator* $\widehat{\mathbf{v}}_{\text{pit}}$*, of the location parameter* $\mathbf{v}$ *is the centroid of the critical set, that is:*

$$\widehat{\mathbf{v}}_{\text{pit}} = \text{centroid}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right). \tag{4.8}$$

Proof of Corollary 4.11. Since we assume that σ is known to the observer we can take WLOG the scale parameter $\sigma = 1$. From Theorem 4.9, we have that

$$\begin{aligned}
\widehat{\mathbf{v}}_{\text{pit}} &= \frac{\int \mathbf{u} f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}}{\int f\left(\mathbf{x}_1 - \mathbf{u}, \ldots, \mathbf{x}_n - \mathbf{u}\right) d\mathbf{u}} && \text{(by Theorem 4.9)} \\
&= \frac{\int \mathbf{u} \prod_{i=1}^n f\left(\mathbf{x}_i - \mathbf{u}\right)) d\mathbf{u}}{\int \prod_{i=1}^n f\left(\mathbf{x}_i - \mathbf{u}\right) d\mathbf{u}} && \text{(by independence of } X_i\text{'s)} \\
&= \frac{\int \mathbf{u} \prod_{i=1}^n \left(\mathbb{I}_{\mathbf{x}_i - K}(\mathbf{u})\right) d\mathbf{u}}{\int \prod_{i=1}^n \left(\mathbb{I}_{\mathbf{x}_i - K}(\mathbf{u})\right) d\mathbf{u}} \\
& && \text{(since each } X_i \sim \mathsf{Unif}[K] \iff f_{X_i}(x) = \tfrac{\mathbb{I}_K(x)}{\text{vol}_d\,(K)}) \\
&= \frac{\int \mathbf{u} \mathbb{I}_{\bigcap_{i=1}^n (\mathbf{x}_i - K)}(\mathbf{u}) d\mathbf{u}}{\int \mathbb{I}_{\bigcap_{i=1}^n (\mathbf{x}_i - K)}(\mathbf{u}) d\mathbf{u}} && \text{(since } \textstyle\prod_{i=1}^n \mathbb{I}_{\mathbf{x}_i - K}(\mathbf{u}) = \mathbb{I}_{\bigcap_{i=1}^n (\mathbf{x}_i - K)}(\mathbf{u})) \\
&= \text{centroid}\left(\bigcap_{i=1}^{n} (\mathbf{x}_i - K)\right) && \text{(per Definition 4.43)}
\end{aligned}$$

As required. ☐

4.C.4 Formal justification for Remark 4.12

Proposition 4.72 (Pitman location estimator for univariate uniform)**.** *Let* $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\lambda, \lambda]$*, with* $\lambda \in \mathbb{R}_{>0}$ *known, and let* $Y_i = X_i + v$ *for each* $i \in [n]$*, so that* $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[v - \lambda, v + \lambda]$*. Then the Pitman location estimator,* $\widehat{v}_{\text{pit}}$*, for* v*, by applying Corollary 4.11, is given by* $\widehat{v}_{\text{pit}} = \frac{Y_{(1)} + Y_{(n)}}{2}$*.*

Proof of Proposition 4.72. We first note that $K \coloneqq [-\lambda, \lambda]$ is symmetric, and thus $-K = [-\lambda, \lambda]$. Therefore $Y_i - K = [Y_i - \lambda, Y_i + \lambda]$, for each $i \in [n]$, i.e., $(Y_i - K)_{i=1}^n$ is a collection of n intervals. Since $v \in Y_i - K$, for each $i \in [n]$, once again $\bigcap_{i=1}^n (Y_i - K)$ is non-empty a.s. and in this case it is again an interval. Now, let $Y_{(1)}, \ldots, Y_{(n)}$ denote the order statistics of our observed sample. Then applying Corollary 4.11 to our setting we observe that since it is

$$\begin{aligned}
\widehat{v}_{\text{pit}} &\coloneqq \text{centroid}\left(\bigcap_{i=1}^{n}(Y_i - K)\right) && \text{(using Corollary 4.11.)}\\
&= \text{centroid}\left(\bigcap_{i=1}^{n}\left(Y_{(i)} - K\right)\right) && \text{(using the order statisics.)}\\
&= \text{centroid}\left(\left[Y_{(n)} - \lambda, Y_{(1)} + \lambda\right]\right) && \text{(since ordered intervals all intersect.)}\\
&= \frac{(Y_{(n)} - \lambda) + (Y_{(1)} + \lambda)}{2}\\
&= \frac{Y_{(1)} + Y_{(n)}}{2}
\end{aligned}$$

As required. □

4.C.5 Proof of Proposition 4.8

Proposition 4.8 (The critical set and its geometric properties)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. We define critical set to be $\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$, for each $n \in \mathbb{N}$. Moreover, the critical set contains the true location vector $\mathbf{v}$ almost surely for each $n \in \mathbb{N}$, and is thus non-empty. Furthermore it is a compact convex set, and is thus closed and bounded.*

Proof of Proposition 4.8. In the following, all statements for each $i \in [n]$ are made almost surely (a.s.):

$$\begin{aligned}
\mathbf{Y}_i \in \mathbf{v} + \sigma K &\iff \mathbf{v} \in \mathbf{Y}_i - \sigma K && \text{(for each } i \in [n].)\\
&\iff \mathbf{v} \in \bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)
\end{aligned}$$

That is, $\mathbb{P}\left(\mathbf{v} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)\right) = 1$, for all $n \in \mathbb{N}$, which proves the first part of the assertion. Now for establishing convexity, we observe that for each $i \in [n]$, that $\mathbf{Y}_i - \sigma K \coloneqq \{\mathbf{Y}_i\} - \sigma K$ is convex by Lemma 4.52. Moreover for each $n \in \mathbb{N}$ we have that $\mathbf{Y}_i - \sigma K \coloneqq f[K]$ is the image of K under the affine map $f \colon \mathbb{R}^d \to \mathbb{R}$, where $f(\mathbf{x}) \coloneqq \mathbf{Y}_i - \sigma \mathbf{x}$. Then by Lemma 4.59 we have that $\mathbf{Y}_i - \sigma K$ is compact. We then have that $\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$ is finite intersection of non-empty (containing $\mathbf{v}$), compact convex sets, and is thus a non-empty, compact convex set. Since it is compact, it is indeed closed and bounded. □

4.C.6 Proof of Proposition 4.13

Proposition 4.13 (Sample mean is consistent and unbiased for location)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer.*

The sample mean $\overline{\mathbf{v}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i$ *is a consistent and unbiased estimator of the location parameter* $\mathbf{v}$*, regardless of the value of the true scale parameter* σ*.*

Proof of Proposition 4.13. Since the centroid (K), is known to the observer, we assume WLOG per Remark 4.6 that centroid $(K) = \mathbf{0}$. By Lemma 4.50, this is equivalent to $\mathbb{E}(\mathbf{X}) = \mathbf{0}$. Additionally since $K \subset \mathbb{R}^d$ is compact, it is indeed bounded. So we have that $\mathbb{E}(\|\mathbf{X}\|_2) < \infty$. We first show that the sample mean $\overline{\mathbf{v}}$, is a consistent estimator for $\mathbf{v}$ as follows:

$$\begin{aligned}
\overline{\mathbf{v}} &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{v} + \sigma\mathbf{X}_i) \\
&= \mathbf{v} + \sigma\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\right)}_{\xrightarrow{p}\mathbf{0},\text{ by WLLN}} && \text{(per Theorem 4.57 and since } \mathbb{E}(\mathbf{X}) = \mathbf{0}) \\
&\xrightarrow{p} \mathbf{v}
\end{aligned}$$

Similarly we can show that the sample mean $\overline{\mathbf{v}}$ is an unbiased estimator of the true location vector $\mathbf{v}$, as below:

$$\begin{aligned}
\mathbb{E}(\overline{\mathbf{v}}) &\coloneqq \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{v} + \sigma\mathbb{E}(\mathbf{X}_i)) \\
&= \mathbf{v} + \sigma(\mathbf{0}) && \text{(since } \mathbb{E}(\mathbf{X}) = \mathbf{0}) \\
&= \mathbf{v} && \square
\end{aligned}$$

4.C.7 Proof of Proposition 4.14

Proposition 4.14 (Sample mean is $\sqrt{\frac{d}{n}}$-consistent)**.** *Let* $(\mathbf{Y}_i)_{i=1}^{n}$ *be generated according to Definition 4.4, with* σ *known to the observer. Then the sample mean estimator, i.e.* $\overline{\mathbf{v}}$*, satisfies* $\|\overline{\mathbf{v}} - \mathbf{v}\|_2 \leq \sigma\,\mathrm{diam}(K)\sqrt{\frac{d}{\gamma n}}$ *with probability at least* $1 - \gamma$*.*

Proof of Proposition 4.14. Our proof will essentially require an application of Chebychev's inequality, though we will need to setup some key definitions and notation. We have here that $\mathbf{Y}_i \overset{a.s.}{=} \mathbf{v} + \sigma \mathbf{X}_i$ for each $i \in [n]$. Let $\widetilde{\mathbf{\Sigma}} := \mathbb{V}(\mathbf{Y}_i) \in \mathbb{R}^{d \times d}$, and $\mathbf{Q} := \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\overline{\mathbf{v}} - \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\mathbf{Y}_i - \mathbf{v}) \in \mathbb{R}^d$. Here $\mathbf{Q}$ is the scaled-centered version of our sample mean location estimator $\overline{\mathbf{v}}$. We then have that $\mathbb{E}(\mathbf{Q}) = \mathbf{0}$, and thus $\mathbb{V}(\mathbf{Q}) = \mathbb{E}\left(\mathbf{Q}\mathbf{Q}^\top\right)$. Moreover we have $\mathbb{V}(\mathbf{Q}) = \frac{1}{n}\mathbf{I}_d$. To see this, we observe the following. For $\mathbb{E}(\mathbf{Q})$ we first note that $\mathbb{E}(\mathbf{Y}_i - \mathbf{v}) = \mathbb{E}(\sigma \mathbf{X}_i) = \sigma \mathbb{E}(\mathbf{X}_i) = \mathbf{0}$. We then have:

$$\begin{aligned}
\mathbb{E}(\mathbf{Q}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\mathbf{Y}_i - \mathbf{v})\right) && \text{(by definition)} \\
&= \mathbb{E}\left(\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\mathbf{Y}_1 - \mathbf{v})\right) \\
& && \text{(by linearity of expectation, and since } \mathbf{Y}_i \text{ are i.i.d.)} \\
&= \mathbf{0} && \text{(since } \mathbb{E}(\mathbf{Y}_1 - \mathbf{v}) = \mathbb{E}(\sigma \mathbf{X}_1) = \mathbf{0}\text{)}
\end{aligned}$$

As required. Now in the case of $\mathbb{V}(\mathbf{Q})$ we proceed as follows.

$$\begin{aligned}
\mathbb{V}(\mathbf{Q}) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\mathbf{Y}_i - \mathbf{v})\right) && \text{(by definition)} \\
&= \left(\frac{1}{n^2}\right)(n)\mathbb{V}\left(\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\mathbf{Y}_1 - \mathbf{v})\right) && \text{(since } \mathbf{Y}_i \text{ are i.i.d.)} \\
&= \frac{1}{n}\left(\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\right)^\top \mathbb{V}(\mathbf{Y}_i - \mathbf{v})\,\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}} \\
&= \frac{1}{n}\left(\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\right)^\top \widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}} \\
&= \frac{1}{n}\mathbf{I}_d && (4.81)
\end{aligned}$$

As required. We also observe that:

$$\begin{aligned}
\mathbb{E}\left(\left\|\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}(\overline{\mathbf{v}} - \mathbf{v})\right\|_2^2\right) &= \mathbb{E}\left(\mathbf{Q}^\top\mathbf{Q}\right) && \text{(by definition of } \mathbf{Q}.\text{)} \\
&= \mathbb{E}\left(\operatorname{tr}\left(\mathbf{Q}^\top\mathbf{Q}\right)\right) && (4.82) \\
&= \mathbb{E}\left(\operatorname{tr}\left(\mathbf{Q}\mathbf{Q}^\top\right)\right) \\
& && \text{(by cyclic permutation invariance of trace.)} \\
&= \operatorname{tr}\left(\mathbb{E}\left(\mathbf{Q}\mathbf{Q}^\top\right)\right) && \text{(linearity of trace and expectation.)} \\
&= \operatorname{tr}(\mathbb{V}(\mathbf{Q})) && (4.83)
\end{aligned}$$

Furthermore, note that $\operatorname{diam}(K) < \infty$, and $\operatorname{diam}(\sigma K) = \sigma \operatorname{diam}(K) < \infty$, using Lemma 4.63. We then observe that $\widetilde{\mathbf{\Sigma}} = \sigma^2 \mathbb{E}\left(\mathbf{X}_1 \mathbf{X}_1^\top\right)$. It follows that

$\mathbf{w}^\top \widetilde{\boldsymbol{\Sigma}} \mathbf{w} \leq \mathbb{E}\left(\|\mathbf{X}_1\|_2^2\right) \lesssim (\operatorname{diam}(\sigma K))^2 < \infty$, for all $\mathbf{w} \in \mathbb{S}^{d-1}$. Additionally we have that $\lambda_{\max}\left(\widetilde{\boldsymbol{\Sigma}}\right) \lesssim (\operatorname{diam}(\sigma K))^2 < \infty$.

Now, let $\varepsilon' > 0$, we then have:

$$
\begin{aligned}
\mathbb{P}\left(\|(\overline{\mathbf{v}} - \mathbf{v})\|_2 > \varepsilon'\right) &= \mathbb{P}\left(\left\|\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} (\overline{\mathbf{v}} - \mathbf{v})\right\|_2 > \varepsilon'\right) \\
&\leq \mathbb{P}\left(\left\|\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right\|_{\mathsf{op}} \left\|\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} (\overline{\mathbf{v}} - \mathbf{v})\right\|_2 > \varepsilon'\right) \\
&= \mathbb{P}\left(\left\|\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} (\overline{\mathbf{v}} - \mathbf{v})\right\|_2 > \varepsilon\right) && \left(\text{where } \varepsilon \coloneqq \frac{\varepsilon'}{\left\|\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right\|_{\mathsf{op}}}\right) \\
&\leq \frac{\mathbb{E}\left(\left\|\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} (\overline{\mathbf{v}} - \mathbf{v})\right\|_2^2\right)}{\varepsilon^2} && (\text{by Markov's inequality}) \\
&= \frac{\operatorname{tr}\left(\mathbb{V}\left(\mathbf{Q}\right)\right)}{\varepsilon^2} && (\text{using Equation (4.83)}) \\
&= \frac{\operatorname{tr}\left(\frac{1}{n}\mathbf{I}_d\right)}{\varepsilon^2} && (\text{using Equation (4.81)}) \\
&= \frac{d}{n\varepsilon^2} \\
&= \frac{d}{n} \frac{\left\|\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right\|_{\mathsf{op}}^2}{(\varepsilon')^2} && \left(\text{using } \varepsilon \coloneqq \frac{\varepsilon'}{\left\|\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right\|_{\mathsf{op}}}\right) \\
&= \frac{d}{n} \frac{\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{\mathsf{op}}}{(\varepsilon')^2} \\
&\leq \frac{d}{n} \frac{(\operatorname{diam}(\sigma K))^2}{(\varepsilon')^2} \\
&\leq \frac{d}{n} \frac{(\sigma \operatorname{diam}(K))^2}{(\varepsilon')^2} && (\text{using Equation (4.63).})
\end{aligned}
$$

It then follows that $\|\overline{\mathbf{v}} - \mathbf{v}\|_2 \leq \sigma \operatorname{diam}(K) \sqrt{\frac{d}{\gamma n}}$ with probability at least $1-\gamma$, as required. $\square$

4.C.8 Formal justification for Remark 4.15

In order to justify Remark 4.15, we want to construct a high probability lower bound for the estimation error, $\|\overline{\mathbf{v}} - \mathbf{v}\|_2$. Given an observation $\mathbf{Y}_i$, we denote its coordinate components as $\mathbf{Y}_i \coloneqq (Y_{i1}, \ldots, Y_{id})^\top$, for each $i \in [n]$. Now

WLOG let us consider the first coordinate of $\overline{\mathbf{v}}$, i.e. $[\overline{\mathbf{v}}]_1 \coloneqq \overline{\mathbf{v}}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_{1j}$. Similarly let $[\mathbf{v}]_1 \coloneqq v_1 = \frac{1}{n}\sum_{k=1}^{n} v_1$. We then have by the Paley-Zygmund inequality:

$$\mathbb{P}\left(n\left(\overline{\mathbf{v}}_1 - v_1\right)^2 > \frac{1}{2}\mathbb{E}\left(n\left(\overline{\mathbf{v}}_1 - v_1\right)^2\right)\right) \geq \frac{1}{4}\frac{\left(\mathbb{E}\left(n\left(\overline{\mathbf{v}}_1 - v_1\right)^2\right)\right)^2}{\mathbb{E}\left(n^2\left(\overline{\mathbf{v}}_1 - v_1\right)^4\right)} \tag{4.84}$$

We now focus on each term in the RHS quotient. For the numerator, using Proposition 4.13 we have that $\mathbb{E}\left(\overline{\mathbf{v}}\right) = \mathbf{v}$, and thus $\mathbb{E}\left(\overline{\mathbf{v}}_1\right) = v_1$, i.e., $\overline{\mathbf{v}}_1$ is an unbiased estimator of $[\mathbf{v}]_1$. So it follows that:

$$\begin{aligned}
\left(\mathbb{E}\left(n\left(\overline{\mathbf{v}}_1 - [\mathbf{v}]_1\right)^2\right)\right)^2 &= n^2\left(\mathbb{E}\left(\left(\overline{\mathbf{v}}_1 - [\mathbf{v}]_1\right)^2\right)\right)^2 \\
&= n^2\left(\mathbb{V}\left(\overline{\mathbf{v}}_1\right)\right)^2 \qquad \text{(by the bias-variance trade-off.)} \\
&=: n^2\left(\mathbb{V}\left(\frac{1}{n}\sum_{k=1}^{n} Y_{1j}\right)\right)^2 \\
&= n^2\frac{1}{n^2}\left(\mathbb{V}\left(Y_{11}\right)\right)^2 \\
&\qquad\qquad \text{(since } Y_{1j} \text{ are i.i.d., for each } j \in [n].\text{)} \\
&= \left(\mathbb{V}\left(Y_{11}\right)\right)^2 \qquad (4.85) \\
&> 0 \qquad \text{(since } Y_{11} \text{ is non-degenerate, so } \mathbb{V}\left(Y_{11}\right) > 0.\text{)}
\end{aligned}$$

For the denominator, we first observe that:

$$\begin{aligned}
\mathbb{E}\left(n^2\left(\overline{\mathbf{v}}_1 - v_1\right)^4\right) &= n^2\mathbb{E}\left(\left(\overline{\mathbf{v}}_1 - v_1\right)^4\right) \\
&= \frac{n^2}{n^4}\mathbb{E}\left(\left(\sum_{j=1}^{n}(Y_{1j} - v_1)\right)^4\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{p,q,r,s\in[n]}(Y_{1p} - v_1)(Y_{1q} - v_1)(Y_{1r} - v_1)(Y_{1s} - v_1)\right)
\end{aligned} \tag{4.86}$$

Since Y_{1j} are i.i.d., for each $j \in [n]$, it follows by linearity of expectation, and unbiasedness of $\overline{\mathbf{v}}_1$, that the expectation in Equation (4.86) is zero, whenever any single index differs from the others. Moreover the expectation is non-zero whenever $p = q = r = s$, and when both pairs of indices match, e.g. $p = q, r = s$.

The number of such cases is of the order n, and n^2, respectively. It then follows that:

$$\frac{1}{n^2}\mathbb{E}\left(\sum_{p,q,r,s\in[n]}(Y_{1p}-v_1)(Y_{1q}-v_1)(Y_{1r}-v_1)(Y_{1s}-v_1)\right) \tag{4.87}$$

$$\asymp (n)\left(\frac{1}{n^2}\right)\mathbb{E}\left((Y_{11}-v_1)^4\right)+(n^2)\left(\frac{1}{n^2}\right)\left(\mathbb{E}\left((Y_{11}-v_1)^2\right)\right)^2 \tag{4.88}$$

$$= \frac{\mathbb{E}\left((Y_{11}-v_1)^4\right)}{n}+\left(\mathbb{V}\left(Y_{11}\right)\right)^2 \tag{4.89}$$

We can then substitute Equations (4.85) and (4.86) into Equation (4.84) to obtain:

$$\mathbb{P}\left(n\left(\overline{\mathbf{v}}_1-v_1\right)^2>\frac{1}{2}\mathbb{E}\left(n\left(\overline{\mathbf{v}}_1-v_1\right)^2\right)\right) \tag{4.90}$$

$$\geq \frac{1}{4}\frac{\left(\mathbb{E}\left(n\left(\overline{\mathbf{v}}_1-v_1\right)^2\right)\right)^2}{\mathbb{E}\left(n^2\left(\overline{\mathbf{v}}_1-v_1\right)^4\right)} \tag{4.91}$$

$$\gtrsim \frac{1}{4}\frac{\left(\mathbb{V}\left(Y_{11}\right)\right)^2}{\frac{\mathbb{E}\left((Y_{11}-v_1)^4\right)}{n}+\left(\mathbb{V}\left(Y_{11}\right)\right)^2} \tag{4.92}$$

$$\gtrsim \frac{1}{4}\frac{\left(\mathbb{V}\left(Y_{11}\right)\right)^2}{\left(\mathbb{V}\left(Y_{11}\right)\right)^2+\left(\mathbb{V}\left(Y_{11}\right)\right)^2}$$

(for sufficiently large n, $\frac{\mathbb{E}\left((Y_{11}-v_1)^4\right)}{n}\leq\left(\mathbb{V}\left(Y_{11}\right)\right)^2$.)

$$\gtrsim \frac{1}{8} \tag{4.93}$$

As required.

4.C.9 Proof of Proposition 4.20

Before proving Proposition 4.20, we need to first prove some mathematical preliminaries. First we quickly prove (for completeness) the univariate uniform Pitman location estimator in Proposition 4.73.

Proposition 4.73 (Pitman location estimator for univariate uniform)**.** *Let* $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\lambda,\lambda]$*, with* $\lambda\in\mathbb{R}_{>0}$ *known, and let* $Y_i = X_i + v$ *for each* $i\in[n]$*, so that* $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[v-\lambda, v+\lambda]$*. Then the Pitman location estimator,* $\widehat{v}_{\text{pit}}$*, for* v*, is given by* $\widehat{v}_{\text{pit}} = \frac{X_{(1)}+X_{(n)}}{2}$.

Proof of Proposition 4.73. This is a standard result, see (Lehmann and Casella, 1998, Example 3.2). To ensure our paper is self-contained we quickly prove the result here. To do this, we observe that:

$$\begin{aligned}
\widehat{v}_{\text{pit}} &= \frac{\int u \frac{1}{(2\lambda)^n} \mathbb{I}\big(X_{(n)} - \lambda \leq u \leq X_{(1)} + \lambda\big) du}{\int \frac{1}{(2\lambda)^n} \mathbb{I}\big(X_{(n)} - \lambda \leq u \leq X_{(1)} + \lambda\big) du} \\
&= \frac{\int_{X_{(n)}-\lambda}^{X_{(1)}+\lambda} u du}{\int_{X_{(n)}-\lambda}^{X_{(1)}+\lambda} du} \\
&= \frac{\frac{1}{2}\left(\big(X_{(1)} + \lambda\big)^2 - \big(X_{(n)} - \frac{1}{2}\big)^2\right)}{X_{(1)} - X_{(n)}} \\
&= \frac{X_{(1)} + X_{(n)}}{2} \qquad (4.94)
\end{aligned}$$

As required. □

Next prove the tail probability concentration bound for the univariate uniform Pitman location estimator in Lemma 4.74.

Lemma 4.74 (Risk bounds for the univariate uniform Pitman location estimator)**.** *Let $\lambda > 0$ be known. Assume further that $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[-\lambda, \lambda]$. And further that $Y_i \overset{a.s.}{=} X_i + v$, for each $i \in [n]$, for some fixed but unknown $v \in \mathbb{R}$. Then the Pitman location estimator, $\widehat{v}_{\text{pit}} = \frac{X_{(1)}+X_{(n)}}{2}$ satisfies*

$$\mathbb{P}\left(|\widehat{v}_{\text{pit}} - v| > \varepsilon\right) \leq \begin{cases} 2\exp\left(-\frac{n\varepsilon}{2\lambda}\right) & \textit{if } \varepsilon \in (0, 2\lambda) \\ 0 & \varepsilon \geq 2\lambda \end{cases}$$

Proof of Lemma 4.74. First, we observe that the CDF of $X \sim \mathsf{Unif}[-\lambda, \lambda]$ is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < -\lambda \\ \frac{x+\lambda}{2\lambda} & x \in [-\lambda, \lambda] \\ 1 & \text{if } x > b \end{cases} \qquad (4.95)$$

Second, we observe that the general CDFs of i.i.d. minimum and maximum order statistics are given by:

$$\begin{aligned}
F_{X_{(1)}}(x) &= 1 - (1 - F_X(x))^n \\
&= \begin{cases} 0 & \text{if } x < -\lambda \\ 1 - \left(\frac{\lambda - x}{2\lambda}\right)^n & x \in [-\lambda, \lambda] \\ 1 & \text{if } x > \lambda \end{cases} \qquad (4.96) \\
F_{X_{(n)}}(x) &= (F_X(x))^n \\
&= \begin{cases} 0 & \text{if } x < -\lambda \\ \left(\frac{x + \lambda}{2\lambda}\right)^n & x \in [-\lambda, \lambda] \\ 1 & \text{if } x > \lambda \end{cases} \qquad (4.97)
\end{aligned}$$

Now let $\varepsilon > 0$ be arbitrary. Since $\lambda > 0$ is fixed, we first note the following:

$$-\lambda + \varepsilon = \begin{cases} (-\lambda, \lambda) & \text{if } \varepsilon \in (0, 2\lambda) \\ [\lambda, \infty) & \varepsilon \geq 2\lambda \end{cases} \qquad (4.98)$$

$$\lambda - \varepsilon = \begin{cases} (-\lambda, \lambda) & \text{if } \varepsilon \in (0, 2\lambda) \\ (-\infty, -\lambda] & \varepsilon \geq 2\lambda \end{cases} \qquad (4.99)$$

We can then directly calculate the tail bound as follows:

$$
\begin{aligned}
\mathbb{P}\left(|\widehat{v}_{\mathrm{pit}} - v| > \varepsilon\right) &= \mathbb{P}\left(\left|\frac{Y_{(1)} + Y_{(n)}}{2} - v\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\left|\frac{Y_{(1)} - (v - \lambda)}{2} + \frac{Y_{(n)} - (v + \lambda)}{2}\right| > \varepsilon\right) \\
&\qquad\qquad \left(\text{since } v = \tfrac{v-\lambda}{2} + \tfrac{v+\lambda}{2}\right) \\
&= \mathbb{P}\left(\left|Y_{(1)} - v + \lambda + Y_{(n)} - v - \lambda\right| > 2\varepsilon\right) \\
&= \mathbb{P}\left(\left|X_{(1)} + \lambda + X_{(n)} - \lambda\right| > \varepsilon\right) \\
&\qquad\qquad \left(\text{since } Y_i \overset{a.s.}{=} v + X_i,\ \forall i \in [n]\right) \\
&\leq \mathbb{P}\left(\left|X_{(1)} + \lambda\right| > \varepsilon\right) + \mathbb{P}\left(\left|X_{(n)} - \lambda\right| > \varepsilon\right) \\
&\qquad\qquad (\text{by union bound}) \\
&\leq \mathbb{P}\left(X_{(1)} > \varepsilon - \lambda\right) + \mathbb{P}\left(X_{(n)} \leq \lambda - \varepsilon\right) \\
&\qquad\qquad (\text{since } X_{(1)} + \lambda \geq 0 \text{ a.s., and } X_{(n)} - \lambda \leq 0 \text{ a.s.}) \\
&= \left[1 - F_{X_{(1)}}\left(\varepsilon - \lambda\right)\right] - F_{X_{(n)}}\left(\lambda - \varepsilon\right) \\
&= \begin{cases} \left(\frac{2\lambda - \varepsilon}{2\lambda}\right)^n + \left(\frac{2\lambda - \varepsilon}{2\lambda}\right)^n & \text{if } \varepsilon \in (0, 2\lambda) \\ 0 & \varepsilon \geq 2\lambda \end{cases} \\
&= \begin{cases} 2\left(1 - \frac{\varepsilon}{2\lambda}\right)^n & \text{if } \varepsilon \in (0, 2\lambda) \\ 0 & \varepsilon \geq 2\lambda \end{cases} \\
&\leq \begin{cases} 2\exp\left(-\frac{n\varepsilon}{2\lambda}\right) & \text{if } \varepsilon \in (0, 2\lambda) \\ 0 & \varepsilon \geq 2\lambda \end{cases} \\
&\qquad\qquad (\text{using } 1 - x \leq \exp(-x),\ \forall x \in \mathbb{R})
\end{aligned}
$$

As required. $\square$

Now, we can conclude with the proof of Proposition 4.20.

Proposition 4.20 (Projection Estimator for hyperrectangles)**.** *Under the setting of Example 4.18, let the marginal projection estimator,* $\widehat{\mathbf{v}}_{\mathrm{marg}}^{\mathrm{rect}}$*, be defined as per* (4.9)*. Then* $\widehat{\mathbf{v}}_{\mathrm{marg}}^{\mathrm{rect}}$ *satisfies* $\left\|\widehat{\mathbf{v}}_{\mathrm{marg}}^{\mathrm{rect}} - \mathbf{v}\right\|_2 \leq \frac{2\sigma\|\boldsymbol{\lambda}\|_2 \log\left(\frac{2d}{\gamma}\right)}{n}$ *with probability at least* $1 - \gamma$*, where* $\boldsymbol{\lambda} \coloneqq (\lambda_1, \ldots, \lambda_d)^\top$*.*

Proof of Proposition 4.20. Let $\boldsymbol{\lambda} \coloneqq (\lambda_1, \ldots, \lambda_d)^\top$. Our proof will be based on applying coordinate-wise union bound on $\left\|\widehat{\mathbf{v}}_{\mathrm{marg}}^{\mathrm{rect}} - \mathbf{v}\right\|_2$, and then applying the univariate uniform Pitman location estimator tail bounds across each coordinate. This can be seen as follows:

$$
\begin{aligned}
\mathbb{P}\left(\left\|\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}-\mathbf{v}\right\|_2>\varepsilon\right) &= \mathbb{P}\left(\left\|\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}-\mathbf{v}\right\|_2^2>\varepsilon^2\right)\\
&\leq \sum_{k=1}^{d}\mathbb{P}\left(\left|\widehat{\pi}_k-v_k\right|^2>\frac{\varepsilon^2\lambda_k^2}{\|\boldsymbol{\lambda}\|_2^2}\right) \quad \text{(by the union bound.)}\\
&\leq \sum_{k=1}^{d}\mathbb{P}\left(\left|\widehat{\pi}_k-v_k\right|>\frac{\varepsilon\lambda_k}{\|\boldsymbol{\lambda}\|_2}\right)\\
&\leq \sum_{k=1}^{d}2\exp\left(-\frac{n\varepsilon\lambda_k}{2\sigma\lambda_k\left\|\boldsymbol{\lambda}\right\|_2}\right) \quad \text{(using Lemma 4.74.)}\\
&= 2d\exp\left(-\frac{n\varepsilon}{2\sigma\left\|\boldsymbol{\lambda}\right\|_2}\right)
\end{aligned}
$$

It then follows that $\left\|\widehat{\mathbf{v}}_{\text{marg}}^{\text{rect}}-\mathbf{v}\right\|_2 \leq \frac{2\sigma\|\boldsymbol{\lambda}\|_2\log\left(\frac{2d}{\gamma}\right)}{n}$ with probability at least $1-\gamma$, as required. $\square$

4.C.10 Proof of Theorem 4.21

Before proving Theorem 4.21, we first need to prove a technical lemma. The main purpose of Lemma 4.75 is to understand the behavior of a perturbed univariate density with compact support. Specifically we want to upper bound the Hellinger distance between the density and it is perturbed counterpart. In order to bound the Hellinger distance appropriately, we need the univariate density to satisfy certain tail-decay and Lipschitz properties. These general conditions, as we will soon see are satisfied in Theorem 4.21 for densities of the marginal projections of the unit ball.

Lemma 4.75. *Suppose $g(t)\colon[a,b]\to\mathbb{R}_{>0}$ is a univariate density with compact support. For $\alpha>0$, let $v\in\left(0,\min\left\{\frac{1}{2},b-a,\frac{b-a}{4},\left(\frac{b-a}{4}\right)^{\alpha}\right\}\right)$ be a perturbation value, both to be determined later. We then have the v-perturbed density $g(t+v)\colon[a-v,b-v]\to\mathbb{R}_{>0}$, and $[a-v,b]=\operatorname{dom}(g(t))\cup\operatorname{dom}(g(t+v))$. Consider the partition of this domain:*

$$
[a-v,b]\coloneqq[a-v,a+v^{\alpha})\sqcup[a+v^{\alpha},b-v-v^{\alpha})\sqcup[b-v-v^{\alpha},b] \tag{4.100}
$$

Let $F_{g(t)}\colon(-\infty,\infty)\to[0,1]$, denote the CDF of $g(t)$. Now assume further that *we have the following:*

$$F_{g(t)}(a+l) \lesssim l^{\gamma}, \text{ for each } l \in (0, b-a) \text{ and a fixed } \gamma \geq 1. \tag{4.101}$$

$$1 - F_{g(t)}(b-l) \lesssim l^{\gamma}, \text{ for each } l \in (0, b-a) \text{ and a fixed } \gamma \geq 1. \tag{4.102}$$

$$\sqrt{g(t)} \text{ is } v^{-\delta\alpha}\text{-Lipschitz over the domain } [a+v^{\alpha}, b-v-v^{\alpha}). \tag{4.103}$$

We then have that $(\mathrm{d_H}(g(t), g(t+v)))^2 \lesssim v^{\frac{2\gamma}{\gamma+2\delta}}$. *Where* $(\mathrm{d_H}(g(t), g(t+v)))^2$ *is the squared Hellinger distance between the perturbed densities.*

Proof of Lemma 4.75. Suppose we consider a component whose marginal density is defined on the compact interval $[a, b]$. Let us denote the marginal density by $g(t) : [a, b] \to \mathbb{R}_{>0}$. We will lower bound the risk of estimating the location parameter in this univariate density. Specifically consider a the density given by perturbing (i.e. translating) this random variable by a scalar $v \in \mathbb{R}$. Let us denote this perturbed density by $g(t+v) : [a-v, b-v] \to \mathbb{R}$. Here we assume $v \in \left(0, \min\left\{\frac{1}{2}, b-a\right\}\right)$ be a perturbation to be specified later. This is best visualized in Figure 4.C.1.

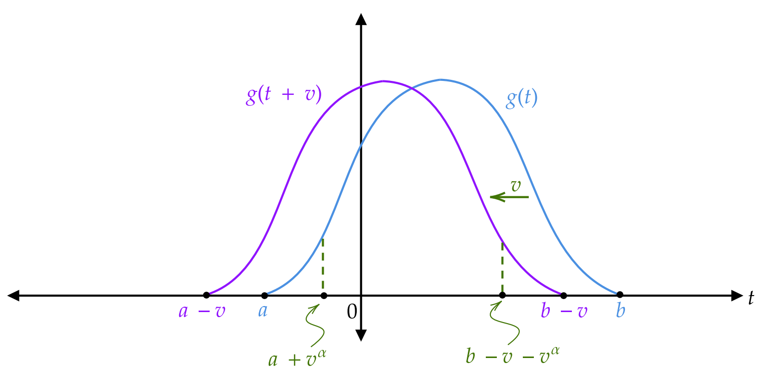

Figure 4.C.1: v-perturbed marginal densities $(g(t), g(t+v))$ for counterexample

Now consider the squared Hellinger distance between the densities $g(t)$ and $g(t+v)$ i.e. $(\mathrm{d_H}(g(t), g(t+v)))^2$. We then have that $(\mathrm{d_H}(g(t), g(t+v)))^2 = 1 - \int_{-\infty}^{\infty} \sqrt{g(t)g(t+v)}\,dt$. However we observe that

$$\begin{aligned}
(\mathrm{d_H}(g(t), g(t+v)))^2 &= 1 - \int_{-\infty}^{\infty} \sqrt{g(t)g(t+v)}\,dt \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt
\end{aligned}$$

In this formulation, the integral is non-zero, over the closed interval $[a - v, b] =: \operatorname{dom}(g(t)) \cup \operatorname{dom}(g(t+v))$, where we assume WLOG that $v \in (0, \frac{1}{2})$. We are given that $\alpha > 0$, which will be determined later. Now we can split this integral over 3 disjoint intervals which form a partition of $[a - v, b]$. Specifically this partition is $\mathrm{I} := [a - v, a + v^{\alpha})$, $\mathrm{II} := [a + v^{\alpha}, b - v - v^{\alpha})$, and $\mathrm{III} := [b - v - v^{\alpha}, b]$. Note that the existince of such a partition with I, II, III being non empty implicitly relies on the fact that $a + v^{\alpha} < b - v - v^{\alpha}$. Sufficient conditions to meet this criteria are $v^{\alpha} < \frac{b-a}{4}$ and $v < \frac{b-a}{4}$. This is not a concern, since the application of this lemma gives us control on the perturbation parameter v and we can always enforce the additional constraints $v \in \left(0, \min\left\{\frac{1}{2}, b - a, \frac{b-a}{4}, \left(\frac{b-a}{4}\right)^{\alpha}\right\}\right)$. Now we have that

$$
\begin{aligned}
\left(\mathrm{d}_{\mathsf{H}}\left(g(t), g(t+v)\right)\right)^2 &= \frac{1}{2}\int_{\mathrm{I} \,\sqcup\, \mathrm{II} \,\sqcup\, \mathrm{III}} \left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt \\
&=: A + B + C
\end{aligned}
$$

Here $A := \frac{1}{2}\int_{\mathrm{I}}\left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt$, $B := \frac{1}{2}\int_{\mathrm{II}}\left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt$, and $C := \frac{1}{2}\int_{\mathrm{III}}\left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt$. Now we upper bound each of these expressions in turn. Firstly we observe that using the identity $\frac{(a-b)^2}{2} \le a^2 + b^2$ for each $a, b \in \mathbb{R}$, we get the upper bound on the squared Hellinger distance as $\left(\mathrm{d}_{\mathsf{H}}\left(g(t), g(t+v)\right)\right)^2 \le \int_{-\infty}^{\infty}(g(t) + g(t+v))dt$. We then have that

$$
\begin{aligned}
A &= \frac{1}{2}\int_{\mathrm{I}}\left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt \\
&\le \int_{\mathrm{I}}(g(t) + g(t+v))dt \\
&= \int_{a-v}^{a+v^{\alpha}} g(t)dt + \int_{a-v}^{a+v^{\alpha}} g(t+v)dt \\
&= \int_{a}^{a+v^{\alpha}} g(t)dt + \int_{a-v}^{a+v^{\alpha}} g(t+v)dt \\
&= F_{g(t)}(a + v^{\alpha}) + F_{g(t+v)}(a + v^{\alpha}) \qquad (\text{since } F_{g(t)}(a) = F_{g(t+v)}(a - v) = 0) \\
&= F_{g(t)}(a + v^{\alpha}) + F_{g(t)}(a + v + v^{\alpha}) \\
&\lesssim v^{\gamma\alpha} + v^{\gamma\alpha} \qquad (\text{per assumption on } F_{g(t)}) \\
&\lesssim v^{\gamma\alpha} \qquad (\text{ignoring lower order terms})
\end{aligned}
$$

A similar calculation shows that $C \lesssim v^{\gamma\alpha}$. More explicitly, this can be seen as follows:

$$
\begin{aligned}
C &= \frac{1}{2}\int_{\text{III}} \left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt \\
&\leq \int_{\text{III}} (g(t) + g(t+v)) dt \\
&= \int_{b-v-v^{\alpha}}^{b} g(t)dt + \int_{b-v-v^{\alpha}}^{b} g(t+v)dt \\
&= \int_{b-v-v^{\alpha}}^{b} g(t)dt + \int_{b-v-v^{\alpha}}^{b-v} g(t+v)dt \\
&= F_{g(t)}(b) - F_{g(t)}(b-v-v^{\alpha}) + F_{g(t+v)}(b-v) - F_{g(t+v)}(b-v-v^{\alpha}) \\
&= F_{g(t)}(b) - F_{g(t)}(b-v-v^{\alpha}) + F_{g(t)}(b) - F_{g(t)}(b-v^{\alpha}) \\
&= 1 - F_{g(t)}(b-v-v^{\alpha}) + 1 - F_{g(t)}(b-v^{\alpha}) \\
&\lesssim v^{\gamma\alpha} + v^{\gamma\alpha} && \text{(per assumption on } F_{g(t)}) \\
&\lesssim v^{\gamma\alpha}. && \text{(ignoring lower order terms)}
\end{aligned}
$$

Finally given that the integral represented by B is on over the interval $\text{II} := [a + v^{\alpha}, b - v - v^{\alpha})$, we are given that $\sqrt{g(t)}$ is $v^{-\delta\alpha}$-Lipschitz over II by assumption. As such it follows that:

$$
\begin{aligned}
B &:= \frac{1}{2}\int_{\text{II}} \left(\sqrt{g(t)} - \sqrt{g(t+v)}\right)^2 dt \\
&\leq \frac{1}{2}\int_{\text{II}} \left(v^{-\delta\alpha}\left|t+v-t\right|\right)^2 dt && \text{(since } \sqrt{g(t)} \text{ is } v^{-\delta\alpha}\text{-Lipschitz over II)} \\
&= \frac{1}{2}\int_{\text{II}} \left(v^{2-2\delta\alpha}\right) dt \\
&= v^{2-2\delta\alpha}\left(b - a - v - 2v^{\alpha}\right) \\
&\lesssim v^{2-2\delta\alpha} && \text{(since } v, v^{\alpha} > 0)
\end{aligned}
$$

Thus we have that $\left(\mathrm{d}_{\mathsf{H}}\left(g(t), g(t+v)\right)\right)^2 = A + B + C \lesssim v^{2-2\delta\alpha} + v^{\gamma\alpha} = \max\left\{v^{2-2\delta\alpha}, v^{\gamma\alpha}\right\}$. Optimizing this over $\alpha > 0$ and $v \in \left(0, \min\left\{\frac{1}{2}, b-a\right\}\right)$, occurs when $\gamma\alpha = 2 - 2\delta\alpha$. This gives us $\alpha = \frac{2}{\gamma+2\delta}$. It then follows that $\left(\mathrm{d}_{\mathsf{H}}\left(g(t), g(t+v)\right)\right)^2 \lesssim v^{\frac{2\gamma}{\gamma+2\delta}}$, as required. □

4.C.11 Final proof of Theorem 4.21

With Lemma 4.75 proved, we are now ready to prove Theorem 4.21.

Theorem 4.21 (Lower bound of $K = \overline{B}_2^3(\mathbf{0}, R)$)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. Furthermore, let $d = 3$, $\sigma = 1$, and $K = \overline{B}_2^3(\mathbf{0}, R) \in \mathcal{K}^3$. Suppose WLOG we are estimating the first coordinate of the location parameter, v_1. Let W denote the class of all such marginal projection estimators for v_1, as per Definition 4.16. Then there exists some $C' \in (0,1)$ such that the following holds:*

$$\inf_{\widetilde{w} \in W} \sup_{v_1 \in \mathbb{R}} \mathbb{P}\left(|\widetilde{w} - v_1| \geq (1 - C')n^{-\frac{3}{4}}\right) \geq C' \tag{4.10}$$

Proof of Theorem 4.21. We can now use this to construct a counterexample in dimension $d = 3$, in the case where $K = B_2^3(\mathbf{0}, R)$ i.e. the closed Euclidean ball of radius R, centered at $\mathbf{0}$.

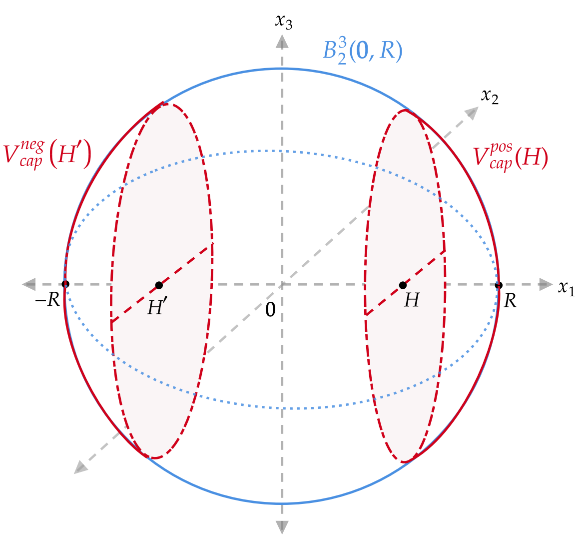

Figure 4.C.2: 3D Spherical caps along a marginal axis

Here we observe by symmetry that there are always 2 caps projected on a marginal axis. We denote them by $V_{\text{cap}}^{\text{neg}}(H')$ and $V_{\text{cap}}^{\text{pos}}(H)$. Where $H' \in [-R, 0)$ and $H \in [0, R]$ as per Figure 4.C.2. Using the well known formulas we have that $V_{\text{cap}}^{\text{neg}}(H') = \frac{\pi(H'+R)^2}{3}(2R - H')$ and $V_{\text{cap}}^{\text{pos}}(H) = \frac{\pi(R-H)^2}{3}(2R + H)$. Since we are sampling uniformly across this sphere, we observe that the probability of lying in either of these caps is simply the ratio of the volume of the spherical cap to the volume of the sphere. In our case $V_{\text{sphere}} = \frac{4\pi}{3}R^3$. Denote the random variable taking values in $[-R, R]$ along the marginal axis by X, when sampling uniformly from the given sphere. We then have that $F_X^{\text{neg}}(H') \coloneqq \mathbb{P}(X \leq H') = \frac{V_{\text{cap}}^{\text{neg}}(H')}{V_{\text{sphere}}}$ and $F_X^{\text{pos}}(H) \coloneqq \mathbb{P}(X \leq H) = 1 - \frac{V_{\text{cap}}^{\text{pos}}(H)}{V_{\text{sphere}}}$. This gives us the following 2 cases:

$$
\begin{aligned}
F_X^{\text{neg}}(H') &= -\frac{H'^3}{4R^3} + \frac{3H'}{4R} + \frac{1}{2} \\
F_X^{\text{pos}}(H) &= \frac{1}{2} - \frac{H^3}{4R^3} + \frac{3H}{4R}
\end{aligned}
$$

That is $F_X^{\text{pos}}(H) = F_X^{\text{neg}}(H)$ for each $H \in [-R, R]$. So the CDF is given by:

$$
F_X(H) = \begin{cases} 0 & \text{if } H < -R \\ \frac{1}{2} - \frac{H^3}{4R^3} + \frac{3H}{4R} & H \in [-R, R] \\ 1 & \text{if } H > R \end{cases}
$$

We then have that $f_X(H) = -\frac{3H^2}{4R^3} + \frac{3}{4R}$. We first need to verify that $F_X(H)$ meets the sufficient conditions from Lemma 4.75. First we observe that $F_X(-R+l) = \frac{l^2(3R-l)}{4R^3} = F_X(R-l) \lesssim l^2$, for each $l \in [0, 2R]$, thus satisfying the first sufficient condition with $\gamma = 2$ in Lemma 4.75. We now need to verify the second (and final) sufficient condition for f_X i.e. $\sqrt{f_X(H)}$ is $v^{-\delta\alpha}$-Lipschitz over $\text{II} \coloneqq [-R+v^\alpha, R-v-v^\alpha)$ Lemma 4.75. For notational convenience, let us denote $h_X(H) \coloneqq \sqrt{f_X(H)} = \sqrt{-\frac{3H^2}{4R^3} + \frac{3}{4R}}$. To show that $h_X(H)$ is Lipschitz over II, we just need to show equivalently that the first derivative is bounded over II. We then have that $h'_X(H) = \frac{-\sqrt{3}H}{2R^3\sqrt{\frac{R^2-H^2}{R^3}}}$. Since this is a continuous decreasing function in H for $H \in \text{II}$. We can bound it by considering the value at $H = -R + v^\alpha$. We then have that

$$
\begin{aligned}
\left|h'_X(-R+v^\alpha)\right| &= \frac{\sqrt{3}\,(R-v^\alpha)}{2\sqrt{R^3 v^\alpha (2R - v^\alpha)}} \\
&\leq \frac{\sqrt{3}R}{2\sqrt{R^4 v^\alpha}} \qquad \text{(since } 0 < v^\alpha \leq R \implies 2R - v^\alpha \geq R) \\
&\lesssim \frac{1}{\sqrt{v^\alpha}} \\
&= v^{-\frac{1}{2}\alpha}
\end{aligned}
$$

As such $v^{-\frac{1}{2}\alpha}$ represents an upper bound on the Lipschitz constant for $h'_X(H)$ over II. So indeed h_X $v^{-\delta\alpha}$-Lipschitz over $\text{II} \coloneqq [-R+v^\alpha, R-v-v^\alpha)$, with $\delta = \frac{1}{2}$ as required for Lemma 4.75. Since both sufficient conditions are for Lemma 4.75 to hold are satisfied by $\sqrt{f_X(H)}$ for $H \in [-R, R]$ we have that $(\mathrm{d}_{\mathsf{H}}(g(t), g(t+v)))^2 \lesssim v^{\frac{2\gamma}{\gamma+2\delta}} = v^{\frac{4}{3}}$, as required. Suppose WLOG we are estimating the first coordinate of the location parameter, v_1. Let $\widetilde{w}$ be any such marginal projection estimator for v_1, projected in any direction and based on n observations. We can apply Le Cam's two-point method (on n observed samples), as per Lemma 4.100, to this process as follows:

$$
\begin{aligned}
\inf_{\widetilde{w} \in \mathbb{R}} \sup_{v_1 \in \mathbb{R}} \mathbb{P}\left(|\widetilde{w} - v_1| \geq v\right) &\geq \left(1 - \frac{1}{2}\mathrm{d}_{\mathsf{TV}}\left(g(t)^{\otimes n}, g(t+v)^{\otimes n}\right)\right) \\
&\geq \left(1 - \frac{1}{2}\mathrm{d}_{\mathsf{H}}\left(g(t)^{\otimes n}, g(t+v)^{\otimes n}\right)\right) \\
&\geq 1 - \sqrt{n}\,\mathrm{d}_{\mathsf{H}}\left(g(t), g(t+v)\right)
\end{aligned}
$$

Now note that $\mathrm{d}_{\mathsf{H}}(g(t), g(t+v)) \leq v^{\frac{2}{3}}$. Setting $1 - \sqrt{n}\, v^{\frac{2}{3}} = C' \in (0,1) \implies v = (1 - C')^{\frac{3}{2}} n^{-\frac{3}{4}}$. It then follows directly that

$$
\inf_{\widetilde{w} \in \mathbb{R}} \sup_{v_1 \in \mathbb{R}} \mathbb{P}\left(|\widetilde{w} - v_1| \geq (1 - C')^{\frac{3}{2}} n^{-\frac{3}{4}}\right) \geq C'
$$

Which indeed shows that we have a worse than n^{-1} rate for estimation of an individual coordinate the location parameter, and certainly for all coordinates in this specific setting. $\square$

Remark 4.76. It is instructive to understand why the marginal projection estimator fails to give the n^{-1} rate when K is the closed three dimensional closed euclidean ball, compared to the case where K is an axis-aligned hyperrectangle. In the case of the axis-aligned hyperrectangle, we observe that projecting along each marginal axis, the resulting marginal projected density is bounded *strictly* away from zero at the boundary points. In the case of the Euclidean ball, we find that each of the projected marginal densities decays to zero at both boundary points. The use of the Euclidean ball as a counterexample has the added property in that it is rotation invariant. As such the marginal projections will have this same boundary decay issue for the marginal projection in *any* direction.

4.C.12 Proof of Proposition 4.22

Proposition 4.22 (Projection Estimator Motivation)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Let $\overline{\mathbf{v}} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$*

denote the sample mean, and $\widehat{\mathbf{v}}_{\text{prj}} = \Pi_{\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)}(\overline{\mathbf{v}})$ *denote the projection location estimator. Then for any* $\mathbf{z} \in \bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$ *we have:*

$$\|\overline{\mathbf{v}} - \mathbf{z}\|_2 \geq \|\widehat{\mathbf{v}}_{\text{prj}} - \mathbf{z}\|_2 \text{ a.s.} \tag{4.12}$$

Proof of Proposition 4.22. We note that the critical set $\bigcap_{i=1}^{n}(\mathbf{Y}_i - \sigma K)$ is a non-empty, closed convex set per Proposition 4.8. Since $\widehat{\mathbf{v}}_{\text{prj}}$ is by definition a euclidean projection onto the critical set, the required result then follows from a direct application of Theorem 4.48. $\square$

4.D Proofs of Section 4.3

4.D.1 Proof of Theorem 4.25

Mathematical Preliminaries

Since $K \in \mathcal{K}^d$ with a centroid $(K) = \mathbf{0} \in \operatorname{int}(K)$, it follows that there is a ball centered at $\mathbf{0}$ which is a proper subset of K. We also have the following important fact of the Lipschitz property of Minkowski gauge functionals on convex bodies as detailed in Theorem 4.77.

Theorem 4.77 (Minkowski Gauge Functionals are Lipschitz)**.** *Let K be a closed, convex set with $\mathbf{0} \in \operatorname{int}(K)$. Then the Minkowski gauge functional $\rho_K : \mathbb{R}^d \to [0, \infty)$ is G-Lipschitz continuous on $\mathbb{R}^d$ with the constant*

$$G \coloneqq \inf \left\{ \frac{1}{r} \,\middle|\, B_2^d(\mathbf{0}, r) \subset K, r > 0 \right\} \tag{4.104}$$

In particular, we have $\rho_K(\mathbf{x}) \leq G \|x\|_2$ for all $\mathbf{x} \in \mathbb{R}^d$.

Proof of Theorem 4.77. See (Mordukhovich and Nam, 2014, Proposition 3.32) for details. □

Final proof of Theorem 4.25

Theorem 4.25 (Minimum is attained in the MLEs)**.** *The* MLEs *for* $\mathbf{v}$ *and* σ *as given by* $\widehat{\mathbf{v}}_{\text{MLE}} \in \arg\min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$, *and* $\widehat{\sigma}_{\text{MLE}} \coloneqq \min_{\boldsymbol{\tau} \in \mathbb{R}^d} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$, *respectively, both exist.*

Proof of Proposition 4.27. We begin by proving our claim that the infimum above is actually a minimum. Since $K \in \mathcal{K}^d$ with a centroid $(K) = \mathbf{0} \in \operatorname{int}(K)$, it follows from Theorem 4.77 that the Minkowski gauge functional of K, $\rho_K(\mathbf{x})$, is a Lipschitz map. Per Lemma 4.61 the pointwise maximum of a finite collection of Lipschitz functions preserves Lipschitzness. As such, for each $n \in \mathbb{N}$, it follows that as $\max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau})$ is continuous. As long as we show that we should search for $\boldsymbol{\tau}$ in a compact set this will show the minimum is achieved. Now we note that

$$\inf_{\boldsymbol{\tau} \in \mathbb{R}^p} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \leq \max_{i \in [n]} \rho_K(\mathbf{Y}_i),$$

with the latter quantity being bounded. Now $\rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \geq \rho_K(\boldsymbol{\tau}) - \rho_K(\mathbf{Y}_i)$. It follows that if $\rho_K(\boldsymbol{\tau}) > 2 \max_i \rho_K(\mathbf{Y}_i)$ we have that

$$\max_i \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) > \max_i \rho_K(\mathbf{Y}_i) \geq \inf_{\boldsymbol{\tau} \in \mathbb{R}^p} \max_{i \in [n]} \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}),$$

which means that such $\boldsymbol{\tau}$ cannot achieve the inf. It follows that one should search for $\boldsymbol{\tau}$ only in the set $\rho_K(\boldsymbol{\tau}) \le 2\max_i \rho_K(\mathbf{Y}_i)$. But K is compact hence the latter is a compact set, therefore completing the proof.
The next claim is that the value of the above problem coincides with $\widehat{\sigma}$, while $\widehat{\boldsymbol{\tau}}$ which achieves the minimum is a point in the set $\bigcap_{i=1}^n (\mathbf{Y}_i - \widehat{\sigma}K)$. Let $\widetilde{\sigma}$ denote the value of the new optimization problem. By the fact that K is closed it follows that $Y_i - \boldsymbol{\tau} \in \widetilde{\sigma}K$ for all i, which implies that $\bigcap_{i=1}^n (\mathbf{Y}_i - \widetilde{\sigma}K) \neq \emptyset$. So $\widetilde{\sigma} \ge \widehat{\sigma}$. On the other hand suppose that $\widehat{\sigma} < \widetilde{\sigma}$. It follows by the definition of $\widetilde{\sigma}$ that there does not exist a $\boldsymbol{\tau}$ such that $\max \rho_K(\mathbf{Y}_i - \boldsymbol{\tau}) \le (\widehat{\sigma} + \widetilde{\sigma})/2$, and therefore the set $\bigcap_{i=1}^n (\mathbf{Y}_i - (\widehat{\sigma} + \widetilde{\sigma})/2K) = \emptyset$. This is a contradiction with the definition of $\widehat{\sigma}$. Thus it has to be the case that $\widehat{\sigma} = \widetilde{\sigma}$ and therefore $\bigcap_{i=1}^n (\mathbf{Y}_i - \widehat{\sigma}K) \neq \emptyset$. ☐

4.D.2 Proof of Proposition 4.27

Proposition 4.27 (Location MLE is contained in the critical set)**.** *Since* $0 < \widehat{\sigma}_{\mathrm{MLE}} \le \sigma$ *and* $\mathbf{0} \in \operatorname{int}(K)$ *(per Remark 4.6), we have that*

$$\bigcap_{i=1}^{n} (\mathbf{Y}_i - \widehat{\sigma}_{\mathrm{MLE}} K) \subseteq \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)$$

Proof of Proposition 4.27. Since $0 < \widehat{\sigma}_{\mathrm{MLE}} \le \sigma$, we have that:

$$\begin{aligned} &(\mathbf{Y}_i - \widehat{\sigma}_{\mathrm{MLE}} K) \subseteq (\mathbf{Y}_i - \sigma K) \qquad (\forall\, i \in [n], \text{ using Lemma 4.54}) \\ \implies &\bigcap_{i=1}^{n} (\mathbf{Y}_i - \widehat{\sigma}_{\mathrm{MLE}} K) \subseteq \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \end{aligned}$$

As required. ☐

4.E Proofs of Section 4.4

4.E.1 Proof of Theorem 4.28

As noted, Theorem 4.28 provides upper bounds for the risk of our projection estimator in the case where $K \in \mathcal{K}^d$ is restricted to the class of convex polytopes. The literature on convex polytopes is vast, e.g., see (Goodman et al., 2018, Chapter 15). In order to keep our proof of Theorem 4.28 largely self-contained we first provide the minimal necessary convex polytope theory in Section 4.E.1.

Required convex polytope theory

Definition 4.78 (Convex Polytope)**.** A convex polytope $K \subset \mathbb{R}^d$ is a bounded subset which is the intersection of a finite number halfspaces.

Remark 4.79. Throughout our paper we will always assume that a convex polytope, say $K \subset \mathbb{R}^d$ is in fact a convex body polytope, i.e. $K \in \mathcal{K}^d$. Thus we assume in addition to Definition 4.78 that K has a non-empty interior.

Definition 4.80 ($\mathcal{V}$-representation of a Polytope)**.** A $\mathcal{V}$-representation of a convex polytope, $K \subset \mathbb{R}^d$ is the convex hull of a finite set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of points in $\mathbb{R}^d$

$$K = \operatorname{co}(X) \coloneqq \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \,\middle|\, \lambda_1, \dots, \lambda_n \geq 0, \sum_{i=1}^{n} \lambda_i = 1 \right\}. \tag{4.105}$$

Definition 4.81 ($\mathcal{H}$-representation of a Polytope)**.** An $\mathcal{H}$-representation of a convex polytope $K \subset \mathbb{R}^d$, is the solution set of a finite system of linear inequalities,

$$K \coloneqq [\mathbf{A} \mid \mathbf{b}] \coloneqq \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{x} \leq b_k, \, \forall\, k \in [m] \right\}, \tag{4.106}$$

with the extra condition that the set of solutions is bounded, that is, such that there is a constant $M > 0$ such that $\|\mathbf{x}\| \leq M$ holds for all $\mathbf{x} \in K$. If this boundedness condition is removed, we say that K is a convex *polyhedron* instead. Here $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a real matrix with rows $\mathbf{a}_k^\top \coloneqq (a_{k1}, \dots, a_{kd}) \in \mathbb{R}^{1 \times d}$, and $\mathbf{b} \coloneqq (b_1, \dots, b_m)^\top \in \mathbb{R}^m$. We say that a polytope K has a *minimal* $\mathcal{H}$-representation, if every inequality in Equation (4.106) is irredundant. That is, removing *any* inequality from the $\mathcal{H}$-representation would violate Definition 4.78. Such a minimal $\mathcal{H}$-representation is unique (up to ordering and scaling).

Remark 4.82. We note that since centroid $(K) = \mathbf{0} \in \operatorname{int}(K)$, that $\mathbf{a}_k \neq \mathbf{0}$ and $b_k \neq 0$ for each $k \in [m]$ i.e. none of the supporting hyperplanes of K pass through $\mathbf{0}$.

Lemma 4.83 ($\mathcal{H}$-representation of a Polytope under affine transformations)**.** *Let $K \in \mathcal{K}^d$ be a convex polytope per Definition 4.78 with (minimal) $\mathcal{H}$-representation $K \coloneqq [\mathbf{A} \mid \mathbf{b}]$. Further, for any fixed $\alpha > 0, \mathbf{c} \in \mathbb{R}^d$, we have that $-\alpha K + \mathbf{c}$, and $\alpha K + \mathbf{c}$ are both convex polytopes. Moreover they have following respective $\mathcal{H}$-representations:*

$$-\alpha K + \mathbf{c} = [-\mathbf{A} \mid -\mathbf{A}\mathbf{c} + \alpha\mathbf{b}] \tag{4.107}$$

$$\alpha K + \mathbf{c} = [\mathbf{A} \mid \mathbf{A}\mathbf{c} + \alpha\mathbf{b}] \tag{4.108}$$

Proof of Lemma 4.83. We prove each of properties specified in Equations (4.107) and (4.108) in turn.

(*Proof of Equation* (4.107)). To see this firstly note that $K \coloneqq [\mathbf{A} \mid \mathbf{b}]$ is the minimal $\mathcal{H}$-representation of K. We then proceed as follows:

$$\begin{aligned}
&\text{Let } \mathbf{w} \in -\alpha K + \mathbf{c} \\
\iff\ & \mathbf{w} = -\alpha\mathbf{k} + \mathbf{c} && (\text{for some } \mathbf{k} \in K.) \\
\iff\ & \mathbf{k} = -\frac{\mathbf{w} - \mathbf{c}}{\alpha} && (\text{since } \alpha > 0.) \\
\iff\ & \mathbf{A}\left(-\frac{\mathbf{w} - \mathbf{c}}{\alpha}\right) \leq \mathbf{b} && (\text{since } \mathbf{k} \in K \iff \mathbf{A}\mathbf{k} \leq \mathbf{b}.) \\
\iff\ & -\mathbf{A}\mathbf{w} + \mathbf{A}\mathbf{c} \leq \alpha\mathbf{b} && (\text{since } \alpha > 0.) \\
\iff\ & -\mathbf{A}\mathbf{w} \leq -\mathbf{A}\mathbf{c} + \alpha\mathbf{b} \\
\iff\ & \mathbf{w} \in [-\mathbf{A} \mid -\mathbf{A}\mathbf{c} + \alpha\mathbf{b}]
\end{aligned}$$

Which indeed proves Equation (4.107). This shows that $-\alpha K + \mathbf{c}$ is a convex polyhedron, given that it has the $\mathcal{H}$-representation $[-\mathbf{A} \mid -\mathbf{A}\mathbf{c} + \alpha\mathbf{b}]$. To prove that it is a polytope, we need to show that it is bounded. But from Lemma 4.59, $-\alpha K + \mathbf{c}$ is in fact compact, and thus bounded, as required. ∎

(*Proof of Equation* (4.108)). The proof is almost identical to the proof of Equation (4.107). The main change is to replace the line:

$$\mathbf{k} = -\frac{\mathbf{w} - \mathbf{c}}{\alpha}, \text{ with } \mathbf{k} = \frac{\mathbf{w} - \mathbf{c}}{\alpha}, \tag{4.109}$$

and continue the proof as before. ∎

Thus all properties specified in Equations (4.107) and (4.108) are now proved. □

Theorem 4.84 (Dual representation of a Polytope)**.** *The definitions of $\mathcal{V}$-polytopes and of $\mathcal{H}$-polytopes are equivalent. That is, every $\mathcal{V}$ polytope has a description by a finite system of inequalities, and every $\mathcal{H}$-polytope can be obtained as the convex hull of a finite set of points (its vertices).*

Proof. See (Ziegler, 1995, Section 1.1) for details. □

Lemma 4.85. *Every convex polytope $K \subset \mathbb{R}^d$ with a non-empty interior is a convex body i.e. $K \in \mathcal{K}^d$.*

Proof. Suppose $K \subset \mathbb{R}^d$ is a convex polytope with a non-empty interior. By Theorem 4.84 and Definition 4.80, K has a $\mathcal{V}$-representation as a convex hull of a finite set of points (vertices). Given that this vertex set is finite, it is indeed compact in $\mathbb{R}^d$. Since taking the convex hull of a compact set preserves compactness (Brøndsted, 1983, Theorem 2.8), it follows that K is compact. We also have by assumption that K has a non-empty interior. Thus K meets the required criteria from Definition 4.41 to be a convex body i.e. $K \in \mathcal{K}^d$. □

Definition 4.86 (Facets of Polytopes)**.** A *face* of a polytope P is the intersection of P and the boundary hyperplane of a halfspace containing P. A facet of P is an inclusion-wise maximal face distinct from P. Equivalently, a face F of P is a facet if and only if $\dim(F) = \dim(P) - 1$.

In what follows, we will assume that K has a *minimal* $\mathcal{H}$-representation (unique up to ordering and scaling) given by $[\mathbf{A} \mid \mathbf{b}]$. In this representation we assume that K is described by m facets i.e. $\{F_1, \ldots, F_m\}$ for some fixed $m \in \mathbb{N}$. By Definition 4.81 we know that such an m is guaranteed to be finite. Here each facet F_m is reduced to a subset of a single (irredundant) inequality in $[\mathbf{A} \mid \mathbf{b}]$.

Lemma 4.87 (Pulling Triangulation of Polytope)**.** *Given a minimal facet representation of a convex (body) polytope $K \subset \mathcal{K}^d$, assuming a fixed $\mathbf{x}^* \in \operatorname{int}(K)$ and having facets $\{F_1, \ldots, F_m\}, m \in \mathbb{N}$, we can subdivide K via the pulling triangulation construction:*

$$K = \bigcup_{k \in [m]} \operatorname{co}\left(F_k \cup \{\mathbf{x}^*\}\right)$$

Proof of Lemma 4.87. See (De Loera et al., 2010, Section 4.2.2) for details. $\square$

Remark 4.88. In this case K is triangulated into m convex polytopes $P_{F_k,\mathbf{x}^*} := \mathrm{co}\,(F_k \cup \{\mathbf{x}^*\})$ which are in fact all *pyramids* associated with a facet F_k as the base and all having a common *apex vector* $\mathbf{x}^* \in \mathrm{int}\,(K)$, for each $k \in [m]$.

Lemma 4.89 (Shortest distance from vector to hyperplane)**.** *Given a fixed point* $\mathbf{x}^* \in \mathbb{R}^d$*, and a hyperplane* $H_{\mathbf{a},b} := \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}^\top\mathbf{x} + b = 0, \mathbf{a} \neq \mathbf{0}, b \neq 0\right\} \subset \mathbb{R}^d$*, the shortest distance from* $\mathbf{x}^*$ *to* $H_{\mathbf{a},b}$ *is* $\frac{|\mathbf{a}^\top\mathbf{x}^*+\mathbf{b}|}{\|\mathbf{a}\|}$.

Proof of Lemma 4.89. We approach this distance minimization using Lagrange multipliers. Consider the Lagrangian $L(\mathbf{x}, \lambda) := \|\mathbf{x} - \mathbf{x}^*\|^2 + \lambda(\mathbf{a}^\top\mathbf{x} + b)$. We then have $\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) = 2(\mathbf{x} - \mathbf{x}^*) + \lambda\mathbf{a}$. By setting $\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) = \mathbf{0}$ we have

$$\lambda\mathbf{a} = 2(\mathbf{x}^* - \mathbf{x}) \tag{4.110}$$

From Equation (4.110) implies 2 additional equations, first by applying the dot product with $\mathbf{a}$ we have:

$$\lambda = \frac{2\mathbf{a}^\top(\mathbf{x}^* - \mathbf{x})}{\|\mathbf{a}\|^2} \tag{4.111}$$

Second, we can apply the dot product with $(\mathbf{x}^* - \mathbf{x})$ to give:

$$\begin{aligned}
2(\mathbf{x}^* - \mathbf{x})^\top(\mathbf{x}^* - \mathbf{x}) &= \lambda(\mathbf{x}^* - \mathbf{x})^\top\mathbf{a} \\
\implies 2\,\|\mathbf{x}^* - \mathbf{x}\|^2 &= \left(\frac{2\mathbf{a}^\top(\mathbf{x}^* - \mathbf{x})}{\|\mathbf{a}\|^2}\right)(\mathbf{x}^* - \mathbf{x})^\top\mathbf{a} \\
&= 2\frac{\left(\mathbf{a}^\top(\mathbf{x}^* - \mathbf{x})\right)^2}{\|\mathbf{a}\|^2} \\
\implies \|\mathbf{x}^* - \mathbf{x}\| &= \frac{\left|\mathbf{a}^\top\mathbf{x}^* + \mathbf{b}\right|}{\|\mathbf{a}\|}
\end{aligned}$$

As required. $\square$

Lemma 4.90. *Every internal pyramid,* $\mathrm{co}\,(F_k \cup \{\mathbf{0}\}) \subset \mathbb{R}^d$, $\forall\, k \in [m]$*, of a triangulated convex polytope* K *(per Lemma 4.87) has a positive volume i.e.* $\mathrm{vol}_d\,(\mathrm{co}\,(F_k \cup \{\mathbf{0}\})) > 0$ *for each* $k \in [m]$.

Proof of Lemma 4.90. We note that a volume of a pyramid in $\mathbb{R}^d$ is given by $\frac{Ah}{d}$ (see (Mathai, 1999, Theorem 1.2.10)). Here A is the area of the base of the pyramid and h is the distance of the orthogonal projection from the apex of the pyramid to the base. In our case consider an arbitrary pyramid $\operatorname{co}(F_k \cup \{\mathbf{0}\})$ with facet F_k as its base. Now facet F_k contains an interior point in dimension $d-1$ (see (Gallier and Quaintance, 2008, Proposition 4.5(ii)) for details). Indeed by Lemma 4.42 it follows that $\operatorname{vol}_{d-1}(F_k) > 0$. Since F_k lies in the hyperplane defined by the equality $\mathbf{a}_k^\top \mathbf{x} = b_k$. We have that the distance h from $\mathbf{0}$ to F_k is given by $h = \frac{|b_k|}{\|\mathbf{a}_k\|}$ by Lemma 4.89. Since $\mathbf{a}_k \neq \mathbf{0}$ and $b_k \neq 0$ (by Remark 4.82) we have that $h > 0$ in our setting. As such we have that the pyramid $\operatorname{co}(F_k \cup \{\mathbf{0}\})$ has positive volume, for each $k \in [m]$. $\square$

Convex Polytopes - mathematical preliminaries

With this preliminary theory of convex polytopes setup, we now proceed to derive an upper bound in the mean squared error for our projection estimator (per Proposition 4.22)

$$\widehat{\mathbf{v}}_{\text{prj}} \coloneqq \underset{\mathbf{w} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)}{\arg\min} \|\overline{\mathbf{v}} - \mathbf{w}\|_2^2$$

In this case the given original convex body, $K \in \mathcal{K}^d$ is assumed to be a convex polytope with $\operatorname{centroid}(K) = \mathbf{0} \in \operatorname{int}(K)$ by Theorem 4.56. The location-scaled convex polytope (with non-empty interior) is then $\mathbf{v} + \sigma K$. However given the translation invariance of the Lebesgue measure in $\mathbb{R}^d$, in what follows we can recenter our polytope by translation $-\mathbf{v}$. We denote this translated polytope by $P \coloneqq \sigma K$, so that $\operatorname{centroid}(P) = \mathbf{0}$. Our observed sample points $\mathbf{Y}_i$ are thus translated to be $\mathbf{Z}_i \coloneqq \mathbf{Y}_i - \mathbf{v}$, $\forall\, i \in [n]$, where $\mathbf{Z}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[P]$ by Lemma 4.51.

Now we generate a sequence of *similar*[11] polytopes to P by applying an increasing scaling sequence $(1-\alpha_n)_{n=1}^{\infty}$ to P. Here $\alpha_n \in (0,1)$ for each $n \in \mathbb{N}$, such that $(\alpha_n) \xrightarrow{n \to \infty} 0$, is a decreasing scaling sequence to be determined later. This sequence of $(1-\alpha_n)$-scaled polytopes is denoted by $\mathcal{P} \coloneqq ((1-\alpha_n)P)_{n=1}^{\infty}$. We further observe that for each $n \in \mathbb{N}$, that $(1-\alpha_n)P \subset P$ (by Lemma 4.54), and that $\mathcal{P}$ is in fact a sequence of monotonically increasing nested polytopes in P.

In what follows, per Definition 4.81 we will assume that P has a *minimal* $\mathcal{H}$-representation (unique up to reordering and scaling) given by $[\mathbf{A} \mid \mathbf{b}]$. In this representation we further assume that P is described by m facets i.e. $\{F_1, \ldots, F_m\}$ for some fixed $m \in \mathbb{N}$. By Definition 4.81 we know that such an m is guaranteed to be finite. Here each facet F_m is reduced to a subset of a single (irredundant) inequality in $[\mathbf{A} \mid \mathbf{b}]$.

[11] Here *similar* means that we have a dilated (similar) version of the polytope P, scaled by the factor $1-\alpha_n$. Note that since $\mathbf{0} \in \operatorname{int}(P)$, this follows per Lemma 4.54 and Remark 4.55.

By Lemma 4.87 we observe that we can triangulate P as a finite union of m *pyramids*, as $P = \bigcup_{k \in [m]} \mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)$. We denote the pyramid with facet F_k as its base (and $\mathbf{0}$ as its apex) by $P_{F_k,\mathbf{0}} \coloneqq \mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)$. For each such pyramid $P_{F_k,\mathbf{0}}$ we *induce the boundary facet shell* S_k as defined by $S_k \coloneqq P_{F_k,\mathbf{0}} \setminus (1-\alpha_n)P_{F_k,\mathbf{0}}$. For clarity these decompositions are geometrically illustrated for a sample polygon $P \subset \mathbb{R}^2$ in Figure 4.E.1. With this constructive set up complete, we are ready to state and prove Lemma 4.91. This demonstrates that if we sample points uniformly i.i.d. on P, we will always sample from S_k, i.e., a single boundary facet shell of P, with positive probability depending on α_n.

Lemma 4.91. *Suppose that we independently uniformly sample points from P, with $S_k \coloneqq P_{F_k,\mathbf{0}} \setminus (1-\alpha_n)P_{F_k,\mathbf{0}}$, denoting the k^{th} boundary facet shell of P. Then the probability of sampling from S_k, i.e., $\mathbb{P}_{S_k}$, is strictly bounded away from $\{0, 1\}$. That is, for each $k \in [m]$ we have:*

$$\mathbb{P}_{S_k} \coloneqq \mathbb{P}\left(\textit{sampling from } S_k, \textit{after sampling uniformly from } P\right) \in (0,1) \tag{4.112}$$

Proof of Lemma 4.91. Let $\mathbb{P}_{S_k}$ be defined per Equation (4.112). We then observe that

$$\begin{aligned}
\mathbb{P}_{S_k} &= \frac{\mathrm{vol}_d\left(S_k\right)}{\mathrm{vol}_d\left(P\right)} && \text{(since we are uniformly i.i.d. sampling from } P.\text{)} \\
&= \frac{\mathrm{vol}_d\left(P_{F_k,\mathbf{0}}\right) - \mathrm{vol}_d\left((1-\alpha_n)P_{F_k,\mathbf{0}}\right)}{\mathrm{vol}_d\left(P\right)} \\
& && \text{(since } S_k \coloneqq P_{F_k,\mathbf{0}} \setminus (1-\alpha_n)P_{F_k,\mathbf{0}}.\text{)} \\
&= \frac{\mathrm{vol}_d\left(P_{F_k,\mathbf{0}}\right) - (1-\alpha_n)^d\, \mathrm{vol}_d\left(P_{F_k,\mathbf{0}}\right)}{\mathrm{vol}_d\left(P\right)} \\
&= \frac{\mathrm{vol}_d\left(\mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)\right) - (1-\alpha_n)^d\, \mathrm{vol}_d\left(\mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)\right)}{\mathrm{vol}_d\left(P\right)} \\
& && \text{(since } P_{F_k,\mathbf{0}} \coloneqq \mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right).\text{)} \\
&= \left(\frac{\mathrm{vol}_d\left(\mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)\right)}{\mathrm{vol}_d\left(P\right)}\right)\left(1-(1-\alpha_n)^d\right)
\end{aligned}$$

We then define the following constants:

$$c_{\min} \coloneqq \min_{k \in [m]} \left\{\frac{\mathrm{vol}_d\left(\mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)\right)}{\mathrm{vol}_d\left(P\right)}\right\} \eqqcolon \min_{k \in [m]} \left\{\frac{\mathrm{vol}_d\left(P_{F_k,\mathbf{0}}\right)}{\mathrm{vol}_d\left(P\right)}\right\} \tag{4.113}$$

$$c_{\max} \coloneqq \max_{k \in [m]} \left\{\frac{\mathrm{vol}_d\left(\mathrm{co}\left(F_k \cup \{\mathbf{0}\}\right)\right)}{\mathrm{vol}_d\left(P\right)}\right\} \eqqcolon \max_{k \in [m]} \left\{\frac{\mathrm{vol}_d\left(P_{F_k,\mathbf{0}}\right)}{\mathrm{vol}_d\left(P\right)}\right\} \tag{4.114}$$

By construction, recall that $P = \bigcup_{k\in[m]} \operatorname{co}(F_k \cup \{\mathbf{0}\})$, i.e., the P_{F_k} form a *finite* subdivision of P. As such the minimum and maximum values are attained as defined by $c_{\min}$ and $c_{\max}$, respectively. Moreover, by Lemma 4.90 it holds that $\operatorname{co}(F_k \cup \{\mathbf{0}\}) > 0$, for each $k \in [m]$. It follows that that $c_{\min}$ and $c_{\max}$ are both strictly positive. Given that $\alpha_n \in (0,1)$ for each $n \in \mathbb{N}$, then for each $k \in [m]$, the probabilities $\mathbb{P}_{S_k}$ satisfy:

$$0 < c_{\min}\left(1-(1-\alpha_n)^d\right) \leq \mathbb{P}_{S_k} \leq c_{\max}\left(1-(1-\alpha_n)^d\right) < 1, \tag{4.115}$$

And thus the $\mathbb{P}_{S_k}$ are always bounded away from $\{0,1\}$ for each $k \in [m]$, as required. □

Remark 4.92. Note that $c_{\min} \leq \frac{1}{m}$ in Lemma 4.91. To see this, first note that by construction that $P = \bigcup_{k\in[m]} \operatorname{co}(F_k \cup \{\mathbf{0}\}) \implies \operatorname{vol}_d(P) = \sum_{k=1}^{m} \operatorname{vol}_d(P_{F_k,\mathbf{0}})$. Then observe that:

$$c_{\min} =: \min_{k\in[m]}\left\{\frac{\operatorname{vol}_d(P_{F_k,\mathbf{0}})}{\operatorname{vol}_d(P)}\right\} \leq \frac{1}{m}\sum_{k=1}^{m}\frac{\operatorname{vol}_d(P_{F_k,\mathbf{0}})}{\operatorname{vol}_d(P)} = \frac{1}{m}\left(\frac{\operatorname{vol}_d(P)}{\operatorname{vol}_d(P)}\right) = \frac{1}{m}. \tag{4.116}$$

As required.

We have thus established that if we sample points uniformly i.i.d. on P, we will always sample from S_k, i.e., a single boundary facet shell of P, with positive probability depending on α_n. Using this result, we now turn our attention to computing the probability of uniformly sampling *at least* one point from *each* facet shell of P. This probability depends on appropriately setting the values for the decreasing sequence $(\alpha_n)_{n=1}^{\infty}$, as defined earlier. This is summarized in Lemma 4.93.

Lemma 4.93. *After uniformly sampling n points from P, let T_k denote the number of these points that are sampled within each boundary facet shell S_k, for each $k \in [m]$. Then the probability of uniformly sampling at least one point from each facet shell S_k of P, is at least $1-\gamma \in (0,1)$ if $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$. Or alternatively, we can write:*

$$\mathbb{P}\left(\bigcap_{k=1}^{m}\{T_k \geq 1\}\right) \geq 1-\gamma, \text{ if } \alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n} \tag{4.117}$$

Proof of Lemma 4.93. Firstly, per Lemma 4.91 we observe that the probabilities $\mathbb{P}_{S_k} \in (0,1)$, for each $k \in [m], n \in \mathbb{N}$. Furthermore we have that

$1-(1-\alpha_n)^d \geq 1-(1-\alpha_n) = \alpha_n$ for each $\alpha_n \in (0,1)$ and $d \geq 1$. From Equation (4.115) this implies that $\mathbb{P}_{S_k} \geq c_{\min}\alpha_n$ for each $k \in [m], n \in \mathbb{N}$. We now observe that sampling each $\mathbf{Z}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[P]$ induces a multinomial distribution on the $(m+1)$ "bins", with the first m bins being the m boundary facet shells of P, i.e., $(S_1, \ldots, S_m)$ with corresponding sampling probabilities $(\mathbb{P}_{S_1}, \ldots, \mathbb{P}_{S_m})$. The final $(m+1)^{\text{th}}$ bin then being the inner scaled polytope $(1-\alpha_n)P$ denoted by $S_{m+1} \coloneqq P \setminus \bigcup_{k\in[m]} S_k$ with the complementary sampling probability $1-\sum_{k=1}^{m}\mathbb{P}_{S_k}$. That is for each $i \in [n]$, $\mathbf{Z}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[P]$ induces an i.i.d. multinomial sample $(T_1, \ldots, T_m)^\top \sim \mathsf{Multi}[n, (\mathbb{P}_{S_1}, \ldots, \mathbb{P}_{S_m}, 1-\sum_{k=1}^{m}\mathbb{P}_{S_k})^\top]$. From this we observe that:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=1}^{m}\{T_k = 0\}\right) &\leq \sum_{k=1}^{m}\mathbb{P}\left(T_k = 0\right) && \text{(by union bound.)}\\
&= \sum_{k=1}^{m}(1-\mathbb{P}_{S_k})^n && \text{(since } \mathbb{P}\left(T_k=0\right) = (1-\mathbb{P}_{S_k})^n.\text{)}\\
&\leq m(1-\min_{k\in[m]}\mathbb{P}_{S_k})^n\\
&\leq m(1-c_{\min}\alpha_n)^n\\
&\qquad \text{(since } \mathbb{P}_{S_k} \geq c_{\min}\alpha_n \text{ for each } k\in[m], n\in\mathbb{N})\\
&\leq m\left(\exp\left(-c_{\min}\alpha_n n\right)\right)\\
&\qquad \text{(since } \mathbb{E}\left((-x\right) \geq 1-x), \text{ for each } x \in \mathbb{R}.)
\end{aligned}$$

It then follows that:

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k=1}^{m}\{T_k \geq 1\}\right) &= 1-\mathbb{P}\left(\bigcup_{k=1}^{m}\{T_k = 0\}\right)\\
&\geq 1-m\left(\exp\left(-c_{\min}\alpha_n n\right)\right)\\
&=: 1-\gamma \in (0,1)\\
\iff \alpha_n &= \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min}n},
\end{aligned}$$

So indeed $\mathbb{P}\left(\bigcap_{k=1}^{m}\{T_k \geq 1\}\right) \geq 1-\gamma$, if $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min}n}$, as required. □

Since we are sampling uniformly and from finitely many such shells, we will eventually sample at least one point from each of them. We can now construct a new sequence of polytopes $\mathcal{P}' \coloneqq (-\alpha_n P)_{n=1}^{\infty}$. Now consider the ray starting at $\mathbf{0}$ and passing through $\mathbf{Z}_i$. Let F_i be the corresponding facet to intersect this ray, where $i \in [m]$. We observe that each facet $F_i \subseteq \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_i^T \mathbf{x} \leq b_i \right\}$, for each $i \in [m]$ by the $\mathcal{H}$-representation of P. For every such inequality we simply rescale and reverse it by $-\alpha_n$ i.e. $\mathbf{a}_i^T \mathbf{x} \geq -\alpha_n b_i$. We claim that this results in the new minimal polytope $[-\mathbf{A} \mid \alpha_n \mathbf{b}]$, and formalize it in Proposition 4.94.

Proposition 4.94. *Let $P \coloneqq [\mathbf{A} \mid \mathbf{b}]$ be the minimal $\mathcal{H}$-representation of P. Then $-\alpha_n P = [-\mathbf{A} \mid \alpha_n \mathbf{b}]$, where $-\alpha_n P \coloneqq \{-\alpha_n \mathbf{z} \mid \mathbf{z} \in P\}$. Moreover $-\alpha_n P$ is a convex polytope.*

Proof of Proposition 4.94. To see this firstly note that $P \coloneqq [\mathbf{A} \mid \mathbf{b}]$ is the minimal $\mathcal{H}$-representation of P. Now, for each $n \in \mathbb{N}$, with $\alpha_n \in (0, 1)$, it follows directly from Lemma 4.83 that $-\alpha_n P = [-\mathbf{A} \mid \alpha_n \mathbf{b}]$, and moreover that it is indeed a convex polytope, as required. $\square$

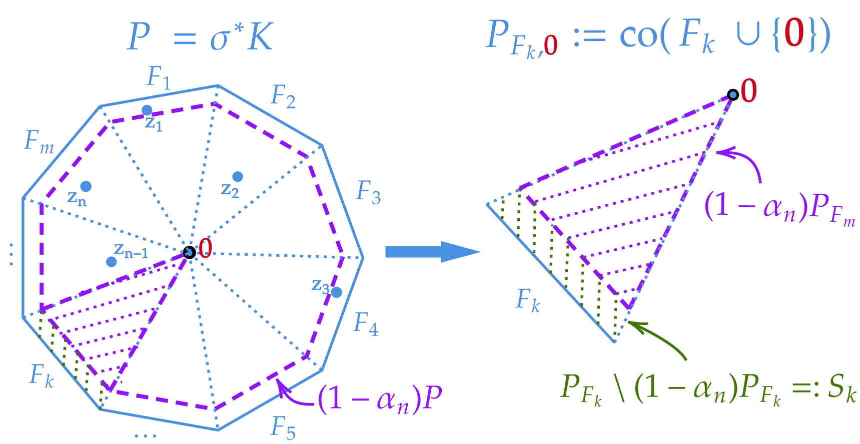

Figure 4.E.1: (Left) Triangulation of P. (Right) Decomposition of pyramid $P_{F_k,\mathbf{0}}$

Proposition 4.95 (Critical set is a subset of $-\alpha_n P + \mathbf{v}$)**.** *We have that for every $n \in \mathbb{N}$ that the critical set, $\bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$ is a subset of the dilated and shifted enveloping polytope $-\alpha_n P + \mathbf{v}$, for each $n \in \mathbb{N}$, with high probability. That is, for each $n \in \mathbb{N}$ we have:*

$$\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \subseteq -\alpha_n P + \mathbf{v}, \tag{4.118}$$

with probability at least $1 - \gamma$, for $\gamma \in (0, 1)$, if $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$.

Proof of Proposition 4.95. Let $P \coloneqq [\mathbf{A} \mid \mathbf{b}]$ be the minimal $\mathcal{H}$-representation of P. We then have:

$$P = \bigcap_{k=1}^{m} \left\{ \mathbf{z} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{z} \leq \mathbf{b}_k \right\}. \tag{4.119}$$

Let $\mathbf{Z}_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[P]$, for each $i \in [n]$. Further let $\gamma \in (0, 1)$, and $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$. Then from Lemma 4.93, we have with probability at least $1 - \gamma$, that at least one point is sampled from each boundary facet shell $S_k \coloneqq P_{F_k,\mathbf{0}} \setminus (1-\alpha_n)P_{F_k,\mathbf{0}}$. Let us denote $\mathbf{Z}_k$ to be such a sampled point from S_k, for each $k \in [m]$. We then have that $\mathbf{Z}_k$ satisfies the following inequalities using the $\mathcal{H}$-representation of P:

$$(1 - \alpha_n)\mathbf{b}_k \leq \mathbf{a}_k^\top \mathbf{Z}_k \leq \mathbf{b}_k. \tag{4.120}$$

Now note that $\bigcap_{i=1}^{n} (\mathbf{z}_i - \sigma K) \subseteq \bigcap_{k=1}^{m} (\mathbf{Z}_k - \sigma K)$. It suffices to show that $\bigcap_{k=1}^{m} (\mathbf{Z}_k - \sigma K) \coloneqq \bigcap_{k=1}^{m} (\mathbf{Z}_k - P) \subseteq -\alpha_n P$, or alternatively that $\mathbf{Z}_k - P \subseteq -\alpha_n P$, for each $k \in [m]$. For each $k \in [m]$, per Equation (4.119) we have the equivalence

$$P \subset \left\{ \mathbf{z} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{z} \leq \mathbf{b}_k \right\} \iff -P \subset \left\{ \mathbf{z} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{z} \geq -\mathbf{b}_k \right\}. \tag{4.121}$$

We then proceed as follows, for each $k \in [m]$.

$$\begin{aligned}
\mathbf{Z}_k - P &\coloneqq \{\mathbf{Z}_k\} + (-P) && \text{(by definition.)} \\
&\subset \{\mathbf{Z}_k\} + \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{x} \geq -\mathbf{b}_k \right\} && \text{(using Equation (4.121))} \\
&= \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top (\mathbf{x} - \mathbf{Z}_k) \geq -\mathbf{b}_k \right\} && (4.122) \\
&= \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{x} \geq \mathbf{a}_k^\top \mathbf{Z}_k - \mathbf{b}_k \right\} \\
&\subset \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{x} \geq (1 - \alpha_n)\mathbf{b}_k - \mathbf{b}_k \right\} && \text{(using Equation (4.120))} \\
&= \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \mathbf{a}_k^\top \mathbf{x} \geq -\alpha_n \mathbf{b}_k \right\} \\
&= -\alpha_n P && (4.123)
\end{aligned}$$

We have thus shown that, for each $n \in \mathbb{N}$:

$$\bigcap_{i=1}^{n} (\mathbf{Z}_i - \sigma K) \subseteq -\alpha_n P \tag{4.124}$$

Now in order to show Equation (4.118), we observe that:

$$\begin{aligned}
& \mathbf{y} \in \bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K) \\
\iff & \mathbf{y} \in \mathbf{Y}_i - P, \text{ for each } i \in [n] && (\text{since } P \coloneqq \sigma K.) \\
\iff & \mathbf{y} \in (\mathbf{Z}_i + \mathbf{v}) - P, \text{ for each } i \in [n] && (\text{since } \mathbf{Y}_i \overset{a.s.}{=} \mathbf{Z}_i + \mathbf{v}.) \\
\iff & \mathbf{y} \in (\underbrace{\mathbf{Z}_i - P}_{\in -\alpha_n P}) + \mathbf{v}, \text{ for each } i \in [n] \\
\iff & \mathbf{y} \in -\alpha_n P + \mathbf{v}
\end{aligned}$$

As required. □

Final proof of Theorem 4.28

Theorem 4.28 (Consistency of location estimators in the critical set, $K \in \mathcal{K}^d$ polytope.)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. Further assume that $K \in \mathcal{K}^d$ is a convex polytope with m facets. Let $\widehat{\mathbf{v}}_{\text{cri}}$ denote any location estimator lying in the critical set, i.e. $\widehat{\mathbf{v}}_{\text{cri}} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$ a.s. We then have that $\widehat{\mathbf{v}}_{\text{cri}}$ satisfies $\|\widehat{\mathbf{v}}_{\text{cri}} - \mathbf{v}\|_2 \leq \sigma\alpha_n(\operatorname{diam}(K))$, with probability at least $1 - \gamma \in (0,1)$, if $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$. Here $c_{\min} \leq \frac{1}{m}$, is a constant that depends on the convex body K.*

Proof. Using the facts from Lemma 4.63, and the fact that both $\mathbf{v}, \widehat{\mathbf{v}}_{\text{prj}} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$ a.s., the following holds with probability γ, and $\alpha_n = \frac{\log\left(\frac{m}{\gamma}\right)}{c_{\min} n}$:

$$\begin{aligned}
\|\widehat{\mathbf{v}}_{\text{prj}} - \mathbf{v}\|_2 &\leq \operatorname{diam}\left(\bigcap_{i=1}^{n} (\mathbf{Y}_i - \sigma K)\right) && (\text{using Equation (4.61)}) \\
&\leq \operatorname{diam}(-\alpha_n P + \mathbf{v}) \\
& && (\text{using Proposition 4.95, and Equation (4.62)}) \\
&= |\alpha_n| (\operatorname{diam}(P)) && (\text{using Equations (4.62) and (4.64)}) \\
&= \sigma\alpha_n(\operatorname{diam}(K)) && (\text{using Equation (4.62), and } \sigma, \alpha_n > 0.)
\end{aligned}$$

So that $\|\widehat{\mathbf{v}}_{\text{prj}} - \mathbf{v}\|_2 \to 0$ at rate $\sigma\alpha_n$ as defined previously. □

4.E.2 Proof of Theorem 4.29

Mathematical preliminaries

As noted, Theorem 4.29 provides upper bounds for the risk of our projection estimator in the case where $K \in \mathcal{K}^d$ is a general convex body. In order to keep our proof of Theorem 4.28 largely self-contained we first provide the minimal necessary convex body theory in below, before describing our proof in Section 4.E.2. We now observe the following simple proposition.

Proposition 4.96 (A convex body is properly nested between 2 Euclidean balls)**.** *For every convex body $P \in \mathcal{K}^d$, with* centroid $(P) = \mathbf{0}$*, there exist* $0 < r < R$ *such that:*

$$B_2^d(\mathbf{0}, r) \subset P \subset B_2^d(\mathbf{0}, R) \tag{4.125}$$

Proof of Proposition 4.96. Since it follows that centroid $(P) = \mathbf{0} \in \operatorname{int}(P)$ (by Theorem 4.56), we know there exists some $r > 0$ such that $B_2^d(\mathbf{0}, r) \subset P$. Since P is a convex body, it is bounded and there also exists an $R > 0$ such that $P \subset B_2^d(\mathbf{0}, R)$. By Jung's Theorem (see (Jung, 1901, 1910), and (Leonard and Lewis, 2016, Theorem 4.3.36) for a proof), such a radius R is guaranteed by fixing $R > \operatorname{diam}(P)\sqrt{\frac{d}{2(d+1)}}$. In sum we have $B_2^d(\mathbf{0}, r) \subset P \subset B_2^d(\mathbf{0}, R)$, where all subset inclusions are proper, as required. ☐

A simple representation of this general convex body setting is shown for the case where $P \subset \mathbb{R}^2$ is an ellipse in Figure 4.E.2.

Remark 4.97. e note that in the proof and discussion that follows we do not impose any additional conditions on the convex body P (e.g. symmetric, or smoothness) other than centroid $(P) = \mathbf{0}$, and satisfying Definition 4.41.

Uniform sampling from closed Euclidean balls

We start with a simple question, namely how do we sample uniformly within the d-dimensional Euclidean closed unit ball (i.e., including boundary and interior), $\overline{B}_2^d(\mathbf{0}, 1) \subset \mathbb{R}^d$? This can be done by first generating a direction by uniformly sampling on the surface of the unit ball. One procedure to generate this direction unit vector is by sampling an isotropic Gaussian vector $\mathbf{g} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and normalizing it i.e. $\frac{\mathbf{g}}{\|\mathbf{g}\|_2}$ per (Muller, 1959).

Then one can *independently* sample a radius $X \in [0, 1]$ with cumulative density function $F_X(x) = x^d$. The latter can be done by first uniformly sampling $U \sim \mathsf{Unif}[0, 1]$, and then taking $U^{\frac{1}{d}}$ as the radius. Similarly, if we want to uniformly sample from the radius-R ball, i.e., $\overline{B}_2^d(\mathbf{0}, R) \subset \mathbb{R}^d$, we again generate a direction uniformly on the surface of the and then we generate a radius $X \in [0, R]$ independently with cumulative density function $F_X(x) = \frac{x^d}{R^d}$. For more details on the above uniform sampling procedures we refer the reader to (Harman and Lacko, 2010), (Fishman, 1996, Section 3.29), and (Blum et al., 2020, Section 2.5).

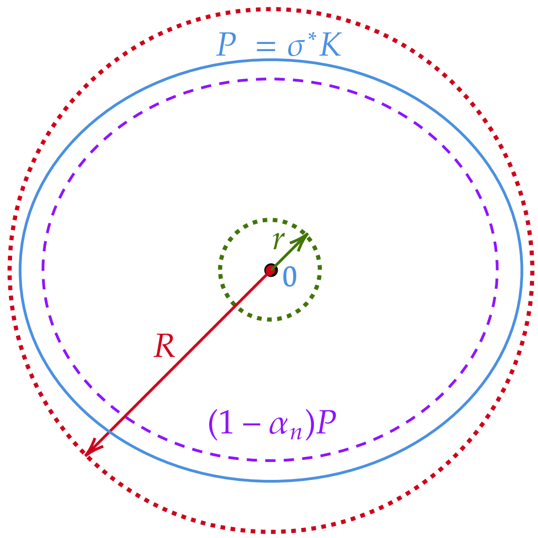

Figure 4.E.2: Geometric view for a general (elliptical) convex body $P = \sigma K \subset \mathbb{R}^2$

Uniform sampling from convex bodies

Next, how can we perform uniform sampling in the convex body P? We know by Proposition 4.96 that there exist $0 < r < R$ such that $B_2^d(\mathbf{0}, r) \subset P \subset B_2^d(\mathbf{0}, R)$. A natural approach is to use acceptance-rejection sampling. Using the methodology described in Section 4.E.2, we can sample a vector in $B_2^d(\mathbf{0}, R)$. If the vector happens to be in $P \subset B_2^d(\mathbf{0}, R)$ we accept it, and we reject it if is not not in P. Next, we generate a point uniformly distributed in the boundary shell set of P i.e. $S_{P,\alpha_n} := P \setminus (1 - \alpha_n)P$. Here we assume that $\alpha_n \in (0, 1)$ for each $n \in \mathbb{N}$. We will later select $(\alpha_n)_{n=1}^{\infty}$ as a decreasing sequence such that $\alpha_n = \frac{\kappa_n}{n}$, where $\kappa_n \to \infty$ and $\kappa_n = o(n)$. Once again, we can use acceptance-rejection sampling. We proceed by sampling a vector uniformly over $B_2^d(\mathbf{0}, R)$. If the vector belongs to the boundary shell set S_{P,α_n} we accept it, and reject it otherwise. We let y denote the length of the segment along the unit-norm random direction $\frac{\mathbf{g}}{\|\mathbf{g}\|_2}$ which connects $\mathbf{0}$ with the boundary of P. That is y is defined such that $y\frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \partial(P)$. Thus the length y is lower bounded by r since $B_2^d(\mathbf{0}, r) \subset P$ is a proper inclusion (from Proposition 4.96). Now, since the radius has cdf $\frac{x^d}{R^d}$, the probability that we will accept a point belonging to the boundary shell set S_{P,α_n} is given by:

$$\begin{aligned}\frac{y^d - (1-\alpha_n)^d y^d}{R^d} &= \frac{y^d}{R^d}\left(1-(1-\alpha_n)^d\right) \\ &\geq \frac{y^d}{R^d}(\alpha_n d) && \text{(By reverse Bernoulli inequality)} \\ &> \frac{r^d}{R^d}(\alpha_n d) && \text{(since } y > r \text{ for all such } y)\end{aligned}$$

Hence, to find points that belong to the boundary shell set S_{P,α_n}, we can generate uniform vectors uniformly in all directions, and then accept with probability at least $\frac{r^d \alpha_n d}{R^d}$. Hence by potentially discarding with a certain probability some of the accepted points, we may assume that the probability of accepting a direction equals precisely to $\frac{r^d \alpha_n d}{R^d}$ and is independent of where the direction has landed. Now the n i.i.d. sampling directions follow a $\mathsf{Binom}\left(n, \frac{r^d \alpha_n d}{R^d}\right)$ distribution i.e. binomial distribution with n trials and success probability $\frac{r^d \alpha_n d}{R^d}$.

This is so since each independently sampled direction has *at least* a probability of $\frac{r^d \alpha_n d}{R^d}$ of being accepted. The mean of this distribution is $\frac{r^d}{2R^d}(n\alpha_n d)$. By the Chernoff bound on the binomial we have $\mathbb{P}\,(\text{number of accepted points } \leq k) \leq \exp\left(-\frac{1}{2p}\left(\frac{(np-k)^2}{n}\right)\right) = \exp(-\frac{np}{8})$, where $k = \frac{r^d}{2R^d}(n\alpha_n d)$, $p = \frac{r^d}{R^d}(\alpha_n d)$. By the choice of α_n this will happen with high probability. Let us denote the sequence of m sampled (unit) direction vectors which are accepted by the above sampling procedure by $\left(\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}\right)_{i=1}^m$, with $m \leq n$. If $\alpha_n = \kappa_n/n$ for any slowly diverging sequence with n κ_n, we have that $m \geq \frac{r^d}{2R^d} d\kappa_n$ with probability at least $1 - \exp(-dr^d\kappa_n/(8R^d))$.

Final proof of Theorem 4.29

With this preliminary theory of convex polytopes setup, we now proceed to derive an upper bound in the mean squared error for our projection estimator projection estimator in the case where $K \in \mathcal{K}^d$ is a general convex body.

Theorem 4.29 (Consistency of location estimators in the critical set, $K \in \mathcal{K}^d$.)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4. Let $\widehat{\mathbf{v}}_{\text{cri}}$ denote any location estimator lying in the critical set, i.e. $\widehat{\mathbf{v}}_{\text{cri}} \in \bigcap_{i=1}^n (\mathbf{Y}_i - \sigma K)$* a.s.. *We then have that $\widehat{\mathbf{v}}_{\text{cri}}$ satisfies $\|\widehat{\mathbf{v}}_{\text{cri}} - \mathbf{v}\|_2 \leq \frac{\sigma C_1 \kappa_n}{n}$, with probability at least $1 - 2\exp(-C_2\kappa_n/\operatorname{polylog}_d(\kappa_n))$, where $C_1 := C_1(d,K)$ and $C_2 := C_2(d,K)$ are constants which depend on the dimension d and the convex body K, κ_n is any slowly diverging sequence with n, and $\operatorname{polylog}_d(\kappa_n)$ is a poly-logarithmic factor of κ_n which also depends on the dimension d.*

Given our sampled sequence of normalized vectors $\left(\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}\right)_{i=1}^m$ a corresponding sequence $\left(c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}\right)_{i=1}^m$ of scaled boundary vectors on $P \coloneqq \sigma K$. Here each *boundary scaling factor* $c_i > 0$ is defined to ensure that $c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2} \in \partial(P)$, for each $i \in [m]$. Indeed, by Proposition 4.96 we additionally have that $0 < r < c_i < R$ for each $i \in [m]$. On the other hand let $\left(c'_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}\right)_{i=1}^m$ be the sequence of points that are actually sampled in the shell S_{P,α_n}. We clearly have $(1-\alpha_n)c_i \leq c'_i \leq c_i$. Now, by the supporting hyperplane theorem on closed convex sets (see (Boyd and Vandenberghe, 2004, Section 2.5.2)), let $\mathbf{x}_i$ be a unit vector such that $\langle c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle = \max_{\mathbf{z}\in P}\langle \mathbf{z}, \mathbf{x}_i\rangle$ (which implies $\langle c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle \geq 0$). It then follows that $\langle c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle \geq \langle c_j\frac{\mathbf{g}_j}{\|\mathbf{g}_j\|_2}, \mathbf{x}_i\rangle$ for all $j \in [m]$. Hence $\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle \geq \frac{r}{R}\langle \frac{\mathbf{g}_j}{\|\mathbf{g}_j\|_2}, \mathbf{x}_i\rangle$.

Consider the (potentially open) polytope, M given by the inequalities $\langle \mathbf{y}, \mathbf{x}_i\rangle \leq \langle c_i\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle$ for each $i \in [m]$. Clearly this polytope contains the set P. Observe that the critical set is a subset of the following set

$$\bigcap_{i=1}^m \left(c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|} - P\right) \subseteq \bigcap_{i=1}^m \left(c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|} - M\right).$$

We will now argue that $\bigcap_{i=1}^m \left(c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|} - M\right) \subseteq -\alpha_n M$. Consider a point $\mathbf{y}' \in \bigcap_{i=1}^m \left(c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|} - M\right)$. We can write $\mathbf{y}' = c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|} - \mathbf{y}_i$ for any i where $\mathbf{y}_i \in M$ for all $i \in [m]$. Hence

$$\langle \mathbf{y}', \mathbf{x}_i\rangle = \langle c'_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|}, \mathbf{x}_i\rangle - \langle \mathbf{y}_i, \mathbf{x}_i\rangle \geq \langle (c'_i - c_i)\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|}, \mathbf{x}_i\rangle \geq -\alpha_n\langle c_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|}, \mathbf{x}_i\rangle,$$

where we used the fact that $c'_i \geq (1-\alpha_n)c_i$ and $\langle c_i\frac{\mathbf{g_i}}{\|\mathbf{g_i}\|}, \mathbf{x}_i\rangle \geq 0$. This shows that the critical set is a subset of $-\alpha_n M$. The remainder of the proof is dedicated to showing that M is a bounded polytope. Once this is established the estimation rate is controlled by $\alpha_n \operatorname{diam}(M)$.
In what follows we are mainly concerned with the maximum of any unit vector $\mathbf{y}$, with a vector sampled from the unit sphere $\mathbf{u}_i = \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}$ i.e. $\inf_{\mathbf{y}\in\mathbb{S}^{d-1}}\max_i\langle \mathbf{y}, \mathbf{u}_i\rangle$ Now we have

$$\frac{R}{r}\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i\rangle \geq \inf_{\mathbf{y}\in\mathbb{S}^{d-1}}\max_{i\in[m]}\langle \mathbf{y}, \mathbf{u}_i\rangle \geq f(m),$$

where $f(m)$ is a high probability lower bound on $\inf_{\mathbf{y}\in\mathbb{S}^{d-1}}\max_{i\in[m]}\langle \mathbf{y}, \mathbf{u}_i\rangle$ which will be established below (see Lemma 4.98 for a precise definition of

$f(m))$. We conclude that $\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i \rangle \geq \frac{r}{R} f(m)$ for each $i \in [m]$. Now take any unit vector $\mathbf{y}$. We now bound

$$\max_i \langle \mathbf{y}, \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2} \rangle \geq \inf_{\mathbf{y} \in \mathbb{S}^{d-1}} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle \geq f(m)$$

Hence there exists $i^* \in [m]$ such that $\langle \mathbf{y}, \frac{\mathbf{g}_{i^*}}{\|\mathbf{g}_{i^*}\|_2} \rangle \geq f(m)$. Consider now

$$\begin{aligned}
\sqrt{2 - 2\langle \mathbf{y}, \mathbf{x}_{i^*} \rangle} &= \|\mathbf{y} - \mathbf{x}_{i^*}\|_2 && (\text{since } \|\mathbf{y}\|_2 = \|\mathbf{x}_{i^*}\|_2 = 1) \\
&\leq \left\| \mathbf{y} - \frac{\mathbf{g}_{i^*}}{\|\mathbf{g}_{i^*}\|_2} \right\| + \left\| \frac{\mathbf{g}_{i^*}}{\|\mathbf{g}_{i^*}\|_2} - \mathbf{x}_{i^*} \right\| \\
& && (\text{by the triangle inequality}) \\
&\leq \sqrt{2 - 2f(m)} + \sqrt{2 - 2\frac{r}{R} f(m)}
\end{aligned}$$

By squaring and expanding both sides, we then conclude that

$$\langle \mathbf{y}, \mathbf{x}_{i^*} \rangle \geq f(m) - 1 + \frac{r}{R} f(m) - \sqrt{2 - 2f(m)} \sqrt{2 - 2\frac{r}{R} f(m)}.$$

As m is substantially large the above can be made bigger than a constant (since $f(m)$ is very close to 1 for large m see the result below). Hence if for $\mathbf{y} \in M$ we have $\|\mathbf{y}\|_2 \geq C$ for a large enough C we will have that $\langle \mathbf{y}, \mathbf{x}_{i^*} \rangle$ will exceed R which is a strict upper bound on $\langle c_{i^*} \frac{\mathbf{g}_{i^*}}{\|\mathbf{g}_{i^*}\|_2}, \mathbf{x}_{i^*} \rangle$, a contradiction. This means that the polytope M will be bounded. Recall that the polytope here is given by the inequalities $\langle \mathbf{y}, \mathbf{x}_i \rangle \leq \langle c_i \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|_2}, \mathbf{x}_i \rangle$ for each $i \in [m]$).

As noted previously, we require control on the quantity $\inf_{\mathbf{y} \in \mathbb{S}^{d-1}} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle$. To this end we have the following:

Lemma 4.98. *Let* $(\mathbf{u}_i)_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[\mathbb{S}^{d-1}]$. *We then have*

$$\mathbb{P}\left(\inf_{\mathbf{y} \in \mathbb{S}^{d-1}} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle \geq f(m) \right) \geq 1 - \exp(-mC(d)/(2 \operatorname{polylog}_d(m))),$$

where $f(m) = 1 - \frac{1}{\sqrt{\log m}} - 3\exp(-mC(d)/(2d \operatorname{polylog}_d(m)))$, $C(d) = \frac{\Gamma((d-1)/2)}{\sqrt{\pi}\Gamma(d/2)}$, *and the precise expression for the polylog factor (which also depends on d) may be gleaned from the proof.*

Proof of Lemma 4.98. Make an ε-net on the unit sphere $\mathbb{S}^{d-1}$, denoted by $\mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)$. It is known that $\left|\mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)\right| \leq (1 + 2/\varepsilon)^d$ (see (Vershynin, 2018, Corollary 4.2.13) for a proof). It follows that $\left|\mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)\right| \sim (3/\varepsilon)^d$, when

$\varepsilon \in (0, 1)$. Take any $\mathbf{y} \in \mathbb{S}^{d-1}$. By construction, there exists a $\mathbf{y}^* \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)$ in the net such that $\|\mathbf{y} - \mathbf{y}^*\|_2 \leq \varepsilon$. It then follows that:

$$\begin{aligned} |\langle \mathbf{y} - \mathbf{y}^*, \mathbf{u}_i^* \rangle| &\leq \|\mathbf{u}_i^*\|_2 \|\mathbf{y} - \mathbf{y}^*\|_2 && \text{(by Cauchy-Schwartz inequality)} \\ &= \|\mathbf{y} - \mathbf{y}^*\|_2 && \text{(since } \|\mathbf{u}_i^*\|_2 = 1) \\ &\leq \varepsilon \\ \implies \langle \mathbf{y} - \mathbf{y}^*, \mathbf{u}_i^* \rangle &\geq -\varepsilon \end{aligned}$$

Hence if $i^* \coloneqq \arg\max_{i \in [m]} \langle \mathbf{y}^*, \mathbf{u}_i \rangle$ it then follows that:

$$\begin{aligned} \langle \mathbf{y}, \mathbf{u}_{i^*} \rangle &= \langle \mathbf{y}^*, \mathbf{u}_{i^*} \rangle + \langle \mathbf{y} - \mathbf{y}^*, \mathbf{u}_{i^*} \rangle \\ &\geq \langle \mathbf{y}^*, \mathbf{u}_{i^*} \rangle - \varepsilon && \text{(by previous inequality)} \end{aligned}$$

We now select any $\mathbf{y} \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)$. We then have the following max tail bound:

$$\begin{aligned} \mathbb{P}\left(\max_{i \in [m]} \langle \mathbf{y}, \mathbf{u}_i \rangle \leq t\right) &= \mathbb{P}\left(\bigcap_i (\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t)\right) \\ &= (\mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t))^m \\ & \text{(by independence of } \mathbf{u}_i \text{, for each } i \in [m].) \end{aligned}$$

Hence in order for us to have that for every element in the net the converse of the above holds we have

$$\begin{aligned} \mathbb{P}\left(\min_{\mathbf{y} \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle > t\right) &= \mathbb{P}\left(\bigcap_{\mathbf{y} \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)} \left(\max_i \langle \mathbf{y}, \mathbf{u}_i \rangle > t\right)\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{\mathbf{y} \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)} \left(\max_i \langle \mathbf{y}, \mathbf{u}_i \rangle \leq t\right)\right) \\ & \text{(by De Morgan's Laws)} \\ &\geq 1 - \sum_{\mathbf{y} \in \mathcal{N}(\mathbb{S}^{d-1}, \varepsilon)} \mathbb{P}\left(\max_i \langle \mathbf{y}, \mathbf{u}_i \rangle \leq t\right) \\ & \text{(applying union bound)} \\ &= 1 - \left(\frac{3}{\varepsilon}\right)^d \mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t)^m \\ & \text{(by previous max tail bound)} \end{aligned}$$

Now for each $\mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t)$ this is the probability that $\mathbf{u}_i$ belongs to a certain complement of a spherical cap. For volumes of spherical caps in higher dimensions we use the results of (Li, 2011). Using these closed form formulas, we can then evaluate for values of $t > 0$:

$$\begin{aligned}\mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t) &= 1 - \mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \geq t) \\ &= 1 - \frac{2\pi^{(d-1)/2}/\Gamma((d-1)/2)\int_0^\phi \sin^{d-2}(\varphi)d\varphi}{2\pi^{d/2}/\Gamma(d/2)},\end{aligned}$$

where ϕ is s.t. $\cos\phi = t$, or in other words $\phi = \operatorname{acos}(t)$. We will now lower bound

$$\int_0^\phi \sin^{d-2}(\varphi)d\varphi \geq \int_{\phi/2}^\phi \sin^{d-2}(\varphi)d\varphi \geq \sin^{d-2}(\phi/2)\phi/2.$$

Now note that $\sin^2\phi/2 + \cos^2\phi/2 = 1$, and by the trigonometric identity $\cos^2\phi/2 = (1+\cos\phi)/2 = (1+t)/2$. Hence $\sin^2\phi/2 = (1-t)/2$. We conclude that

$$\int_0^\phi \sin^{d-2}(\varphi)d\varphi \geq ((1-t)/2)^{d/2-1}\operatorname{acos}(t)/2.$$

Denote with

$$C(d) = \frac{\Gamma((d-1)/2)}{\pi^{1/2}\Gamma(d/2)}.$$

We have that

$$\mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t) \leq 1 - C(d)((1-t)/2)^{d/2-1}\operatorname{acos}(t)/2.$$

Hence

$$\mathbb{P}(\langle \mathbf{y}, \mathbf{u}_i \rangle \leq t)^m \leq \exp(-mC(d)((1-t)/2)^{d/2-1}\operatorname{acos}(t)/2)$$

We conclude that

$$\mathbb{P}\left(\min_{\mathbf{y}\in\mathcal{N}(\mathbb{S}^{d-1},\varepsilon)} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle \geq t\right) \geq 1 - \left(\frac{3}{\varepsilon}\right)^d \exp(-mC(d)((1-t)/2)^{d/2-1}\operatorname{acos}(t)/2).$$

Set $t = 1 - 2/\sqrt{\log m}$. Now we us the fact that $\operatorname{acos}(1-x) \approx \sqrt{2x}$ (in fact $\operatorname{acos}(1-x) \geq \sqrt{2x}$) to conclude that the above expression is bounded as

$$\left(\frac{3}{\varepsilon}\right)^d \exp(-mC(d)/\operatorname{polylog}_d(m)),$$

where $\operatorname{polylog}_d(m) = (\sqrt{\log m})^{(d-1)/2}$. Hence ε can be selected as $3/e^{mC(d)/\operatorname{polylog}_d(m)/(2d)}$ and the above will still converge to 0 while $\varepsilon \to 0$ as well. We conclude that with high probability $\min_{y\in\mathbb{S}^{d-1}} \max_i \langle \mathbf{y}, \mathbf{u}_i \rangle > 1 - 2/\sqrt{\log m} - 3/e^{mC(d)/\operatorname{polylog}_d(m)/(2d)})$ which goes to 1 as m increases. □

Remark 4.99. We observe that in our proof for the upper bound for general convex bodies $K \subset \mathbb{R}^d$, we relied (implicitly) on the fact that $d \geq 2$. We note that in the case $d = 1$, the convex body with known centroid is the compact symmetric interval on the real line. This is simply a convex polytope in $d = 1$, which is already explicitly proved for each $d \geq 1$ in Section 4.4.1.

4.E.3 Proof of Proposition 4.31

Proposition 4.31 (Consistency of the scale parameter MLE, $K \in \mathcal{K}^d$.)**.** *Assume that the same conditions as Theorem 4.29 hold, and let $G > 0$ denote the Lipschitz constant of the Minkowski gauge functional $\rho_K(\mathbf{x})$. We then have that $|\widehat{\sigma}_{\mathrm{MLE}} - \sigma| \leq \frac{\sigma\kappa_n}{n}(GC_1 + 1)$, with probability at least $1 - 2\exp(-C_2\kappa_n/\operatorname{polylog}_d(\kappa_n))$, where C_1, C_2 are as defined in Theorem 4.29.*

Proof of Proposition 4.31. First, we have $\widehat{\sigma}_{\mathrm{MLE}} \coloneqq \max_{i\in[n]} \rho_K(\mathbf{Y}_i - \widehat{\mathbf{v}}_{\mathrm{MLE}})$ and $\widetilde{\sigma} \coloneqq \max_{i\in[n]} \rho_K(\mathbf{Y}_i - \mathbf{v})$, respectively. Our goal is to bound the estimation error for the scale parameter σ, i.e. $|\widehat{\sigma}_{\mathrm{MLE}} - \sigma|$. By the triangle inequality we have $|\widehat{\sigma}_{\mathrm{MLE}} - \sigma| \leq |\widehat{\sigma}_{\mathrm{MLE}} - \widetilde{\sigma}| + |\sigma - \widetilde{\sigma}|$. We proceed by bounding each term separately.

It then follows from the proof of Theorem 4.29 that:

$$|\widehat{\sigma}_{\mathrm{MLE}} - \widetilde{\sigma}| \leq G\,\|\widehat{\mathbf{v}}_{\mathrm{MLE}} - \mathbf{v}\| \leq \frac{G\sigma C_1\kappa_n}{n}, \tag{4.126}$$

with probability at least $1-2\exp(-C_2\kappa_n/\operatorname{polylog}_d(\kappa_n))$, where $C_1 \coloneqq C_1(d, K)$ and $C_2 \coloneqq C_2(d, K)$ are constants which depend on the dimension d and the convex body K, κ_n is any slowly diverging sequence with n, and $\operatorname{polylog}_d(\kappa_n)$ is a poly-logarithmic factor of κ_n which also depends on the dimension d.

In order to bound the second term, we observe that by definition of the Gauge functional, and given that $\mathbf{X}_i$ is sampled from the boundary shell of P, we have with high probability:

$$(1 - \alpha_n)\sigma \leq \rho_K(\mathbf{X}_i) \leq \sigma \tag{4.127}$$

Then using Theorem 4.29 we have that:

$$\begin{aligned}
|\sigma - \widetilde{\sigma}| &= \left|\max_{i\in[n]} \rho_K(\mathbf{X}_i) - \sigma\right| && \text{(by definition of } \widetilde{\sigma}.\text{)} \\
&\leq \sigma - (1 - \frac{\kappa_n}{n})\sigma && (4.128) \\
&\leq \frac{\kappa_n}{n}\sigma, && (4.129)
\end{aligned}$$

Once again this high probability event is contained in the previous one, i.e., by sampling a single boundary point from the shell. Combining the above we obtain:

$$\begin{aligned}|\widehat{\sigma}_{\mathrm{MLE}} - \sigma| &\leq |\widehat{\sigma}_{\mathrm{MLE}} - \widetilde{\sigma}| + |\sigma - \widetilde{\sigma}| &(4.130)\\ &\leq \frac{\sigma \kappa_n}{n}(GC_1 + 1) &(4.131)\end{aligned}$$

□

4.E.4 Proof of Theorem 4.32

Le Cam's Lemma - tail probability form

Lemma 4.100 (Le Cam's Lemma based on Tail Probability Formulation)**.** *Let $\mathcal{P}$ be a set of distributions. For any distributions $P, P_0, P_1 \in \mathcal{P}$*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}\left(d(\widehat{\theta}, \theta(P)) \geq s\right) \geq 1 - \mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{n} P_0, \bigotimes_{i=1}^{n} P_1\right) \tag{4.132}$$

where $\theta(P)$ is some function of P, $\widehat{\theta} := \widehat{\theta}(X_1, \ldots, X_n)$, $s := d(\theta(P_0), \theta(P_1))$, for a specified metric d.

Proof of Lemma 4.100. The proof is based on standard Le Cam based minimax lower bound arguments, e.g., Yu (1997). More specifically, the proof below is directly adapted from Wasserman (2018, Theorem 4, Equation (9)), where it is stated in high expectation form. For completeness, we modify this latter argument slightly, to be based on a high probability bound instead, since this is what we require for our purposes.

Let $\theta_0 := \theta(P_0), \theta_1 := \theta(P_1)$ and $s := d(\theta_0, \theta_1)$. First suppose that $n = 1$ so that we have a single observation X. We then have

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}\left(d(\widehat{\theta}, \theta(P)) \geq s\right) \geq \pi$$

where

$$\pi = \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j)$$

since a maximum is larger than an average,

$$\pi = \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j) \geq \inf_{\psi} \frac{P_0(\psi \neq 0) + P_1(\psi \neq 1)}{2}$$

Define the Neyman-Pearson test

$$\psi_*(x) = \begin{cases} 0 & \text{if } p_0(x) \geq p_1(x) \\ 1 & \text{if } p_0(x) < p_1(x) \end{cases}$$

We show that the sum of the errors $P_0(\psi \neq 0) + P_1(\psi \neq 1)$ is minimized by ψ^*. Now

$$
\begin{aligned}
P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1) &= \int_{p_1 > p_0} p_0(x)\,\mathrm{d}x + \int_{p_0 > p_1} p_1(x)\,\mathrm{d}x \\
&= \int_{p_1 > p_0} [p_0(x) \wedge p_1(x)]\,\mathrm{d}x + \int_{p_0 > p_1} [p_0(x) \wedge p_1(x)]\,\mathrm{d}x = \int [p_0(x) \wedge p_1(x)]\,\mathrm{d}x
\end{aligned}
$$

Thus,

$$
\frac{P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1)}{2} = \frac{1}{2} \int [p_0(x) \wedge p_1(x)]\,\mathrm{d}x
$$

Thus we have shown that

$$
\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}\left(d(\widehat{\theta}, \theta(P)) \geq s\right) \geq \int [p_0(x) \wedge p_1(x)]\,\mathrm{d}x
$$

Now suppose we have n observations. Then, replacing p_0 and p_1 with $p_0^n(x) = \bigotimes_{i=1}^n p_0(x_i)$ and $p_1^n(x) = \bigotimes_{i=1}^n p_1(x_i)$, we have

$$
\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}\left(d(\widehat{\theta}, \theta(P)) \geq s\right) \geq \int [p_0^n(x) \wedge p_1^n(x)]\,\mathrm{d}x,
$$

from which the result follows. □

Total variation bounds on perturbed multivariate uniform distributions

First we note down various set metrics which will be useful in our analysis. Although they are well defined for any pair of compact sets in $\mathbb{R}^d$ we will define them over $\mathcal{K}^d$ since this restricted class is of particular interest

Definition 4.101 (Symmetric Difference Metric on $\mathcal{K}^d$). Let $K_1, K_2 \in \mathcal{K}^d$ be general convex bodies. Then the *Symmetric Set Difference Metric* between $\{K_1, K_2\}$, i.e. $\triangle_d(K_1, K_2)$, is defined by the following equivalent formulations:

$$
\begin{aligned}
\triangle_d(K_1, K_2) &\coloneqq \mathrm{vol}_d(K_1 \triangle K_2) \\
&= \mathrm{vol}_d(K_1 \cup K_2) - \mathrm{vol}_d(K_1 \cap K_2) \\
&= \mathrm{vol}_d(K_1 \setminus K_2) + \mathrm{vol}_d(K_2 \setminus K_1)
\end{aligned}
$$

Remark 4.102. All of these identities are equivalent by first noting the equivalence of the analagous symmetric set difference identities in $\mathbb{R}^d$, i.e. $K_1 \triangle K_2 \coloneqq K_1 \cup K_2 \setminus K_1 \cap K_2 = (K_1 \setminus K_2) \sqcup (K_2 \setminus K_1)$. Here $\sqcup$ denotes a disjoint union of sets. The volume identities in Definition 4.101 then follow directly by applying the finite additivity of $\mathrm{vol}_d(\cdot)$ on these symmetric set difference identities.

In the lower bound discussion that follows our focus will be to consider a convex body $K_1 \in \mathcal{K}^d$, and *perturb* it slightly via a translation vector $\mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, to obtain a $\mathbf{z}$-*translated* convex body $K_2 \coloneqq K_1 + \mathbf{z}$. Since such a pair of $\mathbf{z}$-*translated* convex bodies $\{K_1, K_2\}$ are congruent, i.e. have equal volume, they have additional structure in their symmetric difference metric. This was claimed in (Schymura, 2014) and we formalize this notion in Lemma 4.103. We will see that Lemma 4.103 will later help us to explicitly upper bound the total variation distance between uniform distributions between this pair of $\mathbf{z}$-*translated* convex bodies.

Lemma 4.103 (Symmetric Difference between Translated Convex Bodies)**.** *Let $K_1 \in \mathcal{K}^d$ be a general convex body, and let $K_2 \coloneqq K_1 + \mathbf{z} \in \mathcal{K}^d$ be a translation of it by $\mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Then we have the following:*

$$\operatorname{vol}_d (K_1 \setminus K_2) = \operatorname{vol}_d (K_2 \setminus K_1) \tag{4.133}$$

$$\triangle_d (K_1, K_2) = 2 \operatorname{vol}_d (K_1 \setminus K_2) \tag{4.134}$$

Proof of Lemma 4.103. Let us fix an arbitrary $\mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Since $K_2 \coloneqq K_1 + \mathbf{z}$ we have by the translation invariance of $\operatorname{vol}_d(\cdot)$ in $\mathbb{R}^d$ that:

$$\operatorname{vol}_d (K_2) = \operatorname{vol}_d (K_1 + \mathbf{z}) = \operatorname{vol}_d (K_1) \tag{4.135}$$

Note that such an identity also trivially holds when $\mathbf{z} = \mathbf{0}$. We now proceed by observing the following the disjointification of K_1 and K_2, respectively:

$$K_1 = (K_1 \setminus K_2) \sqcup (K_1 \cap K_2) \tag{4.136}$$

$$K_2 = (K_2 \setminus K_1) \sqcup (K_1 \cap K_2) \tag{4.137}$$

Where $\sqcup$ denotes a disjoint union. Then by finite additivity across disjoint unions of $\operatorname{vol}_d(\cdot)$ in $\mathbb{R}^d$, applied to Equations (4.136) and (4.137) results in the following:

$$\operatorname{vol}_d (K_1) = \operatorname{vol}_d (K_1 \setminus K_2) + \operatorname{vol}_d (K_1 \cap K_2) \tag{4.138}$$

$$\operatorname{vol}_d (K_2) = \operatorname{vol}_d (K_2 \setminus K_1) + \operatorname{vol}_d (K_1 \cap K_2) \tag{4.139}$$

We then have by taking the difference of Equations (4.136) and (4.137) and noting that $\operatorname{vol}_d (K_1) = \operatorname{vol}_d (K_2)$ in our case (from Equation (4.135)), implies that:

$$\operatorname{vol}_d (K_1 \setminus K_2) = \operatorname{vol}_d (K_2 \setminus K_1) \tag{4.140}$$

This proves Equation (4.133). Now we finally have that:

$$\begin{aligned}
&\triangle_d(K_1, K_2) = \mathrm{vol}_d(K_1 \setminus K_2) + \mathrm{vol}_d(K_2 \setminus K_1) && \text{(from Definition 4.101)} \\
\implies &\triangle_d(K_1, K_2) = 2\,\mathrm{vol}_d(K_1 \setminus K_2) && \text{(using Equation (4.140))}
\end{aligned}$$

which proves Equation (4.134) as required. □

Before describing lower bound results, we first derive useful results related to probability metrics on uniform distributions over convex bodies as described in Lemma 4.104.

Lemma 4.104 (Distances Between Translated Uniform Distributions)**.** *Let $K_1 \in \mathcal{K}^d$ be a general convex body, and let $K_2 \coloneqq K_1 + \mathbf{z} \in \mathcal{K}^d$ be a translation of it by $\mathbf{z} \in \mathbb{R}^d$. Further let $X_1 \sim \mathsf{Unif}[K_1]$ and $X_2 \sim \mathsf{Unif}[K_2]$. Then we have the following:*

$$\mathrm{d}_{\mathsf{TV}}(X_1, X_2) = \frac{1}{2}\left(\frac{\mathrm{vol}_d(K_1 \triangle K_2)}{\mathrm{vol}_d(K_1)}\right) \tag{4.141}$$

$$= \frac{\mathrm{vol}_d(K_1 \setminus K_2)}{\mathrm{vol}_d(K_1)} \tag{4.142}$$

Where $\mathrm{d}_{\mathsf{TV}}(X_1, X_2)$ and $\mathrm{d}_{\mathsf{H}}(X_1, X_2)$ are the Total Variation and Hellinger distances between the distributions of X_1 and X_2, respectively.

Proof of Lemma 4.104. We first the Total Variation distance. We are given that $K_1, K_2 \in \mathcal{K}^d$. It then follows that $\mathbf{X}_i \sim \mathsf{Unif}[K_i] \iff f_{\mathbf{X}_i}(\mathbf{x}) = \frac{\mathbb{I}_{K_i}(\mathbf{x})}{\mathrm{vol}_d(K_i)}$ for $i \in \{1, 2\}$. Now we proceed with the total variance distance calculation directly as follows:

$$
\begin{aligned}
\mathrm{d}_{\mathsf{TV}}(X_1, X_2) &= \frac{1}{2}\int_{\mathbb{R}^d} |f_{X_1}(\mathbf{x}) - f_{X_2}(\mathbf{x})| \,\mathrm{d}\lambda_d(\mathbf{x}) && \text{(by Scheffé's lemma)} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\frac{\mathbb{I}_{K_1}(\mathbf{x})}{\mathrm{vol}_d(K_1)} - \frac{\mathbb{I}_{K_2}(\mathbf{x})}{\mathrm{vol}_d(K_2)}\right| \mathrm{d}\lambda_d(\mathbf{x}) \\
& && \text{(since } X_i \sim \mathsf{Unif}[K_i] \text{ for } i \in \{1,2\}) \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\frac{\mathbb{I}_{K_1}(\mathbf{x})}{\mathrm{vol}_d(K_1)} - \frac{\mathbb{I}_{K_2}(\mathbf{x})}{\mathrm{vol}_d(K_1)}\right| \mathrm{d}\lambda_d(\mathbf{x}) \\
& && \text{(since } \mathrm{vol}_d(K_1) = \mathrm{vol}_d(K_2)) \\
&= \frac{1}{2}\frac{1}{\mathrm{vol}_d(K_1)}\int_{\mathbb{R}^d} |\mathbb{I}_{K_1}(\mathbf{x}) - \mathbb{I}_{K_2}(\mathbf{x})| \,\mathrm{d}\lambda_d(\mathbf{x}) \\
& && \text{(since } \mathrm{vol}_d(K_1) > 0) \\
&= \frac{1}{2}\frac{1}{\mathrm{vol}_d(K_1)}\int_{\mathbb{R}^d} \mathbb{I}_{K_1 \triangle K_2}(\mathbf{x}) \,\mathrm{d}\lambda_d(\mathbf{x}) && \text{(since } |\mathbb{I}_A - \mathbb{I}_B| = \mathbb{I}_{A\triangle B}) \\
&= \frac{1}{2}\frac{1}{\mathrm{vol}_d(K_1)}\int_{K_1 \triangle K_2} \mathrm{d}\lambda_d(\mathbf{x}) \\
&= \frac{1}{2}\left(\frac{\mathrm{vol}_d(K_1 \triangle K_2)}{\mathrm{vol}_d(K_1)}\right) \\
&= \frac{\mathrm{vol}_d(K_1 \setminus K_2)}{\mathrm{vol}_d(K_1)} && \text{(by Equation (4.134))}
\end{aligned}
$$

Which proves both Equation (4.141) and Equation (4.142). □

We now come to the critical notion of the *sweep set of a convex body* K by a vector $\mathbf{t}$ from (Schymura, 2014). This notion is formally stated in Definition 4.105 and illustrated in Figure 4.E.3 (adapted from (Schymura, 2014, Figure 1)).

Definition 4.105 (Sweep Set of a Convex Body, (Schymura, 2014))**.** Let $K_1 \in \mathcal{K}^d$, and $\mathbf{z} \in \mathbb{R}^d$ be a translation vector. We define the *sweep set* of a convex body by $\mathbf{z}$, i.e. $K_1 + [0,1]\mathbf{z}$, as follows:

$$K_1 + [0,1]\mathbf{z} \coloneqq \{\mathbf{k}_1 + \lambda\mathbf{z} \mid \lambda \in [0,1], \mathbf{k}_1 \in K_1\}$$

As noted in (Schymura, 2014) we have the following explicit formula for the volume of the sweep set of a convex body under translation summarized in Lemma 4.106.

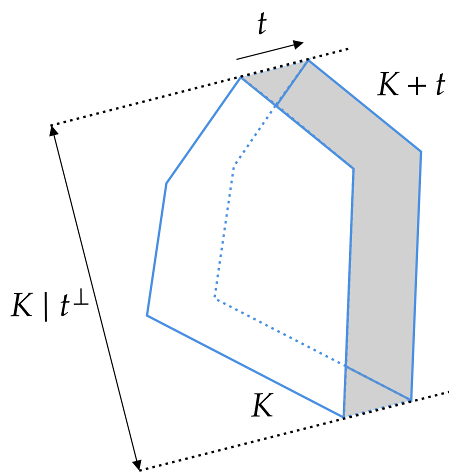

Figure 4.E.3: An illustration of the sweep set (shaded grey) of the convex body K by the vector $\mathbf{t}$

Lemma 4.106 (Volume of a Sweep Set of a Convex Body)**.** *Let $K_1 \in \mathcal{K}^d$, and $\mathbf{z} \in \mathbb{R}^d$ be a translation vector. We then have:*

$$\mathrm{vol}_d\left(K_1 + [0,1]\mathbf{z}\right) = \mathrm{vol}_d\left(K_1\right) + \|\mathbf{z}\|_2 \,\mathrm{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^\perp\right) \tag{4.143}$$

Where $K_1 \mid \mathbf{z}^\perp$ is the image of the orthogonal projection of K_1 onto the orthogonal complement of $\mathbf{z}$, i.e. $\mathbf{z}^\perp := \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \langle \mathbf{x}, \mathbf{z}\rangle = 0\right\}$.

Proof of Lemma 4.106. See (Gardner, 2006, Appendix A.6) for details. □

Moreover this now allows us to have the necessary tools to now upper bound the symmetric set difference of a convex body and it is translation. This idea is mentioned in (Schymura, 2014). In particular, this allows us to upper bound the total variation distance between two convex bodies in $\mathbb{R}^d$. We formalize these statements in Proposition 4.107.

Proposition 4.107 (Upper Bound Distances Between Translated Uniform Distributions)**.** *Let $K_1 \in \mathcal{K}^d$ be a general convex body, and let $K_2 := K_1 + \mathbf{z} \in \mathcal{K}^d$ be a translation of it by $\mathbf{z} \in \mathbb{R}^d$. Further let $X_1 \sim \mathsf{Unif}[K_1]$ and $X_2 \sim \mathsf{Unif}[K_2]$.*

We then have:

$$\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right) \leq \frac{\operatorname{vol}_d\left(K_1 \triangle K_1 + [0,1]\mathbf{z}\right)}{2\operatorname{vol}_d\left(K_1\right)} \tag{4.144}$$

$$= \frac{\|\mathbf{z}\|_2 \operatorname{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^{\perp}\right)}{2\operatorname{vol}_d\left(K_1\right)} \tag{4.145}$$

Proof of Proposition 4.107. Firstly we observe that if for sets $A, B, C \subseteq \mathbb{R}^d$, such that $B \subseteq C$, we have that $A \triangle B \subseteq A \triangle C$. Now since $K_1, K_1 + \mathbf{z}, K_1 + [0,1]\mathbf{z} \in \mathcal{K}^d$ with $K_1 + \mathbf{z} \subseteq K_1 + [0,1]\mathbf{z}$, this implies that $K_1 \triangle (K_1 + \mathbf{z}) \subseteq K_1 \triangle (K_1 + [0,1]\mathbf{z})$. It then follows by the monotonicty of $\operatorname{vol}_d(\cdot)$ in $\mathbb{R}^d$ that:

$$\operatorname{vol}_d\left(K_1 \triangle (K_1 + \mathbf{z})\right) \leq \operatorname{vol}_d\left(K_1 \triangle (K_1 + [0,1]\mathbf{z})\right) \tag{4.146}$$

We also note that since $K_1 = K_1 + (0)\mathbf{z} \implies K_1 \subseteq (K_1 + [0,1]\mathbf{z})$. In fact we then have the disjointification $K_1 + [0,1]\mathbf{z} = K_1 \sqcup (K_1 + (0,1]\mathbf{z})$. By applying finite additivity of $\operatorname{vol}_d(\cdot)$ in $\mathbb{R}^d$, we then have that: $\operatorname{vol}_d\left(K_1 + [0,1]\mathbf{z}\right) = \operatorname{vol}_d\left(K_1\right) + \operatorname{vol}_d\left(K_1 + (0,1]\mathbf{z}\right)$. Comparing this form to Lemma 4.106 and matching terms gives us:

$$\begin{aligned}\operatorname{vol}_d\left(K_1 \triangle (K_1 + [0,1]\mathbf{z})\right) &= \operatorname{vol}_d\left(K_1 + (0,1]\mathbf{z}\right) \\ &= \|\mathbf{z}\|_2 \operatorname{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^{\perp}\right) \end{aligned} \tag{4.147}$$

Combining all of the details gives us:

$$\begin{aligned}\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right) &= \frac{1}{2}\left(\frac{\operatorname{vol}_d\left(K_1 \triangle (K_1 + \mathbf{z})\right)}{\operatorname{vol}_d\left(K_1\right)}\right) && \text{(per Lemma 4.104)} \\ &\leq \frac{\operatorname{vol}_d\left(K_1 \triangle (K_1 + [0,1]\mathbf{z})\right)}{2\operatorname{vol}_d\left(K_1\right)} && \text{(per Equation (4.144))} \\ &= \frac{\|\mathbf{z}\|_2 \operatorname{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^{\perp}\right)}{2\operatorname{vol}_d\left(K_1\right)} && \text{(per Equation (4.147))}\end{aligned}$$

Which proves both Equations (4.144) and (4.145), as required. □

Lemma 4.108 (Subadditivity of Total Variation distance under independence)**.** *We have $\mathbb{P} = \bigotimes_{i=1}^{n} \mathbb{P}_i$ and $\mathbb{Q} = \bigotimes_{i=1}^{n} \mathbb{Q}_i$ probability measures on $(\mathcal{X}, \mathcal{F})$, both being dominated by a σ-finite measure μ. The corresponding μ-densities will be called p, p_i, q, q_i, $\forall i \in [n]$. Given that $\mathrm{d}_{\mathsf{TV}}\left(X_1, X_2\right) = \frac{1}{2}\int_{\mathbb{R}^d} \left|p_1(\mathbf{x}) - p_2(\mathbf{x})\right| \mathrm{d}\lambda_d(\mathbf{x})$ We have that*

$$\mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}, \mathbb{Q}\right) \leq \sum_{i=1}^{n} \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_i, \mathbb{Q}_i\right) \tag{4.148}$$

As a corollary, in the case when $\mathbb{P}_i, \mathbb{Q}_i$ are i.i.d. *for each $i \in [n]$, we have that:*

$$\mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}, \mathbb{Q}\right) \leq n \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1, \mathbb{Q}_1\right) \tag{4.149}$$

Proof of Lemma 4.108. Our proof will proceed by induction on $n \in \mathbb{N}$. We assume the well known fact that $\mathrm{d}_{\mathsf{TV}}\left(\cdot, \cdot\right)$ is a valid metric on the space of probability measures, and satisfies the triangle inequality. First we observe in the case where $n = 1$, we have $\mathbb{P} = \mathbb{P}_1, \mathbb{Q} = \mathbb{Q}_1$, and thus equality holds in Equation (4.148). We will directly prove the case for $n = 2$. Consider:

$$\begin{aligned}
\mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2\right) &\leq \underbrace{\mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{P}_2\right)}_{D_1} + \underbrace{\mathrm{d}_{\mathsf{TV}}\left(\mathbb{Q}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2\right)}_{D_2} \\
&\qquad \text{(by triangle inequality on } \mathrm{d}_{\mathsf{TV}}\left(\cdot, \cdot\right)\text{)} \\
\implies D_1 &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \left|p_1(\mathbf{x}_1)p_2(\mathbf{x}_2) - q_1(\mathbf{x}_1)p_2(\mathbf{x}_2)\right| d\lambda_d(\mathbf{x}_1) d\lambda_d(\mathbf{x}_2) \\
&= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} p_2(\mathbf{x}_2) \left|p_1(\mathbf{x}_1) - q_1(\mathbf{x}_1)\right| d\lambda_d(\mathbf{x}_1) d\lambda_d(\mathbf{x}_2) \\
&= \int_{\mathcal{X}} p_2(\mathbf{x}_2) d\lambda_d(\mathbf{x}_2) \underbrace{\left(\frac{1}{2}\int_{\mathcal{X}} (\mathbf{x}_2) \left|p_1(\mathbf{x}_1) - q_1(\mathbf{x}_1) d\lambda_d(\mathbf{x}_1)\right|\right)}_{\mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1, \mathbb{Q}_1\right)} \\
&= \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1, \mathbb{Q}_1\right) \\
\text{Similarly } D_2 &= \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_2, \mathbb{Q}_2\right) \qquad (4.150) \\
\implies \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2\right) &\leq \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_1, \mathbb{Q}_1\right) + \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_2, \mathbb{Q}_2\right) \qquad (4.151)
\end{aligned}$$

As required. Now we assume our induction hypothesis, i.e. for all $n \leq k \in \mathbb{N}$, the following holds:

$$\mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{k} \mathbb{P}_i, \bigotimes_{i=1}^{k} \mathbb{Q}_i\right) \leq \sum_{i=1}^{k} \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_i, \mathbb{Q}_i\right) \tag{4.152}$$

Now consider the case for $n = k+1$:

$$
\begin{aligned}
\mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{k+1}\mathbb{P}_i, \bigotimes_{i=1}^{k+1}\mathbb{Q}_i\right) &= \mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{k}\mathbb{P}_i \times \mathbb{P}_{k+1}, \bigotimes_{i=1}^{k}\mathbb{Q}_i \times \mathbb{Q}_{k+1}\right) \\
&= \mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{k}\mathbb{P}_i, \bigotimes_{i=1}^{k}\mathbb{Q}_i\right) + \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_{k+1}, \mathbb{Q}_{k+1}\right) \\
&\leq \sum_{i=1}^{k} \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_i, \mathbb{Q}_i\right) + \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_{k+1}, \mathbb{Q}_{k+1}\right) \\
&= \sum_{i=1}^{k+1} \mathrm{d}_{\mathsf{TV}}\left(\mathbb{P}_i, \mathbb{Q}_i\right)
\end{aligned}
$$

As required. □

Lemma 4.109 (Orthogonal Projections on Closed Subspace is Continuous)**.** *The orthogonal projection operator i.e.* $\Pi_K(\mathbf{x}) := \arg\min_{\mathbf{w}\in K} \|\mathbf{x} - \mathbf{w}\|_2$ *on a Closed Convex Subspace in* $K \subseteq \mathbb{R}^d$ *is* 1*-Lipschitz (and thus uniformly continuous) with respect to the Euclidean Metric.*

Proof of Lemma 4.109. See (Deutsch, 2001, Theorem 5.5) for details. □

Lemma 4.110 (Scale invariance of orthogonal projections on Closed Subspace)**.** *Let* $K \in \mathcal{K}^d$*, then for any fixed* $\sigma > 0$*, and any fixed* $\mathbf{z} \in \mathbb{S}^{d-1}$*, the following holds:*

$$
\left(\sigma K \mid \mathbf{z}^{\perp}\right) = \sigma\left(K \mid \mathbf{z}^{\perp}\right) \tag{4.153}
$$

Proof of Lemma 4.110. We will prove the equality of the sets directly as follows.

$$
\begin{aligned}
&\text{Let } \mathbf{x} \in \left(\sigma K \mid \mathbf{z}^{\perp}\right) && \text{(by definition)} \\
&\iff \exists\, \mathbf{k} \in K \,:\, \mathbf{x} = \Pi_{\sigma K|\mathbf{z}^{\perp}}(\sigma k) && \text{(Using Equation (4.49))} \\
&\iff \exists\, \mathbf{k} \in K \,:\, \mathbf{x} = \sigma \Pi_{K|\mathbf{z}^{\perp}}(k) \\
&\iff \mathbf{x} \in \sigma\left(K \mid \mathbf{z}^{\perp}\right)
\end{aligned}
$$

As required. □

Since we are in a finite dimensional setting, we note that the orthogonal projection $\mathbf{z}^{\perp} \coloneqq \left\{\mathbf{y} \in \mathbb{R}^d \,\middle|\, \langle \mathbf{y}, \mathbf{z}\rangle = 0\right\}$ is a closed subspace of $\mathbb{R}^d$ As such $K_1 \mid \mathbf{z}^{\perp}$ is the image of the orthogonal projection of K_1 onto the orthogonal complement of $\mathbf{z}$. Since orthogonal projections are continuous maps, and $K \in \mathcal{K}^d$ is compact set, we have that compactness is preserved by continuous maps.

Final proof of Theorem 4.32

Now we are ready to formulate and prove the lower bound rate. This is summarized below.

Theorem 4.32 (Minimax lower bound for location estimation)**.** *Let $(\mathbf{Y}_i)_{i=1}^n$ be generated according to Definition 4.4, with σ known to the observer. Let $\widehat{\mathbf{v}}$, be any estimator (measurable function) for the location parameter $\mathbf{v}$. We then have that the following holds:*

$$\inf_{\widehat{\mathbf{v}}} \sup_{\mathbf{v} \in \mathbb{R}^d} \mathbb{P}\left(\|\widehat{\mathbf{v}} - \mathbf{v}\|_2 \geq \sup_{\mathbf{z} \in \mathbb{S}^{d-1}} \frac{\sigma \operatorname{vol}_d(K)}{n \operatorname{vol}_{d-1}(K \mid \mathbf{z}^{\perp})}\right) \geq \frac{1}{2}. \tag{4.18}$$

Here $K \mid \mathbf{z}^{\perp}$ is the image of the orthogonal projection of K onto the orthogonal complement of $\mathbf{z}$. Note that $\mathbf{z}^{\perp} \coloneqq \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \langle \mathbf{x}, \mathbf{z}\rangle = 0\right\}$, i.e., the hyperplane through $\mathbf{0} \in \mathbb{R}^d$, with $\mathbf{z} \in \mathbb{R}^d$ as a normal vector.

Let $K_1 = \mathbf{v} + \sigma K$, where $K \in \mathcal{K}^d$. We then consider a perturbation of K_1 by a translation vector $\mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. This results in a $\mathbf{z}$-*translated* convex body $K_2 \coloneqq K_1 + \mathbf{z}$. i.e. $K_2 = \mathbf{v} + \mathbf{z} + \sigma K$. It is clear that $K_2 \in \mathcal{K}^d$. Further we denote $P_1 \sim \mathsf{Unif}[K_1]$ and $P_2 \sim \mathsf{Unif}[K_2]$. With this setup, we are now ready to apply Le Cam's Two Point method to construct our lower bound. This is done as follows:

$$\begin{aligned}
\inf_{\widehat{\mathbf{v}}} \sup_{\mathbf{v} \in \mathbb{R}^d} \mathbb{P}\left(\|\widehat{\mathbf{v}} - \mathbf{v}\|_2 \geq \|\mathbf{z}\|_2\right) &\geq 1 - \mathrm{d}_{\mathsf{TV}}\left(\bigotimes_{i=1}^{n} P_1, \bigotimes_{i=1}^{n} P_2\right) \\
&\qquad \text{(using Equation (4.132))} \\
&\geq 1 - n\mathrm{d}_{\mathsf{TV}}\left(P_1, P_2\right) \quad \text{(using Equation (4.149))} \\
&\geq 1 - \frac{n \|\mathbf{z}\|_2 \operatorname{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^{\perp}\right)}{2 \operatorname{vol}_d\left(K_1\right)} \\
&\qquad \text{(from Equation (4.145))} \\
&= 1 - \left(\frac{n \|\mathbf{z}\|_2}{2}\right)\left(\frac{\operatorname{vol}_{d-1}\left(K_1 \,\middle|\, \mathbf{z}^{\perp}\right)}{\operatorname{vol}_d\left(K_1\right)}\right) \\
&= 1 - \left(\frac{n \|\mathbf{z}\|_2}{2}\right)\left(\frac{\operatorname{vol}_{d-1}\left(\sigma K \,\middle|\, \mathbf{z}^{\perp}\right)}{\operatorname{vol}_d\left(\sigma K\right)}\right) \\
&= 1 - \left(\frac{n \|\mathbf{z}\|_2}{2}\right)\left(\frac{\operatorname{vol}_{d-1}\left(\sigma\left(K \,\middle|\, \mathbf{z}^{\perp}\right)\right)}{\operatorname{vol}_d\left(\sigma K\right)}\right) \\
&= 1 - \left(\frac{n \|\mathbf{z}\|_2}{2}\right)\left(\frac{\sigma^{d-1} \operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)}{\sigma^d \operatorname{vol}_d\left(K\right)}\right) \\
&= 1 - \left(\frac{n \operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right) \|\mathbf{z}\|_2}{2\sigma \operatorname{vol}_d\left(K\right)}\right) \\
&\qquad \text{(for all } \mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\})
\end{aligned}$$

We then observe that for any fixed $\mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, that $\operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)$ is bounded. This follows because the projection operator is (uniformly continuous) using Lemma 4.109, and given $K \in \mathcal{K}^d$ is compact (by definition), it follows that the image under the continuous projection operator preserves compactness. In our finite dimensional setting, this indeed means that $\operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)$ is bounded. Now we finally observe that we can choose $\mathbf{z}$ such that $\operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)$ is minimal across all possible directions. Then by choosing $\mathbf{z}$ such that $\|\mathbf{z}\|_2 = \frac{\sigma \operatorname{vol}_d(K)}{n \operatorname{vol}_{d-1}\left(K \,\middle|\, \mathbf{z}^{\perp}\right)}$ implies that we lower bound our minimax risk away from 0.

4.F Proofs of Section 4.5

In order to ensure to bound the subgradients of the Minkowski gauge functional, we first need Lemma 4.111.

Lemma 4.111 (Bounded dual norm of subgradient of Lipschitz functions). *Let $S \subseteq \mathbb{R}^d$ be non-empty, and $f : S \to \mathbb{R}$ be a convex function. Then, f is L-Lipschitz over S with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in S$ and $\mathbf{z} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{z}\|_\star \leq L$, where $\|\cdot\|_\star$ is the dual norm of $\|\cdot\|$.*

Proof of Lemma 4.111. See Shalev-Shwartz et al. (2012, Lemma 2.6) for details. □

Lemma 4.34 (Bounded subgradient of $\rho_K(\mathbf{x})$). *Let $G > 0$ be the Lipschitz constant for the Minkowski gauge functional, $\rho_K(\mathbf{x})$, as defined in Definition 4.24. Then for any $\mathbf{w} \in \mathbb{R}^d$, and for any $\mathbf{z} \in \partial\left(\rho_K(\mathbf{w})\right)$, we have $\|\mathbf{z}\|_2 \leq G$.*

Proof of Lemma 4.34. We recall that Minkowski functional is convex, and G-Lipschitz, and that the dual norm of $\|\cdot\|_2$ in $\mathbb{R}^d$ is $\|\cdot\|_2$. As such, it follows from Lemma 4.111 that for any $\mathbf{w} \in \mathbb{R}^d$, and for any $\mathbf{z} \in \partial\left(\rho_K(\mathbf{w})\right)$, we have $\|\mathbf{z}\|_2 \leq G$. □

Proposition 4.36 (Convergence of subgradient method). *Suppose that there exist constants $R, G > 0$ with $\left\|\mathbf{x}^{(0)} - \mathbf{x}^\star\right\|_2 \leq R$, $\left\|\mathbf{x}^{(0)} - \mathbf{v}\right\|_2 \leq R$, and $\left\|\mathbf{g}^{(k)}\right\|_2 \leq G$ for all k. Where $\mathbf{x}^\star$ is the unique optimal solution, then running Algorithm 4.1 ensures that*

$$\|\overline{\mathbf{v}} - \mathbf{x}_{\text{best}}^{(L)}\|_2 - \|\overline{\mathbf{v}} - \mathbf{x}^\star\|_2 \leq \frac{R^2 + G^2 \sum_{k=1}^{L} t_k^2}{2\sum_{k=1}^{L} t_k} \tag{4.34}$$

Proof of Proposition 4.36. □

4.F.1 Proof of Proposition 4.35

Proposition 4.35 (Subgradient of $\rho_K(\mathbf{x})$). *Let $K \in \mathcal{K}^d$ with $\mathbf{0} \in \operatorname{int}(K)$. Let $\rho_K(\mathbf{x})$ be it is Minkowski gauge functional, for all $\mathbf{x} \in \mathbb{R}^d$. Further, let $\mathbf{x}'$ be the vector that is parallel to $\mathbf{x}$ and lies on the boundary of K, and let $\mathbf{x}'^*$ be any supporting hyperplane through $\mathbf{x}'$. Then $\frac{\mathbf{x}'^*}{\mathbf{x}'^{*\top}\mathbf{x}'}$ is a subgradient at $\mathbf{x}'$.*

Proof of Proposition 4.35. We now show how to find a subgradient of $\rho_K(\mathbf{x})$ provided that we have a supporting hyperplane oracle to K. By the definition of a subgradient we need to identify $\mathbf{g}$ such that

$$\mathbf{g}^\top(\mathbf{x} - \mathbf{y}) \geq \rho_K(\mathbf{x}) - \rho_K(\mathbf{y})$$

This is implied if $\mathbf{g}$ satisfies the following two properties

$$\mathbf{g}^\top \mathbf{x} = \rho_K(\mathbf{x}), \text{ and } \mathbf{g}^\top \mathbf{y} \leq \rho_K(\mathbf{y}), \text{ for all } \mathbf{y}$$

or equivalently

$$\mathbf{g}^\top \mathbf{x}/\rho_K(\mathbf{x}) = 1, \text{ and } \mathbf{g}^\top \mathbf{y}/\rho_K(\mathbf{y}) \leq 1, \text{ for all } \mathbf{y}$$

Now, let $\mathbf{x}'$ be the vector that is parallel to $\mathbf{x}$ and lies on the boundary of K i.e. $\mathbf{x}' = \mathbf{x}/\rho_K(\mathbf{x})$. We then have $\rho_K(\mathbf{x}') = 1$. Take any supporting hyperplane through that vector, $\mathbf{x}'^*$, and normalize it $g = \mathbf{x}'^*/\mathbf{x}'^{*\top} x'$ (note here that the dot product $\mathbf{x}'^{*\top}\mathbf{x}' > 0$ since $\mathbf{x}'^{*\top}\mathbf{x}' > \mathbf{x}'^{*\top}\mathbf{0} = 0$ (the inequality is strict as $\mathbf{0}$ is an interior point of K hence cannot be on the supporting hyperplane) since $\mathbf{0} \in K$). We argue that this is a subgradient of $\rho_K(\mathbf{x})$. We have

$$\mathbf{g}^\top \mathbf{x}' = 1, \mathbf{g}^\top \mathbf{y}' \leq \mathbf{g}^\top \mathbf{x}' = 1, \text{ for all } \mathbf{y}' \in K.$$

The last inequality includes all boundary points of K which are all points $\mathbf{y}'$ such that $\rho_K(\mathbf{y}') = 1$. This is exactly what we wanted to show. □

4.F.2 Proof of Theorem 4.39

Theorem 4.39. *If you run the subgradient descent in Algorithm 4.2, such that the* RHS *in* (4.36) *is at most* $\frac{C}{n}$ *for some sufficiently small* $C > 0$*, then*

$$\|\mathbf{x}_{\text{best}}^{(L)} - \mathbf{v}\|_2 \lesssim \frac{1}{n}, \tag{4.37}$$

with high probability (say .99).

Proof of Theorem 4.39. In our proof for the general convex body (with known σ) in Section 4.4.2 we look at the thin shell set $\sigma K \setminus (1-\alpha_n)\sigma K$. Suppose instead we consider the positively (slightly) dilated version of our original convex body i.e. $\sigma\left(1+\frac{1}{n}\right)K$. Then the thin strips we look at will be slightly bigger but will be still at the order of $\alpha_n \approx 1/n$, because $1 + 1/n - (1-\alpha_n)$ is still of the order of α_n. So in our convex body proof, we can still sample from the original shell $\sigma K \setminus (1-\alpha_n)\sigma K$. However instead of extending the vector from $\mathbf{0}$ to the boundary of $\partial(\sigma K)$ we extend it to $\partial\sigma\left(1+\frac{1}{n}\right)K$ instead. We draw our supporting hyperplanes to the convex set $\sigma\left(1+\frac{1}{n}\right)K$ and the proof proceeds as for the case of the convex body σK per Section 4.4.2. It follows that since σ is a positive constant (by assumption), we can always find the optimal rate by moving within C/n to the optimal value of the function (i.e. obtain a value that is $\widetilde{\sigma} = \widehat{\sigma} + C/n \leq \sigma(1+1/n)$). □

Five

Conclusion

Recall that the motivating theme of this thesis is to generalize three 'classical' nonparametric, and location-scale (parametric) estimation problems in statistics. We now conclude by summarizing our key findings below, and then describing some exciting future research directions of interest.

Part I Nonparametric Estimation

In Chapter 2 we derived exact (up to constants) minimax rates for density estimation over convex density classes. Our work builds on seminal research of Le Cam (1973); Birgé (1983); Yang and Barron (1999); Wong and Shen (1995). More directly, we non-trivially adapted the techniques of Neykov (2022), who used it for deriving exact rates for the Gaussian sequence model. There, the generating process has a vastly differing geometric structure to our density estimation setting, thus requiring new techniques to be used. Our upper bounds are based on a 'multistage sieve' MLE, which works across *any* convex density class. This estimator can be constructively described via a finite-step (i.e., 'multistage') procedure, where we successively take the MLE on structured subsets (i.e., 'sieves') of our convex density class. Our results demonstrate that the L_2-*local* metric entropy *always* determines that minimax rate under squared L_2-loss in this setting. This provides a unifying minimax density estimation perspective across parametric *and* nonparametric convex density classes. Importantly, our results generalize the seminal work of Yang and Barron (1999), since they are proven under weaker assumptions.

An important open direction that we would like to explore is whether there exists a computationally tractable estimator which is also minimax optimal in our density estimation setting. Recall that our 'multistage sieve' MLE, although provably minimax optimal over any abstract convex density class, is not designed to be practically computable. Another natural question is whether we can apply our techniques to the nonparametric *regression* setting (with Gaussian noise), for estimating a uniformly bounded regression function of interest. Finally, we hope that this research stimulates further activity in approximation theory, i.e., specifically for deriving the L_2-*local* metric entropy for various convex density classes. We leave these promising directions for future work.

In Chapter 3 we considered a *partial* generalization of the classical isotonic regression setup in which the observations can be adversarially sign-corrupted. Under this adversarially sign-corrupted isotonic (ASCI) setting, 'adversarially' refers to the fact that the sign-corruptions may be chosen to have strong dependence with the error terms in the original model. Our motivation was to understand whether robust estimation of the true isotonic regression signal was feasible, under such harsh 'attacks' on the monotonicity of the observations. Our simple three-step estimation procedure, `ASCIFIT`, is easy to implement with existing software. It also has sharp non-asymptotic minimax guarantees on the estimation error, in high probability under squared L_2-loss.

A key restriction of our setting is that that true monotone signal is assumed to be strictly positive for our guarantees to hold. We believe this restriction can be lifted if one uses *unimodal* regression instead of isotonic regression in **Step I** of `ASCIFIT`. Lifting this positivity assumption will ensure that the resulting ASCI setup would be a *complete* generalization of classical isotonic regression, with non-asymptotic minimax estimation guarantees. However, one would need to first establish sharp risk guarantees similar to Zhang (2002) under this unimodal setting. It would also be interesting to see if the moment matching technique could be extended subgaussian error terms. We leave these exciting directions for future work.

Part II **Location-scale estimation**

In Chapter 4 we generalize the classical problem of *univariate* uniform location estimation over an interval, to *multivariate* uniform location estimation over convex bodies, i.e., $K \in \mathcal{K}^d$. We consider both known and unknown scaling regimes, with the latter case being the more challenging estimation setting. Even in the (easier) known scaling regime we demonstrate a fundamental trade-off arising between the statistical optimality and the computational feasibility of estimation in general settings. Motivated by this trade-off we propose projection-based location estimator in the known scaling regime, and also show how to obtain location-scale MLEs in the unknown scaling regimes. All of these proposed estimators lie in region we term the *critical set*. Furthermore, we demonstrate that any location estimator lying in the critical set converges at a rate of $\frac{C(d,K)}{n}$ with high probability, under squared L_2-loss. These rates are supported with matching minimax lower bounds in sample complexity. We also provide feasible algorithms with provable guarantees for our proposed estimators over more general settings compared to known estimators.

This opens up many exciting directions for further exploration in location-scale estimation. For example, our upper bounds hold for *any* estimator in the critical set, and as such are likely suboptimal in the dimension dependent constant $C(d, K)$ compared to the minimax lower bounds. These could be tightened further by exploiting the convex-geometric structure of each individual estimator. Performing inference on these uniform location-scale parameters of interest is another open problem. A promising approach could be to extend the techniques from Wasserman et al. (2020) to our (non-regular) multivariate uniform setting. We also believe that our upper bounds techniques can achieve similar risk convergence rates when the underlying distribution is non-uniform over $K \in \mathcal{K}^d$, provided that the underlying density is suitably bounded away from zero in probability over the entire boundary of the convex body supporting set K. We defer demonstrating this conjecture, and the above open problems to future work.

Overall, our work focused in generalizing various classical models and estimators, and understanding their theoretical properties. The topics spanned density estimation, isotonic regression, and also uniform location (and scale) estimation. Our work demonstrates that we can still continue to gain *new* inferential insights in these classical settings. Importantly we hope it stimulates further research into these (and related) classical topics.

Bibliography

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information . *Ann. Math. Statist.*, 26:641–647.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis . *Ann. Statist.*, 45(1):77–120.

Baldi, P., Kerkyacharian, G., Marinucci, D., and Picard, D. (2009). Adaptive density estimation for directional data using needlets. *Ann. Statist.*, 37(6A):3362–3395.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression* . Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, London-New York-Sydney.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika*, 46(1-2):36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives. II. *Biometrika*, 46:328–335.

Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression . *Ann. Statist.*, 46(2):745–780.

Bellec, P. C. and Tsybakov, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation . *J. Mach. Learn. Res.*, 16:1879–1892.

Best, M. J. and Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Math. Programming*, 47(3, (Ser. A)):425–439.

Bickel, P. J. and Doksum, K. A. (2016). *Mathematical statistics—basic ideas and selected topics. Vol. 2*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, second edition.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2021). Minimax Rates for Conditional Density Estimation via Empirical Entropy. arXiv:2109.10461.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation . *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts . *Probab. Theory Relat. Fields*, 71(2):271–291.

Birgé, L. (2014). Model selection for density estimation with L_2-loss. *Probab. Theory Related Fields*, 158(3-4):533–574.

Birgé, L. and Massart, P. (1993a). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.

Birgé, L. and Massart, P. (1993b). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence . *Bernoulli*, 4(3):329–375.

Birman, M. Š. and Solomjak, M. Z. (1980). *Quantitative analysis in Sobolev imbedding theorems and applications to spectral theory* , volume 114 of *American Mathematical Society Translations, Series 2*. American Mathematical Society, Providence, R.I. Translated from the Russian by F. A. Cezus.

Blum, A., Hopcroft, J., and Kannan, R. (2020). *Foundations of Data Science*. Cambridge University Press.

Boyd, D. W. and Steele, J. M. (1978). Lower bounds for nonparametric density estimation rates. *Ann. Statist.*, 6(4):932–934.

Boyd, S. and Park, J. (2014). Subgradient methods. *lecture notes of EE364b, Stanford University, Spring Quarter*.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

Boyle, J. P. and Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces . In *Advances in order restricted statistical inference (Iowa City, Iowa, 1985)*, volume 37 of *Lect. Notes Stat.*, pages 28–47. Springer, Berlin.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Z. Wahrsch. Verw. Gebiete*, 47(2):119–137.

Brezis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York.

Brøndsted, A. (1983). *An introduction to convex polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381.

Chatterjee, S., Guntuboyina, A., and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems . *Ann. Statist.*, 43(4):1774–1800.

Cleanthous, G., Georgiadis, A. G., Kerkyacharian, G., Petrushev, P., and Picard, D. (2020). Kernel and wavelet density estimators on manifolds and more general metric spaces . *Bernoulli*, 26(3):1832–1862.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods . *Journal of Statistical Software*, 32(5):1–24.

De Loera, J. A., Rambau, J., and Santos, F. (2010). *Triangulations*, volume 25 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin. Structures for algorithms and applications.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Deutsch, F. (1985). Rate of convergence of the method of alternating projections. In *Parametric optimization and approximation (Oberwolfach, 1983)*, volume 72 of *Internat. Schriftenreihe Numer. Math.*, pages 96–107. Birkhäuser, Basel.

Deutsch, F. (2001). *Best approximation in inner product spaces*, volume 7 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, New York.

Devroye, L. (1983). On arbitrarily slow rates of global convergence in density estimation. *Z. Wahrsch. Verw. Gebiete*, 62(4):475–483.

Devroye, L. (1987). *A course in density estimation*, volume 14 of *Progress in Probability and Statistics*. Birkhäuser Boston, Inc., Boston, MA.

Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York. The L_1 view.

Donoho, D. L. (1990). Gelfand n-widths and the method of least squares. *Preprint*.

Efromovich, S. (2008). Adaptive estimation of and oracle inequalities for probability densities and characteristic functions . *Ann. Statist.*, 36(3):1127–1155.

Efroĭmovich, S. Y. and Pinsker, M. S. (1982). Estimation of square-integrable probability density of a random variable . *Problemy Peredachi Informatsii*, 18(3):19–38.

Elandt, R. C. (1961). The folded normal distribution: two methods of estimating parameters from moments . *Technometrics*, 3:551–562.

Ferguson, T. S. (1996). *A course in large sample theory*. Texts in Statistical Science Series. Chapman & Hall, London.

Fishman, G. S. (1996). *Monte Carlo*. Springer Series in Operations Research. Springer-Verlag, New York. Concepts, algorithms, and applications.

Gallier, J. and Quaintance, J. (2008). Notes on Convex Sets, Polytopes, Polyhedra, Combinatorial Topology, Voronoi Diagrams and Delaunay Triangulations .

Gardner, R. J. (2006). *Geometric tomography*, volume 58 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, New York, second edition.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Gillies, S. et al. (2007). Shapely: manipulation and analysis of geometric objects.

Giraud, C. and Verzelen, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Math. Stat. Learn.*, 1(3-4):317–374.

Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields*, 159(3-4):479–543.

Goodman, J. E., O'Rourke, J., and Tóth, C. D., editors (2018). *Handbook of discrete and computational geometry*. Discrete Mathematics and its Applications (Boca Raton). CRC Press, Boca Raton, FL. Third edition of [MR1730156].

Grenander, U. (1981). *Abstract inference*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Grotzinger, S. J. and Witzgall, C. (1984). Projections onto order simplexes. *Applied Mathematics & Optimization*, 12(1):247–270.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statist. Sci.*, 33(4):568–594.

Harman, R. and Lacko, V. (2010). On decompositional algorithms for uniform sampling from n-spheres and n-balls . *J. Multivariate Anal.*, 101(10):2297–2304.

Has'minskiĭ, R. Z. (1978). A lower bound for risks of nonparametric density estimates in the uniform metric . *Teor. Veroyatnost. i Primenen.*, 23(4):824–828.

Has'minskiĭ, R. Z. and Ibragimov, I. A. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory . *Ann. Statist.*, 18(3):999–1010.

Heckbert, P. S. (2013). *Graphics gems*, volume 4. Elsevier.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1977). Estimation of infinite-dimensional parameter in Gaussian white noise. *Dokl. Akad. Nauk SSSR*, 236(5):1053–1055.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1978). Capacity in the transmission of smooth signals. *Dokl. Akad. Nauk SSSR*, 242(1):32–35.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin. Asymptotic theory, Translated from the Russian by Samuel Kotz.

Ibragimov, I. A. and Khas'minskij, R. Z. (1980). On the estimation of the distribution density. *Zap. Nauchn. Semin. Leningr. Otd. Mat. Inst. Steklova*, 98:61–85.

Jones, F. (1993). *Lebesgue integration on Euclidean space*. Jones and Bartlett Publishers, Boston, MA.

Jung, H. (1901). Ueber die kleinste Kugel, die eine räumliche Figur einschliesst. *J. Reine Angew. Math.*, 123:241–257.

Jung, H. W. E. (1910). Über den kleinsten Kreis, der eine ebene Figur einschließt. *J. Reine Angew. Math.*, 137:310–313.

Keener, R. W. (2010). *Theoretical statistics*. Springer Texts in Statistics. Springer, New York. Topics for a core course.

Klemelä, J. S. (2009). *Smoothing of multivariate data: density estimation and visualization*. John Wiley & Sons.

Koldobsky, A., Saroglou, C., and Zvavitch, A. (2016). Estimating volume and surface area of a convex body via its projections or sections .

Laurent, S. (2018). *uniformly: Uniform Sampling*. R package version 0.1.0.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Leonard, I. E. and Lewis, J. E. (2016). *Geometry of convex sets*. John Wiley & Sons, Inc., Hoboken, NJ.

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Stat.*, 4(1):66–70.

Lindahl, L.-A. (2016). Convexity and Optimization.

Maruyama, Y. and Strawderman, W. E. (2021). A Gaussian sequence approach for proving minimaxity: a review. *J. Statist. Plann. Inference*, 211:256–270.

Mathai, A. M. (1999). *An introduction to geometrical probability*, volume 1 of *Statistical Distributions and Models with Applications*. Gordon and Breach Science Publishers, Amsterdam. Distributional aspects with applications.

Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104.

Milbrodt, H. (1987). Minimaxity of Pitman estimators. *J. Multivariate Anal.*, 22(1):126–136.

Mordukhovich, B. S. and Nam, N. M. (2014). *An easy path to convex analysis and applications*, volume 14 of *Synthesis Lectures on Mathematics and Statistics*. Morgan & Claypool Publishers, Williston, VT.

Muller, M. E. (1959). A Note on a Method for Generating Points Uniformly on N-Dimensional Spheres . *Commun. ACM*, 2(4):19–20.

Németh, A. B. and Németh, S. Z. (2012). How to project onto the monotone nonnegative cone using Pool Adjacent Violators type algorithms .

Neykov, M. (2022). On the minimax rate of the Gaussian sequence model under bounded convex constraints . *IEEE Transactions on Information Theory*.

Niculescu, C. P. and Persson, L.-E. (2018). *Convex functions and their applications*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham. A contemporary approach, Second edition of [MR2178902].

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. R package version 1.1.1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pitman, E. J. G. (1939a). The estimation of the location and scale parameters of a continuous population of any given form . *Biometrika*, 30:391–421.

Pitman, E. J. G. (1939b). Tests of hypotheses concerning location and scale parameters. *Biometrika*, 31:200–215.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rademacher, L. (2007). Approximating the centroid is hard. In *Computational geometry (SCG'07)*, pages 302–305. ACM, New York.

Rigollet, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli*, 12(2):351–370.

Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280.

Robbins, H. and Zhang, C.-H. (1986a). Maximum likelihood estimation in regression with uniform errors. In *Adaptive statistical procedures and related topics (Upton, N.Y., 1985)*, volume 8 of *IMS Lecture Notes Monogr. Ser.*, pages 365–385. Inst. Math. Statist., Hayward, CA.

Robbins, H. and Zhang, C.-H. (1986b). Maximum likelihood estimation in regression with uniform errors. *Lecture Notes-Monograph Series*, pages 365–385.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

Royer, M. (2017). Adaptive Clustering through Semidefinite Programming. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1795–1803.

Samarov, A. and Tsybakov, A. (2007). Aggregation of density estimators and dimension reduction. In *Advances in statistical modeling and inference*, volume 3 of *Ser. Biostat.*, pages 233–251. World Sci. Publ., Hackensack, NJ.

Schechtman, E. and Schechtman, G. (1986). Estimating the parameters in regression with uniformly distributed errors . *Journal of Statistical Computation and Simulation*, 26(3-4):269–281.

Schymura, D. (2014). An upper bound on the volume of the symmetric difference of a body and a congruent copy . *Adv. Geom.*, 14(2):287–298.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Shirali, S. and Vasudeva, H. L. (2006). *Metric spaces*. Springer-Verlag London, Ltd., London.

Shrotriya, S. and Neykov, M. (2022a). Revisiting Le Cam's Equation: Exact Minimax Rates over Convex Density Classes.

Shrotriya, S. and Neykov, M. (2022b). Adversarial Sign-Corrupted Isotonic Regression.

Shrotriya, S. and Neykov, M. (2022c). Uniform Location Estimation on Convex Bodies.

Strasser, H. (1982). Local asymptotic minimax properties of Pitman estimates. *Z. Wahrsch. Verw. Gebiete*, 60(2):223–247.

Strawderman, W. E. (2000). Minimaxity. *J. Amer. Statist. Assoc.*, 95(452):1364–1368.

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, 53(1):54–61.

Tsagris, M., Beneki, C., and Hassani, H. (2014). On the Folded Normal Distribution. *Mathematics*, 2(1):12–28.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.*, 18(2):907–924.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators . *Ann. Statist.*, 21(1):14–44.

van Eeden, C. (1956). *Maximum likelihood estimation of partially or completely ordered parameters* . Statist. Afdeling Rep. S 207 (VP 9). Math. Centrum Amsterdam.

van Eeden, C. (1957a). A least squares inequality for maximum likelihood estimates of ordered parameters . *Nederl. Akad. Wetensch. Proc. Ser. A* ***60*** *= Indag. Math.*, 19:513–521.

van Eeden, C. (1957b). Maximum likelihood estimation of partially or completely ordered parameters. I . *Nederl. Akad. Wetensch. Proc. Ser. A.* ***60*** *= Indag. Math.*, 19:128–136.

van Eeden, C. (1957c). Maximum likelihood estimation of partially or completely ordered parameters. II . *Nederl. Akad. Wetensch. Proc. Ser. A.* ***60*** *= Indag. Math.*, 19:201–211.

van Eeden, C. (1958). *Testing and estimating ordered parameters of probability distributions* . Mathematical Centre, Amsterdam.

van Eeden, C. (2006). *Restricted parameter space estimation problems*, volume 188 of *Lecture Notes in Statistics*. Springer, New York. Admissibility and minimaxity properties.

Vershynin, R. (2018). *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. An introduction with applications in data science, With a foreword by Sara van de Geer.

von Neumann, J. (1950). *Functional Operators. II. The Geometry of Orthogonal Spaces*. Annals of Mathematics Studies, No. 22. Princeton University Press, Princeton, N. J.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, S. and Marzouk, Y. (2022). On minimax density estimation via measure transport.

Wang, Y. (1996). The L_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods*, 25(2):281–294.

Wasserman, L. (2004). *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York. A concise course in statistical inference.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

Wasserman, L. (2018). Minimax Theory .

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA*, 117(29):16880–16890.

Whittinghill, D. C. and Hogg, R. V. (2001). A Little Uniform Density with Big Instructional Potential. *Journal of Statistics Education*, 9(2).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.1.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs . *The Annals of Statistics*, pages 339–362.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy . *Ann. Statist.*, 13(2):768–774.

Yi, Y. and Neykov, M. (2021). Non-Asymptotic Bounds for the ℓ_∞ Estimator in Linear Regression with Uniform Noise .

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555.

Ziegler, G. M. (1995). *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

A

Data and Code Availability

A.1 Data and Code Availability

In order to reproduce the simulation results in Chapters 3 and 4, we note the following:

- Data and code used for the analysis in Chapter 3 is available online at: https://github.com/shamindras/ascifit
- Data and code used for the analysis in Chapter 4 is available online at: https://github.com/shamindras/ule