Title
Information Theory, Dimension Reduction and Density Estimation

Author
Saha, Sujayam

Publication Date
2018

Peer reviewed|Thesis/dissertation

Information Theory, Dimension Reduction and Density Estimation

by

Sujayam Saha

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Co-chair
Professor Aditya Guntuboyina, Co-chair
Professor Peter J. Bickel
Professor Lexin Li

Spring 2018

Information Theory, Dimension Reduction and Density Estimation

Abstract

Information Theory, Dimension Reduction and Density Estimation

by

Sujayam Saha

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Co-chair

Professor Aditya Guntuboyina, Co-chair

This thesis documents three different contributions in statistical learning theory. They were developed with careful emphasis on addressing the demands of modern statistical analysis upon large-scale modern datasets. The contributions concern themselves with advancements in information theory, dimension reduction and density estimation - three foundational topics in statistical theory with a plethora of applications in both practical problems and development of other aspects of statistical methodology.

In Chapter 2, I describe the development of an unifying treatment of the study of inequalities between f-divergences, which are a general class of divergences between probability measures which include as special cases many commonly used divergences in probability, mathematical statistics and information theory such as Kullback-Leibler divergence, chi-squared divergence, squared Hellinger distance, total variation distance etc. In contrast with previous research in this area, we study the problem of obtaining sharp inequalities between f-divergences in full generality. In particular, our main results allow m to be an arbitrary positive integer and all the divergences D_f and $D_{f_1}, \ldots, D_{f_m}$ to be arbitrary f-divergences. We show that the underlying optimization problems can be reduced to low-dimensional optimization problems and we outline methods for solving them. We also show that many of the existing results on inequalities between f-divergences can be obtained as special cases of our results and we also improve on some existing non-sharp inequalities.

In Chapter 3, I describe the development of a new dimension reduction technique specially suited for interpretable inference in supervised learning problems involving large-dimensional data. This new technique, Supervised Random Projections (SRP), is introduced with the goal of ensuring that in comparison to ordinary dimension reduction, the compressed data is more relevant to the response variable at hand in a supervised learning problem. By incorporating variable importances, we explicate that the compressed data should still accurately explain the response variable; thus lending more interpretability to the dimension reduction step. Further, variable importances ensure that even in the presence of numerous nuisance parameters, the projected data retains at least a moderate amount of information from the

important variables, thus allowing said important variables a fair chance at being selected by downstream formal tests of hypotheses.

In Chapter 4, I describe the development of several adaptivity properties of the Non-Parametric Maximum Likelihood Estimator (NPMLE) in the problem of estimating an unknown gaussian location mixture density based on independent identically distributed observations. Further, I explore the role of the NPMLE in the problem of denoising normal means, i.e. the problem of estimating unknown means based on observations. This problem has been studied widely. In this problem, I prove that the Generalized Maximum Likelihood Empirical Bayes estimator (GMLEB) approximates the Oracle Bayes estimator at adaptive parametric rates up to additional logarithmic factors in expected squared ℓ_2 norm.

To everyone who made me think

Contents

List of Figures

List of Tables

Acknowledgments

The lion's share of my meager succeses and achievements can directly be attributed to the untiring efforts, guidance, and encouragement set forth by my two stellar and often complementary doctoral advisors, Bin Yu and Aditya Guntuboyina. Through inspiration, example, and curt orders, they have continually shaped and broadened my horizons, both as an academic and as an individual. I will always treasure our interactions and conversations and hope to revisit them through the lens of my memories for years to come, both in times of comfort and of doubt. While I would not deem to belittle their contributions by deigning to enumerate them here, let me assure you that through these two people I have encountered countless thought-provoking human beings, thoughts, and questions.

Besides my advisors, I am extremely grateful to the faculty and department staff at Evans Hall for introducing me to and populating one of the most welcoming, encouraging, and caring environments I have encountered. I can truthfully claim that I have never experienced a disappointing conversation with any faculty and staff member. In particular, I would like to take this moment to acknowledge the time and attention afforded to me by Peter Bickel, David Aldous, Allan Sly, Elchanan Mossel and Ani Adhikari, who deemed me worthy of unassuming dialog. La Shana Porlaris and Mary Melinn are absolute pillars of my continued academic functioning, without them I would have managed to misplace my eyes and limbs within a year of arriving at Berkeley.

I would love to thank Martin Wainwright for agreeing to participate in my qualifying examination and for his engaging and insightful comments and probes. I would further love to thank Peter Bickel and Lexin Li for his continued contributions to my qualifying examination committee and dissertation committee, and for his on-demand encouragement.

At this moment, I would be remiss not to draw attention towards a person whose collaboration and company I have thoroughly enjoyed in the last few years of my stay at Berkeley - Bodhisattva Sen. His infectious optimism and pure glee at mathematical wranglings have helped me persevere through several pesky roadblocks.

I have now come to the juncture, where I must thank all of my friends and peers without whose incessant support everything would have been completely different. But there are so many of you! Know that if you don't find your name pencilled in here in this sorry excuse of a note of gratitude, it is entirely due to my own shortcomings. Perhaps your influences have become so ingrained in my psyche that I barely remember that once I did not think and act this way, that someone had to teach me these ropes.

Thank you Christine Kuang, for everything and everything to come. Thank you Funan Shi, for being my artistic clone. Thank you Siqi Wu and Riddhipratim Basu, for knowing everything. Thank you Sayantan Mukhopadhyay, for knowing everything and then pretending that I could possibly know better. Thank you Sumanta Basu and Nirupam Chakravorty, for being the best elder brothers. Thank you Belle Peng, for your crazed sanity. Thank you Sharmodeep Bhattacharya and Sayak Ray, for being my parents halfway around the world. Thank you Karl Kumbier, for never being afraid. Thank you Adam Bloniarz, for always being calm. Thank you Lisha Li, for classing up all that you touched. Thank you Hyesoo

Choi, for always trusting me. Thank you Katarina Slama, for our shared introversion. Thank you Ryan Giordano, for etching out the balance. Thank you Geoff Schiebinger, for pushing me further. Thank you Angie Zhu, for always making the time. Thank you Rachel Wang, for being a kindred spirit. To Sujatro Chakladar, Sandeepan Parekh, Samit Roy, Tamojoy Ghosh and Kuppili Abhishek, I hope I don't actually have to say anything.

I want to thank each and every member and ex-member of Yu Group, for their unwavering support and insistence on highlighting the best in me; for all the stimulating conversations and company.

I want to thank Douglas Adams for realising that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizzare and inexplicable. And also, for realising that this has already happened.

Lastly, I want to thank my parents for always supporting and believing in me. I will never fully comprehend all that you endured to make all of this possible. I also want to thank my extended family for everything they have done and all the kind thoughts.

Chapter 1

Introduction

This thesis documents three projects I have had the pleasure of developing during the course of my doctoral study. Broadly, these three projects fall under the fold of various facets of statistical learning theory: information theory, dimension reduction and density estimation. These three projects were, from beginning to end, motivated and fueled by a desire to develop statistical methodology addressing the needs and convenience of the modern statistician and modern datasets; coupled with a further urge to establish theoretical properties of such methodology, wherever possible.

Information theory, introduced in Shannon's seminal thesis, is at the core of numerous fields including signal processing, computer science, statistics and mathematics. More specifically and as it relates to my own contribution, the study of divergences between probability distributions arises routinely in statistical learning theory.

Dimension reduction, in its various forms and innovations and sometimes under the guise of data compression, is an indisposable step in most analyses of massive datasets. In many modern domains of practice, the speed of data accumulation and subsequent storage has outstripped the limits of memory and processing power of a standard computing unit. Dimension reduction is a core necessity in such datasets or streams. Often, the efficacy of downstream analyses is heavily influenced by the particulars of data compression or dimension reduction.

Likewise, mixture modelling, and more specifically the study of Gaussian Mixture Models (GMM), is a powerful and staple technique for easily describing populations comprised of multiple subpopulations. Modern datasets frequently exhibit clear heterogeneity which provides a basis for interpreting such populations, be it various kinds of users in a social network, subspecies in a ecology, or objects in an image.

Information Theory

f-divergences are a general class of divergences between probability measures which include as special cases many commonly used divergences in probability, mathematical statistics

and information theory such as Kullback-Leibler divergence, chi-squared divergence, squared Hellinger distance, total variation distance etc. Inequalities between f-divergences are useful in many areas. For example, in mathematical statistics, they are crucial in problems of obtaining minimax bounds Yu [146], Tsybakov [135], Guntuboyina [53], and Guntuboyina [52]. In probability, such inequalities are often used for converting limit theorems proved under a convenient divergence into limit theorems for other divergences Barron [7], Topsøe [133], and Harremoës [54]. They are also helpful for proving results in measure concentration Marton [91, 89, 90]. Some applications in machine learning are described in Reid and Williamson [112]. Further, inequalities involving f-divergences are fundamental to the field of information theory Cover and Thomas [29] and Csiszár and Shields [33].

As such, the study of sharp inequalities between f-divergences is of foundational importance informing research and understanding in the subjects mentioned above. Thus motivated and in collaboration with Aditya Guntuboyina and Geoffrey Schiebinger, I have studied the problem of maximizing or minimizing an f-divergence between two probability measures subject to a finite number of constraints on other f-divergences. We show that these infinite-dimensional optimization problems can all be reduced to optimization problems over small finite dimensional spaces which are tractable. Our results lead to a comprehensive and unified treatment of the problem of obtaining sharp inequalities between f-divergences. We demonstrate that many of the existing results on inequalities between f-divergences can be obtained as special cases of our results and we also improve on some existing non-sharp inequalities. Complete details on this work can be found in Chapter 2.

Dimension Reduction

High-dimensional supervised learning problems are encountered in numerous popular modern-day scientific applications, ranging from genomics, biomedical studies, astronomy and sociological studies. In each of these fields, the core goal of statistical analysis requires inference or uncertainty measures for decision making. As such, a holistic framework for statistical inference in high-dimensional supervised learning problems serves a paramount advantage to practitioners. Under the guidance of my docotoral advisors, I have considered supervised learning problems such as regression, classification and randomized experiments of a high-dimensional nature, viz. in datasets where numerous covariables are available for consideration, often comparable or exceeding the number of samples/observations. To this end, we first develop a new dimension reduction technique called *Supervised Random Projections* (SRP) and further, develop a framework for statistical inference in high-dimensional supervised learning problems based on SRP.

Dimension reduction is a crucial and necessary step towards meaningful and reproducible statistical analyses on modern massive and complex datasets. Ordinarily, dimension reduction techniques such as random projections treat each dimension of the input data with equal importance. However, in supervised learning problems, not all variables/features are equally important. A dimension reduction scheme designed specifically for supervised problems

should attempt to preserve important variables (ones that influence the response strongly) at the expense of less important ones.

Further, in statistical literature, hypothesis testing is widely popular for interpretation and refinement of statistical models in supervised learning problems, for instance identifying statistically significant features. However, hypotheses testing procedures are often encumbered by heavy computational overload and the lack of a holistic approach which can be applied to many varied problems.

Thus motivated and the advisement of my doctoral advisors, I have introduced the idea of supervised dimension reduction, with the goal of ensuring that in comparison to ordinary dimension reduction, the projected data is more relevant to the response variable at hand. By incorporating variable importances, we explicate that the projected data should still accurately explain the response variable (this is in contrast to ordinary dimension reduction, where one only attempts to preserve the geometry between covariables); thus lending more interpretability to the dimension reduction step. Further, variable importances ensure that even in the presence of numerous nuisance parameters, the projected data retains at least a moderate amount of information from the important variables, thus allowing said important variables a fair chance at being selected by downstream formal tests of hypotheses.

Complete details on this subject can be found in Chapter 3 on this thesis.

Density Estimation

In unsupervised learning problems such as clustering, it is ubiquitous to model observed multi-dimensional data as a mixture of random vectors distributed as Gaussian with unknown parameters. Estimating the underlying mixture density of these observations is a core problem leading to downstream analyses such as clustering, classification and denoising. Usual approaches to Gaussian mixture density estimation maximize likelihood over Gaussian mixtures with a fixed number of components. This approach results in a non-convex optimization problem and also needs to know the number of components. An approach to Gaussian mixture density estimation that aims to circumvent these issues is nonparametric maximum likelihood estimation which goes back to [67].

In collaboration with Prof. Adityanand Guntuboyina, I have studied the Nonparametric Maximum Likelihood Estimator (NPMLE) for fitting Gaussian mixture densities. Under the assumption that the covariance matrix of the components is identity, the NPMLE maximizes likelihood over the class of all Gaussian location mixture densities i.e., densities of the form $f_G(x) := \int \phi_d(x-\theta)dG(\theta)$ as G varies over all probability measures on $\mathbb{R}^d$ (this can be modified in the case of well-conditioned unknown covariance provided a lower bound on the eigenvalues of all the component-covariance matrices is available). The above NPMLE was first introduced in [67] and more recently major advancements have been made in the efficient implementation of this estimator via convex optimization ([68]). Several book length treatments are also available on the subject of this NPMLE (see, for example, [80, 18]). In collaboration, I have established several adaptivity properties of the NPMLE. I have proved

that in expected squared Hellinger accuracy, the NPMLE based on n observations estimates the unknown mixture density composed of k components almost at the parametric rate k/n with an additional multiplicative logarithmic factor depending on n, *without a priori knowledge of the number of components* k.

Further, I have explored the role of the NPMLE in the problem of denoising normal means, i.e. the problem of estimating unknown means based on observations. This problem has been studied widely, leading back to [61]. [63] introduced the Generalized Maximum Likelihood Empirical Bayes estimator (GMLEB) for this problem; which is the Bayes estimator where the NPMLE is used as a plug-in estimate for the unknown mixture density. I have proved that the GMLEB approximates the Oracle Bayes estimator at adaptive parametric rates up to additional logarithmic factors in expected squared ℓ_2 norm. Further, the analogous extension to compactly supported mixing density G is also rigorized. Figure 4.1 serves as a short illustration of the accuracy with which the Empirical Bayes estimate (in red) approximates the Oracle Bayes estimate (in blue) whenever G contains some basic structure. The most noteworthy fact here is that the Empirical Bayes estimates require no knowledge of the underlying structure, for instance concentric circles, or triangle or a letter of the alphabet, etc. In fairness, I should also note that the noise distribution was completely specified in these illustrations, including the noise level. Recently, the denoising problem stated here has also been investigated in the field of convex clustering [128, 109, 28]. To my knowledge, no analogue of our results are available for these methods. Complete details on this work are presented in Chapter 4.

Chapter 2

Sharp inequalities for f-divergences

2.1 Overview

f-divergences are a general class of divergences between probability measures which include as special cases many commonly used divergences in probability, mathematical statistics and information theory such as Kullback-Leibler divergence, chi-squared divergence, squared Hellinger distance, total variation distance etc. In this paper, we study the problem of maximizing or minimizing an f-divergence between two probability measures subject to a finite number of constraints on other f-divergences. We show that these infinite-dimensional optimization problems can all be reduced to optimization problems over small finite dimensional spaces which are tractable. Our results lead to a comprehensive and unified treatment of the problem of obtaining sharp inequalities between f-divergences. We demonstrate that many of the existing results on inequalities between f-divergences can be obtained as special cases of our results and we also improve on some existing non-sharp inequalities.

2.2 Introduction

Suppose that the Kullback-Leibler divergence between two probability measures is bounded from above by 2. What then is the maximum possible value of the Hellinger distance between them? Such questions naturally arise in many fields including mathematical statistics and machine learning, information theory, probability, statistical physics etc. and the goal of this paper is to provide a way of answering them. From the variational viewpoint, this problem can be posed as: maximize the Hellinger distance subject to a constraint on the Kullback-Leibler divergence over the space of all pairs of probability measures *over all possible sample spaces*. We shall prove in this paper that the value of this maximization problem remains unchanged if one restricts the sample space to be the three-element set $\{1, 2, 3\}$. In other words, in order to find the maximum Hellinger distance subject to an upper bound on the Kullback-Leibler divergence, one can just restrict attention to pairs of probability measures on $\{1, 2, 3\}$. Thus, the large infinite-dimensional optimization problem is reduced to an

optimization problem over a small finite-dimensional space (of dimension ≤ 4) which makes it tractable.

In this paper, we prove such results in a very general setting. The Kullback-Leibler divergence and the (square of the) Hellinger distance are special instances of a general class of divergences between probability measures called f-divergences (also known as ϕ-divergences). Let $f : (0, \infty) \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. By virtue of convexity, both the limits $f(0) := \lim_{x \downarrow 0} f(x)$ and $f'(\infty) := \lim_{x \uparrow \infty} f(x)/x$ exist, although they may equal $+\infty$. For two probability measures P and Q, the f-divergence (see, for example, Ali and Silvey [3] and Csiszar [30, 31, 32]), $D_f(P||Q)$, is defined by

$$D_f(P||Q) := \int_{q>0} f\left(\frac{p}{q}\right) dQ + f'(\infty)P\{q = 0\}$$

where p and q are densities of P and Q with respect to a common measure λ. The definition does not depend on the choice of the dominating measure λ. Special cases of f lead to, among others, Kullback-Leibler divergence, total variation distance, square of the Hellinger distance and chi-squared divergence.

We are now ready to introduce the general form of the optimization problem we described at the beginning of the paper. Given divergences D_f and $D_{f_i}, i = 1, \ldots, m$ and nonnegative real numbers $D_1, \ldots, D_m$, let

$$A(D_1, \ldots, D_m) := \sup \{D_f(P||Q) : D_{f_i}(P||Q) \leq D_i \ \forall i\}$$

and

$$B(D_1, \ldots, D_m) := \inf \{D_f(P||Q) : D_{f_i}(P||Q) \geq D_i \ \forall i\}$$

where the probability measures on the right hand sides above range over all possible measurable spaces. The goal of this paper is to provide a method for computing these quantities. We show that these large infinite-dimensional optimization problems can all be reduced to optimization problems over small finite-dimensional spaces. Specifically, in Theorem 2.3.1, we show that in order to compute these quantities, one can restrict attention to probability measures on the set $\{1, \ldots, m + 2\}$.

One of the main reasons for studying the quantities $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ is that they yield sharp inequalities for the divergence D_f in terms of the divergences $D_{f_1}, \ldots, D_{f_m}$. Indeed, the inequalities

$$D_f(P||Q) \leq A(D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q)) \tag{2.1}$$

and

$$D_f(P||Q) \geq B(D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q)) \tag{2.2}$$

hold for every pair of probability measures P and Q. Further, the functions A and B satisfy the natural monotonicity inequalities

$$A(D_1, \ldots, D_m) \leq A(D'_1, \ldots, D'_m) \tag{2.3}$$

and

$$B(D_1, \ldots, D_m) \leq B(D'_1, \ldots, D'_m) \tag{2.4}$$

for every $(D_1, \ldots, D_m)$ and $(D'_1, \ldots, D'_m)$ such that $D_i \leq D'_i$ for all i.

The inequalities (2.1) and (2.2) are sharp in the sense that A is the *smallest* function satisfying (2.3) for which (2.1) holds for all probability measures P and Q. Likewise, B is the *largest* function satisfying (2.4) for which (2.2) holds for all probability measures P and Q.

Inequalities between f-divergences are useful in many areas. For example, in mathematical statistics, they are crucial in problems of obtaining minimax bounds Yu [146], Tsybakov [135], Guntuboyina [53], and Guntuboyina [52]. In probability, such inequalities are often used for converting limit theorems proved under a convenient divergence into limit theorems for other divergences Barron [7], Topsøe [133], and Harremoës [54]. They are also helpful for proving results in measure concentration Marton [91, 89, 90]. Some applications in machine learning are described in Reid and Williamson [112]. Further, inequalities involving f-divergences are fundamental to the field of information theory Cover and Thomas [29] and Csiszár and Shields [33].

Because of their widespread use, many papers deal with inequalities between f-divergences (some references being Pinsker [108], Csiszar [30], Kullback [69], Kemperman [66], Vajda [137], Gibbs and Su [50], Fedotov, Harremoës, and Topsoe [44], Topsøe [134], Gilardoni [51], Reid and Williamson [113], and Guntuboyina [52]). However, many of the inequalities presented in previous treatments are not sharp. The few papers which provide sharp inequalities Vajda [137], Fedotov, Harremoës, and Topsoe [44], Gilardoni [51], and Reid and Williamson [113] only deal with certain special f-divergences as opposed to working in full generality. A popular such special case is $m = 1$ and D_{f_1} corresponding to the total variation distance. In this case, sharp inequalities have been derived in Fedotov, Harremoës, and Topsoe [44] for the case when D_f is the Kullback-Leibler divergence and in Gilardoni [51] for the case of general D_f. The case $m > 1$ is comparatively less studied although this has potential applications in the statistical problem of obtaining lower bounds for the minimax risk (see Section 2.7.1 for details). The only paper which deals with sharp inequalities for $m > 1$ is Reid and Williamson [113] but there the authors only study the case when $D_{f_1}, \ldots, D_{f_m}$ are all primitive divergences (see Remark 2.4.2 below for the definition of primitive divergences).

In contrast with all previous papers in the area, we study the problem of obtaining sharp inequalities between f-divergences in full generality. In particular, our main results allow m to be an arbitrary positive integer and all the divergences D_f and $D_{f_1}, \ldots, D_{f_m}$ to be arbitrary f-divergences. We show that the underlying optimization problems can all be reduced to low-dimensional optimization problems and we outline methods for solving them. We also show that many of the existing results on inequalities between f-divergences can be obtained as special cases of our results and we also improve on some existing non-sharp inequalities.

The rest of this paper is structured as follows. Our main result is stated in Theorem 2.3.1. Its three-part proof is given in Section 2.4. The first part is based on a recent representation theorem for f-divergences which implies that the optimization problems for computing $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ can be thought of as maximizing or minimizing an integral functional over a certain class of concave functions satisfying a finite number of integral constraints. In the second part of the proof, we use Choquet's theorem to restrict attention only to the extreme points of the constraint set. Finally, in the third part, we characterize these extreme points and show that they correspond to probability measures over small finite sets.

One possible approach to compute $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ is via joint ranges of f-divergences. Specifically, for $m \geq 1$ and divegences $D_{f_1}, \ldots, D_{f_m}$, their joint range, denoted by $\mathcal{R}(f_1, \ldots, f_m)$ is defined as the set of all vectors in $\mathbb{R}^m$ that equal $(D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q))$ for some pair of probability measures P and Q. If the joint range $\mathcal{R}(f, f_1, \ldots, f_m)$ can be determined, then one can easily calculate the values $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ for every $D_1, \ldots, D_m$. The problem of determining the joint range $\mathcal{R}(f_1, \ldots, f_m)$ was solved for the case $m = 2$ in Harremoes and Vajda [55]. We extend their result to general $m \geq 2$ in Section 2.5.2 by a very simple proof which was communicated to us by an anonymous referee. Unfortunately, it turns out that this approach based on the joint range does not quite prove Theorem 2.3.1. It gives a slightly weaker result. We discuss this in Section 2.5.2.

Also in Section 2.5, we collect some remarks and extensions of our main theorem and, in particular, we show that the theorem is tight in general. In Section 2.6, we consider various special cases and show that many well-known results in the literature can be obtained as simple instances of our main theorem. In Section 2.7, we describe numerical methods for solving the low-dimensional optimization problems that come out of our main theorem. We solve an important subclass of these problems by convex optimization and we also describe heuristic methods for the general case.

2.3 Main Result

For each $n \geq 1$, let $\mathcal{P}_n$ denote the space of all probability measures defined on the finite set $\{1, \ldots, n\}$. Let us define $A_n(D_1, \ldots, D_m)$ to be

$$\sup \{D_f(P||Q) : P, Q \in \mathcal{P}_n \text{ and } D_{f_i}(P||Q) \leq D_i \ \forall i\}$$

and, analogously, $B_n(D_1, \ldots, D_m)$ to be

$$\inf \{D_f(P||Q) : P, Q \in \mathcal{P}_n \text{ and } D_{f_i}(P||Q) \geq D_i \ \forall i\}.$$

Our main theorem is given below. The second part of the theorem requires that $D_{f_1}, \ldots, D_{f_m}$ are finite divergences. We say that a divergence D_f is finite if $\sup_{P,Q} D_f(P||Q) < \infty$. The supremum here is taken over all probability measures over all possible measurable spaces. See Remark 2.4.3 for a detailed explanation of finite divergences.

Theorem 2.3.1. *For every $D_1, \ldots, D_m \geq 0$, we have*

$$A(D_1, \ldots, D_m) = A_{m+2}(D_1, \ldots, D_m). \tag{2.5}$$

Further if $D_{f_1}, \ldots, D_{f_m}$ are all finite, then

$$B(D_1, \ldots, D_m) = B_{m+2}(D_1, \ldots, D_m). \tag{2.6}$$

The conclusions of the above theorem may be better appreciated in the following optimization form. Theorem 2.3.1 states that the quantity $A(D_1, \ldots, D_m)$ equals the optimal value of the following finite-dimensional optimization problem:

$$\begin{aligned} \underset{p,q \in [0,1]^{m+2}}{\text{maximize}} \quad & \sum_{j:q_j>0} q_j f\left(\frac{p_j}{q_j}\right) + f'(\infty) \sum_{j:q_j=0} p_j \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for all } j = 1, \ldots, m+2 \\ & \sum p_j = \sum q_j = 1 \\ & \sum_{j:q_j>0} q_j f_i\left(\frac{p_j}{q_j}\right) + f_i'(\infty) \sum_{j:q_j=0} p_j \leq D_i \end{aligned} \tag{2.7}$$

for $i = 1, \ldots, m$. Similarly, when $D_{f_1}, \ldots, D_{f_m}$ are all finite, $B(D_1, \ldots, D_m)$ equals the optimal value of

$$\begin{aligned} \underset{p,q \in [0,1]^{m+2}}{\text{minimize}} \quad & \sum_{j:q_j>0} q_j f\left(\frac{p_j}{q_j}\right) + f'(\infty) \sum_{j:q_j=0} p_j \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for all } j = 1, \ldots, m+2 \\ & \sum p_j = \sum q_j = 1 \\ & \sum_{j:q_j>0} q_j f_i\left(\frac{p_j}{q_j}\right) + f_i'(\infty) \sum_{j:q_j=0} p_j \geq D_i \end{aligned} \tag{2.8}$$

for $i = 1, \ldots, m$. The proof of Theorem 2.3.1 is provided in the next section. In Section 2.5, we argue that Theorem 2.3.1 is tight in general and also comment on the assumption of finiteness of $D_{f_1}, \ldots, D_{f_m}$ for the validity of identity (2.6). We also describe an attempt to prove this theorem via joint ranges but this only yields a weaker result.

2.4 Proof of the Main Result

2.4.1 Testing Representation

For two probability measures P and Q, let us define the function $\psi_{P,Q} : [0, \infty) \to [0, 1]$ by

$$\psi_{P,Q}(s) := \int \min(p, qs) d\lambda \qquad \text{for } s \in [0, \infty)$$

where p and q denote the densities of P and Q with respect to a common measure λ (which can, for example, be taken to be $P+Q$). This function $\psi_{P,Q}$ is nonnegative, concave, non-decreasing and satisfies the inequality $0 \leq \psi_{P,Q}(s) \leq \min(1,s)$ for all $s \geq 0$. In other words, $\psi \in \mathcal{C}$ where $\mathcal{C}$ denotes the class of all functions ψ on $[0,\infty)$ that are nonnegative, concave, non-decreasing and satisfy the inequality $\psi(s) \leq \min(1,s)$ for all $s \geq 0$. Moreover, it is true (see, for example, Reid and Williamson [113, Corollary 5]) that every function $\psi \in \mathcal{C}$ equals $\psi_{P,Q}$ for some pair of probability measures P and Q.

For each divergence D_f, let us associate the measure ν_f on $(0,\infty)$ defined by

$$\nu_f(a,b] := \partial^r f(b) - \partial^r f(a) \qquad \text{for } 0 < a < b < \infty$$

where ∂^r denotes the right derivative operator (note that by convexity $\partial^r f(x)$ exists for every $x \in (0,\infty)$). We also associate the functional $I_f : \mathcal{C} \to [0,\infty]$ by

$$I_f(\psi) := \int_0^\infty (\min(1,s) - \psi(s))\, d\nu_f(s). \tag{2.9}$$

There is a precise connection between D_f and I_f that is given below:

Lemma 2.4.1. *For every pair of probability measures P and Q, we have*

$$D_f(P||Q) = I_f(\psi_{P,Q}). \tag{2.10}$$

Lemma 2.4.1 is not new although the form in which it is stated above is non-standard. The more standard version simply involves writing the integral in (2.9) over the interval $(0,1)$ by the change of variable $t = s/(1+s)$. In this modified form, Lemma 2.4.1 has been proved in Osterreicher and Vajda [103] in the case when f is twice differentiable and in Liese and Vajda [78] in the general case. A short proof is available in Liese [77, Theorem 2.3].

Remark 2.4.2 (Primitive f-divergences)**.** *For each $s > 0$, let $u_s(t) := \min(1,s) - \min(t,s)$ for $t \in (0,\infty)$. Clearly, u_s is a convex function on $(0,\infty)$ such that $u_s(1) = 0$. Moreover, it is a very simple convex function in the sense that it is piecewise linear with just two linear parts. It is straightforward to check that the divergence corresponding to u_s is given by:*

$$D_{u_s}(P||Q) = \min(1,s) - \psi_{P,Q}(s).$$

Lemma 2.4.1 therefore asserts that any arbitrary f-divergence can be written as an integral of the primitive divergences D_{u_s} with respect to the measure ν_f on $(0,\infty)$. The most well-known of these primitive divergences is the total variation distance which corresponds to $s = 1$. Indeed,

$$D_{u_1}(P||Q) = 1 - \int \min(p,q) d\lambda = \frac{1}{2}\int |p-q| d\lambda =: V(P,Q)$$

Every primitive divergence $D_{u_s}(P||Q)$ is closely related to the smallest weighted average error (Bayes risk) in the problem of statistical testing between the hypotheses P against Q based on an observation X (see, for example, Reid and Williamson [113, Lemma 3]).

Remark 2.4.3 (Finiteness of a divergence)**.** *Lemma 2.4.1 imples that*

$$\sup_{P,Q} D_f(P||Q) = \int_0^\infty \min(1,s)d\nu_f(s) = f(0) + f'(\infty). \tag{2.11}$$

The supremum above is taken over all probability measures P and Q defined on all possible measurable spaces. To see (2.11)*, just note that, by Lemma 2.4.1, we have*

$$\sup_{P,Q} D_f(P||Q) = \sup_{P,Q} I_f(\psi_{P,Q}) = \sup_{\psi\in\mathcal{C}} I_f(\psi) = I_f(0).$$

Intuitively, $\psi_{P,Q}(s) = 0$ for all s implies that P and Q are maximally separated (mutually singular) and thus the maximum value of $I_f(\psi)$ is achieved when ψ is the identically zero function. The definition of I_f gives that

$$I_f(0) = \int_0^\infty \min(1,s)d\nu_f(s)$$

Moreover, for the probability measures $P^* = (1,0)$ and $Q^* = (0,1)$ in $\mathcal{P}_2$, the function $\psi_{P,Q}$ equals 0. Therefore,

$$I_f(0) = D_f(P^*||Q^*) = f(0) + f'(\infty),$$

which proves (2.11)*.*

Recall that an f-divergence is finite if $\sup_{P,Q} D_f(P||Q) < \infty$. By (2.11)*, an f-divergence is finite if and only if*

$$\int_0^\infty \min(1,s)d\nu_f(s) = f(0) + f'(\infty) < \infty. \tag{2.12}$$

Well known examples of finite divergences are the primitive divergences, the square of the Hellinger distance and the capacitory discrimination (which corresponds to the convex function (2.53)*).*

For each f and $D \geq 0$, let us define

$$\mathcal{C}_1(f,D) := \{\psi \in \mathcal{C} : I_f(\psi) \leq D\}$$

and

$$\mathcal{C}_2(f,D) := \{\psi \in \mathcal{C} : I_f(\psi) \geq D\}$$

As a consequence of Lemma 2.4.1, we obtain that

$$A(D_1,\ldots,D_m) = \sup\{I_f(\psi) : \psi \in \cap_{i=1}^m \mathcal{C}_1(f_i,D_i)\} \tag{2.13}$$

and

$$B(D_1,\ldots,D_m) = \inf\{I_f(\psi) : \psi \in \cap_{i=1}^m \mathcal{C}_2(f_i,D_i)\}. \tag{2.14}$$

The following lemma on the derivatives of the function $\psi_{P,Q}$ (the left and right derivative operators are denoted by ∂^l and ∂^r respectively) will be useful in the sequel.

Lemma 2.4.4. *For every function $\psi = \psi_{P,Q}$ in $\mathcal{C}$, we have*

$$\partial^l \psi(s) = Q\{p \geq sq\} \qquad \text{for } s > 0 \tag{2.15}$$

and

$$\partial^r \psi(s) = Q\{p > sq\} \qquad \text{for } s \geq 0. \tag{2.16}$$

Proof. For every $s > 0$,

$$\partial^l \psi(s) = \lim_{\epsilon \downarrow 0} \frac{\psi(s) - \psi(s-\epsilon)}{\epsilon}$$

and

$$\frac{\psi(s) - \psi(s-\epsilon)}{\epsilon} = \int \frac{\min(p, qs) - \min(p, q(s-\epsilon))}{\epsilon} d\lambda$$

It is easy to check that the integrand above is bounded in absolute value by q and converges as $\epsilon \downarrow 0$ to $q\{p \geq qs\}$. The identity (2.15) therefore follows by the dominated convergence theorem. The proof of (2.16) is similar. □

2.4.2 Reduction to Extreme Points

Let us first recall the definition of extreme points. Let S be a subset of a vector space V. A point $a \in S$ is called an extreme point of S if $a = (b+c)/2$ for $b, c \in S$ implies that $a = b = c$. In other words, a cannot be the mid-point of a non-trivial line segment whose end points lie in S. We denote the set of all extreme points of S by $ext(S)$.

An important result about extreme points in infinite dimensional topological vector spaces is Choquet's theorem (see, for example, Phelps [106, Chapter 3]). We shall use the following version of Choquet's theorem in this section:

Theorem 2.4.5 (Choquet)**.** *Let K be a metrizable, compact convex subset of a locally convex space V and let x_0 be an element of K. Then there exists a Borel probability measure μ_0 on K which is concentrated on the extreme points of K and which satisfies $L(x_0) = \int_K L(x) d\mu_0(x)$ for every continuous linear functional L on V.*

The goal of this section is to prove the following:

Lemma 2.4.6. *For every $D_1, \ldots, D_m \geq 0$, we have*

$$A(D_1, \ldots, D_m) = \sup\{I_f(\psi) : \psi \in ext\,(\cap_{i=1}^m \mathcal{C}_1(f_i, D_i))\}$$

and further, if $D_{f_1}, \ldots, D_{f_m}$ are all finite, we have

$$B(D_1, \ldots, D_m) = \inf\{I_f(\psi) : \psi \in ext\,(\cap_{i=1}^m \mathcal{C}_2(f_i, D_i))\}$$

Proof. The proof is based on Theorem 2.4.5. Let $C[0,\infty)$ denote the space of all continuous functions on $[0,\infty)$ equipped with the topology given by the metric:

$$\rho(f,g) := \sum_{k\geq 1} 2^{-k} \min\left(\sup_{0\leq x\leq k} |f(x)-g(x)|, 1\right). \tag{2.17}$$

It is a fact (see, for example, Rudin [120, Chapter 1]) that $C[0,\infty)$ is a locally convex vector space under this topology. We shall apply Choquet's theorem to $V = C[0,\infty)$ and $K = \cap_{i=1}^m \mathcal{C}_1(f_i, D_i)$ for the first identity and $K = \cap_{i=1}^m \mathcal{C}_2(f_i, D_i)$ for the second identity. It is obvious that $\mathcal{C}$ is a subset of $C[0,\infty)$.

Clearly both the sets $\cap_i \mathcal{C}_1(f_i, D_i)$ and $\cap_i \mathcal{C}_2(f_i, D_i)$ are convex. Also, by Fatou's lemma, $\cap_i \mathcal{C}_1(f_i, D_i)$ is closed under pointwise convergence i.e., if $\psi_n \in \cap_i \mathcal{C}_1(f_i, D_i)$ and $\psi_n \to \psi$ pointwise, then $\psi \in \cap_i \mathcal{C}_1(f_i, D_i)$. To see this, observe that by Fatou's lemma, for each $i = 1, \ldots, m$,

$$\begin{aligned} I_{f_i}(\psi) &= \int_0^\infty (\min(1,s) - \psi(s))\, d\nu_{f_i}(s) \\ &= \int_0^\infty \liminf_{n\to\infty} (\min(1,s) - \psi_n(s))\, d\nu_{f_i}(s) \\ &\leq \liminf_{n\to\infty} \int_0^\infty (\min(1,s) - \psi_n(s))\, d\nu_{f_i}(s) \leq D_i. \end{aligned}$$

On the other hand, if each D_{f_i} is a finite divergence, then by the dominated convergence theorem, $\cap_i \mathcal{C}_2(f_i, D_i)$ is also closed under pointwise convergence. Indeed, if $\psi_n \to \psi$ pointwise and D_{f_i} is a finite divergence, then by the dominated convergence (since $0 \leq \min(1,s) - \psi_n(s) \leq \min(1,s)$), we have $I_{f_i}(\psi_n) \to I_{f_i}(\psi)$.

In Lemma 2.4.7 below, we show that $\mathcal{C}$ is a compact subset of $C[0,\infty)$ under the topology given by the metric ρ. Moreover, it is easy to see that convergence in the metric ρ implies pointwise convergence. It follows hence that $\cap_i \mathcal{C}_1(f_i, D_i)$ is a compact, convex subset of $C[0,\infty)$ and if each D_{f_i} is a finite divergence, then $\cap_i \mathcal{C}_2(f_i, D_i)$ is also a compact, convex subset of $C[0,\infty)$.

For each $\epsilon > 0$, let us define the functional Λ_ϵ on $C[0,\infty)$ by

$$\Lambda_\epsilon(\psi) = \int (\min(1,s) - \psi(s))\, \{\epsilon \leq s \leq 1/\epsilon\}\, d\nu_f(s)$$

When restricted to the interval $[\epsilon, 1/\epsilon]$, the measure ν_f is a finite measure. Hence, Λ_ϵ is a continuous, linear functional on $C[0,\infty)$. Thus, by Theorem 2.4.5, we get that for every $\psi_0 \in \cap_i \mathcal{C}_1(f_i, D_i)$, there exists a Borel probability measure τ_0 that is concentrated on the set of extreme points, $ext(\cap_i \mathcal{C}_1(f_i, D_i))$, of $\cap_i \mathcal{C}_1(f_i, D_i)$ such that

$$\Lambda_\epsilon(\psi_0) = \int \Lambda_\epsilon(\psi) d\tau_0(\psi),$$

for every $\epsilon > 0$. Now, by the monotone convergence theorem,

$$\Lambda_\epsilon(\psi) \uparrow I_f(\psi) \qquad \text{as } \epsilon \downarrow 0$$

for every $\psi \in \mathcal{C}$. As a result, we can use the monotone convergence theorem again to assert that

$$\int \Lambda_\epsilon(\psi) d\tau_0(\psi) \uparrow \int I_f(\psi) d\tau_0(\psi) \qquad \text{as } \epsilon \downarrow 0.$$

We therefore obtain

$$I_f(\psi_0) = \int I_f(\psi) d\tau_0(\psi). \tag{2.18}$$

Since this is true for all functions ψ_0 in $\cap_i \mathcal{C}_1(f_i, D_i)$, we obtain

$$\sup_{\psi \in \cap_i \mathcal{C}_1(f_i, D_i)} I_f(\psi) = \sup_{\psi \in ext(\cap_i \mathcal{C}_1(f_i, D_i))} I_f(\psi)$$

The proof of the first assertion of Lemma 2.4.6 is now complete by (2.13). Similarly, when each divergence D_{f_i} is finite, we can prove that

$$\inf_{\psi \in \cap_i \mathcal{C}_2(f_i, D_i)} I_f(\psi) = \inf_{\psi \in ext(\cap_i \mathcal{C}_2(f_i, D_i))} I_f(\psi)$$

and this, together with (2.14), completes the proof of Lemma 2.4.6. □

In the above proof, we used the fact that $\mathcal{C}$ is compact in $C[0, \infty)$, the space of all continuous functions on $[0, \infty)$. We prove this fact below.

Lemma 2.4.7. *The class $\mathcal{C}$ is compact in $C[0, \infty)$ equipped with the topology given by the metric* (2.17).

Proof. We show that $\mathcal{C}$ is sequentially compact. Consider a sequence $\{\psi_n\}$ in $\mathcal{C}$. For every fixed $s_0 \in [0, \infty)$, the sequence $\{\psi_n(s_0)\}$ is a sequence of real numbers in $[0, 1]$ and hence has a convergent subsequence. By a standard diagonalization argument, we assert the existence of a subsequence $\{\phi_k\}$ of $\{\psi_n\}$ that converges pointwise over the set of all nonnegative rational numbers (denoted by $\mathbb{Q}_+$).

Let us now fix $\epsilon > 0$ and a real number $s_0 \in [0, \infty)$. Choose $r_1, r_2 \in \mathbb{Q}_+$ such that $r_1 \leq s_0 \leq r_2$ and such that $r_2 - r_1 < \epsilon/4$. Also, let $N \geq 1$ be large enough so that

$$|\phi_k(r_i) - \phi_l(r_i)| < \epsilon/4 \qquad \text{for } k, l \geq N$$

and for $i = 1, 2$. Using properties of functions in $\mathcal{C}$, we get that

$$\begin{aligned} |\phi_k(s_0) - \phi_l(s_0)| &< |\phi_k(r_1) - \phi_l(r_2)| + |\phi_k(r_2) - \phi_l(r_1)| \\ &< 2|\phi_k(r_1) - \phi_l(r_1)| + 2|r_1 - r_2| < \epsilon. \end{aligned}$$

In the last inequality above, we have used the fact that functions in $\mathcal{C}$ are Lipschitz with constant 1 (this can be proved for instance using the derviatives given by Lemma 2.4.4). It therefore follows that the sequence $\{\phi_k\}$ converges pointwise on $[0, \infty)$. The proof is now complete by the observation that convergence in the metric ρ is equivalent to pointwise convergence on $[0, \infty)$. □

2.4.3 Characterization of Extreme Points

Lemma 2.4.6 asserts that for the purposes of finding the supremum or infimum of I_f subject to constraints on I_{f_i}, it is enough to focus on the extreme points of the constraint set. In the next theorem, we provide a necessary condition for a function in the constraint set to be an extreme point of the constraint set.

Theorem 2.4.8. *Let ψ be a function in $\cap_i \mathcal{C}_1(f_i, D_i)$ and let k be the number of indices i for which $I_{f_i}(\psi) = D_i$. Then a necessary condition for ψ to be extreme in $\cap_i \mathcal{C}_1(f_i, D_i)$ is that ψ equals $\psi_{P,Q}$ for two probability measures $P, Q \in \mathcal{P}_{k+2}$. The same conclusion also holds for extreme functions in $\cap_i \mathcal{C}_2(f_i, D_i)$ provided all the involved divergences $D_{f_1}, \ldots, D_{f_m}$ are finite.*

Remark 2.4.9. *When $m = k = 0$, the sets $\cap_i \mathcal{C}_1(f_i, D_i)$ and $\cap_i \mathcal{C}_2(f_i, D_i)$ can both be taken to be equal to $\mathcal{C}$. As will be clear from the proof, the above theorem will also be true in this case where it states that a necessary condition for a function ψ to be extreme in $\mathcal{C}$ is that ψ equals $\psi_{P,Q}$ for two probability measures $P, Q \in \mathcal{P}_2$.*

The proof of Theorem 2.4.8 relies on the following lemma whose proof is provided after the proof of Theorem 2.4.8.

Lemma 2.4.10. *Let P and Q be two probability measures on a space $\mathcal{X}$ having densities p and q with respect to λ. Let $l \geq 1$ be fixed. Suppose that for every decreasing sequence $s_1 > \cdots > s_l$ of positive real numbers, the following condition holds:*

$$\min_{1 \leq j \leq l+1} (P(B_j) + Q(B_j)) = 0$$

where $B_1 = \{p \geq q s_1\}, B_i = \{q s_i \leq p < q s_{i-1}\}$ for $i = 2, \ldots, l$ and $B_{l+1} = \{p < q s_l\}$. Then $\psi_{P,Q}$ can be written as $\psi_{P',Q'}$ for two probability measures $P', Q' \in \mathcal{P}_l$.

Proof of Theorem 2.4.8. Let ψ be a function in $ext(\cap_i \mathcal{C}_1(f_i, D_i))$. Since $\psi \in \mathcal{C}$, we can write $\psi(s) = \psi_{P,Q}(s) = \int \min(p, sq) d\lambda$ for some probability measures P and Q on a measurable space $\mathcal{X}$ having densities p and q with respect to a common sigma finite measure λ. Without loss of generality, we assume that

$$I_{f_i}(\psi) = D_{f_i}(P||Q) = D_i \qquad \text{for } i = 1, \ldots, k \tag{2.19}$$

and

$$I_{f_i}(\psi) = D_{f_i}(P||Q) < D_i \qquad \text{for } i = k+1, \ldots, m. \tag{2.20}$$

Let $\alpha : \mathcal{X} \to (-1, 1)$ be a function satisfying

$$\int \alpha p d\lambda = \int \alpha q d\lambda = 0. \tag{2.21}$$

Note that $(1+\alpha)p, (1-\alpha)p, (1+\alpha)q$ and $(1-\alpha)q$ are all probability densities with respect to λ. Let P^+, P^-, Q^+, Q^- be probability measures having densities $p_+ := (1+\alpha)p$, $p_- := (1-\alpha)p$, $q_+ := (1+\alpha)q$, $q_- := (1-\alpha)q$ respectively with respect to λ. Also, let

$$\psi_+(s) := \psi_{P^+,Q^+}(s) = \int (1+\alpha)\min(p, sq)d\lambda$$

and

$$\psi_-(s) := \psi_{P^-,Q^-}(s) = \int (1-\alpha)\min(p, sq)d\lambda$$

so that $\psi = (\psi_+ + \psi_-)/2$. For every $i = 1, \ldots, m$, we observe that

$$\begin{aligned} I_{f_i}(\psi_+) &= D_{f_i}(P_+||Q_+) \\ &= \int q_+ f_i\left(\frac{p_+}{q_+}\right) d\lambda + f_i'(\infty)P_+\left\{q_+ = 0\right\}. \end{aligned}$$

Writing $(1+\alpha)p$ and $(1+\alpha)q$ for p_+ and q_+ respectively and noting that $1+\alpha > 0$ because α takes values in $(-1, 1)$, we obtain

$$I_{f_i}(\psi_+) = I_{f_i}(\psi) + \int \alpha r_i d\lambda \tag{2.22}$$

where

$$r_i := q f_i\left(\frac{p}{q}\right) + f_i'(\infty)p\left\{q = 0\right\}.$$

It follows similarly that

$$I_{f_i}(\psi_-) = I_{f_i}(\psi) - \int \alpha r_i d\lambda \tag{2.23}$$

We observe that $\int r_i d\lambda \le D_i$ for each $i = 1, \ldots, m$ which implies that

$$\int |\alpha r_i| d\lambda < \infty \tag{2.24}$$

for every function α that takes values in $(-1, 1)$ and $i = 1, \ldots, m$.

From (2.19), (2.22) and (2.23), it follows that the two inequalities:

$$I_{f_i}(\psi_+) \le D_i \quad \text{and} \quad I_{f_i}(\psi_-) \le D_i \tag{2.25}$$

will be satisfied for $i = 1, \ldots, k$ if and only if

$$\int \alpha r_i d\lambda = 0 \qquad \text{for } i = 1, \ldots, k. \tag{2.26}$$

Moreover, from (2.20), (2.22) and (2.23), it follows that if $\sup_{x\in\mathcal{X}} |\alpha(x)|$ is sufficiently small, then (2.25) will be satisfied also for $i = k+1, \ldots, m$. Let us say that α is a good function

if it satisfies (2.21) and (2.26) and if $\sup_x |\alpha(x)|$ is sufficiently small. We have thus proved that if α is a good function, then both ψ_+ and ψ_- belong to $\cap_i \mathcal{C}_1(f_i, D_i)$. Because ψ is extreme and $\psi = (\psi_+ + \psi_-)/2$, we assert that $\psi = \psi_+ = \psi_-$ for every good function α. As a result, $\partial^l \psi(s) = \partial^l \psi_+(s)$ for every $s > 0$ and $\partial^r \psi(s) = \partial^r \psi_+(s)$ for every $s \geq 0$. Because of Lemma 2.4.4 and the relations $p_+ = (1+\alpha)p$ and $q_+ = (1+\alpha)q$, we get that

$$\int_{p \geq sq} \alpha q d\lambda = 0 \quad \text{and} \quad \int_{p > sq} \alpha q d\lambda = 0 \tag{2.27}$$

for every $s > 0$. On the other hand, the equality $s\psi(1/s) = s\psi_+(1/s)$ for every $s > 0$ implies that $\psi_{Q,P}(s) = \psi_{Q^+,P^+}(s)$. Reversing the role of q and p in the argument that led to equation (2.27), we equate derivatives and use $\int \alpha p d\lambda = 0$ to get

$$\int_{p \geq sq} \alpha p d\lambda = 0 \quad \text{and} \quad \int_{p > sq} \alpha p d\lambda = 0 \tag{2.28}$$

for every $s > 0$. We have therefore shown that both (2.27) and (2.28) hold for every $s > 0$ whenever α is a good function. We now show that for every decreasing sequence $s_1 > \cdots > s_{k+2}$ of positive real numbers, the following condition must hold

$$\min_{1 \leq j \leq k+3} (P(B_j) + Q(B_j)) = 0 \tag{2.29}$$

where $B_1 = \{p \geq q s_1\}$, $B_i = \{q s_i \leq p < q s_{i-1}\}$ for $i = 2, \ldots, k+2$, and $B_{k+3} = \{p < q s_{k+2}\}$. The proof would then be completed by Lemma 2.4.10.

We prove (2.29) via contradiction. Suppose that the condition (2.29) does not hold for some $s_1 > \cdots > s_{k+2}$. Let $\alpha = \sum_{j=1}^{k+3} \alpha_j I_{B_j}$ where $\alpha_1, \ldots, \alpha_{k+3}$ are real numbers in $(-1, 1)$ and I_{B_j} denotes the indicator function of B_j. We claim that for this α, the conditions (2.27) and (2.28) cannot hold unless $\alpha_1 = \cdots = \alpha_{k+3} = 0$. To see this, note that (2.27) and (2.28) for $s = s_1$ give $\alpha_1(P(B_1) + Q(B_1)) = 0$. But since $P(B_1) + Q(B_1)$ is strictly positive (we are assuming that (2.29) does not hold), it follows that $\alpha_1 = 0$. We now use (2.27) and (2.28) for $s = s_2$ to obtain $\alpha_2 = 0$. Continuing this argument, we get that (2.27) and (2.28) cannot hold unless $\alpha_1 = \cdots = \alpha_{k+3} = 0$. As a result, it follows that $\alpha = \sum_{j=1}^{k+3} \alpha_j I_{B_j}$ is not a good function for every non-zero vector $(\alpha_1, \ldots, \alpha_{k+3})$ in $\mathbb{R}^{k+3}$.

On the other hand, as can be easily seen by writing down the conditions (2.21) and (2.26), for $\alpha = \sum_{j=1}^{k+3} \alpha_j I_{B_j}$ to be a good function, $\max_j |\alpha_j|$ needs to be sufficiently small and the following equalities need to be satisfied:

$$\sum_{j=1}^{k+3} \alpha_j P(B_j) = 0 = \sum_{j=1}^{k+3} \alpha_j Q(B_j)$$

and

$$\sum_{j=1}^{k+3} \alpha_j \int_{B_j} r_i d\lambda = 0 \qquad \text{for } i = 1, \ldots, k.$$

If (2.29) is not satisfied, then the above represent $k+2$ linear equalities for the $k+3$ variables $\alpha_1, \ldots, \alpha_{k+3}$. Therefore, a solution exists where $\alpha_1, \ldots, \alpha_{k+3}$ are non-zero (and where $\max_j |\alpha_j|$ is small) for which $\alpha = \sum_{j=1}^{k+3} \alpha_j I_{B_j}$ becomes a good function. Since this is a contradiction, we have established (2.29).

By Lemma 2.4.10, it follows that ψ can be written as $\psi_{P',Q'}$ for two probability measures P' and Q' on $\{1, \ldots, k+2\}$. This proves the first part of the theorem. The case of $\cap_i \mathcal{C}_2(f_i, D_i)$ is very similar. In the above argument, the only place where we used the fact that the constraints in $\cap_i \mathcal{C}_1(f_i, D_i)$ are of the $\leq$ form is in asserting (2.24). In the case of $\cap_i \mathcal{C}_2(f_i, D_i)$, the statement (2.24) still holds under the assumption that each divergence D_{f_i} is finite. The rest of the proof proceeds exactly as before. □

Below, we provide the proof of Lemma 2.4.10 which was used in the above proof.

Proof of Lemma 2.4.10. Let η denote the probability measure $(P+Q)/2$. Suppose

$$N := \{x \in (0,1) : x = \eta\{p \geq qs\} \text{ for some } s \in (0, \infty)\}.$$

We claim that N is a finite set having cardinality at most $l-1$. To see this, suppose, if possible, that there exist points $0 < x_1 < \cdots < x_l < 1$ in N. Then, we can write $x_i = \eta\{p \geq qs_i\}$ for some $s_1 > \cdots > s_l > 0$. But then $\eta(B_1) = x_1$, $\eta(B_i) = x_i - x_{i-1} > 0$ for $i = 2, \ldots, l$ and $\eta(B_{l+1}) = 1 - x_l > 0$ which contradicts the condition given in the lemma. Let us therefore assume that the cardinality of N equals $k \leq l-1$ and let $N = \{x_1, \ldots, x_k\}$ where $0 < x_1 < \cdots < x_k < 1$. Let

$$s_i^* := \sup\{s > 0 : \eta\{p \geq qs\} = x_i\}$$

for $i = 1, \ldots, k$. Also let

$$s_{k+1}^* := \sup\{s > 0 : \eta\{p \geq qs\} = 1\}$$

if there exists $s > 0$ with $\eta\{p \geq qs\} = 1$. If there exists no such $s > 0$, we define $s_{k+1}^* = 0$. It is easy to see that $s_1^* \in (0, \infty]$ and $s_{k+1}^* \in [0, \infty)$ while $s_2^*, \ldots, s_k^* \in (0, \infty)$. Let us first consider the case when $s_1^* < \infty$ and $s_{k+1}^* > 0$. In this case, for each $i = 1, \ldots, k+1$, there exists a sequence $\{t_n(i)\}$ with $0 < t_n(i) \uparrow s_i^*$ such that $\eta\{p \geq qt_n(i)\} = x_i$ (we take $x_{k+1} = 1$). Because the sets $\{p \geq qt_n(i)\}$ decrease to $\{p \geq qs_i^*\}$ as $n \to \infty$, it follows that $\eta\{p \geq qs_i^*\} = x_i$ for each $i = 1, \ldots, k+1$. Also it is easy to see that

$$\eta\{p > qs_i^*\} = \lim_{s \downarrow s_i^*} \eta\{p \geq qs\} = x_{i-1}$$

for each $i = 1, \ldots, k+1$ where we take $x_0 = 0$. It follows therefore that $\eta\{p = qs_i^*\} = x_i - x_{i-1}$ for $1 \leq i \leq k+1$. Because $\sum_{i=1}^{k+1}(x_i - x_{i-1}) = x_{k+1} - x_0 = 1$, it follows that

$$\sum_{i=1}^{k+1} \eta\{p = qs_i^*\} = 1. \tag{2.30}$$

It can be checked that the above statement is also true in the case when $s_1^* = \infty$ and/or $s_{k+1}^* = 0$ provided we interpret

$$\{p = q \cdot \infty\} = \{q = 0\} \text{ and } \{p = q \cdot 0\} = \{p = 0\}.$$

The equality (2.30) is the same as

$$\sum_{i=1}^{k+1} P\{p = qs_i^*\} = 1 \text{ and } \sum_{i=1}^{k+1} Q\{p = qs_i^*\} = 1. \tag{2.31}$$

Let $p_i = P\{p = qs_i^*\}$ and $q_i = Q\{p = qs_i^*\}$ for $i = 1, \ldots, k+1$ so that $P' = (p_1, \ldots, p_{k+1})$ and $Q' = (q_1, \ldots, q_{k+1})$ are probability measures on $\{1, \ldots, k+1\}$. For each $i = 1, \ldots, k+1$, we have

$$p_i = P\{p = qs_i^*\} = \int_{p=qs_i^*} p d\lambda = s_i^* \int_{p=qs_i^*} q d\lambda = s_i^* q_i$$

where the above statement is to be interpreted as $q_1 = 0$ if $s_1^* = \infty$ and as $p_{k+1} = 0$ if $s_{k+1}^* = 0$. Also

$$\int_{p=qs_i^*} \min(p, qs) d\lambda = \min(s_i^*, s) Q\{p = qs_i^*\} = \min(p_i, q_i s)$$

for every $s \geq 0$ and $i = 1, 2, \ldots, k+1$. Therefore,

$$\begin{aligned} \psi_{P,Q}(s) &= \int \min(p, qs) d\lambda \\ &= \sum_{i=1}^{k+1} \int_{p=qs_i^*} \min(p, qs) d\lambda = \psi_{P',Q'}(s). \end{aligned}$$

The proof is complete because $k + 1 \leq l$. □

2.4.4 Completion of the Proof

We shall prove (2.5). The proof of (2.6) is entirely analogous. Theorem 2.4.8 states that every function in $\cap_i \mathcal{C}_1(f_i, D_i)$ that is extreme equals $\psi_{P,Q}$ for some $P, Q \in \mathcal{P}_{m+2}$. Therefore, by Lemma 2.4.6, we get that $A(D_1, \ldots, D_m)$ equals

$$\sup \left\{ I_f(\psi_{P,Q}) : \psi_{P,Q} \in \cap_{i=1}^m \mathcal{C}_1(f_i, D_i) \text{ and } P, Q \in \mathcal{P}_{m+2} \right\}.$$

Because $I_{f_i}(\psi_{P,Q})$ equals $D_{f_i}(P||Q)$, the constraint $\psi \in \cap_i \mathcal{C}_1(f_i, D_i)$ is equivalent to $D_{f_i}(P||Q) \leq D_i$ for all $i = 1, \ldots, m$. The proof is therefore complete.

2.5 Remarks and Extensions

2.5.1 Stronger Version

The proof of Theorem 2.3.1 actually yields a smaller expression for $A(D_1, \ldots, D_m)$ than $A_{m+2}(D_1, \ldots, D_m)$ and a larger expression for $B(D_1, \ldots, D_m)$ than $B_{m+2}(D_1, \ldots, D_m)$. For each subset J of $\{1, \ldots, m\}$, let $A^J(D_1, \ldots, D_m)$ denote the supremum of $D_f(P||Q)$ over all probability measures $P, Q \in \mathcal{P}_{k+2}$ (where k is the cardinality of J) for which $D_{f_i}(P||Q) = D_i$ for $i \in J$ and $D_{f_i}(P||Q) < D_i$ for $i \notin J$. It is clear that

$$A^J(D_1, \ldots, D_m) \leq A_{m+2}(D_1, \ldots, D_m)$$

for each $J \subseteq \{1, \ldots, m\}$. The following is therefore a stronger version of Theorem 2.3.1:

$$A(D_1, \ldots, D_m) = \max_{J \subseteq \{1,\ldots,m\}} A^J(D_1, \ldots, D_m) \tag{2.32}$$

An analogous statement also holds for $B(D_1, \ldots, D_m)$. Let us now show that our proof of Theorem 2.3.1 given in Section 2.4.4 results in (2.32). By Theorem 2.4.8, every function ψ in $\cap_i \mathcal{C}_1(f_i, D_i)$ that is extreme equals $\psi_{P,Q}$ for some $P, Q \in \mathcal{P}_{k+2}$ where k is the number of indices i for which $I_{f_i}(\psi) = D_{f_i}(P||Q) = D_i$. Therefore, if J denotes these indices, then

$$\begin{aligned} I_f(\psi) = D_f(P||Q) &\leq A^J(D_1, \ldots, D_m) \\ &\leq \max_{J \subseteq \{1,\ldots,m\}} A^J(D_1, \ldots, D_m) \end{aligned}$$

for every $\psi \in ext(\cap_i \mathcal{C}_1(f_i, D_i))$. The equality (2.32) therefore follows from Lemma 2.4.6.

2.5.2 Joint Ranges

Recall that the joint range of divergences $D_{f_1}, \ldots, D_{f_m}$ is denoted by $\mathcal{R}(f_1, \ldots, f_m)$ and is defined as the set of all vectors in $\mathbb{R}^m$ that equal $(D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q))$ for some P and Q. The quantities $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ can easily be calculated from knowledge of $\mathcal{R}(f, f_1, \ldots, f_m)$. It therefore makes sense to try to prove Theorem 2.3.1 by trying to determine the joint range $\mathcal{R}(f, f_1, \ldots, f_m)$. We argue here that this approach is not good enough to prove Theorem 2.3.1; it results in the weaker identities (2.36) and (2.37).

In the following theorem, we characterize the joint range $\mathcal{R}(f_1, \ldots, f_m)$ for every arbitrary set of m divergences. We show that it suffices to restrict attention to pairs of probability measures in $\mathcal{P}_{m+2}$. For each $k \geq 1$, let

$$\mathcal{R}_k(f_1, \ldots, f_m) := \{(D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q)) : P, Q \in \mathcal{P}_k\}.$$

Theorem 2.5.1. *For every $m \geq 1$ and divergences $D_{f_1}, \ldots, D_{f_m}$, we have*

$$\mathcal{R}(f_1, \ldots, f_m) = \mathcal{R}_{m+2}(f_1, \ldots, f_m).$$

For the special case $m = 2$, this theorem has already been proved by Harremoes and Vajda [55]. The short proof given below uses the Caratheodory theorem and was communicated to us by an anonymous referee. In contrast, the proof given in Harremoes and Vajda [55] for $m = 2$ is much more elaborate. The counterexamples in Harremoes and Vajda [55] show the tightness of this theorem. After the proof, we describe an attempt to prove Theorem 2.3.1 via Theorem 2.5.1.

Proof. We just need to prove that $\mathcal{R}(f_1, \ldots, f_m) \subseteq \mathcal{R}_{m+2}(f_1, \ldots, f_m)$. Let $u \in \mathcal{R}(f_1, \ldots, f_m)$. Then $u = (D_{f_1}(P||Q), \ldots, D_{f_m}(P||Q))$ for some pair of probability measures P and Q. If p and q denote the densities of P and Q with respect to a common measure λ, then

$$u = \int_{\{q>0\}} \left(f_1\left(\frac{p}{q}\right), \ldots, f_m\left(\frac{p}{q}\right)\right) dQ + P\{q = 0\} \left(f_1'(\infty), \ldots, f_m'(\infty)\right). \tag{2.33}$$

Let $S \subseteq \mathbb{R}^{m+1}$ be defined by $S := \{(s, f_1(s), \ldots, f_m(s)) : s \geq 0\}$. Then clearly the vector

$$\int_{\{q>0\}} \left(\frac{p}{q}, f_1\left(\frac{p}{q}\right), \ldots, f_m\left(\frac{p}{q}\right)\right) dQ$$

lies in the convex hull of S. Because S is a connected subset of $\mathbb{R}^{m+1}$, we can use Caratheodory's theorem (see, for example, Bárány and Karasev [6]) to assert that any point in its convex hull can be written as a convex combination of at most $m + 1$ points in S. As a result, we can write

$$\int_{\{q>0\}} \left(\frac{p}{q}, f_1\left(\frac{p}{q}\right), \ldots, f_m\left(\frac{p}{q}\right)\right) dQ = \sum_{i=1}^{m+1} \alpha_i \left(s_i, f_1(s_i), \ldots, f_m(s_i)\right) \tag{2.34}$$

for some $\alpha_1, \ldots, \alpha_{m+1} \geq 0$ with $\sum_i \alpha_i = 1$ and $s_1, \ldots, s_{m+1} \geq 0$. One consequence of this representation is that

$$\sum_{i=1}^{m+1} \alpha_i s_i = \int_{q>0} \left(\frac{p}{q}\right) dQ = P\{q > 0\}. \tag{2.35}$$

We now define two probability measures P' and Q' in $\mathcal{P}_{m+2}$ as follows: $P'\{i + 1\} = \alpha_i s_i$ for $1 \leq i \leq m + 1$ and $P'\{1\} = P\{q = 0\}$; and $Q'\{i + 1\} = \alpha_i$ for $1 \leq i \leq m + 1$ and $Q'\{1\} = 0$. The fact that $\sum_{i=1}^{m+2} P'\{i\} = 1$ follows from (2.35). The equalities (2.33) and (2.34) together clearly imply that $u = (D_{f_1}(P'||Q'), \ldots, D_{f_m}(P'||Q'))$. Thus $u \in \mathcal{R}_{m+2}(f_1, \ldots, f_m)$ and this completes the proof. □

Clearly $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ can be written as functions of the joint range $\mathcal{R}(f, f_1, \ldots, f_m)$. Theorem 2.5.1 immediately therefore implies

$$A(D_1, \ldots, D_m) = A_{m+3}(D_1, \ldots, D_m) \tag{2.36}$$

and

$$B(D_1, \ldots, D_m) = B_{m+3}(D_1, \ldots, D_m). \tag{2.37}$$

These results are clearly weaker than those given by Theorem 2.3.1. Strictly speaking, one can deduce a slightly stronger conclusion than (2.36) and (2.37) from Theorem 2.5.1. A probability measure on $\{1, \ldots, m+3\}$ is determined by $m+2$ real numbers. Therefore, a pair of probability measures in $\mathcal{P}_{m+3}$ are determined by $2m+4$ real numbers. The inequalities (2.36) and (2.37) therefore reduce the optimization problems for $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ into optimization problems over $2m+4$ variables. A closer inspection at the proof of Theorem 2.5.1 shows that one actually gets a reduction to $2m+3$ variables. This is because the probability measure Q' in the proof satisfies $Q'\{1\} = 0$. Therefore, by an argument based solely on the joint range of $D_f, D_{f_1}, \ldots, D_{f_m}$, one can reduce the optimization problems for $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ into optimization problems over $2m+3$ variables. Because of the tightness of Theorem 2.5.1, this is the best reduction that one can hope for the quantities $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ via an argument based on the joint range alone. On the other hand, Theorem 2.3.1 achieves a reduction to $2m+2$ variables.

2.5.3 Tightness

The conclusion of Theorem 2.3.1 is tight in the sense that, in general, one cannot reduce the optimization problems to pairs of probability measures on spaces of cardinality strictly smaller than $m+2$. We shall demonstrate this fact in this section by means of an example. We also explain this fact numerically in Example 2.7.6.

Consider the problem of maximizing an f-divergence subject to a upper bound on the total variation distance. In other words, let

$$A(V) := \sup\{D_f(P||Q) : V(P,Q) \leq V\}$$

where D_f is an arbitrary f-divergence. In this case, Theorem 2.3.1 asserts that $A(V)$ equals $A_3(V)$ where, as before,

$$A_k(V) := \sup\{D_f(P||Q) : P, Q \in \mathcal{P}_k, V(P,Q) \leq V\}.$$

We shall show below that when D_f is a finite divergence and when f is strictly convex on $(0,\infty)$, the quantity $A_3(V)$ is strictly larger than $A_2(V)$ for all $V \in (0,1)$.

The quantity $A_3(V) = A(V)$ can be determined precisely. The easiest way is to use Lemma 2.4.1. Because

$$V(P,Q) = D_{u_1}(P||Q) = 1 - \psi_{P,Q}(1),$$

the constraint $V(P,Q) \leq V$ is equivalent to $\psi_{P,Q}(1) \geq 1 - V$. Therefore, by Lemma 2.4.1, we get

$$A(V) = \sup\{I_f(\psi) : \psi \in \mathcal{C} \text{ and } \psi(1) \geq 1 - V\}.$$

It is obvious that the supremum above is achieved for $\psi(s) = (1-V)\min(1,s)$ which equals $\psi_{P',Q'}$ for $P' = (1-V, V, 0)$ and $Q' = (1-V, 0, V)$. Thus

$$A(V) = D_f(P'||Q') = V\left(f(0) + f'(\infty)\right).$$

In other words, by Remark 2.4.3, the quantity $A(V)$ equals V times the maximum possible value of the divergence D_f.

Let us now consider the quantity $A_2(V)$. By compactness and the form of the constraint, it follows that there exist two probability measures P^* and Q^* in $\mathcal{P}_2$ with $V(P^*, Q^*) = V$ and $D_f(P^*||Q^*) = A_2(V)$. We can then, without loss of generality, parametrize P^* and Q^* by $P^* = (\rho, 1-\rho)$ and $Q^* = (\rho + V, 1-\rho-V)$ for some $0 \leq \rho \leq 1 - V$. Consider now the probability measures

$$\tilde{P} = \left(\frac{\rho}{2}, \frac{\rho}{2}, 1-\rho\right) \text{ and } \tilde{Q} = \left(\frac{\rho}{2} + \frac{V}{4}, \frac{\rho}{2} + \frac{3V}{4}, 1-\rho-V\right)$$

in $\mathcal{P}_3$. If $V \in (0,1)$, by strict convexity of the function f, it is easy to see that

$$D_f(\tilde{P}||\tilde{Q}) > D_f(P^*||Q^*) = A_2(V).$$

On the other hand, it is easy to see that $V(\tilde{P}, \tilde{Q})$ equals V and hence $A_3(V) > D_f(\tilde{P}||\tilde{Q})$. Therefore, $A_3(V) > A_2(V)$. Thus, Theorem 2.3.1 is tight in general. However, in some special cases, one can obtain stronger conclusions, see Sections 2.6.1 and 2.6.2.

2.5.4 Finiteness assumption for $B(D_1, \ldots, D_m)$

In order to prove (2.6), we required that all the divergences $D_{f_1}, \ldots, D_{f_m}$ are finite. The reason is mainly technical and the finiteness assumption was crucially used in the proof of Lemma 2.4.6. The set $\cap_i \mathcal{C}_2(f_i, D_i)$ will not be closed (in $C[0,\infty)$ equipped with the metric ρ) if some of the divergences D_{f_i} were non-finite (closedness of $\cap_i \mathcal{C}_2(f_i, D_i)$ was critical in the application of Choquet's theorem in Lemma 2.4.6). To illustrate this non-closedness, let us consider $m = 1$ and the set $\mathcal{C}_2(f_1, D_1)$ for some non-finite divergence D_{f_1} and $D_1 > 0$. By (2.12), because D_{f_1} is non-finite, we have

$$\int_0^\infty \min(1, s) d\nu_{f_1}(s) = \infty.$$

The function $\psi_0(s) = \min(1, s)$ clearly does not belong to $\mathcal{C}_2(f_1, D_1)$ because $I_{f_1}(\psi_0) = 0$. But we shall show that ψ_0 belongs to the closure of $\mathcal{C}_2(f_1, D_1)$. Indeed, if

$$\psi_n(s) := \left(1 - \frac{1}{n}\right) \min(1, s) \qquad \text{for } s \geq 0,$$

then clearly ψ_n converges to ψ in the metric ρ. Moreover, for each n, $\psi_n \in \mathcal{C}$ and

$$I_{f_1}(\psi_n) = \frac{1}{n} \int_0^\infty \min(1, s) d\nu_{f_1}(s) = \infty.$$

Thus $\psi_n \in \mathcal{C}_2(f_1, D_1)$ for each $n \geq 1$ which implies that ψ_0 belongs to the closure of $\mathcal{C}_2(f_1, D_1)$. Therefore, $\mathcal{C}_2(f_1, D_1)$ is not closed.

The quantity $B(D_1, \ldots, D_m)$ behaves strangely when some of the divergences D_{f_i} are non-finite and when D_f is finite. Indeed, in this case, one can simply drop the constraints corresponding to the non-finite divergences and reduce the problem to the case when all divergences are finite. This is the content of the next lemma.

Lemma 2.5.2. *Let $D_f, D_{f_1}, \ldots, D_{f_m}$ be finite divergences and let $D_{f_{m+1}}, \ldots, D_{f_{m+l}}$ be non-finite divergences. Then*

$$B(D_1, \ldots, D_{m+l}) = B(D_1, \ldots, D_m)$$

Proof. We shall work with (2.14). Because $\cap_{i=1}^{m+l} \mathcal{C}_2(f_i, D_i)$ is contained in $\cap_{i=1}^{m} \mathcal{C}_2(f_i, D_i)$, it follows that $B(D_1, \ldots, D_{m+l})$ is larger than or equal to $B(D_1, \ldots, D_m)$. To prove the other inequality, let $\psi \in \cap_{i=1}^{m} \mathcal{C}_2(f_i, D_i)$. For each $n \geq 1$, define

$$\psi_n(s) = \min\left[\left(1 - \frac{1}{n}\right)\min(1, s), \psi(s)\right]$$

It is easy to check that $\psi_n \in \mathcal{C}$. Note that for $1 \leq i \leq m$,

$$\begin{aligned} I_{f_i}(\psi_n) &= \int_0^\infty (\min(1, s) - \psi_n(s))\, d\nu_{f_i}(s) \\ &\geq \int_0^\infty (\min(1, s) - \psi(s))\, d\nu_{f_i}(s) = I_{f_i}(\psi) \geq D_i. \end{aligned}$$

Moreover, for $m < i \leq m + l$, we have

$$\begin{aligned} I_{f_i}(\psi_n) &= \int_0^\infty (\min(1, s) - \psi_n(s))\, d\nu_{f_i}(s) \\ &\geq \frac{1}{n}\int_0^\infty \min(1, s) d\nu_{f_i}(s) = \infty \geq D_i. \end{aligned}$$

It therefore follows that $\psi_n \in \cap_{i=1}^{m+l} \mathcal{C}_2(f_i, D_i)$ for every $n \geq 1$. Consequently,

$$I_f(\psi_n) \geq B(D_1, \ldots, D_{m+l}) \qquad \text{for every } n \geq 1.$$

Observe that $\psi_n(s)$ converges to $\psi(s)$ for every $s \geq 0$. Thus, because D_f is a finite divergence, it follows by the dominated convergence theorem that $I_f(\psi_n)$ converges to $I_f(\psi)$ which results in

$$I_f(\psi) \geq B(D_1, \ldots, D_{m+l}).$$

Finally, because $\psi \in \cap_{i=1}^{m} \mathcal{C}_2(f_i, D_i)$ is arbitrary, we have proved that $B(D_1, \ldots, D_m)$ is larger than or equal to $B(D_1, \ldots, D_{m+l})$ which completes the proof of the lemma. □

Remark 2.5.3. *If D_f is finite and if all the divergences $D_{f_1}, \ldots, D_{f_m}$ are non-finite, then Lemma 2.5.2 gives that*

$$B(D_1, \ldots, D_m) = 0 \tag{2.38}$$

for all values of $D_1, \ldots, D_m$. Here is a special instance of this result. Suppose that D_f denotes the total variation distance, $m = 1$ and that D_{f_1} is the Kullback-Leibler divergence. Then (2.38) *shows that the smallest value of the total variation distance over all probability measures with Kullback-Leibler divergence at least 5 (say) equals 0. The same conclusion holds for multiple non-finite divergence constraints as well.*

Theorem 2.3.1 gives a formula for $B(D_1, \ldots, D_m)$ for arbitrary D_f and for finite $D_{f_1}, \ldots, D_{f_m}$. In Lemma 2.5.2, we showed that when D_f is finite, then the case when one of more of $D_{f_1}, \ldots, D_{f_m}$ are non-finite can be reduced to the case where all the constraint divergences are finite which is handled by Theorem 2.3.1. The case that we are unable to resolve is $B(D_1, \ldots, D_m)$ when D_f is non-finite and when one or more of $D_{f_1}, \ldots, D_{f_m}$ are non-finite. This case is neither covered by Theorem 2.3.1 nor by Lemma 2.5.2.

2.5.5 Sufficiency of the extreme point characterization

In Theorem 2.4.8, we gave a necessary condition for functions in the classes $\cap_i \mathcal{C}_1(f_i, D_i)$ and $\cap_i \mathcal{C}_2(f_i, D_i)$ to be extreme. As we have seen, this necessary condition was enough to prove Theorem 2.3.1. For the sake of completeness, in this section, we investigate whether the condition in Theorem 2.4.8 is sufficient as well for extremity.

Let $j \in \{1, 2\}$ and let ψ be a function in $\cap_i \mathcal{C}_j(f_i, D_i)$. Suppose ψ satisfies the condition given in Theorem 2.4.8 i.e., let $\psi = \psi_{P,Q}$ for two probability measures $P, Q \in \mathcal{P}_{k+2}$ where k is the number of indices where $I_{f_i}(\psi) = D_i$. Here, we explore the question of extremity of ψ in $\cap_i \mathcal{C}_j(f_i, D_i)$.

Let $l \leq k + 2$ be the size of the (finite) support set of the measure $P + Q$ and let $P = \{p_1, \ldots, p_l\}$ and $Q = \{q_1, \ldots, q_l\}$, then $\psi(s) = \sum_{i=1}^{l} \min(p_i, q_i s)$. Because the size of the support set of $P + Q$ is l, it follows that $\max(p_i, q_i) > 0$ for every i. It is easy to check that ψ is piecewise linear with knots at p_i/q_i (this ratio can equal 0 or ∞ as well).

Suppose that $\psi = (\psi_1 + \psi_2)/2$ for two functions ψ_1 and ψ_2 in $\cap_i \mathcal{C}_j(f_i, D_i)$. Because ψ_1 and ψ_2 are both concave, it follows that they both have to be linear in the regions where ψ is linear. As a result, one can write

$$\psi_1(s) = \sum_{i=1}^{l} (1 + \alpha_i) \min(p_i, q_i s)$$

and

$$\psi_2(s) = \sum_{i=1}^{l} (1 - \alpha_i) \min(p_i, q_i s)$$

for some $\alpha_1, \ldots, \alpha_n \in [-1, 1]$ satisfying

$$\sum_{i=1}^{l} \alpha_i p_i = \sum_{i=1}^{l} \alpha_i q_i = 0. \tag{2.39}$$

Now, whenever $I_{f_i}(\psi) = D_i$, because of the above, we must have $I_{f_i}(\psi_1) = D_i$. This latter equality can be written as a linear equality in $\alpha_1, \ldots, \alpha_l$. Because $I_{f_i}(\psi) = D_i$ for k indices i, we obtain k linear equations for $\alpha_1, \ldots, \alpha_l$. These, together with (2.39), give rise to $k+2$ linear equations for the $l \leq k+2$ variables $\alpha_1, \ldots, \alpha_l$. Under appropriate linear independence conditions on the measures ν_{f_i}, these would imply that $\alpha_i = 0$ for every $1 \leq i \leq l$ which would further imply that $\psi_1 = \psi = \psi_2$ and that ψ is extreme.

In the case when $m \leq 1$ however, no such explict linear independence conditions are necessary and, moreover, one can also give a geometric proof of the sufficiency characterization of the extreme points. We do this below in two parts: Lemma 2.5.4 deals with $m = 0$ (i.e., extreme points of $\mathcal{C}$) and Lemma 2.5.5 deals with the $m = 1$ case.

Lemma 2.5.4. *For every $P, Q \in \mathcal{P}_2$, the function $\psi_{P,Q}$ is extreme in $\mathcal{C}$.*

Proof. Fix two probability measures P and Q on $\{1, 2\}$ and let J denote the smallest open interval (possibly infinite) such that $\psi_{P,Q}(s) = \min(1, s)$ for $s \notin J$. By explicitly writing down the expression for ψ in terms of $P\{1\}$ and $Q\{1\}$, it is easy to see that if J is non-empty, then $\psi_{P,Q}$ is linear on J.

Suppose now that $\psi_{P,Q}$ equals the convex combination $(\psi_1 + \psi_2)/2$ for two functions ψ_1 and ψ_2 in $\mathcal{C}$. If J is empty, then $\psi_{P,Q}$ equals the function $\min(1, s)$ for all s and since all functions in $\mathcal{C}$ and bounded from above by $\min(1, s)$, it follows that

$$\psi_{P,Q}(s) = \psi_1(s) = \psi_2(s) = \min(1, s) \tag{2.40}$$

for all $s \geq 0$.

Let us therefore assume that J is non-empty. In this case, again it is obvious that (2.40) holds for $s \notin J$. Concavity of functions in $\mathcal{C}$ and linearity of ψ in J would then imply that $\psi_1 \geq \psi_{P,Q}$ and $\psi_2 \geq \psi_{P,Q}$. Since $\psi_{P,Q}$ is the average of ψ_1 and ψ_2, this can happen only when $\psi_{P,Q} = \psi_1 = \psi_2$. The proof is complete. □

Lemma 2.5.5. *Let $j \in \{1, 2\}$ and consider the class $\mathcal{C}_j(f_1, D_1)$ for $D_1 > 0$. For every $P, Q \in \mathcal{P}_3$ with $D_{f_1}(P||Q) = D_1$, the function $\psi_{P,Q}$ is extreme in $\mathcal{C}_j(f_1, D_1)$.*

Proof. Fix two probability measures P and Q in $\mathcal{P}_3$ with $D_{f_1}(P||Q) = D_1$ so that $I_{f_1}(\psi_{P,Q}) = D_1$. For notational convenience, let us denote $\psi_{P,Q}$ by ψ. As in the proof of Lemma 2.5.4, let J denote the smallest interval outside which $\psi(s)$ equals $\min(1, s)$. If J is empty, then ψ equals the function $\min(1, s)$ which is obviously extreme. So let us assume that J is non-empty. In that case, because $P, Q \in \mathcal{P}_3$, it can be checked that ψ is piecewise linear with at most two segments in J.

Suppose that $\psi = (\psi_1 + \psi_2)/2$ for two functions $\psi_1, \psi_2 \in \mathcal{C}_j(f_1, D_1)$. Because, $I_{f_1}(\psi) = D_{f_1}(P||Q) = D_1$, it follows that

$$I_{f_1}(\psi_1) = I_{f_1}(\psi_2) = I_{f_1}(\psi) = D_{f_1}(P||Q) = D_1. \tag{2.41}$$

If ψ has exactly one segment in J, then, by concavity, the inequalities $\psi_1(s) \geq \psi(s)$ and $\psi_2(s) \geq \psi(s)$ hold for all s. Because ψ_1 and ψ_2 average out to ψ, we must then have $\psi = \psi_1 = \psi_2$.

Now suppose that ψ has exactly two segments in J_ψ. Let a be the point in J such that ψ is linear on both $J \cap [0, a]$ and $J \cap [a, \infty)$. We shall show that $\psi(a) = \psi_1(a) = \psi_2(a)$. Concavity of ψ_1 and ψ_2 and linearity of ψ on $J \cap [0, a]$ and $J \cap [a, \infty)$ can then be used to show that $\psi = \psi_1 = \psi_2$. Suppose, if possible, that $\psi_1(a) > \psi(a)$. Using the concavity of ψ_1, it then follows that $\psi_1(s) > \psi(s)$ for all $s \in J$. Because of (2.41), it follows that

$$\int_J (\psi_1(s) - \psi(s))\, d\nu_{f_1}(s) = \int_0^\infty (\psi_1(s) - \psi(s))\, d\nu_{f_1}(s) = 0$$

This implies that $\nu_{f_1}(J) = 0$. But then

$$D_1 = I_{f_1}(\psi) = \int_J (\min(1, s) - \psi(s))\, d\nu_{f_1}(s) = 0$$

which contradicts the fact that $D_1 > 0$. We have thus obtained $\psi_1(a) \leq \psi(a)$. Similarly, $\psi_2(a) \leq \psi(a)$ and since $\psi(a)$ is an average of $\psi_1(a)$ and $\psi_2(a)$, it follows that $\psi(a) = \psi_1(a) = \psi_2(a)$. The proof is complete. □

2.6 Applications and Special Cases

2.6.1 Primitive Divergences

In this section, we consider the case of the quantity $B(D_1, \ldots, D_m)$ where all the divergences $D_{f_1}, \ldots, D_{f_m}$ are primitive divergences (see Remark 2.4.2). In Theorem 2.6.1 below, we show that, in this case, $B(D_1, \ldots, D_m)$ actually equals $B_{m+1}(D_1, \ldots, D_m)$ as opposed to $B_{m+2}(D_1, \ldots, D_m)$.

The problem of minimizing an f-divergence subject to constraints on primitive divergences and the related problem of obtaining inequalities between f-divergences and primitive divergences has received much attention in the literature and has a long history. Let us briefly mention some important works in this area. The most well-known such inequality is Pinsker's inequality which states that $D_{KL}(P||Q) \geq 2V^2(P, Q)$ where D_{KL} is the Kullback-Leibler divergence which corresponds to $f(x) = x \log x$ and V is the total variation distance. Pinsker Pinsker [108] proved this inequality with the constant 2 replaced by 1. The inequality with the constant 2 (which cannot be improved further) has been proved independently almost at the same time by Csiszar Csiszar [30], Kemperman Kemperman [66] and Kullback Kullback [69].

Although Pinsker's inequality is very useful, it is not sharp in the sense that

$$\inf \{D_{KL}(P||Q) : V(P, Q) \geq V\} > 2V^2$$

for every $V \neq 0$. The problem of finding sharp inequalities between $D_{KL}(P||Q)$ and $V(P, Q)$ was solved in Fedotov, Harremoës, and Topsoe [44] where an implicit expression for the infimum in the left hand side above was provided.

The more general problem of finding the best lower bound for an arbitrary f-divergence given a lower bound on total variation distance was solved by Gilardoni in Gilardoni [51]. The problem of finding lower bounds for f-divergences given constraints on a finite number of primitive divergences was studied by Reid and Williamson [113]. In Remark 2.6.2, we explain how our theorem below gives an equivalent but simpler solution compared to the solution of Reid and Williamson [113].

Theorem 2.6.1. *Suppose that D_f is an arbitrary divergence and that all divergences $D_{f_1}, \ldots, D_{f_m}$ are primitive divergences. Then*

$$B(D_1, \ldots, D_m) = B_{m+1}(D_1, \ldots, D_m).$$

Proof. Theorem 2.3.1 states that $B(D_1, \ldots, D_m)$ equals $B_{m+2}(D_1, \ldots, D_m)$. We shall show therefore that $B_{m+2}(D_1, \ldots, D_m)$ equals $B_{m+1}(D_1, \ldots, D_m)$.

It is obvious that

$$B_{m+2}(D_1, \ldots, D_m) \leq B_{m+1}(D_1, \ldots, D_m)$$

because we have a minimization problem and the constraint set is larger in the case of $B_{m+2}(D_1, \ldots, D_m)$. It is therefore enough to prove that

$$B_{m+2}(D_1, \ldots, D_m) \geq B_{m+1}(D_1, \ldots, D_m).$$

Fix two probability measures $P = (p_1, \ldots, p_{m+2})$ and $Q = (q_1, \ldots, q_{m+2})$ in $\mathcal{P}_{m+2}$ with $D_{f_i}(P||Q) \geq D_i$ for every $i = 1, \ldots, m$. We show below that

$$D_f(P||Q) \geq B_{m+1}(D_1, \ldots, D_m)$$

which will complete the proof.

Without loss of generality, we assume that $p_i + q_i > 0$ for each i and that the likelihood ratios $r_i := p_i/q_i \in [0, \infty]$ satisfy $r_1 \leq \cdots \leq r_{m+2}$. Because each divergence D_{f_i} is assumed to be primitive, the convex function f_i is piecewise linear with exactly two linear parts. As a result, there exists some index $j \in \{1, \ldots, m+1\}$ such that all the functions $f_1, \ldots, f_m$ are linear in the interval $[r_j, r_{j+1}]$.

Now consider the two probability measures P^* and Q^* in $\mathcal{P}_{m+1}$ defined by

$$P^* := (p_1, \ldots, p_{j-1}, p_j + p_{j+1}, p_{j+2}, \ldots, p_{m+2})$$

and

$$Q^* := (q_1, \ldots, q_{j-1}, q_j + q_{j+1}, q_{j+2}, \ldots, q_{m+2})$$

Because of the linearity of $f_1, \ldots, f_m$ on $[r_j, r_{j+1}]$, it is easy to check that

$$D_{f_i}(P^*||Q^*) = D_{f_i}(P||Q) \geq D_i \qquad \text{for all } 1 \leq i \leq m.$$

As a result, we have

$$D_f(P^*||Q^*) \geq B_{m+1}(D_1, \ldots, D_m).$$

On the other hand, by convexity or as a consequence of the data processing inequality for f-divergences (see, for example, Csiszár and Shields [33, Lemma 4.1]), it follows that

$$D_f(P||Q) \geq D_f(P^*||Q^*) \geq B_{m+1}(D_1, \ldots, D_m).$$

The proof is complete. □

Remark 2.6.2. *Let $0 < s_1 < \cdots < s_m < \infty$ and let D_{f_i} be the primitive divergence corresponding to $f_i = u_{s_i}$ (the functions u_{s_i} are defined in Remark 2.4.2). Then the optimization problem corresponding to $B_{m+1}(D_1, \ldots, D_m)$ can be written as:*

$$\begin{aligned} \underset{p,q\in[0,1]^{m+1}}{\text{minimize}} \quad & \sum_{j:q_j>0} q_j f\left(\frac{p_j}{q_j}\right) + f'(\infty) \sum_{j:q_j=0} p_j \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for all } j = 1, \ldots, m+1 \\ & \sum p_j = \sum q_j = 1 \\ & \sum_j \min\left(p_j, q_j s_i\right) \leq \min(1, s_i) - D_i \end{aligned} \tag{2.42}$$

for $i = 1, \ldots, m$. According to Theorem 2.6.1, the optimal value of this problem equals $B(D_1, \ldots, D_m)$. As we mentioned before, the problem of determining $B(D_1, \ldots, D_m)$ when the divergences D_{f_i} are all primitive divergences has been studied by Reid and Williamson [113]. Their main result Reid and Williamson [113, Theorem 6] gives a characterization of $B(D_1, \ldots, D_m)$ that is much more complicated than (2.42)*. However, the two forms are essentially equivalent. To understand the equivalence, observe that, by Lemma 2.4.1, $D_f(P||Q)$ can be written as an integral functional of $\psi_{P,Q}$. It is possible to precisely characterize the form of the function $\psi_{P,Q}$ when $P, Q \in \mathcal{P}_{m+1}$. As a result, the optimization problem* (2.42) *can be reformulated in terms of such concave functions ψ. This, after some tedious algebra, leads to the formula for $B(D_1, \ldots, D_m)$ given in Reid and Williamson [113, Theorem 6]. Our formula* (2.42) *is much simpler and, moreover, is conceptually easier to understand.*

The special case of $m = 1$ in Theorem 2.6.1 asserts that in order to determine $B(D)$ when D_{f_1} is a primitive divergence, one only needs to consider probabilities on $\{1, 2\}$. This fact is well-known at least in the case when D_{f_1} is the total variation distance (see, for example, Gilardoni [51, Proposition 2.1]). It is then possible to give a more direct expression for $B(D)$ which is the content of the following lemma, whose special case for $s = 1$ appears in Gilardoni [51, Proposition 2.1].

Lemma 2.6.3. *Let $m = 1$ and consider the quantity $B(D)$ where D_f is an arbitrary f-divergence and D_{f_1} is the primitive divergence corresponding to $f_1 = u_s$ for a fixed $s > 0$. Then, for every $0 \leq D \leq \min(1, s)$, the quantity $B(D)$ equals*

$$\inf_{0\leq q\leq H/s} \left[(1-q) f\left(\frac{H - qs}{1-q}\right) + q f\left(\frac{1 + qs - H}{q}\right)\right] \tag{2.43}$$

where $H := \min(1, s) - D$.

Proof. We shall now show that $B_2(D)$ equals (2.43). Note that $B_2(0) = 0$ and (2.43) also equals 0 when $D = 0$. To see this, note that it is trivially zero (because $f(1) = 0$) when $s = 1$ and when $s \neq 1$, then it is zero because the value at $q = (1 - \min(1, s))/(1 - s)$ equals 0. So we shall assume below that $D > 0$. The optimization problem corresponding to $B_2(D)$ is:

$$\begin{aligned} \underset{p,q\in[0,1]^2}{\text{minimize}} \quad & \sum_{j:q_j>0} q_j f\left(\frac{p_j}{q_j}\right) + f'(\infty) \sum_{j:q_j=0} p_j \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for } j = 1, 2 \\ & p_1 + p_2 = q_1 + q_2 = 1 \\ & \min(p_1, q_1 s) + \min(p_2, q_2 s) = H. \end{aligned} \tag{2.44}$$

Note that we have equality as opposed to $\leq$ in the last constraint above. This is because of the fact that for every (p_1, p_2) and (q_1, q_2) lying in the constraint set for which the last constraint is not tight, we can get $(\tilde{p}_1, \tilde{p}_2)$ and $(\tilde{q}_1, \tilde{q}_2)$ still lying in the constraint set with the last constraint satisfied with an equality sign and for which the objective function is reduced.

We will now finish the proof by showing that the optimal value of the optimization problem (2.44) is (2.43). Let (p_1, p_2) and (q_1, q_2) satisfy the constraint set with $p_1/q_1 \leq 1 \leq p_2/q_2$. If $s \notin [p_1/q_1, p_2/q_2]$, then clearly $\min(p_1, q_1 s) + \min(p_2, q_2 s) = \min(1, s)$ and such (p_1, p_2) and (q_1, q_2) do not satisfy the constraint set because $D > 0$. So we assume that $s \in [p_1/q_1, p_2/q_2]$. In this case, the final constraint gives $p_1 = H - q_2 s$. We can therefore write each of p_1, p_2 and q_1 in terms of q_2. Plugging these values in the objective function leads to the function in (2.43) (with q replaced by q_2). The fact that each of p_1, p_2, q_1 and q_2 need to lie between 0 and 1 gives the constraint $0 \leq q_2 \leq H/s$. The proof is complete. □

For completeness, let us note the special case of the above lemma in the case of the total variation distance, which corresponds to $s = 1$. This result is due to Gilardoni Gilardoni [51, Proposition 2.1].

Corollary 2.6.4 (Gilardoni)**.** *Let $m = 1$ and consider the quantity $B(V)$ where D_f is an arbitrary f-divergence and $D_{f_1}(P||Q)$ equals $V(P, Q)$, the total variation distance between P and Q. Then, for every $0 \leq V \leq 1$,*

$$B(V) := \inf \{T(q, V) : 0 \leq q \leq 1 - V\} \tag{2.45}$$

where

$$T(q, V) := (1 - q) f\left(\frac{1 - V - q}{1 - q}\right) + q f\left(\frac{q + V}{q}\right).$$

Consequently, for every pair of probability measures P and Q, we have the inequality

$$D_f(P||Q) \geq \inf \{T(q, V(P, Q)) : 0 \leq q \leq 1 - V(P, Q)\} \tag{2.46}$$

Moreover, this represents the sharpest possible inequality between D_f and total variation distance.

Although the expression (2.45) cannot be simplified further in general, one can get much simpler expressions for $B(V)$ in certain special cases. One such special case of interest corresponds to *symmetric* f-divergences. An f-divergence is said to be symmetric if the underlying convex function f satisfies the identity: $f(x) = xf(1/x)$ for all $x \in (0, \infty)$. It is easy to check that under this condition, one has $D_f(P||Q) = D_f(Q||P)$ for all P and Q. Examples of symmetric divergences include the total variation distance, squared Hellinger distance, triangular discrimination and the Jensen-Shannon divergence. The following result is due to Gilardoni Gilardoni [51]. We include it here for completeness and also because our proof is more direct than that in Gilardoni [51].

Corollary 2.6.5 (Gilardoni)**.** *Let $m = 1$ and consider the quantity $B(V)$ where D_f is a symmetric f-divergence and $D_{f_1}(P||Q)$ equals $V(P, Q)$, the total variation distance between P and Q. Then, for every $0 \leq V \leq 1$,*

$$B(V) = (1 - V)f\left(\frac{1 + V}{1 - V}\right). \tag{2.47}$$

Consequently, for every pair of probability measures P and Q, we have

$$D_f(P||Q) \geq (1 - V(P, Q))f\left(\frac{1 + V(P, Q)}{1 - V(P, Q)}\right). \tag{2.48}$$

Moreover, this represents the sharpest possible inequality between the symmetric divergence D_f and total variation distance.

Proof. We shall show that the right hand side of (2.45) equals the right hand side of (2.47) when D_f is a symmetric divergence. Consider the quantity $T(q, V)$ defined in Corollary 2.6.4. Because $f(x) = xf(1/x)$, it can be easily checked that

$$T(q, V) = T(1 - q - V, V) \qquad \text{for all } q \in [0, 1 - V].$$

In other words, the function $q \mapsto T(q, V)$ is symmetric in the interval $[0, 1 - V]$ about the mid-point $(1-V)/2$. Moreover, as can be checked by taking derivatives (one-sided derivatives if f is not differentiable), $q \mapsto T(q, V)$ is convex on $[0, 1 - V]$ (this fact does not require f to be symmetric). These two facts clearly imply that

$$\inf_{0 \leq q \leq 1-V} T(q, V) = T\left(\frac{1 - V}{2}, V\right) = (1 - V)f\left(\frac{1 + V}{1 - V}\right)$$

which completes the proof. □

2.6.2 Chi-squared divergence

In this section, we describe another situation where the conclusion of Theorem 2.3.1 can be further simplified.

Theorem 2.6.6. *Let $m = 1$ and consider the quantity $A(D)$ where D_f is the chi-squared divergence, $\chi^2(P||Q)$ which corresponds to $f(x) := x^2 - 1$. Also let the function f_1 be such that the function $h : (0, \infty) \to (0, \infty)$ defined by $h(x) := (1 + f_1(x))/x$ is a strictly increasing, strictly convex, twice differentiable bijective mapping. Then $A(D) = h^{-1}(D + 1) - 1$, where h^{-1} denotes the inverse function of h on $(0, \infty)$.*

Proof. By Theorem 2.3.1, $A(D)$ equals the optimal value of the problem:

$$\begin{aligned} \underset{p,q\in[0,1]^3}{\text{maximize}} \quad & \sum_{j:q_j>0} \frac{p_j^2}{q_j} - 1 + \infty \cdot \sum_{j:q_j=0} p_j \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for all } j = 1, 2, 3 \\ & \sum p_j = \sum q_j = 1 \\ & \sum_{j:q_j>0} q_j f_1\left(\frac{p_j}{q_j}\right) + f_1'(\infty) \sum_{j:q_j=0} p_j \leq D \end{aligned}$$

By convexity of h, we have

$$h(x) \geq h(a) + h'(a)(x - a) \tag{2.49}$$

for every $x > 0$ and $a > 0$. One consequence of this and the fact that h is strictly increasing is that

$$h(1) + h'(1)(x - 1) \leq h(x) \leq h(1)$$

for all $x \in (0, 1)$. This implies that $\lim_{x\downarrow 0} xh(x) = 0$ and as a result

$$f_1(0) = \lim_{x\downarrow 0} f_1(x) = \lim_{x\downarrow 0} (xh(x) - 1) = -1$$

Further, because h is strictly increasing, we have $h'(a) > 0$ and thus

$$f_1'(\infty) = \lim_{x\to\infty} h(x) = \infty$$

which implies that we only need to consider P and Q for which $\sum_{j:q_j=0} p_j = 0$. Writing (2.49) in terms of $f_1(x)$, we obtain

$$1 + f_1(x) \geq x\left(h(a) - ah'(a)\right) + x^2 h'(a).$$

for every $x > 0$ and also at $x = 0$ (because $f_1(0) := \lim_{x\downarrow 0} f_1(x)$). Applying this inequality to $x = p_j/q_j$ for $q_j > 0$ and then multiplying by q_j, we obtain

$$q_j + q_j f_1(p_j/q_j) \geq p_j\left(h(a) - ah'(a)\right) + \frac{p_j^2}{q_j} h'(a)$$

for each $j = 1, 2, 3$. As a result, we get

$$h'(a) \sum_{j:q_j>0} \frac{p_j^2}{q_j} \leq \sum_{j:q_j>0} q_j f_1\left(\frac{p_j}{q_j}\right) + 1 - h(a) + ah'(a)$$

Because P and Q satisfy the constraint, we have

$$\sum_{j:q_j>0} q_j f_1\left(\frac{p_j}{q_j}\right) \leq D$$

and hence

$$\sum_{j:q_j>0} \frac{p_j^2}{q_j} - 1 \leq \left[\frac{D+1-h(a)+ah'(a)}{h'(a)}\right] - 1.$$

Because $a > 0$ is arbitrary, we get

$$A(D) \leq \inf_{a>0} \left[\frac{D+1-h(a)+ah'(a)}{h'(a)}\right] - 1.$$

By elementary algebra, the above infimum is achieved at $a^* = h^{-1}(D+1)$ and we then obtain $A(D) \leq h^{-1}(D+1)-1$. To see that $A(D)$ is exactly equal to $h^{-1}(D+1)-1$, observe that the probabilities $P = (1, 0, 0)$ and $Q = (1/a^*, 1-1/a^*, 0)$ satisy $D_{f_1}(P||Q) = D$ and $\chi^2(P||Q) = h^{-1}(D+1) - 1$. □

The function $f_1(x) = x^l - 1$ for $l > 2$ clearly satisfies the conditions of the above theorem. We therefore obtain the following result as a simple corollary.

Corollary 2.6.7. *Let $m = 1$ and consider the quantity $A(D)$ where $D_f(P||Q) = \chi^2(P||Q)$ and D_{f_1} is the power divergence, $D^{(l)}(P||Q)$, corresponding to $f_1(x) = x^l - 1$ for $l > 2$. Then $A(D) = (1+D)^{1/(l-1)} - 1$. This yields the sharp inequality*

$$\chi^2(P||Q) + 1 \leq \left(1 + D^{(l)}(P||Q)\right)^{1/(l-1)}$$

between the chi-squared divergence and power divergence for $l > 2$.

2.7 Numerical Computation

In this section we explore numerical methods for solving the optimization problems (2.7) and (2.8) in order to compute $A(D_1, \ldots, D_m)$ and $B(D_1, \ldots, D_m)$ respectively. In Section 2.7.1, we consider the special case when D_f is a primitive divergence. This special case is motivated by the statistical problem of obtaining lower bounds for the minimax risk and we show that the quantity $A(D_1, \ldots, D_m)$ can be computed exactly via convex optimization for every $m \geq 1$ and every arbitrary choice of $D_{f_1}, \ldots, D_{f_m}$. In Section 2.7.2, we consider the special case $m = 1$ and demonstrate that (2.7) and (2.8) can be solved for practically any pair of f-divergences by a gridded search over the low-dimensional parameter speace. We verify several known inequalities and also improve on some existing inequalities that are not sharp.

2.7.1 Maximizing Primitive Divergences

In this subsection we consider maximizing a primitive divergence subject to upper bounds on arbitrary f-divergences. While this optimization problem is not a-priori convex, we reduce it to a collection of convex problems.

The optimization problem (2.7) where D_f is a primitive divergence is, of course, closely related to the problem of bounding from above a primitive divergence subject to upper bounds on other f-divergences. This latter problem arises in obtaining lower bounds for the minimax risk in nonparametric statistical estimation (see, for example, Guntuboyina [52], Guntuboyina [53], Yu [146], and Tsybakov [135]). For example, Le Cam's inequality, which is a popular technique for obtaining minimax lower bounds, says that the minimax risk is bounded from below by a multiple of the L_1 affinity between two probability measures P and Q, where the L_1 affinity is defined as $1 - V(P, Q)$. The L_1 affinity also appears in Assouad's Lemma, another technique for obtaining minimax lower bounds. Evaluating $V(P, Q)$ is hard because P and Q are typically product distributions of the form $P = \otimes_{i=1}^{n} P_i$ (or mixtures of such distributions), so it is difficult to express $V(P, Q)$ in terms of $V(P_i, Q_i)$ (which can be easier to compute).

Application of Le Cam's inequality in practice, therefore, requires one to obtain a good upper bound on the total variation, $V(P, Q)$. One typically first bounds $D_f(P||Q)$ for an f-divergence that decouples for product distributions such as squared Hellinger, chi-squared, or Kullback-Leibler divergence and then translates this into a bound on $V(P, Q)$. It is common to use crude bounds like Pinsker's inequality for this purpose and we believe there is room for improvement by using tight bounds. Also, one typically uses only one f-divergence to bound $V(P, Q)$; but we shall argue here that one gets better bounds (Figure 2.3) when using multiple divergences simultaneously. This is one of our motivations for studying the case $m \geq 2$ as opposed to just $m = 1$. The constants underlying minimax lower bounds might be improved by the use of these better bounds addressing a common criticism of minimax lower bound techniques.

Theorem 2.7.1 below solves the problem of maximizing a primitive divergence D_{u_s} given constraints on m other divergences D_{f_i} exactly via convex optimization. This leads to a fast algorithm with well-studied convergence properties.

For each $m \geq 1$, let

$$\mathcal{S}_m = \{\sigma \in \{-1, 1\}^{m+2} : \sigma_i \leq \sigma_j \text{ for } i \leq j\}$$

For each $\sigma \in \mathcal{S}_m$, let us consider the following *convex* optimization problem and denote its

optimal value by $V_\sigma(D_1, \ldots, D_m)$.

$$\begin{aligned} \underset{p,q\in[0,1]^{m+2}}{\text{maximize}} \quad & \sum_{j=1}^{m+2} \sigma_j \left(p_j - s q_j\right) \\ \text{subject to} \quad & p_j \geq 0,\ q_j \geq 0 \text{ for all } j = 1, \ldots, m+2 \\ & \sum p_j = \sum q_j = 1 \\ & \sum_{j:q_j>0} q_j f_i\left(\frac{p_j}{q_j}\right) + f_i'(\infty) \sum_{j:q_j=0} p_j \leq D_i \end{aligned} \tag{2.50}$$

for $i = 1, \ldots, m$. Note that this problem is convex because the objective function is linear and the constraint set is convex in $p_1, \ldots, p_{m+2}, q_1, \ldots, q_{m+2}$. The fact that the constraint set is convex is a consequence of the convexity of $D_{f_i}(P||Q)$ in (P, Q) (see, for example, Csiszár and Shields [33, Lemma 4.1]). It is also clear that this is a $2m+2$-dimensional optimization problem because there are $2m+4$ variables in all which satisfy two linear equality constraints.

Theorem 2.7.1. *Let D_f denote the primitive f-divergence corresponding to $f = u_s$ for some $s > 0$. Then*

$$A(D_1, \ldots, D_m) = -\frac{|s-1|}{2} + \max_{\sigma\in\mathcal{S}_m} V_\sigma(D_1, \ldots, D_m) \tag{2.51}$$

Consequently, $A(D_1, \ldots, D_m)$ can be computed by solving the $|\mathcal{S}_m| = m+3$ convex optimization problems (2.50).

Proof. Theorem 2.3.1 asserts that $A(D_1, \ldots, D_m)$ equals the optimal value of the optimization problem (2.7). Note that the constraint sets of the problems (2.7) and (2.50) are the same. Let us denote this constraint set by $\mathbb{C}_m$ so that

$$A(D_1, \ldots, D_m) = \max_{P,Q\in\mathbb{C}_m} D_{u_s}(P||Q).$$

The objective of (2.7) can be written as

$$\begin{aligned} D_{u_s}(P||Q) &= \min(1, s) - \sum_{j=1}^{m+2} \min(p_j, s q_j) \\ &= \min(1, s) - \frac{1}{2} \sum_{j=1}^{m+2} p_j + s q_j - |p_j - s q_j| \\ &= -\frac{|s-1|}{2} + \max_{\sigma\in\{-1,1\}^{m+2}} \sum_{j=1}^{m+2} \sigma_j \left(p_j - s q_j\right). \end{aligned}$$

Because two maxima can always be interchanged, we have

$$\max_{P,Q\in\mathbb{C}_m} \left[\max_{\sigma\in\{-1,1\}^{m+2}} \sum_{j=1}^{m+2} \sigma_j (p_j - s q_j) \right] = \max_{\sigma\in\{-1,1\}^{m+2}} \left[\max_{P,Q\in\mathbb{C}_m} \sum_{j=1}^{m+2} \sigma_j (p_j - s q_j) \right].$$

Note that the inner maximization in the right hand side above is precisely the convex problem (2.50).

Because the optimal value of (2.50) is invariant to permuting the indices of σ, we have the reduction

$$\max_{\sigma\in\{-1,1\}^{m+2}} \max_{P,Q\in\mathbb{C}_m} \sigma^T(P - sQ) = \max_{\sigma\in\mathcal{S}_m} \max_{P,Q\in\mathbb{C}_m} \sigma^T(P - sQ).$$

This shows that we can restrict attention only to those problems (2.50) for $\sigma \in \mathcal{S}_m$. It is obvious that $|\mathcal{S}_m| = m + 3$. The proof is complete. □

Example 2.7.2. *Consider the special case of Theorem 2.7.1 when $m = 1$, $s = 1$ and when D_{f_1} is the squared Hellinger distance which corresponds to $f_1(x) = (\sqrt{x} - 1)^2/2$. In other words, we consider the problem of maximizing the total variation distance subject to an upper bound on the Hellinger distance. The solution to this problem given by Theorem 2.7.1 is plotted in Figure 2.1(a). Each red dot shows $A(H) =: A_H^{TV}(H)$ computed by solving the four 4-dimensional convex optimization problems* (2.50) *(each corresponding to a $\sigma \in \mathcal{S}_1$).*

Note that the quantity $A_H^{TV}(H)$ can be obtained analytically in a closed form. Indeed, since f_1 is a symmetric divergence, the sharp inequality bounding the total variation distance by the squared Hellinger distance is given by (2.48) *with $f(x) = (\sqrt{x} - 1)^2$ (this inequality is usually attributed to Le Cam [72]) which implies that*

$$A_H^{TV}(H) = \sqrt{2H}\sqrt{1 - \frac{H}{2}}.$$

We have plotted this function analytically by the solid cyan line in Figure 2.1(a). It is clear that our numerical optimization method given by Theorem 2.7.1 agrees with the known analytical bound.

Example 2.7.3. *For another simple application of Theorem 2.7.1, consider maximizing the total variation subject to an upper bound on the Kullback-Leibler divergence. In other words, we take $m = 1$, $s = 1$ and $f_1(x) = x \log x$ and plot the solution given by Theorem 2.7.1 in Figure 2.1(b). Each black dot shows $A(K) =: A_{KL}^{TV}(K)$ for a different value of K, computed by solving the four 4-dimensional convex optimization problems* (2.50)*. The solid green line shows Pinsker's analytic upper bound $\sqrt{2K}$ which is not sharp for any $K > 0$.*

Example 2.7.4. *We now consider maximizing the total variation subject to constraints on both the Hellinger distance and Kullback-Leibler divergence. In other words, we take $m = 2$, $s = 1$, $f_1(x) = (\sqrt{x} - 1)^2/2$ and $f_2(x) = x \log x$. To the best of our knowledge, there does not exist a closed form analytical solution to this problem. However, numerical solution is straightforward by Theorem 2.7.1 as shown below.*

According to Theorem 2.7.1, for fixed $H, K \geq 0$ we can compute $A(H, K) =: A_{HKL}^{TV}(H, K)$ by solving five 6-dimensional convex programs (2.50)*. Figure 2.2 shows the function $A_{HKL}^{TV}(H, K)$*

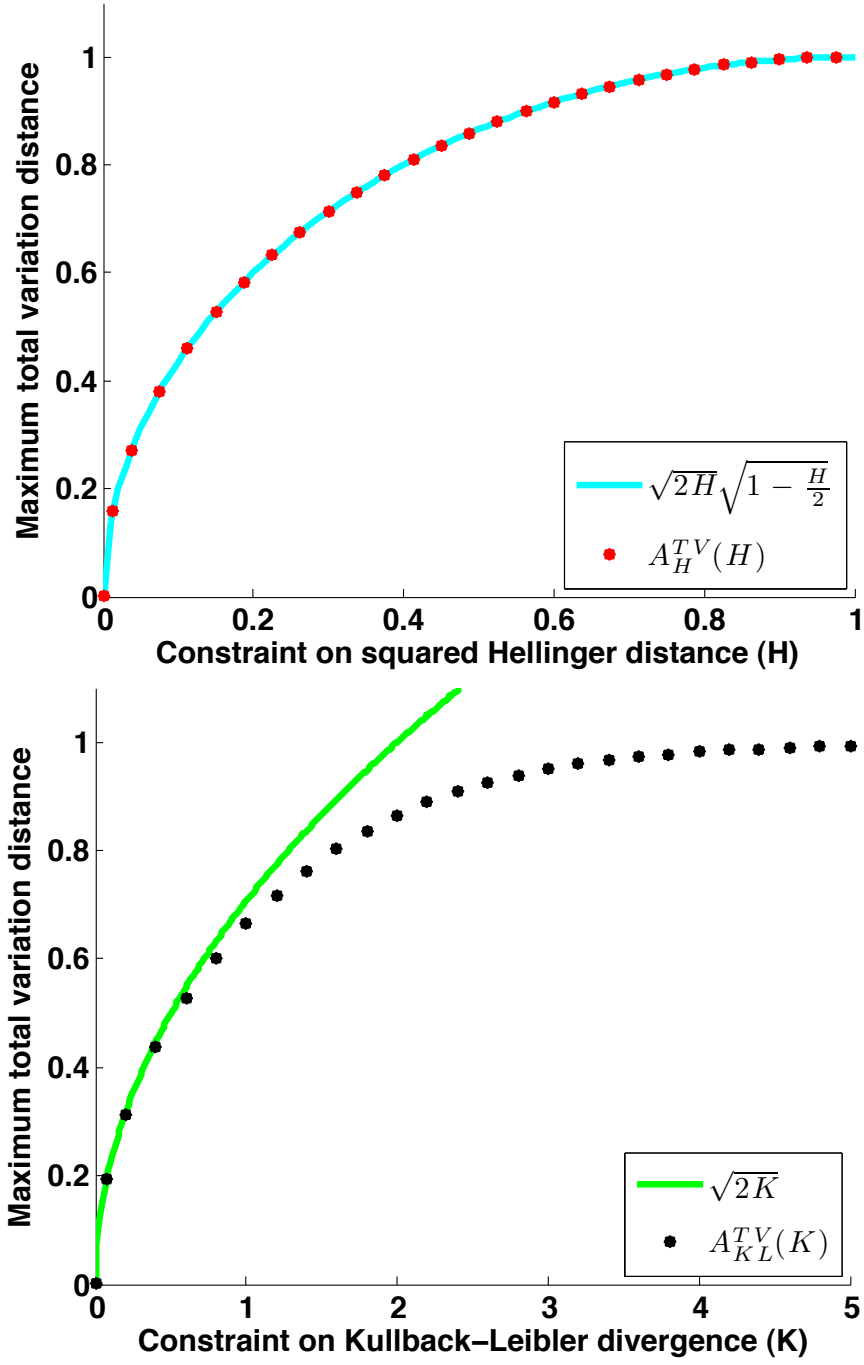


Figure 2.1: Two simple applications of Theorem 2.7.1 discussed in examples 2.7.2 and 2.7.3. Here and in all subsequent plots we set the axis limits to the maximum value of the relevant f-divergence and to 5 in the case of the Kullback-Leibler divergence (which has no maximum value).

interpolated from 14884 (H, K) pairs. We used CVX in MATLAB to solve the convex programs. The height of each point in the surface shows how large the total variation can be when the squared Hellinger distance and Kullback-Leibler divergence are bounded by H and K respectively. As expected, the total variation is zero when either $H = 0$ or $K = 0$, and it approaches 1 for large values of H and K. Next, observe that the surface $A^{TV}_{HKL}(H, K)$ is flat as K varies for small H, and vice-versa flat as H varies for small K. This is because only one constraint is tight in these regions. In other words, the surface $A^{TV}_{HKL}(H, K)$ is approximately the point-wise minimum of the two surfaces $A^{TV}_{H}(H)$ and $A^{TV}_{KL}(K)$, with a diagonal ridge at the intersection of these two surfaces. But, as can be seen in Figure 2.3, our bound that simultaneously leverages both single-coordinate bounds is strictly better than the simple minimum of those two individual bounds for some (H, K). In other words, there exist (H, K) such that

$$\min\left(A^{TV}_{H}(H), A^{TV}_{KL}(K)\right) - A^{TV}_{HKL}(H, K) > 0 \tag{2.52}$$

The left hand side above is positive when both single-coordinate bounds are informative, i.e. when both constraints in the optimization problem (2.7) *are active. We will explain later (see Example 2.7.5 and Figure 2.4) that the location of this ridge is predicted by an inequality between $D_H(P||Q)$ and $D_{KL}(P||Q)$.*

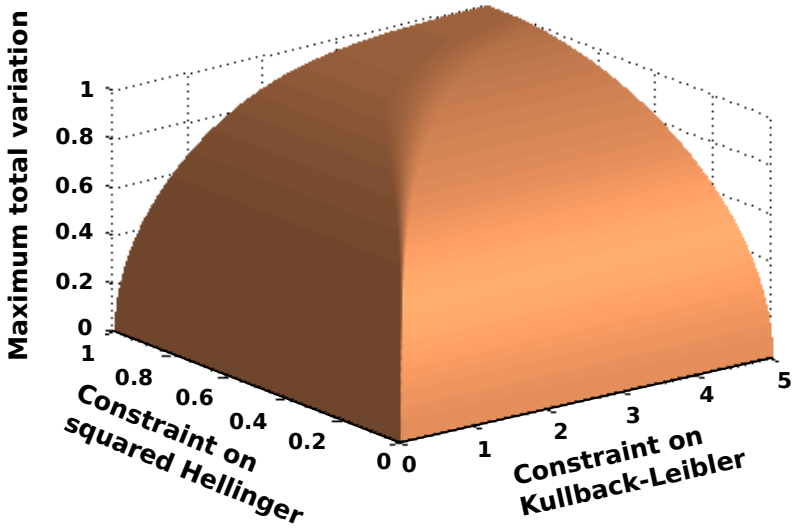


Figure 2.2: The height of each point in the surface above shows $A^{TV}_{HKL}(H, K)$ for a different (H, K) pair–the the least upper bound on total variation when squared Hellinger distance and Kullback-Leibler divergence are bounded by H and K respectively (see example 2.7.4).

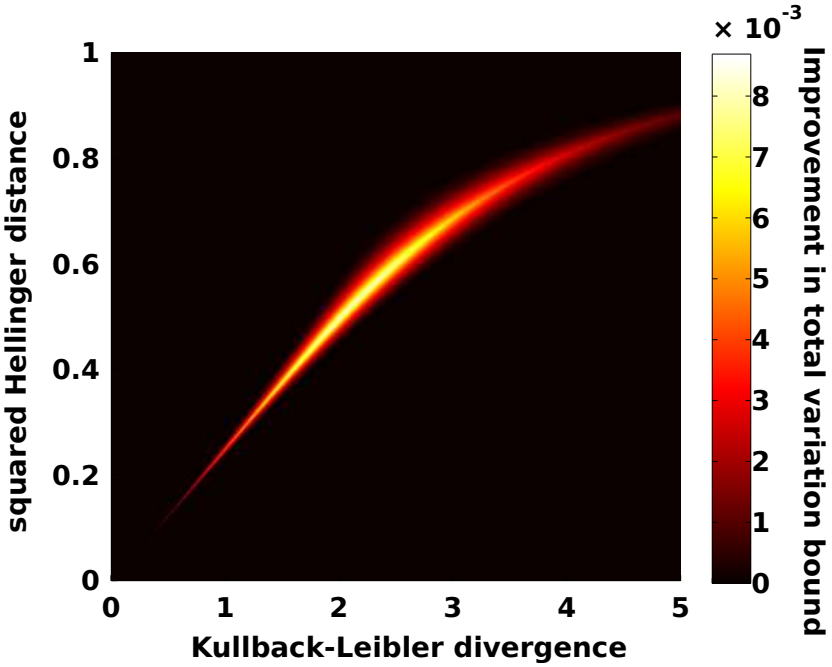


Figure 2.3: Improvement over simple point-wise minimum of single-coordinate bounds. The color of the pixel at (H, K) represents the magnitude of the left hand side of (2.52). The bright region corresponds to (H, K) for which the bound displayed in Figure 2.2 is a strict improvement over the simple pointwise minimum of the two bounds shown in Figure 2.1.

2.7.2 The General Case

Theorem 2.7.1 requires D_f to be a primitive divergence. We do not know if, in general, the optimization problems (2.7) and (2.8) can be solved by convex optimization algorithms. However, if m is not too large, heuristic optimization techniques can be used. We demonstrate this in this subsection for $m = 1$.

Example 2.7.5. *Consider the optimization problem* (2.7) *for* $m = 1$, $f(x) = (\sqrt{x}-1)^2/2$ *and* $f_1(x) = x \log x$. *In other words, we consider the problem of maximizing the squared Hellinger distance subject to an upper bound on the Kullback-Leibler divergence. The optimization problem* (2.7) *is clearly 4-dimensional (there are six variables in all* p_1, p_2, p_3 *and* q_1, q_2, q_3 *but they satisfy two linear constraints as they sum to one). Because the variable space is only 4-dimensional, there was no trouble solving this by gridding the parameter space. We plot the solution in Figure 2.4(a) where each blue dot shows* $A(K) =: A^H_{KL}(K)$ *for a different value of* K.

The quantity $A^H_{KL}(K)$ *can be used to better understand the inequality* (2.52). *Indeed, when we overlay the curve* $(K, A^H_{KL}(K))$ *on Figure 2.3 (see Figure 2.4(b)), we see that the curve* $(K, A^H_{KL}(K))$ *(plotted by the blue line) lies above the region where the inequality* (2.52) *holds. Only the constraint on* $D_{KL}(P||Q)$ *is active in the optimization problem considered*

in Example 2.7.4 when $H > A_{KL}^H(K)$. For such (H, K), therefore, the inequality (2.52) does not hold.

Example 2.7.6. *Consider maximizing the squared Hellinger distance between P and Q with the total variation between P and Q, $V(P,Q)$, bounded by V. In other words, we consider the special case of the problem (2.7) for $m = 1$, $f(x) = (\sqrt{x} - 1)^2/2$ and $f_1(x) = |x - 1|/2$. This is a special case of the problem we considered in section 2.5.3 where we proved that $A_2(V) < A_3(V)$ for all $V \in (0, 1)$. Here we confirm this fact numerically.*

We compute both the quantities $A_2(V)$ and $A_3(V)$ by a gridded search over pairs of probabilities satisfying the constraint in $\mathcal{P}_2$ and $\mathcal{P}_3$ respectively. These functions are plotted in Figure 2.5. Each red triangle in Figure 2.5 shows $A_3(V)$ for a different V. Each point in the dotted blue line shows $A_2(V)$ for a different V. It is evident that the inequality $A_2(V) < A_3(V)$ holds for all $V \in (0, 1)$. In other words, when we restrict the constraint set to probability measures in $\mathcal{P}_2$, the maximum Hellinger distance is strictly smaller for all $V \in (0, 1)$. Therefore, Theorem 2.3.1 is in general tight and cannot be improved.

Note also that the plot $A_3(V)$ agrees with the form $A_3(V) = A(V) = V(f(0)+f'(\infty)) = V$ derived in Section 2.5.3. This gives rise to the sharp inequality $H^2(P,Q) \leq V(P,Q)$ which is again attributed to Le Cam [72].

Example 2.7.7. *The capacitory discrimination between two probability measures P and Q is defined by*

$$C(P,Q) = D_{KL}\left(P||\frac{P+Q}{2}\right) + D_{KL}\left(Q||\frac{P+Q}{2}\right).$$

It is easy to check that $C(P,Q)$ is an f-divergence that corresponds to the convex function:

$$x \log x - (x+1)\log(x+1) + 2\log 2. \tag{2.53}$$

The triangular discrimination $\Delta(P,Q)$ is another f-divergence that corresponds to the convex function

$$\frac{(x-1)^2}{x+1}. \tag{2.54}$$

Topsøe proved the following inequality between these two f-divergences Topsøe [134]:

$$\frac{1}{2}\Delta(P,Q) \leq C(P,Q) \leq (\log 2)\Delta(P,Q). \tag{2.55}$$

Let us investigate here the sharpness of these inequalities. Let

$$A(D_1) := \sup\{C(P,Q) : \Delta(P,Q) \leq D_1\}$$

and

$$B(D_1) := \inf\{C(P,Q) : \Delta(P,Q) \geq D_1\}.$$

We solved the optimization problems (2.7) and (2.8) for $m = 1$, $f(x)$ given by (2.53) and $f_1(x)$ given by (2.54) by a gridded search. The resulting solutions for $A(D_1)$ and $B(D_1)$ are

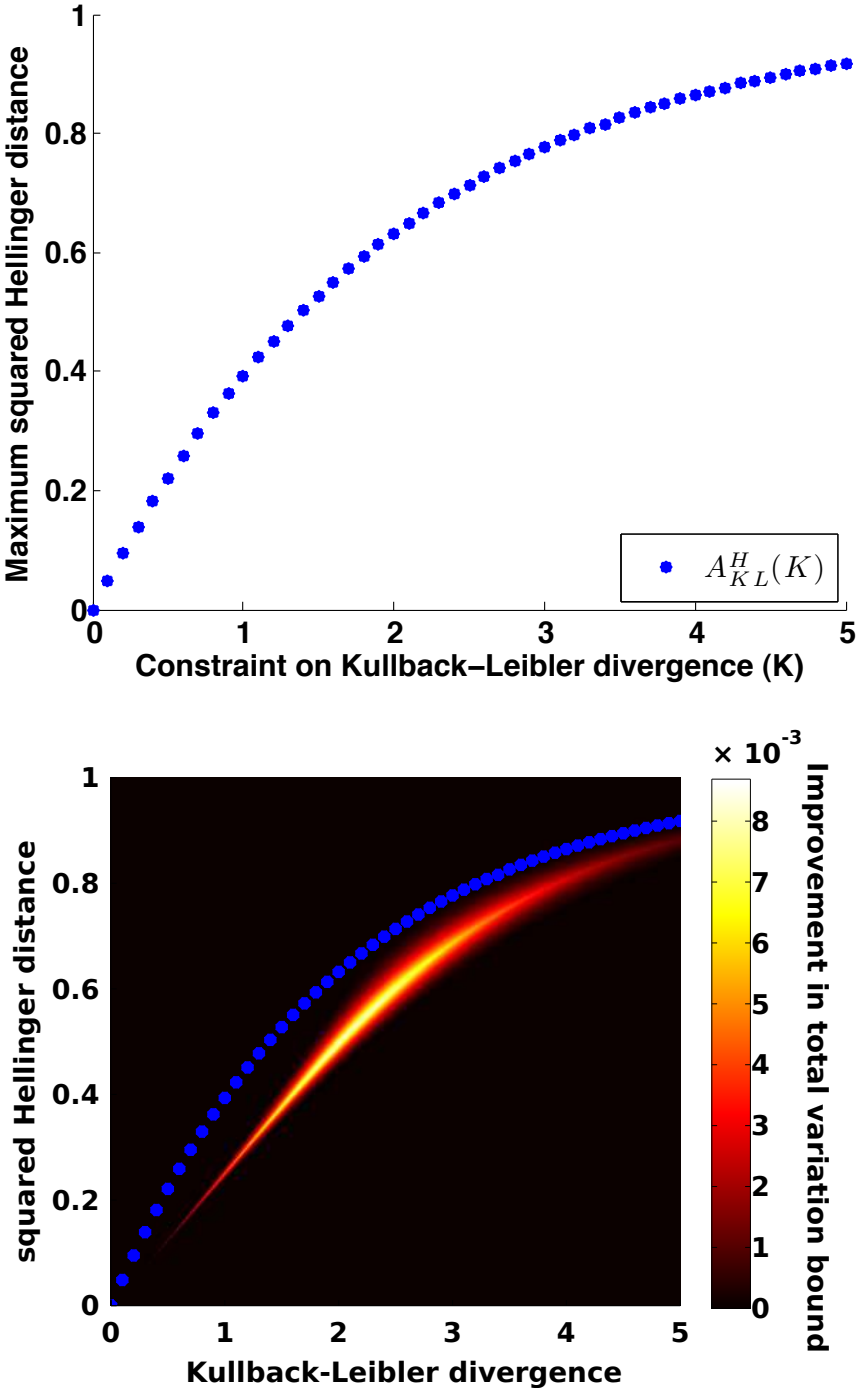


Figure 2.4: A sharp inequality between squared Hellinger distance and Kullback-Leibler divergence bounds the support of the ridge. The upper panel displays a sharp inequality between squared Hellinger and Kullback-Leibler divergence. The height of each blue dot represents the optimal value $A^H_{KL}(K)$ with a different constraint, K, on the Kullback-Leibler divergence. The lower panel shows the same blue curve overlaid on Figure 2.3. Observe that the region with positive improvement is bounded by the blue curve from the upper panel.

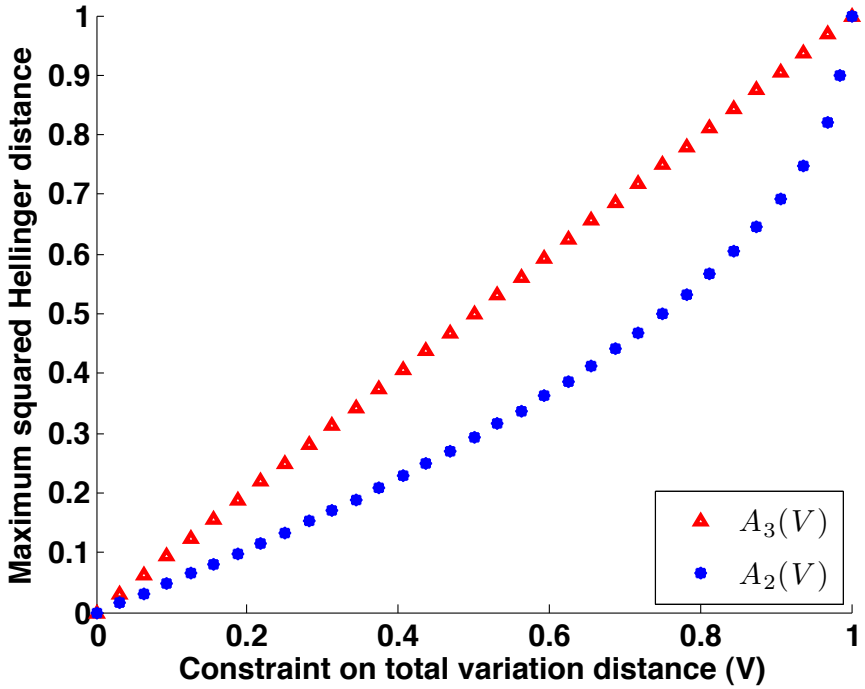


Figure 2.5: Three point measures strictly improve on two point measures. Each red triangle shows $A_3(V)$ computed by a gridded search over pairs of probability measures in $\mathcal{P}_3$. Each blue dot shows $A_2(V)$ computed by a gridded search over pairs of probability measures in $\mathcal{P}_2$. The simulation over three point measures is exactly a straight line with slope one–agreeing with Le Cam's bound $H^2 \leq V$. And $A_2(V) < A_3(V)$ for all $V \in (0, 1)$.

plotted in Figure 2.6, with red triangles corresponding to $A(D_1)$ and blue dots corresponding to $B(D_1)$. We have also plotted the bounds given by (2.55) *in Figure 2.6 with the green line corresponding to* $(\log 2)D_1$ *and the blue line to* $D_1/2$*. It is clear from the figure that the upper bound in* (2.55) *is sharp while the lower bound is not sharp. The sharp lower bound is given by $B(D_1)$. We are unaware of an analytic formula for $B(D_1)$, but we conjecture that $B_2(D_1) = B_3(D_1)$ because this equality holds numerically. It may be possible to use this fact to find an analytic formula for $B(D_1)$.*

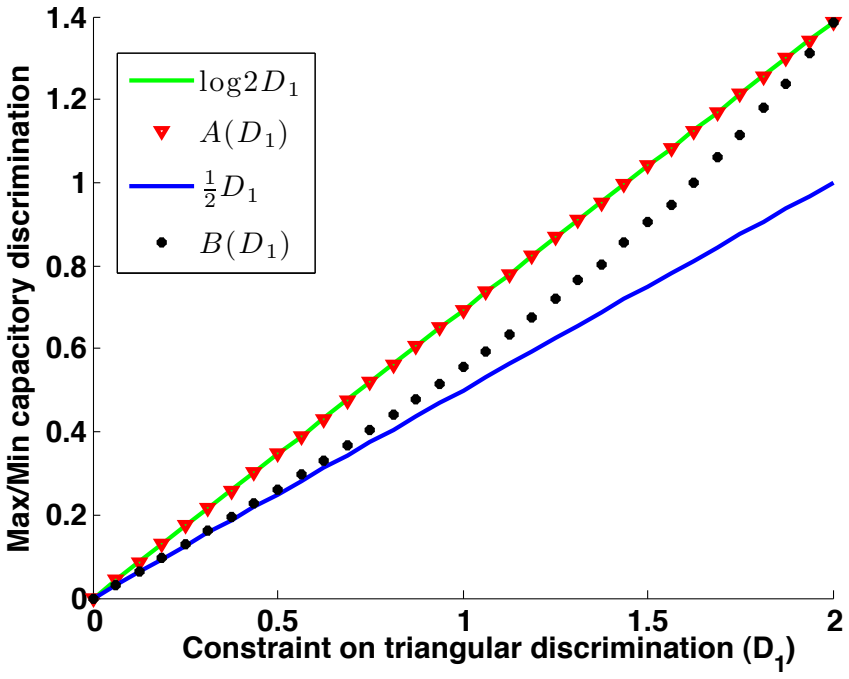


Figure 2.6: The green line with slope $\log 2$ and the blue line with slope $\frac{1}{2}$ trace the bounds in (2.55), while the red triangles and the black dots display $A(D_1)$ and $B(D_1)$ respectively.

Chapter 3

Supervised random projections and their role in high-dimensional inference

3.1 Introduction

High-dimensional supervised learning problems are encountered in numerous popular modern-day scientific applications, ranging from genomics, biomedical studies, astronomy and sociological studies. In each of these fields, the core goal of statistical analysis requires inference or uncertainty measures for decision making. As such, a holistic framework for statistical inference in high-dimensional supervised learning problems serves a paramount advantage to practitioners. In more precise terms, in the present article we consider supervised learning problems such as regression, classification and randomized experiments of a high-dimensional nature, viz. in datasets where numerous covariables are available for consideration, often comparable or exceeding the number of samples/observations. To this end, we first develop a new dimension reduction technique called *Supervised Random Projections* (SRP) and further, develop a framework for statistical inference in high-dimensional supervised learning problems based on SRP.

A crucial advantage of our proposal is modularity. Our method can be broken into three easy modules: (i) estimation of variable importances, (ii) supervised dimension reduction (iii) low-dimensional inference. Each of these steps are decoupled in their implementations; allowing practitioners immense freedom in choosing and combining the most effective and creative data-scientific methods and domain knowledge to reach the desired balance of speed and accuracy. We emphasize that this modularity ensures ease of communication of concepts and implementability of our solution.

Dimension reduction is a crucial and necessary step towards inference on large datasets (Bingham and Mannila [13], Blum [16], and Benner, Mehrmann, and Sorensen [10]). Alongwith the immediate advantages of computational and storage economy in statistical problems

dealing with very large datasets, dimension reduction has also been employed to remove ambient noise (Indyk and Motwani [60] and Kambhatla and Leen [65]); thus promoting interpretability. Random projections are one powerful dimensional reduction technique especially known for their simplicity and computational efficiency (see Vempala [138] for a book reference) . They have been successfully applied to many supervised learning tasks in the machine learning literature. An incomplete listing of the relevant papers here is: Fradkin and Madigan [46], Rahimi and Recht [110], Maillard and Munos [87], Maillard and Munos [88], Paul et al. [105], Pilanci and Wainwright [107], Heinze, McWilliams, and Meinshausen [56], Li [76], Zhou, Wasserman, and Lafferty [151], and Zhang et al. [150, 149].

While employing dimension reduction in a supervised learning problem, not all variables are equally important; instead the dimension reduction scheme should attempt to preserve important variables (ones that influence the response strongly) at the expense of less important ones.

In the context of supervised problems, we introduce the idea of using variable importance indices in the dimension reduction/data compression step to gain higher accuracy in downstream inference. We demonstrate that if important variables are allowed more influence in the compression step, the resultant projected data retains more information about the response variable, as evidenced by the higher accuracy of downstream inference (as a first example, consider the MSE of estimates based on supervised and unsupervised dimension reduction in the problem of estimating a linear loading in a high-dimensional linear regression model as shown in Figure 3.1).

Our idea of supervised dimension reduction is exceedingly simple: given a set of covariates, Random Projections constructs a smaller set of projected covariates simply by taking random linear combinations. In a supervised problem, some covariates are more important than others. Given an additional index of the importance of each variable, we propose to construct projected variables by taking random linear combinations *weighted* by variable importances. We call this technique *Supervised* Random Projections (SRP).

Supervised dimension reduction enables a new and modular approach to high-dimensional statistical inference. An overriding difficulty in many such problems is the presence of a large number of covariables. In contrast, if only a few covariables are considered, commonly such problems can be solved to satisfaction by classical statistical methodology dating back years and decades. Our proposal then becomes evident: using supervised dimension reduction we project the inference problem at hand onto a small number of covariables (while ensuring that the parameter of interest is preserved as much as possible); subsequently we tap into the immense wealth of 'low-dimensional' inferential procedures available in statistical literature and implemented in popular statistical programming languages.

We illustrate the efficacy of the above blueprint via applications in three popular high-dimensional statistical inference problems which have exhibited noticeable activity in recent years, viz. (i) identifying statistically significant variables in a (non-linear) regression model with many covariables, (ii) with the additional assumption of linearity, (iii) quantifying uncertainty in estimated Average Treatment Effect (ATE) in a randomized experiment with many covariates. In each problem, we present a simple fast approximate *silver bullet* which

is compared against the need to redevelop specialized inferential procedure separately for each high-dimensional problem whose low-dimensional analog is already well understood.

Specifically, our main contributions are as follows:

(a) We introduce the idea of supervised dimension reduction, with the goal of ensuring that in comparison to ordinary dimension reduction, the projected data is more relevant to the response variable at hand. By incorporating variable importances, we explicate that the projected data should still accurately explain the response variable (this is in contrast to ordinary dimension reduction, where one only attempts to preserve the geometry between covariables); thus lending more interpretability to the dimension reduction step. Further, variable importances ensure that even in the presence of numerous nuisance parameters, the projected data retains at least a moderate amount of information from the important variables, thus allowing said important variables a fair chance at being selected by downstream formal tests of hypotheses.

(b) We rigorize the idea of Supervised Random Projections (SRP) in Section 3.2 and demonstrate its advantages in estimation of a linear regression coefficient. Further, we discuss the role of specific variable importance indices used in the projection step and a satisfactory choice of projected dimension.

(c) Through formal tests of hypotheses, we identify significant variables in a high-dimensional non-linear regression problem by applying SRP in conjuction with Random Forests (Breiman [20]) in Section 3.3. We demonstrate that SRP is faster and slightly better than the non-parametric permutation based test proposed in Altmann et al. [4] when interactions are absent; though it fails to control Type I error when interactions are present. In Figure 3.2(c) we demonstrate that in the presence of pairwise interactions and gaussian design, SRP finds $\approx 30\%$ insignificant variables to be significant, even though it was nominally calibrated to make such errors at the level of 10%; in the same experiment, the permutation test described in Altmann et al. [4] manages to keep such errors below 10% and maintain the same proportion of true discoveries.

(d) Under the additional assumption of linearity of regressor, we study the applicability of SRP in conjuction with Lasso in Section 3.4. We demonstrate that SRP performs just as well and in the same runtime as all methods except debiasing (Zhang and Zhang [148]); which requires a much higher computational overhead. In comparison with debiasing, SRP does well in terms of coverage but is more conservative, leading to only approximately 60% as many discoveries as found by the significantly (10-30$\times$) slower debiasing method.

(e) Finally, in the problem of statistical inference on Average Treatment Effect (ATE) in the presence of many covariates we demonstrate the applicability of SRP in Section 3.5. We demonstrate that SRP is only slightly inferior ($\approx 10\%$ higher variance) compared to the Lasso-adjusted method presented in Bloniarz et al. [15]. SRP is much

faster than the refitting based (Lasso+OLS) procedure and requires the same computational overhead as the Lasso procedure, both proposed in Bloniarz et al. [15]. The (Lasso+OLS) procedure requires (6-8×) as much computational time as SRP or the Lasso procedure.

3.1.1 Literature Review

Inference on covariate significance (via hypothesis testing) in the high-dimensional non-linear model via the Random Forests framework has been discussed in Strobl et al. [126], Mentch and Hooker [101], Altmann et al. [4] and Paul, Verleysen, Dupont, et al. [104].

Inference in the high-dimensional linear model under sparsity has been subject to intense research in the statistics community in the past few years. Notable existing approaches include (a) methods based on sample splitting Wasserman and Roeder [141], Meinshausen, Meier, and Bühlmann [100], and Meinshausen and Bühlmann [99], (b) debiasing methods Zhang and Zhang [148], Bühlmann [23], Geer et al. [47], and Javanmard and Montanari [62], (c) post-selection inference Berk et al. [11], Lockhart et al. [86], Lee et al. [74], and Lee and Taylor [73], (d) bootstrap Chatterjee and Lahiri [26] and Liu and Yu [85], and (e) strict bounds Stark [124] and Meinshausen [98]. The recent survey paper Dezeure et al. [37] discusses the merits and demerits of these methods in detail.

Inference for the ATE under the Neyman model has also received significant attention in the literature. Let us first mention here that the Neyman model for completely randomized experiments is also known as the Neyman-Rubin model; standard references include Splawa-Neyman, Dabrowska, and Speed [123], Rubin [119, 118], and Holland [58]. A simple estimator for the ATE was proposed in Splawa-Neyman, Dabrowska, and Speed [123] which does not use any information on the covariates. In the low dimensional setting, it is known that covariates can help in improving estimation and inference for the ATE via regression adjustments (see, for example, Lin [79]). Inference for the ATE in the presence of high-dimensional covariates was studied in Bloniarz et al. [15] who proposed a LASSO based regression adjustment method for the ATE. A more recent work, Wager et al. [139], also deals with the high-dimensional setting but they impose more assumptions on the data-generating mechanism that are usually not part of the Neyman model.

3.2 Adaptive Random Projections

In this section, we introduce the idea of *supervised* random projection (SRP). The main contribution of SRP (as applicable to supervised problems) over ordinary random projections (ORP) is in employing a flexible and more informative dimension reduction step which incorporates additional information about which covariates have higher influence on the response. In practice, this ensures that the projected data contains more information pertinent to the response variable compared to ORP. In the following, we describe SRP in this general setting:

Let $\{(Y_i, X_i)\}_{i=1}^n$ be n independent observations arising from the model

$$Y_i = f(X_i) + \epsilon_i, \qquad \mathbb{E}(\epsilon_i | X_i) = 0 \tag{3.1}$$

where $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^p$. In many such problems, we have at our disposal standard regression methods which provide an *information* index for each variable X^j. As an illustration, Random Forests (Breiman [20]) provides internal estimates of variable importance; in the regression context this importance index measures the increase in residual variance of the response if a certain variable is deleted. Additionally, if we were privy to the structure of f, for instance if f is known to be linear in variables X^j, we also have at our disposal myriad linear regression estimates (such as Lasso (Tibshirani [129]) and Ridge (Tihonov [131]) estimators) where we can interpret the estimated linear loading (in absolute value) of each variable as an estimate of variable importance. This interpretation also extends to the case of generalized linear models with a known link function.

We are now ready to describe SRP. Let $\hat{v}_j$ denote any user-chosen estimate of variable importance of X^j; and m the number of random projections or projected variables.

Algorithm 1 `SRP-dimReduce`: SRP based dimension reduction

Input $X_{n\times p}, \hat{\mathbf{v}}, m$
Draw supervised random projections matrix $A_{p\times m}$ with independent entries $a_{ij} \sim N(0, \hat{v}_j/m)$
Return compressed variable matrix $U_{n\times m} \leftarrow XA$

This simple SRP-dimReduction algorithm (Algorithm 1) provides an universal approach to supervised data compression and is our main contribution. As can be gleaned from Algorithm 1 the only difference from ORP is in the incorporation of variable importances in the sizes of random projection coefficients. The algorithm above requires an additional tuning parameter m. We shall discuss the role of m shortly.

Example 1 We now illustrate the advantages of this supervised algorithm in comparison to ORP in the linear regression setting. For this small demonstration, let us draw n i.i.d. observations from the linear model

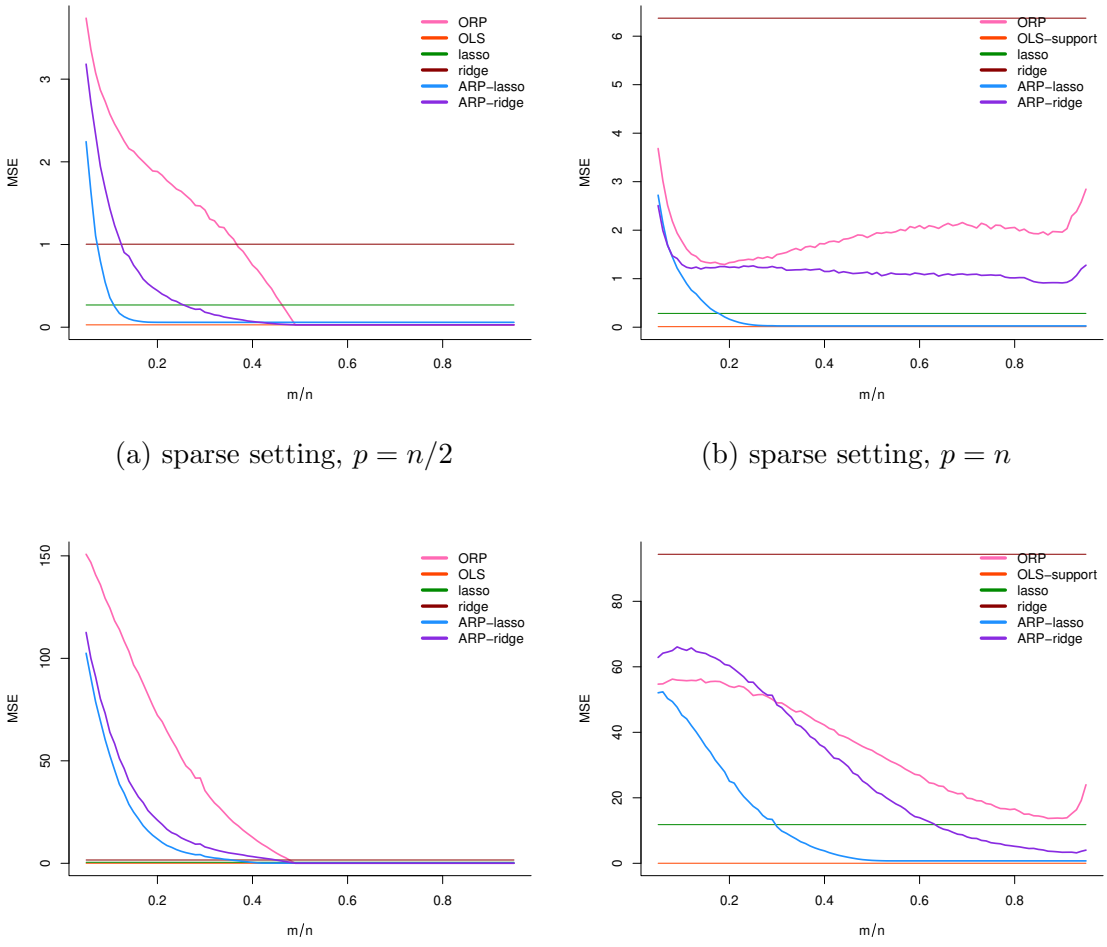
$$Y_i = X_i^\top \beta^* + \epsilon_i, \qquad \epsilon_i \sim \mathbb{N}(0, \sigma^2). \tag{3.2}$$

We will compare the MSEs of the SRP and ORP estimates of β_1^* for various values of m. For our SRP estimate, we consider two different choices of $\hat{v}$: provided by cross-validated Lasso and cross-validated Ridge estimates. To be precise, two possible variable importances of X^j are $|\hat{\beta}_j^{(lasso)}(\lambda_{cv,lasso})|$ and similarly, $|\hat{\beta}_j^{(ridge)}(\lambda_{cv,ridge})|$. We now describe the SRP estimate of β_1^*, and in general β_k^* for an integer k. We should clarify that we only need to preserve the parameter of interest, in this case the scalar β_1^*.

Description of Figure 3.1: Plots of resulting MSE for different values of m. In each plot, the x-axis plots the fraction m/n. For this small study, we set $n = 100$. For the

Algorithm 2 `SRP`: SRP estimate of β_k^* in high-dim linear regression

Input $Y_n, X_{n\times p}, \hat{\mathbf{v}}, m, k$
Construct supervised compressed matrix $U \leftarrow$ `SRP-dimReduce`$(X_{-k}, \hat{\mathbf{v}}_{-k}, m)$
Perform linear regression of Y on $[X_k, U]$ and return estimated linear loading corresponding to X_k



(a) sparse setting, $p = n/2$

(b) sparse setting, $p = n$

(c) dense setting, $p = n/2$

(d) dense setting, $p = n$

Figure 3.1: Comparison between two different dimension reduction techniques - ORP and SRP in a linear regression problem.

left hand plots, we set $p = n/2 = 50$ and for the right hand plots we set $p = n = 100$. For the top plots we set the first $s = 20$ entries of β^* to be non-zero (sparse setting). For the bottom plots we set the first $s = 50$ entries of β^* to be non-zero (dense setting). The design matrix X is constructed following the gaussian random design with equicorrelated covariance matrix (correlation between any pair of variables is 0.2). Non-zero entries of β^* are drawn independently from $N(0, 4^2)$. ϵ is set to be distributed as $N(0, 1)$. $10,000$ replications are performed in each setting keeping X fixed. The lines show MSE (lower is better) in estimating β_1^* achieved by each of the estimators mentioned in the legend. Note that OLS, Lasso and Ridge estimators are not tuned using m and thus describe horizontal lines in the plots. For the right hand plots where $p = n$, the OLS estimate, while defined, is ill-behaved and is replaced by OLS-support; i.e. performing linear regression on only the relevant variables.

The resulting MSE in estimating β_1^* for varying m is showing in Figure 3.1. From Figure 3.1, our main observation is that SRP is markedly more accurate in each of the four settings compared to ORP and the Lasso and Ridge estimators. This validates our proposal of using supervised dimension reduction over naive dimension reduction. Additionally, we note that `SRP-lasso` is more accurate than `SRP-ridge`, and the Lasso and Ridge estimates. Finally, we draw attention to choice of m. We should note that the computational complexity of SRP directly increases with m. This reveals a tradeoff in Figure 3.1, concentrating on `SRP-lasso`, accuracy demands a large value of m while computational speed demands a small value of m. Indeed, the optimal choice for `SRP-lasso` seems simple and intuitive: $m \approx s$; where s is the number of non-zero entries in β^*. In practice, s is unknown; we recommend the universal choice of setting $m = n/2$. In the following sections, we shall employ the idea of sample splitting (to be explained in detail in the subsequent algorithms) in which case the number of available observations for compressed regression is lower than n. In these cases, we recommend choosing m to be half the number of observations available in the current split.

3.3 Inference in non-linear regression problems

In this section, we propose a new method for statistical inference in non-linear models by drawing upon the respective advantages of non linear regression via Random Forests and dimension reduction via Supervised Random Projections. We consider a standard non-linear regression setting as described in (3.1). In this model, we address the problem of identifying statistically significant variables (via formal tests of hypotheses). In the context of Random Forests, this problem has been addressed by Paul, Verleysen, Dupont, et al. [104], Mentch and Hooker [101] and Altmann et al. [4]. We build our SRP based significance test upon the **R** package **ranger** (Wright and Ziegler [144]). **ranger** ships with an implementation of the non-parametric permutation based test proposed in Altmann et al. [4] (`PIMP`), which we use as our main comparison in this section.

In the regression context, we define the support S of a function $f(x_1, \ldots, x_p) : \mathbb{R}^p \mapsto \mathbb{R}$

as the *minimal* set of indices so that for any vector z_S of appropriate length, $f\big|_{x_S=z_S}$ is a constant function (its range is a singleton). The support S then precisely determines the set of significant variables, intuitively, the set of variables that appear explicitly in the definition of f. In the case of random design, the support is well-defined whenever no two variables are perfectly collinear. Since, in this extreme case we can not identify the 'truly' significant variable from the insignificant one based only on the data, we exclude this possibility from the rest of our discussions.

SRP targets the linear projection of the non-linear model described in (3.1). Denote by S, the support of f. Define,

$$\beta^*_S = \underset{\beta}{\operatorname{argmin}}\, \mathbb{E}_{X_S}[X_S\beta - f(X_S)]^2, \qquad \beta^*_j = 0 \quad \text{for any } j \notin S.$$

Definition: Linearly Important Variable (LIV) X^j is called a linearly important variable (LIV) if $\beta^*_j \neq 0$.

Consequently, we choose to answer the statistical test of the linear significance of X^j by querying whether the confidence interval of β^*_j based on data (Y, X) contains 0 or not. Note that due to the above implication this modification can only err on the side of caution; at the population level some significant variables may be deemed insignificant due ot loss of information in the linear projection of (3.1), but an insignificant variable can not become significant due to the linear projection. Using the notion of sample splitting, we propose Algorithm 3 for constructing SRP confidence intervals.

Algorithm 3 SRP confidence interval for β^*_k using variable importances computed via Random Forests

Input $Y_n, X_{n\times p}, m, \alpha$

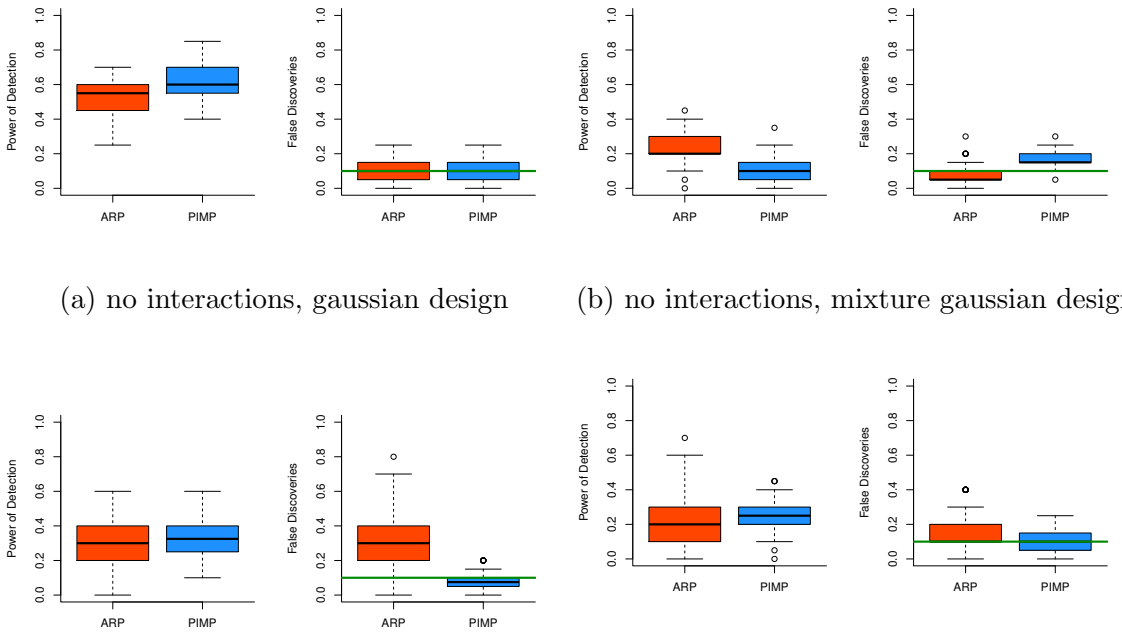
Split available data randomly into two sets of equal size $\{Y^{(1)}, X^{(1)}\}$ and $\{Y^{(2)}, X^{(2)}\}$

Using Random Forests on the first subsample $\{Y^{(1)}, X^{(1)}\}$, compute variable importances $\hat{v}_j$ corresponding to each variable X^j

Construct supervised compressed matrix $U^{(2)} \leftarrow \texttt{SRP-dimReduce}(X^{(2)}_{-k}, \hat{\mathbf{v}}_{-k}, m)$

Perform linear regression of $Y^{(2)}$ on $[X^{(2)}_k, U^{(2)}]$ and return confidence interval of linear loading corresponding to $X^{(2)}_k$

Description of Figure 3.2: Boxplots of rate of false discoveries (lower is better) and detection (higher is better). Green horizontal line indicates that the nominal level of significance tests are set at 0.1. For this study, we set $n = p = 500$. For the left hand plots, X is constructed following the gaussian random design with equicorrelated covariance matrix (correlation between any pair of variables is 0.2). For the right hand plots, X is constructed as mixture of gaussians with same covariance matrix as in the previous sentence but with means drawn from $0.5\delta_{\mathbf{2}}+0.5\delta_{-\mathbf{2}}$. For the top plots, $f(x) = \sum_{i=1}^{20} \beta_j \mathbf{1}(x^j \geq c_j)$ where $c_j \sim N(0, 0.5^2)$ and $\beta_j \sim |N(0, 2^2)|$. Thus for the top plots, f contains non-linear threshold functions but no interactions. For the bottom plots, $f(x) = \sum_{i=1}^{10} \beta_j \mathbf{1}(x^{2j-1} \leq c_{2j-1}, x^{2j} \geq c_{2j})$, where $c_\cdot$ and



(a) no interactions, gaussian design

(b) no interactions, mixture gaussian design

(c) pairwise interactions, gaussian design

(d) pairwise interactions, mixture gaussian design

Figure 3.2: Comparison between SRP and PIMP in a non-linear regression setting.

$\beta_\cdot$ are drawn as before. Thus for the bottom plots, f contains both non-linear terms and pairwise interactions. ϵ is set to be distributed as $N(0, 1)$. 100 replications were performed in each setting.

As shown in Figure 3.2, SRP is either comparable or better than PIMP whenever f does not contain interactions. When f does contain pairwise interactions, SRP seems to exceed the nominal level. Specially in the case of gaussian design and pairwise interactions, exceedance seems to be quite severe. We should note here, identification of interactions using Random Forests is not yet a completely understood topic. In a very recent paper, Basu et al. [9] provides a deeper discussion and specialized methodology on this issue.

To conclude this section, we note that in the settings described in Figure 3.2, SRP requires on average 358 seconds to conclude while PIMP requires on average 524 seconds; implying that supervised dimension reduction has resulted in a 30% gain in computational overhead.

3.4 Inference in linear regression problems

In this section, we discuss the problem of constructing confidence intervals for the linear loadings of each variable (β_j^* corresponding to X^j) in the linear model (3.2). Our methodology here is near identical to Algorithm 3, with the variable importance being estimated

by cross-validated Lasso instead of Random Forests. We choose to implement SRP in conjunction with Lasso since this performs better than the Ridge analog for the closely related problem of estimating β_j^*, as exhibited in Section 3.2. The simplicity of modifying Algorithm 3 to arrive at Algorithm 4 underlines the wide applicability and modular nature of SRP.

Algorithm 4 SRP confidence interval for β_k^* using variable importances computed via LASSO

Input $Y_n, X_{n\times p}, m, \alpha$

Split available data randomly into two sets of equal size $\{Y^{(1)}, X^{(1)}\}$ and $\{Y^{(2)}, X^{(2)}\}$

Using cross-validated Lasso on the first subsample $\{Y^{(1)}, X^{(1)}\}$, compute variable importances

$$\hat{v}_j = |\hat{\beta}_j(\lambda_{cv})|$$

Construct supervised compressed matrix $U^{(2)} \leftarrow \texttt{SRP-dimReduce}(X^{(2)}_{-k}, \hat{\mathbf{v}}_{-k}, m)$

Perform linear regression of $Y^{(2)}$ on $[X_k^{(2)}, U^{(2)}]$ and return confidence interval of linear loading corresponding to $X_k^{(2)}$

The problem of constructing confidence intervals (CIs) for each entry of β^*, and equivalently testing for the significance of each variable X^j is a widely studied problem under various sparsity and regularity assumptions. Dezeure et al. [37] provides a good survey and **R** implementation of various popular methods in this field of study. SRP does not explicitly make opaque assumptions on the structure of the problem beyond linearity, but performs better whenever such regularity (such as sparsity of the vector β^*, restricted isometry of X) are present. In the sequel, we present some simulation studies comparing SRP with popular methods for statistical inference in high-dimensional linear models.

For the purpose of all simulation studies presented in this section, we choose to compare ordinary random projections (`ORP`) and SRP based on Lasso variable importance (`SRP-Lasso`) with three popular existing methods: debiasing (`LDPE`) as formulated in Zhang and Zhang [148], `LPR` (Liu and Yu [85]) which is based on the bootstrap followed by preferential regularization and finally `ssLasso`, which modifies the sample splitting based method proposed in Wasserman and Roeder [141] (as published, this method splits the available sample into two halves, uses a model selection procedure, viz. Lasso, in the first half and proceeds to compute CIs via OLS on the selected model) to preclude the possibility of singleton confidence intervals by always including the variable X^j in the selected model.

Figures 3.3, 3.4, 3.5, 3.6 and 3.7 present our findings based on a myriad of different simulation settings within the purview of high-dimensional linear models. For the sake of precision, we note that a CI counts towards coverage if its target is inside the CI; a CI of a non-zero target counts towards power if it does not contain 0.

Description of Figure 3.3: Linear model with gaussian errors. Performance of methods in a linear model with gaussian errors. Random projections methods are shown in solid lines while other methods are shown in dashed lines. For this study we set $n = 200, p = 500$. X is constructed as a gaussian random design with covariance Σ satisfying Σ^{-1} is sparse

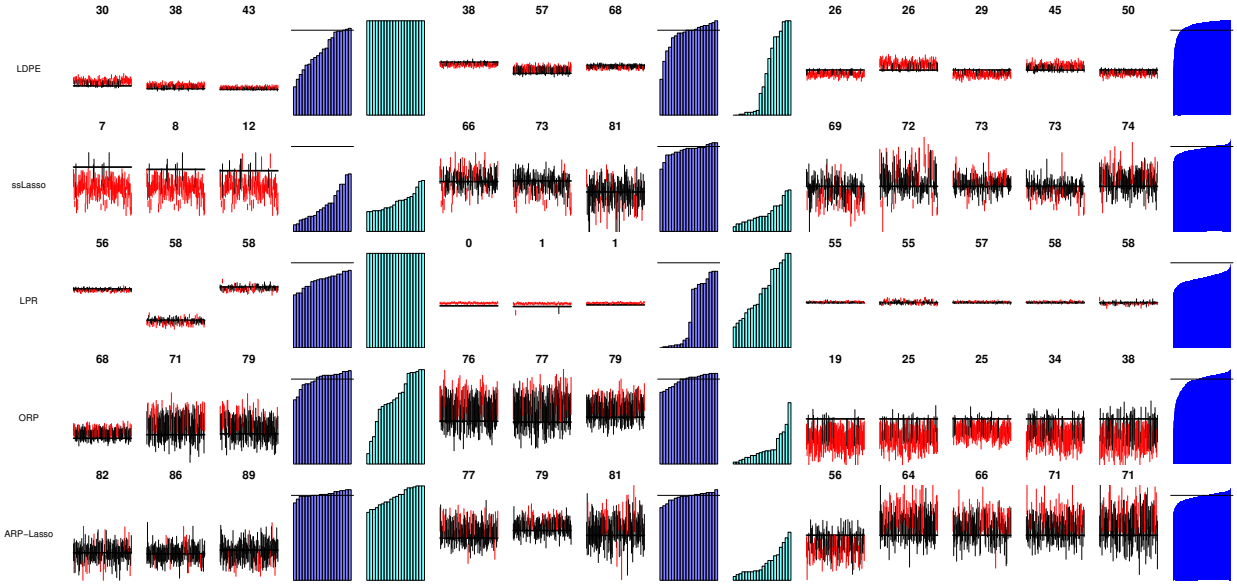


Figure 3.3: Performance in data with gaussian distributions.

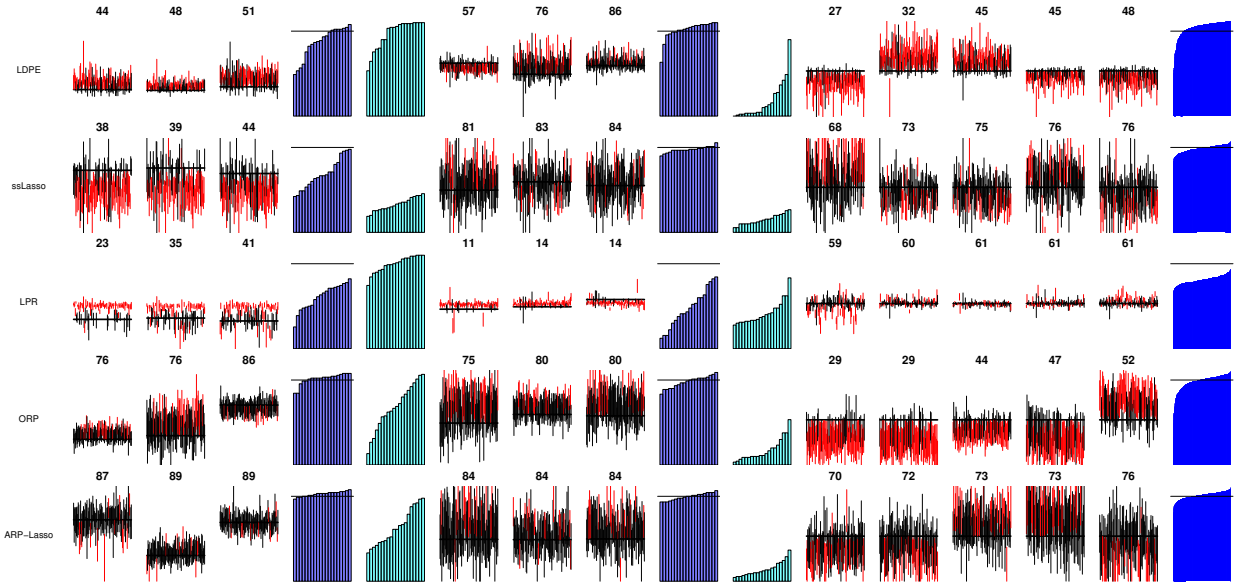


Figure 3.4: Performance in data with heavy-tailed distributions.

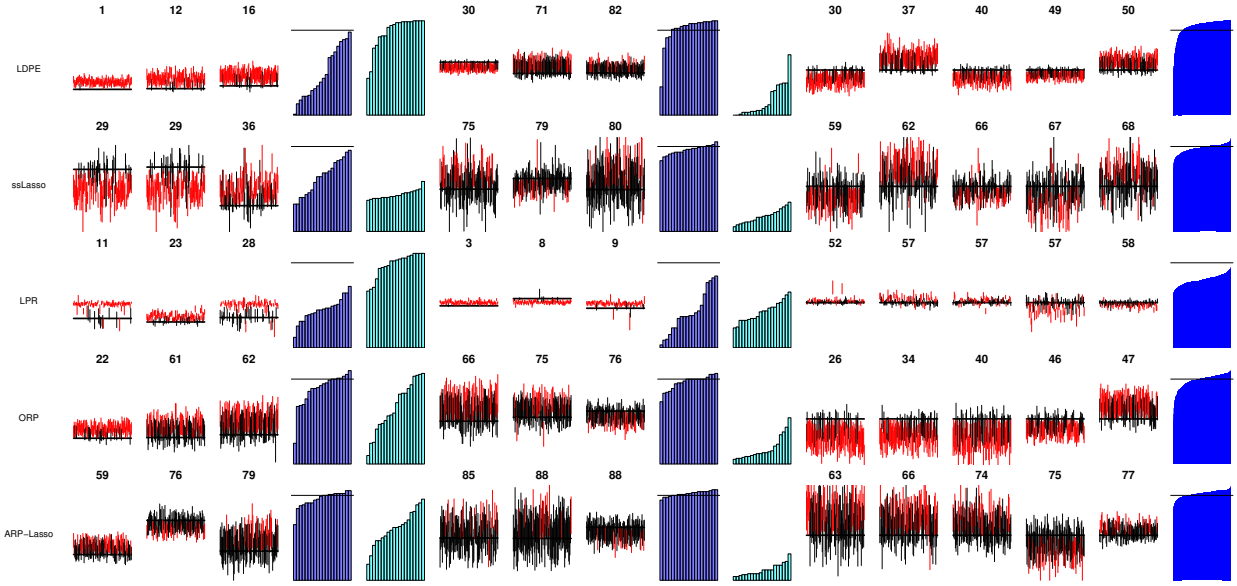


Figure 3.5: Performance in data with outliers.

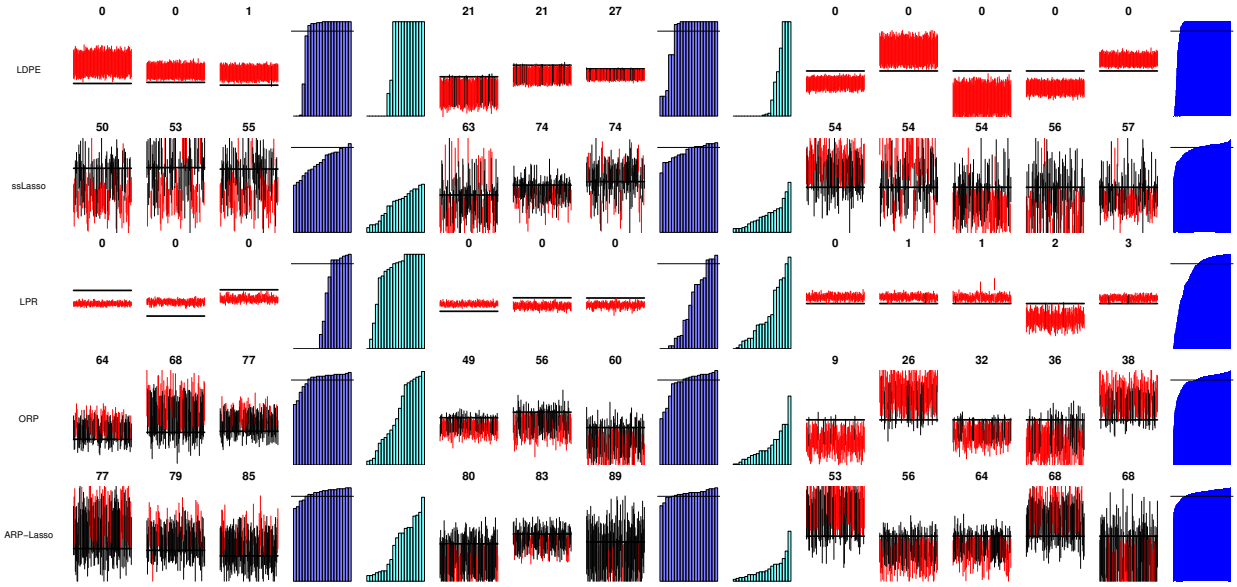


Figure 3.6: Performance in data with missing covariates.

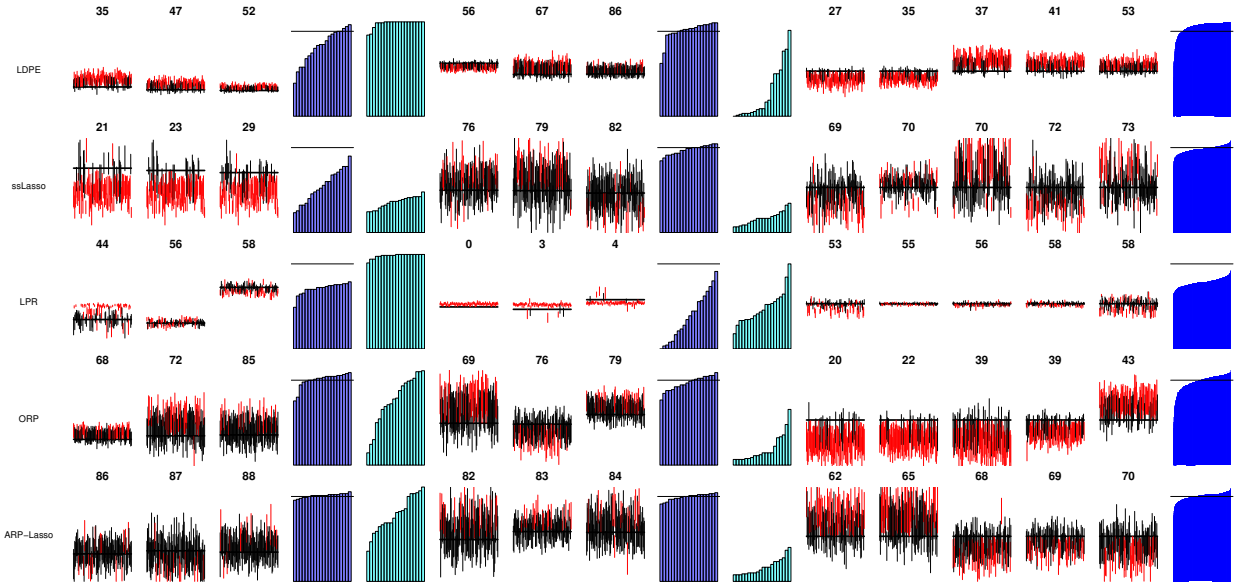


Figure 3.7: Performance in data with heterogenous errors.

and all eigenvalues of Σ are between $1/5$ and 5. 20 indices of β^* are chosen at random and populated with independent copies of $\text{Unif}((-2,-1) \cup (1,2))$; these constitute 'big' entries of β^0. 20 disjoint indices in β^* are chosen at random and populated with independent copies of $\text{Unif}(-0.75, 0.75)$; these constitute 'small' entries of β^*. The rest of β^* is filled with zeros. Within the course of one experiment, all of these objects are fixed. Nominal level of all confidence intervals are 90%. 100 replicates are performed where ϵ is drawn from $N(0, 2^2)$.

The plot in Figure 3.3 is a modification of those available in Dezeure et al. [37]. Each row of the above plot is dedicated to a different method. Subplots containing red and black lines represents an individual coefficient of β^*. The horizontal line depicts the true value of this coefficient and each vertical line depicts a confidence interval (CI). CIs are colored black if they cover the true value and are otherwise colored red. Overall coverage in 100 replicates is printed at the top of these subplots. Each row from left to right shows for the current method, the three 'big' coefficients with worst coverage, coverage of all 20 'big' coefficients (in sorted order), power for all 20 'big' coefficients, the three 'small' coefficients with worst coverage, coverage of all 20 'small' coefficients, power for all 20 'small' coefficients, the three zero coefficients with worst coverage, coverage of all 460 zero coefficients.

Description of Figure 3.4: Heavy-tailed errors. The specifics of this study are same as in Figure 3.3 with the only exception that ϵ is now distributed as the Student's t-distribution with 2 degrees of freedom.

Description of Figure 3.5: Data containing outliers. The specifics of this study are same as in Figure 3.3 with the only exception that each entry of Y can be replaced by an independent $N(0, 1)$ with 20% chance.

Description of Figure 3.6: Data with missing variables. In this study, the response Y

depends on 20 further covariates which are not supplied to the methods under study. The linear loadings corresponding to missing variables are drawn independently from Unif$(-2, 2)$

Description of Figure 3.7: Data with heterogenous errors. The specifics of this study are same as in Figure 3.3 with the only exception that $\epsilon_i \sim N(0, 4(1 + |X_i^1|))$

3.4.1 Discussion on the role of SRP in inference in high-dimensional linear models

In Figures 3.3, 3.4, 3.5, 3.6 and 3.7 we have investigated the properties of `ORP`, `SRP-Lasso`, `LDPE`, `LPR` and `ssLasso` in the high-dimensional linear model and some practically crucial variants of it. Below we itemize a few important observations from this extensive study which shed light on the worth of Random Projections based methods in this field.

To clarify somewhat further, blue subplots show overall performance of each method while the subplots in red and black highlight worst case behavior. For the plots in blue, coverages (plotted in dark blue) should ideally match the flat line at 0.90 while power(plotted in light blue) should ideally be as high as possible. For the worst case plots, it is desirable to have more black lines and less red lines, the number on top of these plots counts the number of black lines out of 100.

(i) Both the random projections based methods, `ORP` and `SRP-Lasso`, exhibit conservative behavior in comparison to `LDPE` and `LPR`. They enjoy valid coverage across the board (demonstrating resilience to the various forms of misspecification, or corruption, we have introduced in Figures 3.4 - 3.7) while their power of detection is lower than those of `LDPE` and `LPR`. In contrast, `LPR` seems to be more aggressive, choosing to trade in coverage for higher power of detection. This behavior is exhibited by the below nominal coverage; the only exception to this rule being the coverage of zero coefficients in Figure 3.6.

(ii) In all of the settings, except in Figure 3.3, there is little to differentiate between the behavior of `ORP` and `SRP-Lasso` in terms of overall metrics. In Figure 3.3, `SRP-Lasso` shows better power of detection for 'big' coefficients in comparison to `ORP`. At a more subtle level, all figures show that the worst case behavior of `SRP-Lasso` is better than `ORP`, demonstrating that `SRP-Lasso` is able to often produce valid confidence intervals even in the worst case.

This leads us to conclude that SRP is reliant upon the quality of variable importances. Figure 3.3 refers to the case of an well-behaved high-dimensional linear model with no further bells and whistles. In this setting, we expect the cross-validated Lasso estimates to be of high fidelity which is leveraged by `SRP-Lasso` to push ahead of `ORP`. In contrast, in the presence of heavy-tailed errors, outliers, missing covariates or heterogeneous errors the estimated variable importances suffer in quality; which in turn diminishes the gap between `ORP` and `SRP-Lasso`.

(iii) `ssLasso`, while a simple method, shows the weakest performance among the methods studied here. In all of the figures, we observe that in terms of overall metrics, `ssLasso` exhibits lower coverage (worse) and lower power (worse) for non-zero coefficients in comparison to `ORP` and `SRP-Lasso`. We should note that `ssLasso` exhibits valid coverage of zero coefficients across all settings.

(iv) We would be amiss not to comment upon the robustness demonstrated by `LDPE` in this study. With the only exception of subpar worst case behavior in Figure 3.5, `LDPE` achieves the right balance between reserve and aggression. It should also be noted that `LDPE` is also, by a fat margin, the most expensive procedure under study here. In comparison to `LDPE`, `LPR` provides an aggressive alternative while `ORP` and `SRP-Lasso` provide conservative alternatives; in addition, `SRP-Lasso` is able to demonstrate higher power compared to `ORP` if accurate variable importances are available.

3.4.2 Comparison of Computational Complexities

For practical applicability, the observations made in Section 3.4.1 need to be weighed against the computational cost of each method. Recall that n is the number of observations, p is the number of variables and m is the number of random projections (equivalently, the dimension of the compressed covariable matrix). We denote by B, the number of bootstrap replicates used by `LPR`. Using these notation, the theoretical computational complexity of each of the methods discussed above is tabulated below (assuming $n \leq p$) in Table 3.1: The theoretical

Method	Order of Computation
`LDPE`	p^4
`ORP`	pnm^2
`SRP-Lasso`	$\max\{p^3, pnm^2\}$
`ssLasso`	$\max\{p^3, pn\hat{s}^2_{\texttt{ssLasso}}\}$
`LPR`	Bp^3

Table 3.1: Theoretical leading order of computation of all methods discussed in Section 3.4. $\mathcal{O}$ notation is suppressed for readability.

complexities are not always accurate reflections of empirical running times, specially for iterative algorithms. In Figure 3.10, we visualize the empirically observed running times. Figure 3.10 clearly exhibits that `LDPE` is a very expensive procedure requiring $\approx 10 - 30\times$ runtime compared to the next most expensive method (`ORP`). The other four methods are arranged as `ORP`, `SRP-Lasso`, `LPR`, `ssLasso` in order of decreasing time complexity; though practically these four methods require approximately similar time to completion.

3.5 Inference in randomized experiments

We now apply SRP to the problem of constructing CI for the Average Treatment Effect (ATE) in the Neyman randomized experiments model in the presence of covariates, with particular attention to the case where the number of variables presented is comparable or larger than the number of experiments conducted. Following our modus operandi in the previous sections, SRP compresses all the variables available while considering variable importances and then computes the OLS adjusted CI for ATE (Lin [79]). This is explained in Algorithm 5; In the following paragraph we set down some notation to describe the Neyman model with two randomized groups (treatment and control).

n subjects indexed by $i = 1, \ldots, n$ are divided via simple random sampling into two groups $\mathcal{A}$ (treatment) and $\mathcal{B}$ (control) of sizes $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ respectively ($n_{\mathcal{A}} + n_{\mathcal{B}} = n$). For each subject, an outcome value Y_i is observed along with p covariate values, $x_{ij}, 1 \leq j \leq p$. As in regression, we shall denote the vector of outcome values $y_1, \ldots, y_n$ by Y and the matrix of explanatory variable values by $X = ((x_{ij}))_{n \times p}$. Neyman model assumes that each subject has two potential outcomes: $\mathfrak{a}_i$ (potential outcome if assigned to treatment group $\mathcal{A}$) and $\mathfrak{b}_i$ (potential outcome if assigned to control group $\mathcal{B}$). The observed outcome Y_i would equal $\mathfrak{a}_i$ if the i^{th} subject is assigned to group $\mathcal{A}$ and equal $\mathfrak{b}_i$ if the i^{th} subject is assigned to group $\mathcal{B}$. The model also assumes that the quantities $\mathfrak{a}_1, \ldots, \mathfrak{a}_n$ and $\mathfrak{b}_1, \ldots, \mathfrak{b}_n$ are fixed and non-random and that the only randomness in the observations $Y_1, \ldots, Y_n$ comes from the random sampling assignment. ATE (Average Treatment Effect) is then defined as

$$\text{ATE} := \bar{\mathfrak{a}} - \bar{\mathfrak{b}} = \frac{1}{n}\sum_{i=1}^{n} \mathfrak{a}_i - \frac{1}{n}\sum_{i=1}^{n} \mathfrak{b}_i. \tag{3.3}$$

When covariates are present and observed, it is implied that the potential outcomes are influenced by covariate values but no structural assumptions as to the analytical form of this influence is made in the Neyman model.

We would like to remark here that the ATE is one of the most important estimands of interest in randomized experiments (see, for example, Imbens and Rubin [59]). This inference problem is also well understood in the low-dimensional setting (i.e., where p is small compared to n, Lin [79]). Recently, Bloniarz et al. [15] presented a Lasso adjustment based estimate of ATE in the case where p is comparable to or larger than n.

Description of Figure 3.8: In each setting, the boxplots on the left represent the bias $(\widehat{\text{ATE}} - \text{ATE})$ incurred by each method and the numbers on top of each boxplot report the standard deviations. The barplots on the right represent the power and coverage for each method in each setting (nominal $\alpha = 0.1$). The specific simulation settings are described in full detail in Bloniarz et al. [15]. $n \in \{250, 1000\}$ and the covariate matrix X is drawn from gaussian random design with Toeplitz covariance with parameter $\rho \in \{0, 0.6\}$; thus resulting in four configurations in total. The potential outcomes exhibit non-linear dependency on the covariates and are influenced by 10 observed and 10 unobserved covariates.

In Figure 3.8, we compare `SRP` (as defined in Algorithm (5)) and the ordinary random projections estimate (`RP`) with the unadjusted estimator (`Unadj`) which does not take covari-

Algorithm 5 SRP confidence interval for ATE using variable importances computed via LASSO

Input $Y_{\mathcal{A}}, X_{\mathcal{A}}, Y_{\mathcal{B}}, X_{\mathcal{B}}, m_{\mathcal{A}}, m_{\mathcal{B}}, \alpha$.
On group $\mathcal{A}$ compute variable importances $\hat{\mathbf{v}}_{\mathcal{A},j}$ using cross-validated Lasso on $(Y_{\mathcal{A}}, X_{\mathcal{A}})$
On group $\mathcal{A}$ construct supervised compressed matrix $U_{\mathcal{A}} \leftarrow \texttt{SRP-dimReduce}(X_{\mathcal{A}}, \hat{\mathbf{v}}_{\mathcal{A}}, m_{\mathcal{A}})$
On group $\mathcal{A}$ compute $\hat{\gamma}_{\mathcal{A}}$, the linear regression coefficients of $Y_{\mathcal{A}}$ on $U_{\mathcal{A}}$; and $\hat{\sigma}^2_{\mathcal{A}}$, the residual sum of squares divided by residual degrees of freedom $= n_{\mathcal{A}} - m_{\mathcal{A}}$
On group $\mathcal{B}$ repeat the previous three steps
Compute

$$\widehat{\text{ATE}}_{\text{SRP}} := \left[\bar{Y}_{\mathcal{A}} - (\bar{U}_{\mathcal{A}} - \bar{U})^\top \hat{\gamma}_{\mathcal{A}}\right] - \left[\bar{Y}_{\mathcal{B}} - (\bar{U}_{\mathcal{B}} - \bar{U})^\top \hat{\gamma}_{\mathcal{B}}\right]$$

and

$$\hat{\sigma}^2_{\text{SRP}} := \frac{1}{n_{\mathcal{A}}}\hat{\sigma}^2_{\mathcal{A}} + \frac{1}{n_{\mathcal{B}}}\hat{\sigma}^2_{\mathcal{B}}.$$

Return $100(1-\alpha)\%$ CI of ATE,

$$\left[\widehat{\text{ATE}}_{\text{SRP}} - z_{\alpha/2}\hat{\sigma}_{\text{SRP}}, \widehat{\text{ATE}}_{\text{SRP}} + z_{\alpha/2}\hat{\sigma}_{\text{SRP}}\right]$$

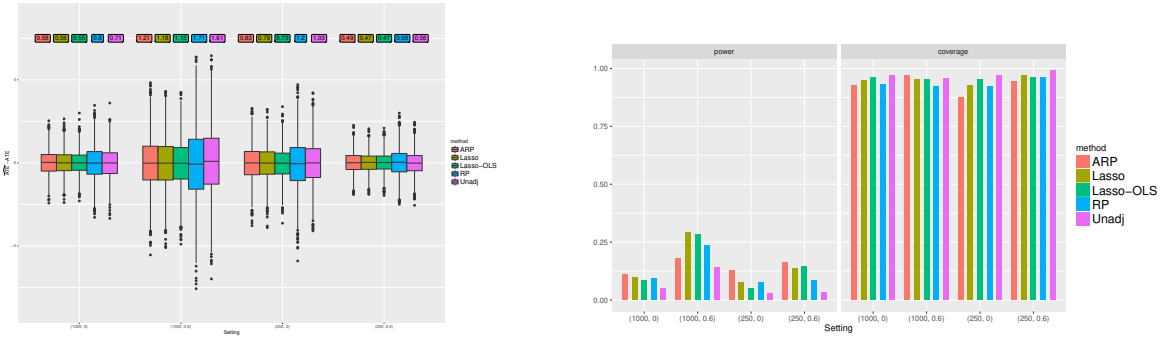


Figure 3.8: Performance inference of Average Treatment Effect

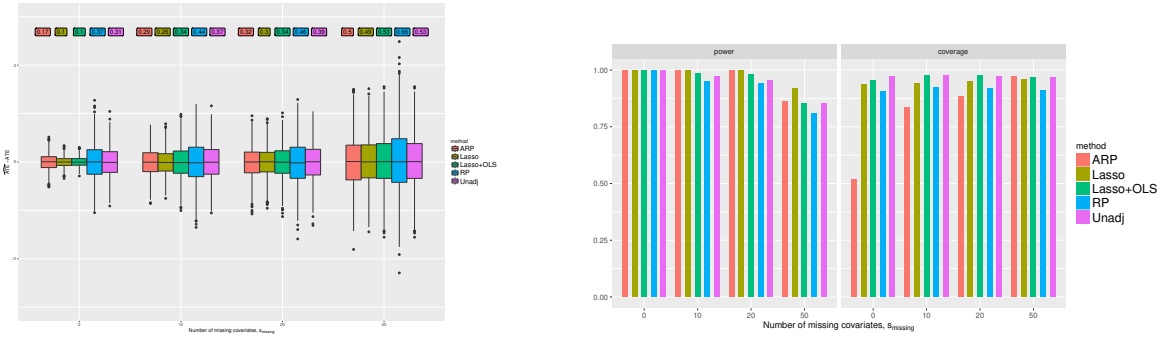


Figure 3.9: A simulation study modified from Bloniarz et al. [15] with higher signal-to-noise ratio, which accentuates the characteristics of each method.

ate information into account, and the methods `Lasso` and `Lasso+OLS` established in Bloniarz et al. [15]. For the purposes of this study, we replicate 2500 datasets from each of the four configurations described in detail in Bloniarz et al. [15].

Figure 3.9 show the results of a second simulation study performed in a setting with higher signal-to-noise ratio.

3.5.1 Discussion on the role of SRP in inference in randomized experiments

(i) From the left hand plots of Figures 3.8 and 3.9 (which show the distribution of $\widehat{\text{ATE}} - \text{ATE}$ for various methods), all methods are shown to be unbiased. In terms of standard deviation, both visually from the boxplots and as noted in the readouts above them, we realize that `RP` is subpar compared to `Unadj`. On the other hand, `SRP` enjoys a lower (better) standard deviation than `Unadj` and is only slightly higher those of `Lasso` and `Lasso+OLS` in Figure 3.8.

(ii) The left hand plots of Figure 3.9 make this hierarchy clearer by increasing the signal strength. `Unadj` performs better than `RP` while the other three methods perform better than `Unadj`, where `Lasso` and `Lasso+OLS` have lower variance than `SRP`. This allows us to conclude that the ordinary dimension reduction procedure fails in this problem but the supervised procedure doesn't; thus, the only avenue to employ the advantages of dimension reduction (such as, leaner memory and computational footprint) in this problem is through an supervised method.

(iii) The right hand plots show coverage and power for each of the methods. We should note the low power across the board in Figure 3.8, which motivated us to investigate the setting in Figure 3.9. We should draw attention to the lower than nominal coverage of `SRP` in the left-most setting in Figure 3.9. Other than this lone blemish, we conclude that `SRP` provides a viable alternative to `Lasso` and `Lasso+OLS` in the problem of statistical inference on ATE in the Neyman model with many covariables.

(iv) Echoing a note from our conclusions in Section 3.4.1, while the refitting based procedure `Lasso+OLS` is the best performer in these studies, it is also a much slower method.

Comparison of Computational Complexities

The unadjusted method (`Unadj`), is clearly computationally the fastest with $\mathcal{O}(n)$ operations. The ordinary random projections method, denoted as `RP` requires $\mathcal{O}(nm^2)$ operations. The other three methods, `Lasso`, `Lasso + OLS` and `SRP` all possess a theoretical computational complexity of $\mathcal{O}(p^3)$ operations. But in practice, we observe a clearly heavier computational overhead to `Lasso + OLS`. This becomes clear from the empirical running times we present in Figure 3.11. To clarify, `Lasso` and `SRP` share practically the same runtime, while the same

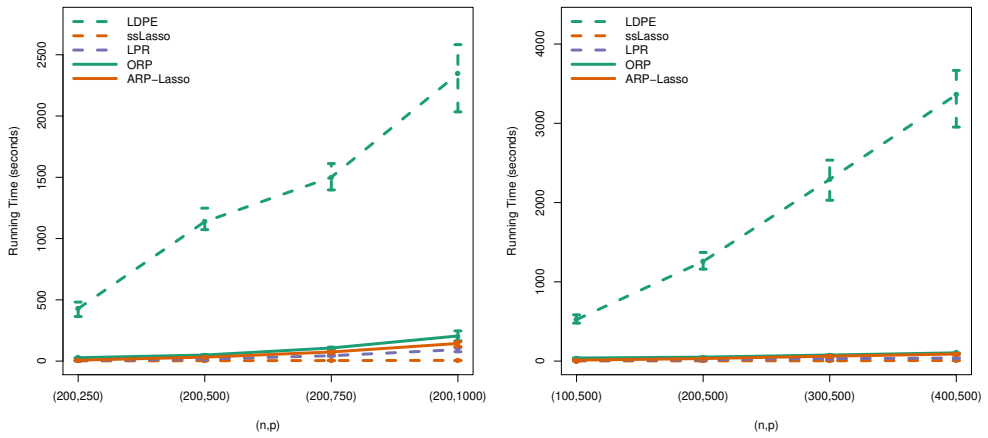


Figure 3.10: Runtime. The left hand plot shows a comparison for different values of p keeping n fixed. The right hand plot shows a comparison for different values of n keeping p fixed.

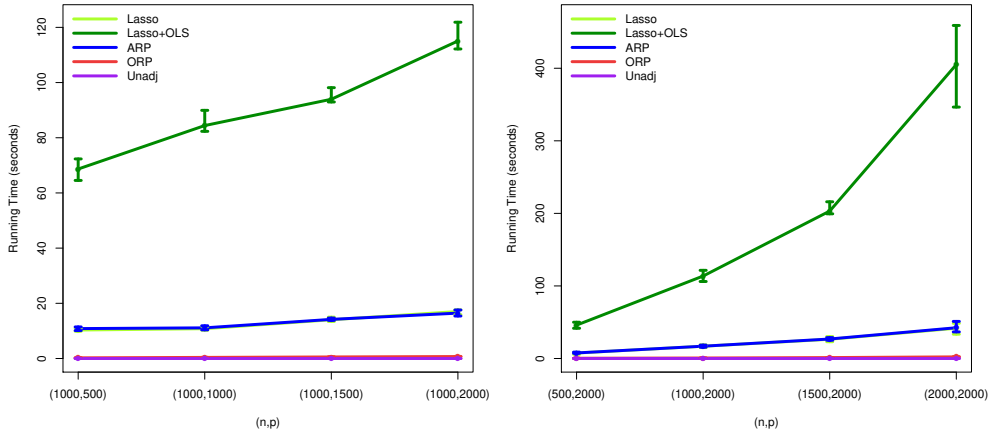


Figure 3.11: Runtime. The left hand plot shows a comparison for different values of p keeping n fixed. The right hand plot shows a comparison for different values of n keeping p fixed.

is true for `Unadj` and `RP`. We observe that `Lasso + OLS` is $\approx 6-8\times$ slower than `Lasso` and `SRP`. The other four methods enjoy reasonably fast implementations.

3.6 Discussion

Dimension reduction shares the twin advantages of leaner computational footprint and increased interpretability in high-dimensional inferential procedures. In this chapter, we proposed an universal dimension reduction technique which leverages variable importance in-

dices, thus improving the accuracy of the dimension reduction step while preserving the computational economy which makes dimension reduction, equivalently compression, based methodology vastly desirable. Our supervised framework is highly customizable and modular in nature; we allow the practitioner to employ any black-box method of computing variable importance as he/she deems fit. Our judicial use of sample splitting further bolsters the validity of subsequent inference.

We have shown that SRP provides a valid alternative to computationally pricey non-parametric permutation tests for identifying significant variables in non-linear regression using Random Forests. We have further shown that in the context of inference in high-dimensional linear regression problems, SRP and ORP provide robust, faster and conservative alternative to the debiasing method and other existing methods. We have lastly shown that, supervised dimension reduction is a viable if slightly inferior alternative to Lasso based adjustments in the problem of statistical inference on Average Treatment Effect in the presence of a large number of covariates.

To conclude, the universality of our supervised dimension reduction approach may lead to suboptimal performance compared to specialized techniques in specific problems; but in our experience this specialized techniques are often much slower. While in many of the problems we have discussed here, other fast methods are available, the supervised dimension reduction technique is able to match the performance and computational overload of such methods. In contrast, SRP is an universal idea which is applicable in a wide variety of high-dimensional inference problems. Thus, we propose SRP as an universal alternative which is approximate but computationally vastly more appealing.

Chapter 4

Nonparametric Maximum Likelihood Estimator for Gaussian Location Mixture Densities with Application to Gaussian Denoising

4.1 Overview

We study the Nonparametric Maximum Likelihood Estimator (NPMLE) for estimating Gaussian location mixture densities in d-dimensions from independent observations. Unlike usual likelihood-based methods for fitting mixtures, NPMLEs are based on convex optimization. We prove finite sample results on the Hellinger accuracy of every NPMLE. Our results imply, in particular, that every NPMLE achieves near parametric risk (up to logarithmic multiplicative factors) when the true density is a discrete Gaussian mixture without any prior information on the number of mixture components. NPMLEs can naturally be used to yield empirical Bayes estimates of the Oracle Bayes estimator in the Gaussian denoising problem. We prove bounds for the accuracy of the empirical Bayes estimate as an approximation to the Oracle Bayes estimator. Here our results imply that the empirical Bayes estimator performs at nearly the optimal level (up to logarithmic multiplicative factors) for denoising in clustering situations without any prior knowledge of the number of clusters.

4.2 Introduction

In this chapter, we study the performance of the Nonparametric Maximum Likelihood Estimator (NPMLE) for estimating a Gaussian location mixture density in multiple dimensions. We also study the performance of the empirical Bayes estimator based on the NPMLE for estimating the Oracle Bayes estimator in the problem of Gaussian denoising.

By a Gaussian location mixture density in $\mathbb{R}^d, d \geq 1$, we refer to a density of the form

$$f_G(x) := \int \phi_d(x-\theta) dG(\theta) \tag{4.1}$$

for some probability measure G on $\mathbb{R}^d$ where $\phi_d(z) := (2\pi)^{-d/2} \exp(-\|z\|^2/2)$ is the standard d-dimensional normal density ($\|z\|$ is the usual Euclidean norm of z). Note that f_G is the density of the random vector $X = \theta + Z$ where θ and Z are independent d-dimensional random vectors with θ having distribution G (i.e., $\theta \sim G$) and Z having the Gaussian distribution with zero mean and identity covariance matrix (i.e., $Z \sim N(0, I_d)$). We let $\mathcal{M}$ to be the class of all Gaussian location mixture densities i.e., densities of the form f_G as G varies over all probability measures on $\mathbb{R}^d$.

Given n independent d-dimensional data vectors $X_1, \ldots, X_n$ (throughout the chapter, we assume that $n \geq 2$) generated from an unknown Gaussian location mixture density $f^* \in \mathcal{M}$, we study the problem of estimating f^* from $X_1, \ldots, X_n$. This problem is fundamental to the area of estimation in mixture models to which a number of books (see, for example, Everitt and Hand [43], Titterington, Smith, and Makov [132], Lindsay [80], Böhning [18], McLachlan and Peel [97], and Schlattmann [121]) and papers have been devoted. We focus on the situation where d is small or moderate, n is large and where no specific prior information is available about the mixing measure corresponding to f^*. Consistent estimation in the case where d is comparable in size to n needs simplifying assumptions on f^* (such as that the mixing measure is discrete with a small number of atoms and that it is concentrated on a set of sparse vectors in $\mathbb{R}^d$) which we do not make in this chapter. Let us also note here that we focus on the problem of estimating f^* and not on estimating the mixing measure corresponding to f^*.

There are two well-known likelihood-based approaches to the estimation of Gaussian location mixture densities: (a) the first approach involves fixing an integer k and performing maximum likelihood estimation over the class $\mathcal{M}_k$ which is the collection of all densities $f_G \in \mathcal{M}$ where G is discrete and has at most k atoms, and (b) the second approach involves performing maximum likelihood estimation over the entire class $\mathcal{M}$. This results in the Nonparametric Maximum Likelihood Estimator (NPMLE) for f^* and is the focus of this chapter.

The first approach (maximum likelihood estimation over $\mathcal{M}_k$ for a fixed k) is quite popular. However, it suffers from the two well-known issues: choosing k is non-trivial and, moreover, maximizing likelihood over $\mathcal{M}_k$ results in a non-convex optimization problem. This non-convex algorithm is usually approximately solved by the EM algorithm (see, for example, Dempster, Laird, and Rubin [35], McLachlan and Krishnan [96], and Watanabe and Yamaguchi [142]). Recent progress on obtaining a theoretical understanding of the behaviour of the non-convex EM algorithm has been made by Balakrishnan, Wainwright, Yu, et al. [5]. For the issue of choosing k, one can adapt standard model selection methodology such as those based on the AIC Akaike [2] or BIC Schwarz [122]. However theoretical properties of the resulting estimator are not well understood because the usual regularity conditions that are required for AIC or BIC to work do not hold in this mixture model setting. More

recently, Maugis and Michel [93] (see also Maugis-Rabusseau and Michel [95]) proposed a penalization likelihood criterion to choose k by suitably employing the general theory of non-asymptotic model selection via penalization due to Birgé and Massart [14], Barron, Birgé, and Massart [8] and Massart [92]. Maugis and Michel [93] also established nonasymptotic risk properties of the resulting estimator. The computational aspects of their estimator are quite involved however (see Maugis and Michel [94]) as their estimators are based on solving multiple non-convex optimization problems.

The present chapter concentrates on second likelihood-based approach involving non-parametric maximum likelihood estimation of f^*. This method is not affected by the issues of non-convexity and the need for choosing k. Formally, by an NPMLE, we refer to any maximizer $\hat{f}_n$ of $\sum_{i=1}^n \log f(X_i)$ as f varies over $\mathcal{M}$ i.e.,

$$\hat{f}_n \in \operatorname*{argmax}_{f \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i). \tag{4.2}$$

Note that because the maximization is done over the entire class $\mathcal{M}$ of all Gaussian location mixture densities (and not on any non-convex subset such as $\mathcal{M}_k$), the optimization in (4.2) is a convex optimization problem. Indeed, the objective function in (4.2) is concave in f and the constraint set $\mathcal{M}$ is a convex class of densities.

The idea of using NPMLEs for estimating mixture densities has a long history (see, for example, the classical references Kiefer and Wolfowitz [67], Lindsay [81, 82], Lindsay [80], and Böhning [18]). The optimization problem (4.2) and its solutions have been studied by many authors. It is known that maximizers of $f \mapsto \sum_{i=1}^n \log f(X_i)$ exist over $\mathcal{M}$ which implies that NPMLEs exist. Maximizers are non-unique however so there exist multiple NPMLEs. Nevertheless, for every NPMLE $\hat{f}_n$, the values $\hat{f}(X_i)$ for $i = 1, \ldots, n$ are unique (this is essentially because the objective function in the optimization (4.2) only depends on f through the values $f(X_1), \ldots, f(X_n)$). Proofs of these basic facts can be found, for example, in Böhning [18, Chapter 2].

There exist many algorithms in the literature for approximately solving the optimization (4.2) (note that though (4.2) is a convex optimization problem, it is infinite-dimensional which is probably why exact algorithms seem to be unavailable). These algorithms range from: (a) vertex direction methods and vertex exchange methods (see the review papers: Böhning [17], Lindsay and Lesperance [83] and the references therein), (b) EM algorithms (see Laird [70] and Jiang and Zhang [63]), and (c) modern large-scale interior point methods (see Koenker and Mizera [68] and Feng and Dicker [45]). Most of these methods focus on the case $d = 1$ and involve maximizing the likelihood over mixture densities where the mixing measure is supported on a fixed fine grid in the range of the data. The algorithm of Koenker and Mizera [68] is highly scalable (relying on the commercial convex optimization library Mosek [102]) and can obtain an approximate NPMLE efficiently even for large sample sizes (n of the order $100,000$). See Section 4.5 for more algorithmic and implementation details as well as some simulation results.

Let us now describe the main objectives and contributions of the current chapter. Our first goal is to investigate the theoretical properties of NPMLEs. In particular, we study the accuracy of $\hat{f}_n$ as an estimator of the density f^* from which the data $X_1, \ldots, X_n$ are generated. We shall use, as our loss function, the squared Hellinger distance:

$$\mathfrak{H}^2(f, g) := \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx, \tag{4.3}$$

which is one of the most commonly used loss functions for density estimation problems. We present a detailed analysis of the risk, $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*)$, of every NPMLE (the expectation here is taken with respect to $X_1, \ldots, X_n$ distributed independently according to f^*). The other common loss function used in density estimation is the total variation distance. The total variation distance is bounded from above by a constant multiple of $\mathfrak{H}$ so that upper bounds for risk under the squared Hellinger distance automatically imply upper bounds for risk in squared total variation distance.

Our results imply that, for a large class of true densities $f^* \in \mathcal{M}$, the risk of every NPMLE $\hat{f}_n$ is parametric (i.e., n^{-1}) up to multiplicative factors that are logarithmic in n. In particular, our results imply that when the true $f^* \in \mathcal{M}_k$ for some $1 \leq k \leq n$, then every NPMLE has risk k/n up to a logarithmic multiplicative factor in n. It is not hard to see that the minimax risk over $\mathcal{M}_k$ is bounded from below by k/n which implies therefore that every NPMLE is nearly minimax over $\mathcal{M}_k$ (ignoring logarithmic factors in n) for every $k \geq 1$. This is interesting because NPMLEs do not use any a priori knowledge of k. The price in squared Hellinger risk that is paid for not knowing k in advance is only logarithmic in n. Our results are non-asymptotic and the bounds for risk over $\mathcal{M}_k$ hold even when k grows with n. Our results also imply that NPMLEs have parametric risk (again up to multiplicative logarithmic factors) when the mixing measure of f^* is supported on a fixed compact subset of $\mathbb{R}^d$. Note that we have assumed that the covariance matrix of every Gaussian component of mixture densities in the class $\mathcal{M}$ is the identity matrix. Our results can be extended to the case of arbitrary covariance matrices provided a lower bound on the eigenvalues is available (see Proposition 4.3.5) (on the other hand, when no *a priori* information on the covariance matrices is available, it is well-known that likelihood based approaches are infeasible). These results are described in detail in Section 4.3.

Previous results on the Hellinger accuracy of NPMLEs were due to Ghosal and Vaart [49] and Zhang [147] who dealt with the univariate ($d = 1$) case. They studied the Hellinger accuracy under conditions on the moments of the mixing measure corresponding to f^*. The accuracy of NPMLEs in the interesting case when $f^* \in \mathcal{M}_k$ does not appear to have been studied previously even in $d = 1$. We study the Hellinger risk of NPMLEs for all $d \geq 1$ and also under a much broader set of assumptions on f^* compared to existing papers.

We would like to mention here that numerous papers have appeared in the theoretical computer science community establishing rigorous theoretical results for estimating densities in $\mathcal{M}_k$. For example, the papers Daskalakis and Kamath [34], Suresh et al. [127], Bhaskara, Suresh, and Zadimoghaddam [12], Chan et al. [25, 24], Acharya et al. [1], and Li and Schmidt

[75] have results on estimating densities in $\mathcal{M}_k$ with rigorous bounds on the error in estimation. The estimation error is mostly measured in terms of the total variation distance which is smaller (up to constant multiplicative factors) compared to the Hellinger distance used in the present chapter. Their sample complexity results imply rates of estimation of k/n up to logarithmic factors in n for densities in $\mathcal{M}_k$ in terms of the squared total variation distance and hence these results are comparable to our results for the NPMLE. The estimation procedures used in these papers range from (a) hypothesis selection over a set of candidate estimators via an improved version of the Scheffé estimate (Daskalakis and Kamath [34] and Suresh et al. [127]; see Devroye and Lugosi [36, Chapter 6] for background on the Scheffé estimate), (b) reduction to finding sparse solutions to a non-negative linear systems (Bhaskara, Suresh, and Zadimoghaddam [12]), and (c) fitting piecewise polynomial densities (Chan et al. [25, 24], Acharya et al. [1], and Li and Schmidt [75]; these papers have the sharpest results). These methods are very interesting and, remarkably, come with precise time complexity guarantees. They are not based on likelihood maximization however and, in our opinion, conceptually more involved compared to the NPMLE studied in this chapter. An additional minor difference between our work and this literature is that k is taken to be a constant (and sometimes even known) in these papers while we allow k to grow with n and the NPMLE does not need prior knowledge of k.

Let us now describe briefly the proof techniques underlying our risk results for the NPMLEs. Our technical arguments are based on standard ideas from the literature on empirical processes for assessing the performance of maximum likelihood estimators (see Vaart and Wellner [136], Wong and Shen [143], and Zhang [147]). These techniques involve bounding the covering numbers of the space of Gaussian location mixture densities. For each compact subset $S \subseteq \mathbb{R}^d$, we prove covering number bounds for $\mathcal{M}$ under the supremum distance (L_∞) on S. Our bounds can be seen as extensions of the one-dimensional covering number results of Zhang [147] (which are themselves enhancements of corresponding results in Ghosal and Vaart [49]). The covering number results of Zhang [147] can be viewed as special instances of our bounds for the case when $S = [-M, M]$. The extension to arbitrary compact sets S is crucial for dealing with rates for densities in $\mathcal{M}_k$. For proving the final Hellinger risk bounds of $\hat{f}_n$ from these L_∞ covering numbers, we use appropriate modifications of tail arguments from Zhang [147].

The second goal of the present chapter is to use NPMLEs to yield empirical Bayes estimates in the Gaussian denoising problem. By Gaussian denoising, we refer to the problem of estimating vectors $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$ from independent d-dimensional observations $X_1, \ldots, X_n$ generated as

$$X_i \sim N(\theta_i, I_d) \qquad \text{for } i = 1, \ldots, n. \tag{4.4}$$

The naive estimator in this denoising problem simply estimates each θ_i by X_i. It is well-known that, depending on the structure of the unknown $\theta_1, \ldots, \theta_n$, it is possible to achieve significant improvement over the naive estimator by using information from $X_j, j \neq i$ in addition to X_i for estimating θ_i. An ideal prototype for such information sharing across observations is given by the *Oracle Bayes* estimator which will be denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ and

is defined in the following way:

$$\hat{\theta}_i^* := \mathbb{E}\left(\theta | X = X_i\right) \qquad \text{where } \theta \sim \bar{G}_n \text{ and } X|\theta \sim N(\theta, I_d)$$

and $\bar{G}_n$ is the empirical measure corresponding to the true set of parameters $\theta_1, \ldots, \theta_n$. In other words, $\hat{\theta}_i^*$ is the posterior mean of θ given $X = X_i$ under the model $X|\theta \sim N(\theta, I_d)$ and the prior $\theta \sim \bar{G}_n$. This is an Oracle estimator that is infeasible in practice as it uses information on the unknown parameters $\theta_1, \ldots, \theta_n$ via their empirical measure $\bar{G}_n$. It is well-known (see, for example, Robbins [115], Brown [21], Stein [125], and Efron [40]) that $\hat{\theta}_i^*$ has the following alternative expression as a consequence of Tweedie's formula:

$$\hat{\theta}_i^* = X_i + \frac{\nabla f_{\bar{G}_n}(X_i)}{f_{\bar{G}_n}(X_i)} \tag{4.5}$$

where $f_{\bar{G}_n}$ is the Gaussian location mixture density with mixing measure $\bar{G}_n$ (defined as in (4.1)). From the above expression, it is clear that the Oracle Bayes estimator can be estimated from the data $X_1, \ldots, X_n$ provided one can estimate the Gaussian location mixture density, $f_{\bar{G}_n}$, from the data $X_1, \ldots, X_n$. For this purpose, as insightfully observed in Jiang and Zhang [63], any NPMLE, $\hat{f}_n$, as in (4.2) can be used. Indeed, if $\hat{f}_n$ denotes any NPMLE based on the data $X_1, \ldots, X_n$, then Jiang and Zhang [63] argued that $\hat{f}_n$ is a good estimator for $f_{\bar{G}_n}$ under (4.4) so that $\hat{\theta}_i^*$ is estimable by

$$\hat{\theta}_i := X_i + \frac{\nabla \hat{f}_n(X_i)}{\hat{f}_n(X_i)}. \tag{4.6}$$

This yields a completely tuning-free solution to the Gaussian denoising problem (note however that the noise distribution is assumed to be completely known as $N(0, I_d)$). This is the General Maximum Likelihood empirical Bayes estimator of Jiang and Zhang [63] who proposed it and studied its theoretical properties in detail for estimating sparse univariate normal means. To the best of our knowledge, the properties of the estimator (4.6) for multidimensional denoising problems have not been previously explored. More generally, the empirical Bayes approach to the Gaussian denoising problem goes back to Robbins [114, 116, 117]. The effectiveness of nonparametric empirical Bayes estimators for estimating sparse normal means has been explored by many authors including Johnstone and Silverman [64], Brown and Greenshtein [22], Jiang and Zhang [63], Donoho and Reeves [39], and Koenker and Mizera [68] but most work seems restricted to the univariate setting. On the other hand, there exists prior work on parametric empirical Bayes methods in the multivariate Gaussian denoising problem (see, for example, Efron and Morris [41, 42]) but the role of nonparametric empirical Bayes methods in multivariate Gaussian denoising does not seem to have been explored previously.

We perform a detailed study of the accuracy of $\hat{\theta}_i$ in (4.6) as an estimator of the Oracle Bayes estimator $\hat{\theta}_i^*$ for $i = 1, \ldots, n$ in terms of the following squared error risk measure:

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \|\hat{\theta}_i - \hat{\theta}_i^*\|^2\right] \tag{4.7}$$

where the expectation is taken with respect to $X_1, \ldots, X_n$ generated independently according to (4.4). The risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ depends on the configuration of the unknown parameters $\theta_1, \ldots, \theta_n$ and we perform a detailed study of the risk for natural configurations of the points $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$. Our results imply that, under natural assumptions on $\theta_1, \ldots, \theta_n$, the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded by the parametric rate $1/n$ up to logarithmic multiplicative factors. For example, when the number of distinct vectors among $\theta_1, \ldots, \theta_n$ equals k for some $k \leq n$ (an assumption which makes sense in clustering situations), we prove that the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded from above by the parametric rate k/n up to logarithmic multiplicative factors in n. This result is especially remarkable because the estimator (4.6) is tuning free and does not have knowledge of k. We also prove that the analogous minimax risk over this class is bounded from below by k/n implying that the empirical Bayes estimate is minimax up to logarithmic multiplicative factors. Our result also implies that when $\theta_1, \ldots, \theta_n$ take values in a bounded region on $\mathbb{R}^d$, then also the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is nearly parametric. Summarizing, our results imply that, under a wide range of assumptions on $\theta_1, \ldots, \theta_n$, the empirical Bayes estimator $\hat{\theta}_i$ performs comparably to the Oracle Bayes estimator $\hat{\theta}_i^*$ for denoising. These results are in Section 4.4. The results and the proof techniques are inspired by the arguments of Jiang and Zhang [63] who studied the univariate denoising problem under sparsity assumptions. We generalize their arguments to multidimensions.

In addition to theoretical results, we also present simulation evidence for the effectiveness of $\hat{\theta}_i$ in the Gaussian denoising problem in Section 4.5 (where we also present some implementation and algorithmic details for computing approximate NPMLEs). Here, we illustrate the performance of (4.6) for denosing when the true parameter vectors $\theta_1, \ldots, \theta_n$ take values in certain natural regions in $\mathbb{R}^2$. We also numerically analyze the performance of (4.6) in clustering situations when $\theta_1, \ldots, \theta_n$ take k distinct values for some small k. Here we compare the performance of (4.6) to other natural procedures such as k-means with k selected via the gap statistic (see Tibshirani, Walther, and Hastie [130]). We argue that (4.6) performs very efficiently in terms of the risk measure $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$. In terms of a purely clustering based comparison index (such as the Adjusted Rand Index), we argue that the performance of (4.6) is still reasonable.

The rest of the chapter is organized in the following manner. In Section 4.3, we state our results on the Hellinger accuracy of NPMLEs for estimating Gaussian location mixture densities. Section 4.4 has statements of our results on the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ in the denoising problem. Section 4.5 has algorithmic details and simulation evidence for the effectiveness of (4.6) for denoising. Proofs for results in Section 4.3 are given in Section 4.6 while proof for Section 4.4 are in Section 4.7. Metric entropy results for multivariate Gaussian location mixture densities play a crucial rule in the proofs of the main results; these results are stated and proved in Section 4.8. Section 4.9 contains the statement and proof for a crucial ingredient for the proof of the main denoising theorem. Finally, additional technical results needed in the proofs of the main results are collected in Section A together with their proofs.

4.3 Hellinger Accuracy of NPMLE

Given data $X_1, \ldots, X_n$, let $\hat{f}_n$ be any NPMLE defined as in (4.2). In this section, we shall study the accuracy of $\hat{f}_n$ in terms of the squared Hellinger distance (defined in (4.3)). All the results in this section are proved in Section 4.6.

For investigations into the performance of $\hat{f}_n$, it is most natural to assume that the data $X_1, \ldots, X_n$ are independent observations having common density $f^* \in \mathcal{M}$ in which case we seek bounds on $\mathfrak{H}^2(\hat{f}_n, f^*)$. However, following Zhang [147], we work under the more general assumption that $X_1, \ldots, X_n$ are independent but not identically distributed and that each X_i has a density that belongs to the class $\mathcal{M}$. This additional generality will be used in Section 4.4 for proving results on the Empirical Bayes estimator (4.6) for the Gaussian denoising problem.

Specifically, we assume that $X_1, \ldots, X_n$ are independent and that each X_i has density f_{G_i} for some probability measures $G_1, \ldots, G_n$ on $\mathbb{R}^d$. This distributional assumption on the data $X_1, \ldots, X_n$ includes the following two important special cases: (a) $G_1, \ldots, G_n$ are all identically equal to G (say): in this case, the observations $X_1, \ldots, X_n$ are identically distributed with common density $f^* = f_G \in \mathcal{M}$, and (b) Each G_i is degenerate at some $\theta_i \in \mathbb{R}^d$: here each data point X_i is normal with $X_i \sim N_d(\theta_i, I_d)$.

We let $\bar{G}_n := (G_1 + \cdots + G_n)/n$ to be the average of the probability measures $G_1, \ldots, G_n$. In the case when $G_1, \ldots, G_n$ are all identically equal to G, then clearly $\bar{G}_n = G$. On the other hand, when each G_i is degenerate at some $\theta_i \in \mathbb{R}^d$, then $\bar{G}_n$ equals the empirical measure corresponding to $\theta_1, \ldots, \theta_n$.

Under the above *independent but not identically distributed* assumption on $X_1, \ldots, X_n$, it has been insightfully pointed out by Zhang [147] that every NPMLE $\hat{f}_n$ based on $X_1, \ldots, X_n$ (defined as in (4.2)) is really estimating $f_{\bar{G}_n}$. Note that $f_{\bar{G}_n}$ denotes the average of the densities of $X_1, \ldots, X_n$.

In this section, we shall prove bounds for the accuracy of any NPMLE $\hat{f}_n$ as an estimator for $f_{\bar{G}_n}$ under the Hellinger distance i.e., for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$. For every compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$, we shall prove an upper bound for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ in terms of S and M. As will be seen later in this section, under some simplifying assumptions on $\bar{G}_n$, our bound for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ can be optimized over S and M to produce an explicit bound.

In order to state our main theorem, we need to introduce the following notation. For nonempty sets $S \subseteq \mathbb{R}^d$, we define the function $\mathfrak{d}_S : \mathbb{R}^d \to [0, \infty)$ by

$$\mathfrak{d}_S(x) := \inf_{u \in S} \|x - u\| \qquad \text{for } x \in \mathbb{R}^d \tag{4.8}$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Also for $S \subseteq \mathbb{R}^d$, we let

$$S^1 := \{x : \mathfrak{d}_S(x) \leq 1\}. \tag{4.9}$$

Our bound on $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ will be controlled by the following quantity. For every non-empty

compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$, let $\epsilon_n(M, S)$ be defined via

$$\epsilon_n^2(M, S) := \text{Vol}(S^1) \frac{M^d}{n} \left(\sqrt{\log n}\right)^{(4-d)_+} + (\log n) \inf_{p \geq \frac{d+1}{2 \log n}} \left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p \tag{4.10}$$

where S^1 is defined in (4.9), $(4-d)_+$ is defined by way of $x_+ := \max(x, 0)$ and $\mu_p(\mathfrak{d}_S)$ is defined as the moment

$$\mu_p(\mathfrak{d}_S) := \left(\int_{\mathbb{R}^d} (\mathfrak{d}_S(\theta))^p \, d\bar{G}_n(\theta)\right)^{1/p} \qquad \text{for } p > 0.$$

Note that the moments $\mu_p(\mathfrak{d}_S)$ quantify how the probability (under $\bar{G}_n$) decays as one moves away from the set S.

The next theorem proves that $\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n})$ is bounded (with high probability and in expectation) by a constant (depending on d) multiple of $\epsilon_n^2(M, S)$ for every estimator $\hat{f}_n$ having the property that the likelihood of the data at $\hat{f}_n$ is not too small compared to the likelihood at $f_{\bar{G}_n}$ (made precise in inequality (4.11)). Every NPMLE trivially satisfies this condition (as it maximizes likelihood) but the theorem also applies to certain approximate likelihood maximizers.

Theorem 4.3.1. *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Fix $M \geq \sqrt{10 \log n}$ and a non-empty compact set $S \subseteq \mathbb{R}^d$ and let $\epsilon_n(M, S)$ be as defined in (4.10). Then there exists a positive constant C_d (depending only on d) such that for every estimator $\hat{f}_n$ based on the data $X_1, \ldots, X_n$ satisfying*

$$\prod_{i=1}^n \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \geq \exp\left[\frac{C_d(\beta - \alpha)}{\min(1-\alpha, \beta)} n \epsilon_n^2(M, S)\right] \qquad \text{for some } 0 < \beta \leq \alpha < 1, \tag{4.11}$$

we have

$$\mathbb{P}\left\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq \frac{t \epsilon_n(M, S) \sqrt{C_d}}{\sqrt{\min(1-\alpha, \beta)}}\right\} \leq 2n^{-t^2} \qquad \text{for every } t \geq 1. \tag{4.12}$$

and

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq \frac{4C_d}{\min(1-\alpha, \beta)} \epsilon_n^2(M, S). \tag{4.13}$$

Theorem 4.3.1 asserts that the risk $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, \bar{f}_{\bar{G}_n})$ is bounded from above by a constant (depending on d, α and β) multiple of $\epsilon_n^2(M, S)$ for every $M \geq \sqrt{10 \log n}$ and compact subset $S \subseteq \mathbb{R}^d$. This is true for every estimator $\hat{f}_n$ satisfying (4.11). Every NPMLE satistfies (4.11) with $\alpha = \beta = 0.5$ (note that the right hand side of (4.11) is always less than or equal to one because $\beta \leq \alpha$).

Theorem 4.3.1 is novel to the best of our knowledge. When $d = 1$ and S is taken to be $[-R, R]$ for some $R \geq 0$, then the conclusion given by Theorem 4.3.1 appears implicitly in Zhang [147, Proof of Theorem 1]. The advantages of allowing S to be an arbitrary compact

set will be clear from the special cases of Theorem 4.3.1 that are given below. Our proof of Theorem 4.3.1 (given in Section 4.6) is greatly inspired by Zhang [147, Proof of Theorem 1].

To get the best rate for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ from Theorem 4.3.1, we need to choose M and S so that $\epsilon_n(M, S)$ is small. These choices depend on $\bar{G}_n$ and in the next result, we describe how to choose M and S based on reasonable assumptions on $\bar{G}_n$. This leads to explicit rates for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$. For simplicity, we shall assume, for the next result, that $\hat{f}_n$ is an NPMLE so that (4.11) is satisfied with $\alpha = \beta = 0.5$. We shall also only state the results on the risk $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n})$.

Corollary 4.3.2. *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $\hat{f}_n$ be an NPMLE based on $X_1, \ldots, X_n$ defined as in (4.2). Below C_d denotes a positive constant depending on d alone.*

1. *Suppose that $\bar{G}_n$ is supported on a compact subset S of $\mathbb{R}^d$. Then*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \frac{\mathrm{Vol}(S^1)}{n} \left(\sqrt{\log n}\right)^{d+(4-d)_+}. \tag{4.14}$$

2. *Suppose there exist a compact subset $S \subseteq \mathbb{R}^d$ and real numbers $0 < \alpha \leq 2$ and $K \geq 1$ such that*

$$\mu_p(\mathfrak{d}_S) \leq K p^{1/\alpha} \qquad \text{for all } p \geq 1. \tag{4.15}$$

Then

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \frac{\mathrm{Vol}(S^1)(Ke^{1/\alpha})^d}{n} \left(\sqrt{\log n}\right)^{(2d/\alpha)+(4-d)_+}. \tag{4.16}$$

3. *Suppose there exists a compact set $S \subseteq \mathbb{R}^d$ and real numbers $\mu > 0$ and $p > 0$ such that $\mu_p(\mathfrak{d}_S) \leq \mu$. Then there exists a positive constant $C_{d,\mu,p}$ (depending only on d, μ and p) such that*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_{d,\mu,p} \left(\frac{\mathrm{Vol}(S^1)}{n}\right)^{p/(p+d)} \left(\sqrt{\log n}\right)^{(2d+p(4-d)_+)/(p+d)}. \tag{4.17}$$

Corollary 4.3.2 is a generalization of Zhang [147, Theorem 1] as the latter result can be seen as a special case of Corollary 4.3.2 for $d = 1$ and $S = [-R, R]$ for some $R \geq 0$. The fact that S can be arbitrary in Corollary 4.3.2 allows us to deduce the following important adaptation results of NPMLEs for estimating Gaussian mixtures whose mixing measures are discrete. These results are, to the best of our knowledge, novel.

Proposition 4.3.3 (Near parametric risk for discrete Gaussian mixtures)**.** *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $\hat{f}_n$ be an NPMLE based on $X_1, \ldots, X_n$ defined as in (4.2). Then there exists a positive constant C_d depending only on d such that whenever $\bar{G}_n$ is a discrete probability measure that is supported on a set of cardinality k, we have*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \left(\frac{k}{n}\right) \left(\sqrt{\log n}\right)^{d+(4-d)_+}. \tag{4.18}$$

The significance of Proposition 4.3.3 is the following. Note that the right hand side of (4.18) is the parametric risk k/n up to an additional multiplicative factor that is logarithmic in n. This inequality shows important adaptation properties of NPMLEs. When the true unknown Gaussian mixture $f_{\bar{G}_n}$ is a discrete mixture having k Gaussian components, then every NPMLE nearly (up to logarithmic factors) achieves the parametric squared Hellinger risk k/n. For a fixed k, it is well-known that fitting a k-component Gaussian mixture via maximum likelihood is a non-convex problem that is usually solved by the EM algorithm. On the other hand, NPMLE is given by a convex optimization algorithm, does not require any prior specification of k and still achieves the k/n rate (up to logarithmic factors) when the truth is a k-component Gaussian mixture.

Note that Proposition 4.3.3 applies to the case of independent but not identically distributed $X_1, \ldots, X_n$ which is more general compared to the i.i.d assumption. This implies, in particular, that (4.18) also applies to the case when $X_1, \ldots, X_n$ are i.i.d having density $f^* \in \mathcal{M}$. In this case, we have

$$\sup_{f^* \in \mathcal{M}_k} \mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*) \leq C_d \left(\frac{k}{n}\right) \left(\sqrt{\log n}\right)^{d+(4-d)_+}. \tag{4.19}$$

The interesting aspect of this inequality is that it holds for every $k \geq 1$ and that the estimator $\hat{f}_n$ does not know or use any information about k.

It is straightforward to prove a minimax lower bound over $\mathcal{M}_k$ that complements Proposition 4.3.3. The following result proves that the minimax risk over $\mathcal{M}_k$ is bounded from below by a constant multiple of k/n. This implies that the NPMLE is minimax optimal over $\mathcal{M}_k$ ignoring logarithmic factors of n. Moreover, this optimality is adaptive since MLE does not require knowledge of k. This minimax lower bound is stated for the i.i.d case which implies that it holds for the more general independent but not identically distributed case as well.

Lemma 4.3.4. *For $k \geq 1$, let*

$$\mathcal{R}(\mathcal{M}_k) := \inf_{\tilde{f}} \sup_{f \in \mathcal{M}_k} \mathbb{E}_f \mathfrak{H}^2(\tilde{f}, f)$$

where $\mathbb{E}_f$ denotes expectation when the data $X_1, \ldots, X_n$ are independent observations drawn from the density f. Then there exists a universal positive constant C such that

$$\mathcal{R}(\mathcal{M}_k) \geq C\frac{k}{n} \qquad \text{for every } 1 \leq k \leq n. \tag{4.20}$$

Inequality (4.19) and Lemma 4.3.4 together imply that every NPMLE $\hat{f}_n$ is minimax optimal up to logarithmic factors in n over the class $\mathcal{M}_k$ for every $k \geq 1$. This optimality is adaptive since the NPMLE requires no information on k. The logarithmic terms in (4.19) are likely suboptimal but we are unable to determine the exact power of $\log n$ in (4.19).

So far we have studied estimation of Gaussian location mixture densities where the covariance matrix of each Gaussian component is fixed to be the identity matrix. We next

show that the same estimator (NPMLE defined as in (4.2)) can be modified to estimate arbitrary Gaussian mixtures (where the covariance matrices can be different from identity) provided a lower bound on the eigenvalues of the covariance matrices is available. Suppose that h^* is the Gaussian mixture density

$$h^*(x) := \sum_{j=1}^{k} w_j \phi_d(x; \mu_j, \Sigma_j) \qquad \text{for } x \in \mathbb{R}^d \tag{4.21}$$

where $k \geq 1$, $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and $\Sigma_1, \ldots, \Sigma_k$ are $d \times d$ positive definite matrices. Here $\phi_d(\cdot; \mu, \Sigma)$ denotes the d-variate normal density with mean μ and covariance matrix Σ. Suppose $\sigma^2_{\min}$ and $\sigma^2_{\max}$ are two positive numbers that are, respectively, smaller and larger than all the eigenvalues of $\Sigma_1, \ldots, \Sigma_k$ i.e.,

$$\sigma^2_{\min} \leq \min_{1 \leq j \leq k} \lambda_{\min}(\Sigma_j) \leq \max_{1 \leq j \leq k} \lambda_{\max}(\Sigma_j) \leq \sigma^2_{\max} \tag{4.22}$$

Consider the problem estimating h^* from i.i.d observations $Y_1, \ldots, Y_n$. It turns out that for every NPMLE $\hat{f}_n$ computed as in (4.2) based on the data $X_1 := Y_1/\sigma_{\min}, \ldots, X_n := Y_n/\sigma_{\min}$ can be coverted to a very good estimator for h^* via

$$\hat{h}_n(x) := \sigma^{-d}_{\min} \hat{f}_n(\sigma^{-1}_{\min} x) \qquad \text{for } x \in \mathbb{R}^d. \tag{4.23}$$

Our next result shows that the squared Hellinger risk of $\hat{h}_n$ is bounded from above by (k/n) up to a logarithmic factor in n provided that $\sigma_{\max}/\sigma_{\min}$ is bounded by a constant.

Proposition 4.3.5. *Let $Y_1, \ldots, Y_n$ be independent and identically distributed observations having density h^* defined in* (4.21)*. Consider the estimator $\hat{h}_n$ for h^* defined in* (4.23)*. Then*

$$\mathbb{E}\mathfrak{H}^2(\hat{h}_n, h^*) \leq C_d \left(\frac{k}{n}\right) (\max(1, \tau))^d \left(\sqrt{\log n}\right)^{d + (4-d)_+} \qquad \text{where } \tau := \sqrt{\frac{\sigma^2_{\max}}{\sigma^2_{\min}} - 1}. \tag{4.24}$$

Proposition 4.3.5 shows that the estimator $\hat{h}_n$ achieves near parametric risk k/n (up to logarithmic factors in n) provided τ is bounded from above by a constant. Note that this estimator $\hat{h}_n$ uses knowledge of $\sigma^2_{\min}$ but does not use knowledge of any other feature of h^* including the number of components k. In particular, this is an estimation procedure which (without knowing the value of k) achieves nearly the k/n rate for k-component well-conditioned Gaussian mixtures provided a lower bound $\sigma^2_{\min}$ on eigenvalues is known *a priori*.

It is natural to compare Proposition 4.3.5 to the main results in Maugis and Michel [93] where an adaptive procedure is developed for estimating k-component Gaussian mixtures at the rate k/n (up to a logarithmic factor) without prior knowledge of k. The estimator of Maugis and Michel [93] is very different from ours. They first fit m-component Gaussian mixtures for different values of m and then select one of these estimators by optimizing a penalized model-selection criterion. Thus, their procedure is based on solving multiple

non-convex optimization problems. Also, Maugis and Michel [93] impose upper and lower bounds on the means and the eigenvalues of the covariance matrices of the components of the mixture densities. On the contrary, our method is based on convex optimization and we only need a lower bound on the eigenvalues of the covariance matrices (no bounds on the means are necessary). On the flip side, the result of Maugis and Michel [93] has much better logarithmic factors compared to Proposition 4.3.5 and it is also stated in the form of an Oracle inequality.

4.4 Application to Gaussian Denoising

In this section, we explore the role of the NPMLE for estimating the Oracle Bayes estimator in the Gaussian denoising problem. The goal is to estimate unknown vectors $\theta_1, \ldots, \theta_n$ in $\mathbb{R}^d$ from independent random vectors $X_1, \ldots, X_n$ such that $X_i \sim N(\theta_i, I_d)$ for $i = 1, \ldots, n$. The Oracle estimator is $\hat{\theta}_i^*, i = 1, \ldots, n$ which is given by (4.5) where $\bar{G}_n$ is the empirical measure corresponding to $\theta_1, \ldots, \theta_n$.

It is natural to estimate the Oracle Bayes estimator by the Empirical Bayes estimator $\hat{\theta}_i$ which is defined as in (4.6) for $i = 1, \ldots, n$. Here $\hat{f}_n$ is any NPMLE based on $X_1, \ldots, X_n$ (defined as in (4.2)). We will gauge the performance of $\hat{\theta}_i$ as an estimator for $\hat{\theta}^*$ in terms of the squared error risk measure $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ defined in (4.7).

The main theorem of this section is given below. This is stated in a form that is similar to the statement of Theorem 4.3.1. It proves that, for every compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$, the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded from above by $\epsilon_n^2(M, S)$ up to an additional logarithmic multiplicative factor in n. Recall the form of $\epsilon_n^2(M, S)$ from (4.10).

Theorem 4.4.1. *Let $X_1, \ldots, X_n$ with independent random vectors with $X_i \sim N(\theta_i, I_d)$ for $i = 1, \ldots, n$. Let $\bar{G}_n$ denote the empirical measure corresponding to $\theta_1, \ldots, \theta_n$. Let $\hat{f}_n$ denote an NPMLE based on $X_1, \ldots, X_n$ defined as in (4.2). Let $\hat{\theta}_1, \ldots, \hat{\theta}_n$ be as defined in (4.6) and let $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ be as in (4.5). Also, let $\mathfrak{R}(\hat{\theta}, \hat{\theta}^*)$ be as in (4.7). Fix a non-empty compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$ and let $\epsilon_n(M, S)$ be defined as in (4.10). Then there exists a positive constant C_d (depending only on d) such that*

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_d \epsilon_n^2(M, S) \left(\log n\right)^{\max(d,3)}.$$

Remark 4.4.2. *For the case of $d = 1$, Jiang and Zhang [63, Theorem 5] established a related result on the risk of $\hat{\theta}_i$ in comparison to $\hat{\theta}_i^*$. The risk used therein is*

$$\left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\theta}_i - \theta_i|^2\right)\right]^{1/2} - \left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\theta}_i^* - \theta_i|^2\right)\right]^{1/2}$$

Jiang and Zhang [63] investigated the above risk in the case where $d = 1$ and $S = [-R, R]$ for some $R \geq 0$. The statement of Theorem 4.4.1 and its proof as well as the following corollary are inspired by Jiang and Zhang [63, Proof of Theorem 5].

Under specific reasonable assumptions on $\bar{G}_n$, it is possible to choose M and S explicitly which leads to the following result that is analogous to Corollary 4.3.2.

Corollary 4.4.3. *Consider the same setting and notation as in Theorem 4.4.1. Below C_d denotes a positive constant depending on d alone.*

1. *Suppose that $\bar{G}_n$ is supported on a compact subset S of $\mathbb{R}^d$. Then*

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_d \frac{\mathrm{Vol}(S^1)}{n} \left(\sqrt{\log n}\right)^{d+(4-d)_+} (\log n)^{\max(d,3)}. \tag{4.25}$$

2. *Suppose there exist a compact subset $S \subseteq \mathbb{R}^d$ and real numbers $0 < \alpha \leq 2$ and $K \geq 1$ such that (4.15) holds. Then*

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_d \frac{\mathrm{Vol}(S^1)(Ke^{1/\alpha})^d}{n} \left(\sqrt{\log n}\right)^{(2d/\alpha)+(4-d)_+} (\log n)^{\max(d,3)}. \tag{4.26}$$

3. *Suppose there exists a compact set $S \subseteq \mathbb{R}^d$ and real numbers $\mu > 0$ and $p > 0$ such that $\mu_p(\mathfrak{d}_S) \leq \mu$. Then there exists a positive constant $C_{d,\mu,p}$ (depending only on d, μ and p) such that*

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_{d,\mu,p} \left(\frac{\mathrm{Vol}(S^1)}{n}\right)^{p/(p+d)} \left(\sqrt{\log n}\right)^{(2d+p(4-d)_+)/(p+d)} (\log n)^{\max(d,3)}. \tag{4.27}$$

Corollary 4.4.3 has interesting consequences. Inequality (4.25) states that when $\bar{G}_n$ is supported on a fixed compact set S, then the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is parametric upto logarithmic multiplicative factors in n. This is especially interesting because $\hat{\theta}_1, \ldots, \hat{\theta}_n$ do not use any knowledge of S.

Corollary 4.4.3 also leads to the following result with gives an upper bound for $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ when $\theta_1, \ldots, \theta_n$ are clustered into k groups.

Proposition 4.4.4. *Consider the same setting and notation as in Theorem 4.4.1. Suppose that $\theta_1, \ldots, \theta_n$ satisfy*

$$\max_{1 \leq i \leq n} \min_{1 \leq j \leq k} \|\theta_i - a_j\| \leq R \tag{4.28}$$

for some $a_1, \ldots, a_k \in \mathbb{R}^d$ and $R \geq 0$. Then

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_d \, (1 + R)^d \left(\frac{k}{n}\right) \left(\sqrt{\log n}\right)^{d+(4-d)_+} (\log n)^{\max(d,3)}. \tag{4.29}$$

The assumption (4.28) means that $\theta_1, \ldots, \theta_n$ can be grouped into the k balls each of radius R centered at the points $a_1, \ldots, a_k$. When R is not large, this implies $\theta_1, \ldots, \theta_n$ can be clustered into k groups. In particular, when $R = 0$, the assumption (4.28) implies that $\theta_1, \ldots, \theta_n$ take only k distinct values. In words, Proposition 4.4.4 states that when

$\theta_1, \ldots, \theta_n$ are clustered into k groups, then $\hat{\theta}_1, \ldots, \hat{\theta}_n$ estimate $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ in squared error loss with accuracy k/n up to logarithmic multiplicative factors in n. The notable aspect about this result is that the estimator does not use any knowledge of k and is tuning-free. It is well-known in the clustering literature that choosing the optimal number of clusters is a challenging task (see, for example, Tibshirani, Walther, and Hastie [130]). It is therefore helpful that the estimator $\hat{\theta}_1, \ldots, \hat{\theta}_n$ achieves nearly the k/n rate in (4.28) without explicitly getting into the pesky problem of estimating k. Moreover, $\hat{\theta}_1, \ldots, \hat{\theta}_n$ is given by convex optimization (on the other hand, one usually needs to deal with non-convex optimization problems for solving clustering-type problems even if the number of clusters k is known).

There exist techniques for estimating the number of clusters and subsequently employing algorithms for minimizing the k-means objective (notably, the "gap statistic" of Tibshirani, Walther, and Hastie [130]). However, we are not aware of any result analogous to Proposition 4.28 for such techniques. There also exist other techniques for clustering based on convex optimization such as the method of convex clustering (see, for example, Lindsten, Ohlsson, and Ljung [84], Hocking et al. [57], and Chen et al. [27]) which is based on a fused lasso-type penalized optimization. This method requires specification tuning parameters. While interesting theoretical development exists for convex clustering (see, for example, Radchenko and Mukherjee [109], Zhu et al. [152], Tan and Witten [128], Wu et al. [145], and Wang et al. [140]), to the best of our knowledge, a result similar to Proposition 4.28 is unavailable.

It is straightforward to see that it is impossible to devise estimators that achieve a rate that is faster than k/n for the risk measure $\mathfrak{R}_n$. We provide a proof of this via a minimax lower bound in the following lemma. The logarithmic factors can probably be improved in Proposition 4.28 but we are unable to do so at the present moment. For the lower bound, let $\Theta_{n,d,k}$ denote the class of all n-tuples $(\theta_1, \ldots, \theta_n)$ with each $\theta_i \in \mathbb{R}^d$ and such that the number of distinct vectors among $\theta_1, \ldots, \theta_n$ is equal to k. Equivalently, $\Theta_{n,d,k}$ consists of all n-tuples $(\theta_1, \ldots, \theta_n)$ whose empirical measure is supported on a set of cardinality k. The minimax risk for estimating $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ with $(\theta_1, \ldots, \theta_n) \in \Theta_{n,d,k}$ in squared error loss from the observations $X_1, \ldots, X_n$ can be defined as

$$\mathcal{R}^*(\Theta_{n,d,k}) := \inf_{\tilde{\theta}_1, \ldots, \tilde{\theta}_n} \sup_{(\theta_1, \ldots, \theta_n) \in \Theta_{n,d,k}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\tilde{\theta}_i - \hat{\theta}_i^*\right\|^2\right]$$

The following result proves that $\mathcal{R}^*(\Theta_{n,d,k})$ is at least Ck/n for a universal positive constant C.

Lemma 4.4.5. *Let $\Theta_{n,d,k}$ and $\mathcal{R}^*(\Theta_{n,d,k})$ be defined as above. There exists a universal positive constant C such that*

$$\mathcal{R}^*(\Theta_{n,d,k}) \geq C\frac{k}{n} \qquad \text{for every } 1 \leq k \leq n. \tag{4.30}$$

Lemma 4.4.5, together with Proposition 4.4.4, implies that $\hat{\theta}_1, \ldots, \hat{\theta}_n$ is nearly minimax optimal (up to logarithmic multiplicative factors) for estimating $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ over the class

$\Theta_{n,d,k}$. Moreover, this optimality is adaptive over k because the estimator does not use any knowledge of k.

4.5 Implementation Details and Some Simulation Results

In this section, we shall discuss some computational details concerning the NPMLE and also provide numerical evidence for the effectiveness of the estimator (4.6) based on the NPMLE for denoising.

For the optimization problem (4.2), it can be shown that $\hat{f}_n$ exists and is non-unique. However $\hat{f}_n(X_1), \ldots, \hat{f}_n(X_n)$ are unique and they solve the finite dimensional optimization problem:

$$\begin{aligned} &\text{argmax} \sum_{i=1}^{n} \log f_i \\ &\text{subject to } (f_1, \ldots, f_n) \in \text{ConvexHull}\left\{(\phi(X_1 - \theta), \ldots, \phi(X_n - \theta)) : \theta \in \mathbb{R}^d\right\}. \end{aligned} \tag{4.31}$$

The constraint set in the above problem however involves every $\theta \in \mathbb{R}^d$. A natural way of computing an approximate solution is to fix a finite data-driven set $\{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d$ and restrict the infinite convex hull to the convex hull over θ belonging to this set. This leads to the problem:

$$\begin{aligned} &\text{argmax} \sum_{i=1}^{n} \log f_i \\ &\text{subject to } (f_1, \ldots, f_n) \in \text{ConvexHull}\left\{(\phi_d(X_1 - a_j), \ldots, \phi_d(X_n - a_j)) : j = 1, \ldots, m\right\}. \end{aligned} \tag{4.32}$$

This can also be seen as an approximation to (4.2) where the densities $f \in \mathcal{M}$ are restricted to have atoms in $\{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d$. (4.32) is a convex optimization problem over the probability simplex in m dimensions and can be solved using many algorithms (for example, standard interior point methods as implemented in the software, Mosek, can be used here).

The effectiveness of (4.32) as an approximation to (4.2) depends crucially on the choice of $\{a_1, \ldots, a_m\}$. For $d = 1$, Koenker and Mizera [68] propose the use of a uniform grid within the range of the observations $[\min_{1 \le i \le n} X_i, \max_{1 \le i \le n} X_i]$. Dicker and Zhao [38] discuss this approach in more detail and recommend the choice $m := [\sqrt{n}]$. They also prove (see Dicker and Zhao [38, Theorem 2]) that the resulting approximate MLE, $\tilde{f}_n$, has a squared Hellinger accuracy, $\mathfrak{H}^2(\tilde{f}_n, f_0)$, of $O_p((\log n)^2/n)$ when the mixing measure corresponding to f_0 has bounded support. For $d \ge 1$, Feng and Dicker [45] recommend taking a regular grid in a compact region containing the data. They also mention that empirical results seem "fairly insensitive" to the choice of m.

A proposal for selecting $\{a_1, \ldots, a_m\}$ that is different from gridding is the so called "exemplar" choice where one takes $m = n$ and $a_i = X_i$ for $i = 1, \ldots, n$. This choice is

proposed in Böhning [18] for $d = 1$ and in Lashkari and Golland [71] for $d \geq 1$. This avoids gridding which can be problematic in multiple dimensions. Also, this method is computationally feasible as long as n is moderate (up to a few thousands) but becomes expensive for larger n . In such instances, a reasonable strategy is to take $a_1, \ldots, a_m$ as a random subsample of the data $X_1, \ldots, X_n$. For fast implementations, one can also extend the idea of Koenker and Mizera [68] by binning the observations and weighting the likelihood terms in (4.2) by relative multinomial bin counts.

We shall now provide some graphical evidence of the effectiveness of the NPMLE for denoising. In all these plots, the NPMLE is approximately computed via the algorithm (4.32) where the $a_1, \ldots, a_m$ are chosen to be the data points $X_1, \ldots, X_n$ with $m = n$ (i.e., we follow the exemplar recommendation of Böhning [18] and Lashkari and Golland [71]). We use the software, Mosek, to solve the problem (4.32). The theorems of this chapter do not apply directly to these approximate NPMLEs and extending them is the subject of future work. However, we shall argue via simulations that these approximate NPMLEs work well for denoising.

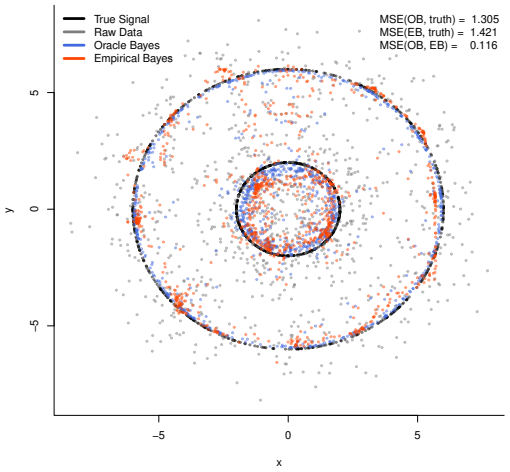
In Figure 4.1, we illustrate the performance of $\hat{\theta}_1, \ldots, \hat{\theta}_n$ (defined as in (4.6)) for denoising when the true vectors $\theta_1, \ldots, \theta_n$ take values in a bounded region of $\mathbb{R}^2$. The plots refer to these estimates as the Empirical Bayes estimates and the quantities (4.5) as the Oracle Bayes estimates. In each of the four subfigures in Figure 4.1, we generate n vectors $\theta_1, \ldots, \theta_n$ from a bounded region in $\mathbb{R}^d$ for $d = 2$: they are generated from two concentric circles in the first subfigure, a triangle in the second subfigure, the digit 8 in the third subfigure and the uppercase letter A in the last subfigure. Note that, in each of these cases, the empirical measure $\bar{G}_n$ is supported on a bounded region so that Corollary 4.4.3 yields the near parametric rate $1/n$ up to logarithmic multiplicative factors in n for every NPMLE. In each of the subfigures in Figure 4.1, we plot the true parameter values $\theta_1, \ldots, \theta_n$ in black, the data $X_1, \ldots, X_n$ (generated independently according to $X_i \sim N(\theta_i, I_2)$) are plotted in gray, the Oracle Bayes estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ are plotted in blue while the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_n$ are plotted in red. The mean squared discrepancies:

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\theta}_i^* - \theta_i\right\|^2, \quad \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\theta}_i - \theta_i\right\|^2 \text{ and } \quad \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\theta}_i^* - \hat{\theta}_i\right\|^2$$

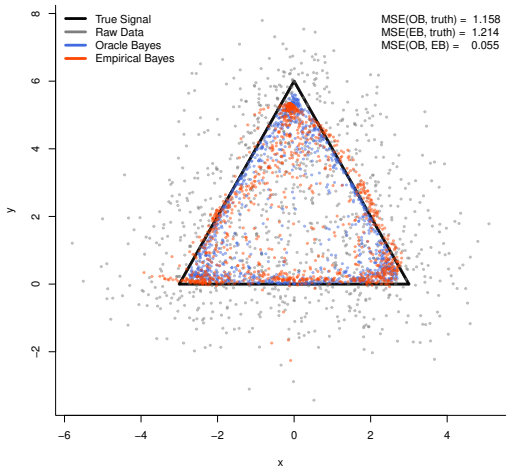
are given in each figure in the legend at the upper-right corner. Note that the third MSE is much smaller than the other two in each subfigure.

As can be observed from Figure 4.1, the Empirical Bayes estimates (4.6) approximate their targets (4.5) quite well. The most noteworthy fact is that the estimates (4.6) do not require any knowlege of the underlying structure in $\bar{G}_n$, for instance, concentric circles, or triangle or a letter of the alphabet etc. We should also note here that the noise distribution here is completely known to be $N(0, I_d)$ which implies, in particular, that there is no unknown scale parameter representing the noise variance.

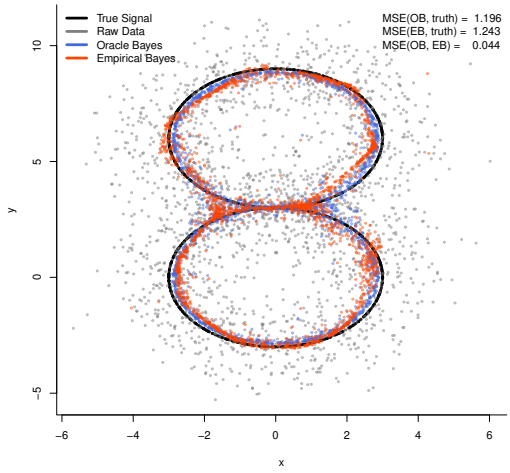
We shall now illustrate the denoising performance when the true vectors $\theta_1, \ldots, \theta_n$ have a clustering structure. Here we take $d = 2$ and consider the following four simulation settings:



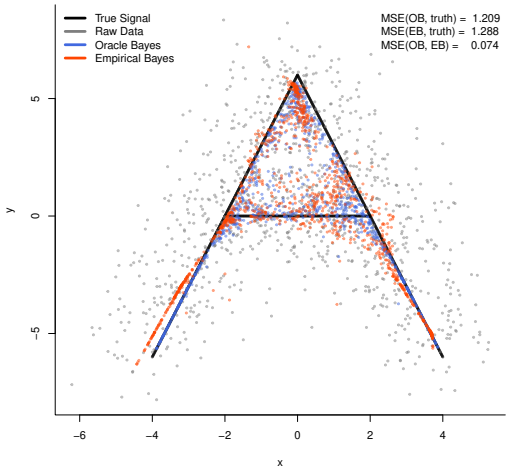
(a) **Two circles:** $n = 1000$. Half of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the concentric circles of radii 2 and 6 respectively.



(b) **Triangle:** $n = 999$. A third of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each edge of the triangle with vertices $(-3, 0)$, $(0, 6)$ and $(3, 0)$



(c) **Digit** 8**:** $n = 1000$. Half of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the circles of radii 3 cnetered at $(0, 0)$ and $(0, 6)$ respectively.



(d) **Letter A:** $n = 1000$. A fifth of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the line segments joining the points $(-4, -6), (-2, 0), (0, 6), (2, 0)$ and $(4, 6)$ so as to form the letter A.

Figure 4.1: Illustrations of denoising using the Empirical Bayes estimates (4.6)

1. Setting One: We generate $\theta_1, \ldots, \theta_n$ as i.i.d from the distribution which puts equal probability (0.5) at $(0, 0)$ and $(2, 2)$.

2. Setting Two: We generate $\theta_1, \ldots, \theta_n$ as i.i.d from the distribution which puts $1/4$ probability at $(0, 0)$ and $3/4$ probability at $(2, 2)$.

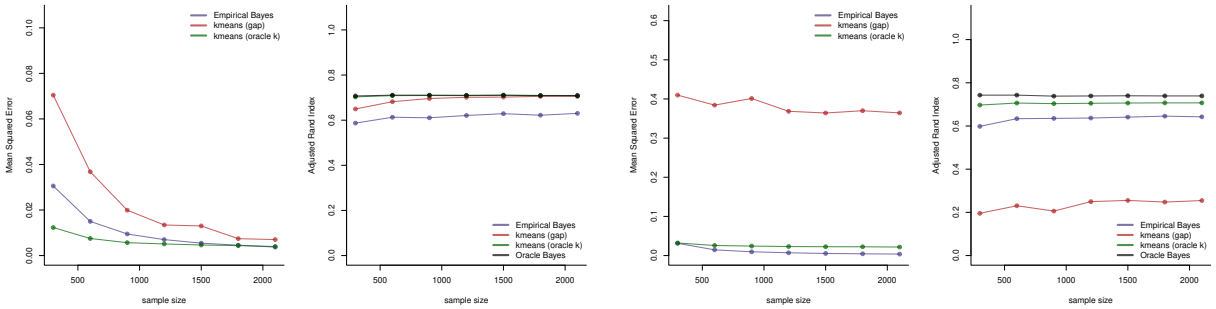
3. Setting Three: We generate $\theta_1, \ldots, \theta_n$ as i.i.d from the distribution which puts $1/4$ probability each at $(0, 0)$ and $(0, 2)$ and $1/2$ probability at $(2, -2)$.

4. Setting Four: We generate a random probability vector $(\alpha_1, \alpha_2, .\alpha_3, \alpha_4)$ from the Dirichlet distribution with parameters $(1, 1, 1, 1)$ and then generate $\theta_1, \ldots, \theta_n$ as i.i.d from the probability distribution with puts probabilities $\alpha_1, \alpha_2, \alpha_3$ and α_4 at the four points $(0, 0)$, $(0, 3)$, $(3, 0)$ and $(3, 3)$.

The observed data $X_1, \ldots, X_n$ are, as usual, generated independently as $X_i \sim N(\theta_i, I_d)$. We allow the sample size n to range in the set $\{300, 600, 900, 1200, 1500, 1800, 2100\}$. For each n, we perform 1000 replicates to get accurate estimates of mean squared error. For each dataset, we compute the Empirical Bayes estimates (4.6). For comparison, we also computed k-means estimates based on the true (Oracle) value of k and those based on the gap statistic (from Tibshirani, Walther, and Hastie [130]). These estimates will be referred to, in the sequel, as `kmeans-Oracle` and `kmeans-gap` respectively. For k-means, we used the standard Lloyd's algorithm based on 10 random starts and the best solution is considered of the random starts. Note that because of non-convexity, no implementation of k-means can provably reach global optimum.

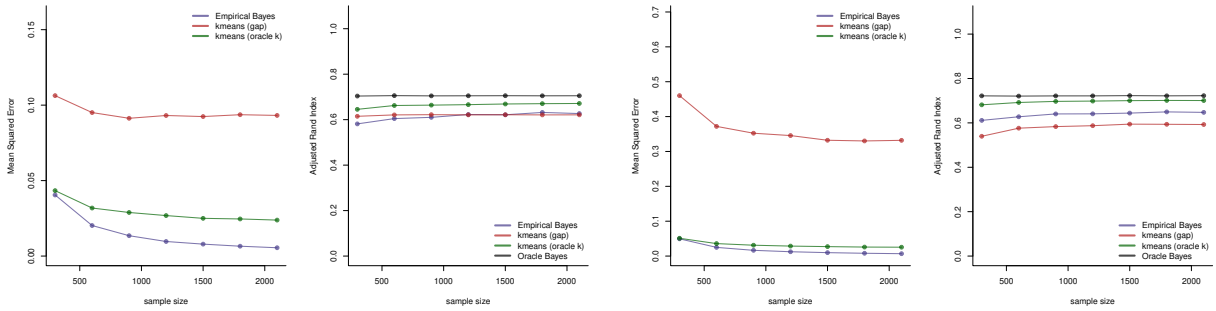
For each of the these three estimates, we plotted the mean squared errors in Figure 4.2 (see the first plot in each pair of plots for the different settings). From these MSE plots, it is clear that the Empirical Bayes estimates based on the NPMLE are more accurate than `kmeans-gap`. In fact, with the exception of the first setting, the Empirical Bayes estimates are even more accurate than `kmeans-Oracle`. This is probably because of the non-convexity of k-means.

In addition to the MSE, we also compared the clusterings given by the different methods based on the Adjusted Rand Index (ARI) [111]. The Empirical Bayes is designed to work well for the squared error objective and not quite for the ARI. We plotted the average ARI of each of the three methods as well as the average ARI of the Oracle Bayes estimate. Higher ARIs are preferred to lower values. Here the Oracle Bayes estimate is the best; the `kmeans-Oracle` method is superior to the Empirical Bayes estimate as well as `kmeans-gap`. The comparison between the Empirical Bayes and the `kmeans-gap` estimates in terms of ARI can be summarized as follows. In the first setting, the performance of `kmeans-gap` is very good and is indistinguishable from `kmeans-Oracle`. In more complicated settings with more than two clusters and/or with imbalanced cluster proportions, a distinction between the two methods becomes apparent. In the second and fourth settings, the Empirical Bayes method outperforms `kmeans-gap`. In the third setting, the performances of the two methods start to coincide for larger sample sizes.



(a) Setting 1. Two equally sized clusters centered at $(0, 0)$ and $(2, 2)$. For clarification, in the ARI plot the red and green curves coincide.

(b) Setting 2. Two clusters centered at $(0, 0)$ and $(2, 2)$ with cluster proportions $1/4$ and $3/4$.For clarification, in the ARI plot the red and green curves coincide.



(c) Setting 3. Three clusters centered at $(0, 0), (0, 2), (2, -2)$ with cluster proportions $1/4, 1/4, 1/2$ respectively.

(d) Setting 4. Four cluster centers centered at $(0, 0), (0, 3), (3, 0), (3, 3)$ with cluster proportions drawn from Dirichlet distribution with parameters $(1, 1, 1, 1)$

Figure 4.2: Empirical performance of methods in the denoising problem in four different clustering settings. A method with lower MSE is preferred over one with higher MSE. In contrast, a method with higher ARI is preferred over one with lower ARI. The lines show mean of the metric in question over 1000 replicates.

4.6 Proofs of results in Section 4.3

The following notation will be used in the proofs in the sequel.

1. For $x \in \mathbb{R}^d$ and $a > 0$, let

$$B(x, a) := \{u \in \mathbb{R}^d : \|u - x\| \leq a\}$$

denote the closed ball of radius a centered at x.

2. For a subset $S \subseteq \mathbb{R}^d$ and $a > 0$, we denote the set S^a by

$$S^a := \cup_{x \in S} B(x, a) = \{y : \mathfrak{d}_S(y) \leq a\} \tag{4.33}$$

where $\mathfrak{d}_S(\cdot)$ is defined as in (4.8).

3. For a compact subset S of $\mathbb{R}^d$ and $\epsilon > 0$, we denote the ϵ-covering number of S in the usual Euclidean distance by $N(\epsilon, S)$ i.e., $N(\epsilon, S)$ stands for the smallest number of closed balls of radius ϵ whose union contains S.

4. Given a pseudometric ϱ on $\mathcal{M}$, let $N(\epsilon, \mathcal{M}, \varrho)$ denote the ϵ-covering number of $\mathcal{M}$ under the pseudometric ϱ by $N(\epsilon, \mathcal{M}, \varrho)$ i.e., $N(\epsilon, \mathcal{M}, \varrho)$ denotes the smallest positive integer N for which there exist densities $f_1, \ldots, f_N \in \mathcal{M}$ satisfying

$$\sup_{f \in \mathcal{M}} \inf_{1 \leq i \leq N} \varrho(f, f_i) \leq \epsilon.$$

In the proof below, we will be concerned with $N(\epsilon, \mathcal{M}, \varrho)$ for the following choice of ϱ. For a compact set S, let $\|\cdot\|_{\infty,S}$ denote the pseudonorm on $\mathcal{M}$ defined by

$$\|f\|_{\infty,S} := \sup_{x \in S} |f(x)|. \tag{4.34}$$

This pseudonorm naturally induces a pseudometric on $\mathcal{M}$ given by $\varrho(f, g) := \|f - g\|_{\infty,S}$. The covering number for this pseudometric will be denoted by $N(\epsilon, \mathcal{M}, \|\cdot\|_{\infty,S})$. In the proofs for the results in Section 4.4, we will need to deal with covering numbers for other pseudometrics ϱ on $\mathcal{M}$ as well. These pseudometrics will be introduced in Section 4.7.

With the above notation in place, we are now ready to give the proof of Theorem 4.3.1. This proof uses additional ingredients which are proved in later sections. Arguably the most important ingredient for the proof of this theorem is a bound on the covering numbers $N(\epsilon, \mathcal{M}, \|\cdot\|_{\infty,S})$. These bounds are given in Section 4.8 (specifically inequality (4.76) in Theorem 4.8.1 will be used). Other ingredients include inequality (A.13) (which is a consequence of Lemma A.0.4) and a standard fact (Lemma A.0.9) giving a volumetric upper bound for Euclidean covering numbers.

4.6.1 Proof of Theorem 4.3.1

Proof of Theorem 4.3.1. We shall prove inequalities (4.12) and (4.13) under the assumption that the sample size n satisfies

$$n \geq \max\left(\exp\left(\frac{d+1}{2}\right), \frac{1}{2}(2\pi)^{(d-1)/2}\right). \tag{4.35}$$

If (4.35) is not satisfied, then $\epsilon_n(M,S)$ (and also the larger quantity $\epsilon_n(M,S)/\min(1-\alpha,\beta)$) will be bounded from below by a positive constant κ_d. We can then therefore choose C_d in (4.12) and (4.13) large enough so that $\epsilon_n(M,S)\sqrt{C_d} > \sqrt{2\min(1-\alpha,\beta)}$. Because the Hellinger distance $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ is always bounded from above by $\sqrt{2}$, the probability on the left hand side of (4.12) will then equal zero so that (4.12) holds trivially. Inequality (4.13) will also be trivial because its right hand side will then be larger than 2.

Let us therefore fix n satisfying (4.35). Fix a positive sequence $\{\gamma_n\}$ and assume that $\hat{f}_n$ satisfies

$$\prod_{i=1}^{n} \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \geq \exp\left((\beta-\alpha)n\gamma_n^2\right) \qquad \text{for some } 0 < \beta \leq \alpha < 1. \tag{4.36}$$

We shall then bound the probability

$$\mathbb{P}\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n\}$$

for $t \geq 1$.

Fix a non-empty compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10\log n}$. We shall work with the set S^M (defined as in (4.33)) and the pseudometric given by the pseudonorm $\|\cdot\|_{\infty,S^M}$ (defined as in (4.34)).

Let $\eta := 1/n^2$ and let $\{h_1,\ldots,h_N\} \subseteq \mathcal{M}$ denote an η-covering set of $\mathcal{M}$ in the pseudometric given by $\|\cdot\|_{\infty,S^M}$ where $N = N(\eta,\mathcal{M},\|\cdot\|_{\infty,S^M})$ i.e.,

$$\sup_{h\in\mathcal{M}} \inf_{1\leq j\leq N} \|h - h_j\|_{\infty,S^M} \leq \eta.$$

Inequality (4.76) in Theorem 4.8.1 gives an upper bound for N that will be crucially used in this proof.

Let J denote the set of all $j \in \{1,\ldots,N\}$ for which there exists a density $h_{0j} \in \mathcal{M}$ satisfying

$$\|h_{0j} - h_j\|_{\infty,S^M} \leq \eta \quad \text{and} \quad \mathfrak{H}(h_{0j}, f_{\bar{G}_n}) \geq t\gamma_n.$$

Because $h_1,\ldots,h_N$ cover $\mathcal{M}$, there will exist $1 \leq j \leq N$ such that $\|h_j - \hat{f}_n\|_{\infty,S^M} \leq \eta$. If $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n$, then $j \in J$ and consequently

$$\|\hat{f}_n - h_{0j}\|_{\infty,S^M} \leq 2\eta. \tag{4.37}$$

We now define a function $v := v_{S,M} : \mathbb{R}^d \to (0, \infty)$ via

$$v(x) := \begin{cases} \eta & \text{if } x \in S^M \\ \eta \left(\frac{M}{\mathfrak{d}_S(x)} \right)^{d+1} & \text{otherwise} \end{cases} \tag{4.38}$$

where $\mathfrak{d}_S : \mathbb{R}^d \to [0, \infty)$ is defined as in (4.8).

Inequality (4.37) clearly implies that $\hat{f}_n(X_i) \leq h_{0j}(X_i) + 2\eta = h_{0j}(X_i) + 2v(X_i)$ whenever $X_i \in S^M$ which allows us to write

$$\prod_{i=1}^n \hat{f}_n(X_i) \leq \prod_{i:X_i \in S^M} \{h_{0j}(X_i) + 2v(X_i)\} \prod_{i:X_i \notin S^M} (2\pi)^{-d/2}$$

where we used the bound $\hat{f}_n(X_i) \leq \sup_x \hat{f}_n(x) \leq (2\pi)^{-d/2}$ for $X_i \notin S^M$ (the bound $\sup_x f(x) \leq (2\pi)^{-d/2}$ holds for every $f \in \mathcal{M}$ as can easily be seen). From here, we deduce

$$\begin{aligned} \prod_{i=1}^n \hat{f}_n(X_i) &\leq \prod_{i=1}^n \{h_{0j}(X_i) + 2v(X_i)\} \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{h_{0j}(X_i) + 2v(X_i)} \\ &\leq \prod_{i=1}^n \{h_{0j}(X_i) + 2v(X_i)\} \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \end{aligned}$$

We have therefore proved that the inequality

$$\prod_{i=1}^n \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \leq \max_{j \in J} \prod_{i=1}^n \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)}$$

holds on the event $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n$.

Because $\hat{f}_n$ satisfies (4.36), we obtain

$$\begin{aligned} \mathbb{P}\left(\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n \right) &\leq \mathbb{P}\left\{ \max_{j \in J} \prod_{i=1}^n \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \geq \exp\left((\beta - \alpha) n t^2 \gamma_n^2 \right) \right\} \\ &\leq \mathbb{P}\left\{ \max_{j \in J} \prod_{i=1}^n \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \geq e^{-\alpha n t^2 \gamma_n^2} \right\} \\ &\quad + \mathbb{P}\left\{ \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \geq e^{\beta n t^2 \gamma_n^2} \right\}. \end{aligned} \tag{4.39}$$

We shall bound the two probabilities above separately. For the first probability:

$$\begin{aligned}\mathbb{P}\left\{\max_{j\in J}\prod_{i=1}^{n}\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}\geq e^{-\alpha nt^2\gamma_n^2}\right\} &\leq \sum_{j\in J}\mathbb{P}\left\{\prod_{i=1}^{n}\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}\geq e^{-\alpha nt^2\gamma_n^2}\right\}\\ &\leq e^{\alpha nt^2\gamma_n^2/2}\sum_{j\in J}\mathbb{E}\prod_{i=1}^{n}\sqrt{\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}}\\ &= e^{\alpha nt^2\gamma_n^2/2}\sum_{j\in J}\prod_{i=1}^{n}\mathbb{E}\sqrt{\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}}.\end{aligned}$$

Now for each fixed $j\in J$, we have

$$\begin{aligned}\prod_{i=1}^{n}\mathbb{E}\sqrt{\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}} &= \exp\left(\sum_{i=1}^{n}\log\mathbb{E}\sqrt{\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}}\right)\\ &\leq \exp\left(\sum_{i=1}^{n}\mathbb{E}\sqrt{\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}}-n\right)\\ &\leq \exp\left(\sum_{i=1}^{n}\int\sqrt{\frac{h_{0j}+2v}{f_{\bar{G}_n}}}f_{G_i}-n\right)=\exp\left(n\int\sqrt{(h_{0j}+2v)\,f_{\bar{G}_n}}-n\right).\end{aligned}$$

Because of $\sqrt{\alpha+\beta}\leq\sqrt{\alpha}+\sqrt{\beta}$ and the Cauchy-Schwartz inequality (along with $\int f_{\bar{G}_n}=1$), we obtain

$$\begin{aligned}\int\sqrt{(h_{0j}+2v)\,f_{\bar{G}_n}} &\leq \int\sqrt{h_{0j}f_{\bar{G}_n}}+\sqrt{2}\int\sqrt{vf_{\bar{G}_n}}\\ &\leq \int\sqrt{h_{0j}f_{\bar{G}_n}}+\sqrt{2}\sqrt{\int v}=1-\frac{1}{2}\mathfrak{H}^2(h_{0j},f_{\bar{G}_n})+\sqrt{2}\sqrt{\int v}.\end{aligned}$$

We now use Lemma A.0.8 which gives an upper bound on $\int v$. This (along with the fact that $\mathfrak{H}(h_{0j},f_{\bar{G}_n})\geq t\gamma_n$) allows us to deduce:

$$\int\sqrt{(h_{0j}+2v(X_i))\,f_{\bar{G}_n}}\leq 1-\frac{t^2}{2}\gamma_n^2+C_d\sqrt{2\eta\mathrm{Vol}(S^M)}.$$

We have therefore proved that

$$\begin{aligned}\mathbb{P}\left\{\max_{j\in J}\prod_{i=1}^{n}\frac{h_{0j}(X_i)+2v(X_i)}{f_{\bar{G}_n}(X_i)}\geq e^{-\alpha nt^2\gamma_n^2}\right\} &\leq \exp\left(\frac{\alpha}{2}nt^2\gamma_n^2+\log|J|-\frac{1}{2}nt^2\gamma_n^2+nC_d\sqrt{\eta\mathrm{Vol}(S^M)}\right)\\ &\leq \exp\left(\frac{\alpha-1}{2}nt^2\gamma_n^2+\log N+C_d\sqrt{\mathrm{Vol}(S^M)}\right)\end{aligned}\tag{4.40}$$

because $\eta := n^{-2}$ and $|J| \leq N$ (as $J \subseteq \{1, \ldots, N\}$).

We now use the upper bound on N from inequality (4.76) in Theorem 4.8.1. Because $\eta = 1/n^2$ and $n \geq 2$, the quantity a appearing in Theorem 4.8.1 satisfies

$$a = \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\eta}} = \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}} + 4 \log n} \leq \sqrt{6 \log n}.$$

Also because of (4.35), we have $2n \geq (2\pi)^{(d-1)/2}$ so that

$$a = \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\eta}} = \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}} + 4 \log n} \geq \sqrt{2 \log(1/n) + 4 \log n} = \sqrt{2 \log n}.$$

Thus Theorem 4.8.1 gives

$$\log N \leq C_d N(a, (S^M)^a)(\log n)^2 \leq C_d N(\sqrt{2 \log n}, S^{M+\sqrt{6 \log n}})(\log n)^2.$$

Using Lemma A.0.9 to bound the Euclidean covering number appearing in the right hand side above, we deduce that

$$\begin{aligned} N(\sqrt{2 \log n}, S^{M+\sqrt{6 \log n}}) &\leq C_d (\sqrt{2 \log n})^{-d} \mathrm{Vol}(S^{M+\sqrt{6 \log n}+\sqrt{2 \log n}/2}) \\ &\leq C_d (\log n)^{-d/2} \mathrm{Vol}(S^{M+\sqrt{10 \log n}}) \leq C_d (\log n)^{-d/2} \mathrm{Vol}(S^{2M}) \end{aligned}$$

as $M \geq \sqrt{10 \log n}$. Thus

$$\log N \leq C_d (\log n)^{2-(d/2)} \mathrm{Vol}(S^{2M}).$$

Using the above in (4.40), we obtain

$$\begin{aligned} \mathbb{P}\left\{ \max_{j \in J} \prod_{i=1}^{n} \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \geq e^{-\alpha n t^2 \gamma_n^2} \right\} \leq \exp\Bigg(& \frac{\alpha - 1}{2} n t^2 \gamma_n^2 \\ & + C_d (\log n)^{2-(d/2)} \mathrm{Vol}(S^{2M}) + C_d \sqrt{\mathrm{Vol}(S^M)} \Bigg). \end{aligned} \tag{4.41}$$

We shall now bound the second probability in (4.39). First observe, by Markov's inequality, that

$$\mathbb{P}\left\{ \prod_{i: X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \geq e^{\beta n t^2 \gamma_n^2} \right\} \leq \exp\left(-\frac{\beta n t^2 \gamma_n^2}{2 \log n} \right) \mathbb{E}\left(\prod_{i: X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \right)^{1/(2 \log n)}$$

The expectation above can be bounded as (recall the formula for $v(\cdot)$ from (4.38))

$$\begin{aligned}\mathbb{E}\left(\prod_{i:X_i\notin S^M}\frac{(2\pi)^{-d/2}}{2v(X_i)}\right)^{1/(2\log n)} &\le \mathbb{E}\left(\prod_{i:X_i\notin S^M}\frac{1}{v(X_i)}\right)^{1/(2\log n)}\\ &= \mathbb{E}\left(\prod_{i:X_i\notin S^M}\frac{\mathfrak{d}_S(X_i)}{M\eta^{1/(d+1)}}\right)^{(d+1)/(2\log n)}\\ &= \mathbb{E}\left[\prod_{i=1}^n\left(\frac{\mathfrak{d}_S(X_i)}{M\eta^{1/(d+1)}}\right)^{I\{\mathfrak{d}_S(X_i)\ge M\}}\right]^{(d+1)/(2\log n)}\end{aligned}$$

The above term will be controlled below by using inequality (A.13) (which is a consequence of Lemma A.0.4) with

$$a := \frac{1}{M\eta^{1/(d+1)}} \qquad \text{and} \qquad \lambda := \frac{d+1}{2\log n} \tag{4.42}$$

to obtain

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^n\left(\frac{\mathfrak{d}_S(X_i)}{M\eta^{1/(d+1)}}\right)^{I\{\mathfrak{d}_S(X_i)\ge M\}}\right]^{(d+1)/(2\log n)} &\le \exp\Bigg\{C_d a^\lambda M^{\lambda+d-2}\\ &\quad + (aM)^\lambda n\left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p\Bigg\}.\end{aligned} \tag{4.43}$$

We need to assume here that

$$\log n \ge \frac{d+1}{2\min(1,p)}$$

to ensure that $\lambda \le \min(1,p)$ as required for inequality (A.13). This is satisfied as long as $p \ge (d+1)/(2\log n)$ because under the assumption (4.35), we have $\log n \ge \frac{d+1}{2}$. Thus (4.43) holds for all $p \ge (d+1)/(2\log n)$.

For notational convenience, we write $\mu_p := \mu_p(\mathfrak{d}_S)$ in the rest of the proof. With the choices (4.42) (and $\eta = 1/n^2$), the first term in the exponent of the right hand side of (4.43) is calculated as

$$a^\lambda M^{\lambda+d-2} = M^{d-2}\eta^{-\lambda/(d+1)} = M^{d-2}n^{1/(\log n)} = eM^{d-2}.$$

On the other hand, the second term in the exponent in (4.43) becomes

$$(aM)^\lambda n\left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p = en\left(\frac{2\mu_p}{M}\right)^p.$$

Therefore the second probability in (4.39) satisfies the inequality:

$$\mathbb{P}\Bigg\{\prod_{i:X_i\notin S^M}\frac{(2\pi)^{-d/2}}{2v(X_i)} \ge e^{\beta nt^2\gamma_n^2}\Bigg\} \le \exp\left(-\frac{\beta nt^2\gamma_n^2}{2\log n} + C_d M^{d-2} + en\left(\frac{2\mu_p}{M}\right)^p\right).$$

This is true for all $p \geq (d+1)/(2\log n)$ so we can also write

$$\mathbb{P}\Big\{ \prod_{i:X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \geq e^{\beta n t^2 \gamma_n^2} \Big\} \leq \exp\left(-\frac{\beta n t^2 \gamma_n^2}{2\log n} + C_d M^{d-2} + en \inf_{p \geq (d+1)/(2\log n)} \left(\frac{2\mu_p}{M}\right)^p \right).$$

We have proved therefore that for every $t > 0$

$$\begin{aligned}\mathbb{P}\left\{ \mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n \right\} &\leq \exp\left(\frac{\alpha - 1}{2} n t^2 \gamma_n^2 + C_d (\log n)^{2-(d/2)} \mathrm{Vol}(S^{2M}) + C_d \sqrt{\mathrm{Vol}(S^M)} \right) \\ &\quad + \exp\left(-\frac{\beta n t^2 \gamma_n^2}{2\log n} + C_d M^{d-2} + en \inf_{p \geq (d+1)/(2\log n)} \left(\frac{2\mu_p}{M}\right)^p \right).\end{aligned}$$

We now note that

$$\mathrm{Vol}(S^M) \leq \mathrm{Vol}(S^{2M}) \leq C_d M^d \mathrm{Vol}(S^1)$$

which follows from inequality (A.27) in Lemma A.0.9. This, along with the definition of $\epsilon_n^2(M, S)$ in (4.10), gives

$$\max\left((\log n)^{2-(d/2)} \mathrm{Vol}(S^{2M}), \sqrt{\mathrm{Vol}(S^M)}, M^{d-2}, n \inf_{p \geq (d+1)/(2\log n)} \left(\frac{2\mu_p}{M}\right)^p \right) \leq C_d n \epsilon_n^2(M, S).$$

As a result,

$$\mathbb{P}\left\{ \mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n \right\} \leq \exp\left(\frac{\alpha - 1}{2} n t^2 \gamma_n^2 + C_d n \epsilon_n^2(M, S) \right) + \exp\left(-\frac{\beta n t^2 \gamma_n^2}{2\log n} + C_d n \epsilon_n^2(M, S) \right).$$

Now suppose that

$$\gamma_n^2 = C_d' \frac{\epsilon_n^2(M, S)}{\min(1 - \alpha, \beta)} \tag{4.44}$$

for some $C_d' \geq 4C_d$. We deduce then that, for every $t \geq 1$,

$$\begin{aligned}\mathbb{P}\left\{ \mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n \right\} &\leq \exp\left(-\frac{1 - \alpha}{2} n t^2 \gamma_n^2 + \frac{1 - \alpha}{4} n \gamma_n^2 \right) + \exp\left(-\frac{\beta}{2\log n} n t^2 \gamma_n^2 + \frac{\beta}{4\log n} n \gamma_n^2 \right) \\ &\leq 2\exp\left(-\frac{\min((1 - \alpha), \beta)}{4\log n} n t^2 \gamma_n^2 \right) \qquad (4.45)\end{aligned}$$

Observe now that (because $M \geq \sqrt{10 \log n}$)

$$\epsilon_n^2(M, S) \geq \mathrm{Vol}(S^1) \frac{M^d}{n} \left(\sqrt{\log n} \right)^{(4-d)_+} \geq \mathrm{Vol}(B(0, 1)) \frac{(\log n)^2}{n}$$

so that we can choose the constant C_d' such that

$$n \min(1 - \alpha, \beta) \gamma_n^2 \geq C_d' n \epsilon_n^2(M, S) \geq 4(\log n)^2.$$

This gives, via (4.45),

$$\mathbb{P}\left\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n\right\} \leq 2n^{-t^2}.$$

We have therefore proved the above inequality for γ_n as chosen in (4.44) (provided C'_d is chosen sufficiently large) for every estimator $\hat{f}_n$ satisfying (4.36). This completes the proof of (4.12).

For (4.13), we multiply both sides of (4.12) by t and then integrate from $t = 1$ to $t = \infty$ to obtain

$$\mathbb{E}\left(\frac{\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n})\min(1-\alpha,\beta)}{C_d\epsilon_n^2(M,S)}\right) \leq 1 + 4\int_1^\infty tn^{-t^2} \leq 1 + \frac{2}{n\log n} \leq 4$$

which proves (4.13) and completes the proof of Theorem 4.3.1. □

4.6.2 Proof of Corollary 4.3.2

Proof of Corollary 4.3.2. To prove (4.14), assume that $\bar{G}_n$ is supported on a compact set S. We then apply Theorem 4.3.1 to this S and $M = \sqrt{10\log n}$. Because $\bar{G}_n$ is supported on S, we have $\mu_p(\mathfrak{d}_S) = 0$ for every $p > 0$ so that $\epsilon_n^2(M,S)$ (defined in (4.10)) becomes

$$\epsilon_n^2(M,S) = \mathrm{Vol}(S^1)\frac{M^d}{n}\left(\sqrt{\log n}\right)^{(4-d)_+} = \frac{\mathrm{Vol}(S^1)}{n}\left(\sqrt{\log n}\right)^{d+(4-d)_+}.$$

Inequality (4.14) then immediately follows from Theorem 4.3.1.

We next prove (4.16) assuming the condition (4.15). Let

$$M := 4K(e\log n)^{1/\alpha}. \tag{4.46}$$

This quantity $M \geq \sqrt{10\log n}$ because $K \geq 1$ and $\alpha \leq 2$. We shall apply (4.13) with this M. Let

$$T_2(M,S) := (\log n)\inf_{p \geq \frac{d+1}{2\log n}}\left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p$$

be the second term on the right hand side of (4.10) in the definition of $\epsilon_n^2(M,S)$. The infimum over p above is easily seen to be achieved at $p = (M/(2K))^\alpha(1/e)$. By the expression (4.46) for M, it is easy to see that $p \geq (d+1)/(2\log n)$ provided

$$n \geq \exp\left(\sqrt{(d+1)/2}\right). \tag{4.47}$$

We then deduce that

$$T_2(M,S) \leq (\log n)\exp\left(\frac{-1}{\alpha e}\left(\frac{M}{2K}\right)^\alpha\right).$$

It follows from here that $T_2(M,S) \le (\log n)/n$ because $M \ge (4K)(e\log n)^{1/\alpha} \ge (2K)(\alpha e \log n)^{1/\alpha}$. Thus

$$\begin{aligned}\epsilon_n^2(M,S) &= \mathrm{Vol}(S^1)\frac{M^d}{n}\left(\sqrt{\log n}\right)^{(4-d)_+} + T_2(M,S) \\ &\le \mathrm{Vol}(S^1)\frac{(4Ke^{1/\alpha})^d}{n}(\log n)^{d/\alpha}\left(\sqrt{\log n}\right)^{(4-d)_+} + \frac{\log n}{n}\end{aligned}$$

and hence (4.16) readily follows as a consequence of Theorem 4.3.1. When the assumption (4.47) does not hold, inequality (4.16) becomes trivially true when C_d is chosen sufficiently large.

We now turn to (4.17). Assume that S is such that $\mu_p(\mathfrak{d}_S) \le \mu$ for fixed $\mu > 0$ and $p > 0$. Then Theorem 4.3.1 gives

$$\begin{aligned}\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) &\le C_d \inf_{M \ge \sqrt{10\log n}} \epsilon_n^2(M,S) \\ &= C_d \inf_{M \ge \sqrt{10\log n}} \left(\mathrm{Vol}(S^1)\frac{M^d}{n}\left(\sqrt{\log n}\right)^{(4-d)_+} + (\log n)\left(\frac{2\mu}{M}\right)^p\right)\end{aligned}$$

where we assumed that n is large enough so that $p \ge (d+1)/(2\log n)$. Taking

$$M = \left(\sqrt{\log n}\right)^{(2-(4-d)_+)/(p+d)}\left(\frac{n\mu^p}{\mathrm{Vol}(S^1)}\right)^{1/(p+d)}$$

results in (4.17). When n is large enough, M chosen as above exceeds $\sqrt{10\log n}$. For smaller n, the inequality (4.17) trivially holds provided $C_{d,\mu,p}$ is chosen large enough. □

4.6.3 Proof of Proposition 4.3.3

This directly follows from inequality (4.14). Suppose that $\bar{G}_n$ is supported on a finite set S of cardinality k. We then apply inequality (4.14) to this S. It is easy to see then that $\mathrm{Vol}(S^1) \le C_d k$ which proves (4.18).

4.6.4 Proof of Lemma 4.3.4

The following uses standard ideas involving Assouad's lemma (see, for example, Tsybakov [135, Chapter 2].

Proof of Lemma 4.3.4. Fix $\delta > 0$ and $M > 0$. Let $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ be points in $\mathbb{R}^d$ such that

$$\min\left(\min_{i\ne j}\|a_i - a_j\|, \min_{i\ne j}\|b_i - b_j\|, \min_{i\ne j}\|a_i - b_j\|\right) \ge M \tag{4.48}$$

and such that

$$\|a_i - b_i\| = \delta \qquad \text{for every } 1 \le i \le k. \tag{4.49}$$

Now for every $\tau \in \{0,1\}^k$, let

$$f_\tau(x) = \frac{1}{k}\sum_{i=1}^{k} \phi_d(x - a_i(1-\tau_i) - b_i\tau_i)$$

where $\phi_d(\cdot)$ is the standard normal density on $\mathbb{R}^d$. Clearly $f_\tau \in \mathcal{M}_k$ for every $\tau \in \{0,1\}^k$. We shall now employ Assouad's lemma which gives

$$\mathcal{R}(\mathcal{M}_k) \geq \frac{k}{8} \min_{\tau \neq \tau'} \frac{\mathfrak{H}^2(f_\tau, f_{\tau'})}{\Upsilon(\tau,\tau')} \min_{\Upsilon(\tau,\tau')=1} \left(1 - \left\|P_{f_\tau} - P_{f_{\tau'}}\right\|_{TV}\right)$$

where $\Upsilon(\tau,\tau') := \sum_{i=1}^k I\{\tau_j \neq \tau_j'\}$ denotes Hamming distance and P_f (for $f \in \mathcal{M}$) denotes the joint distribution of $X_1, \ldots, X_n$ which are independently distributed according to f.

We now fix $\tau \neq \tau' \in \{0,1\}^k$ and bound $\mathfrak{H}^2(f_\tau, f_{\tau'})$ from below. For simplicity, let $f = f_\tau$ and $g = f_{\tau'}$. Also, for $i = 1, \ldots, k$, let

$$f_i(x) := \phi_d(x - a_i(1-\tau_i) - b_i\tau_i) \quad \text{and} \quad g_i(x) := \phi_d(x - a_i(1-\tau_i') - b_i\tau_i')$$

so that $f = \sum_{i=1}^k f_i/k$ and $g = \sum_{i=1}^k g_i/k$. This gives

$$\frac{1}{2}\mathfrak{H}^2(f,g) = 1 - \int \sqrt{f(x)g(x)}dx = 1 - \int \sqrt{\frac{1}{k^2}\sum_{i,j} f_i(x)g_j(x)}dx \geq 1 - \frac{1}{k}\sum_{i,j}\int \sqrt{f_i(x)g_j(x)}dx$$

Because f_i and g_j are normal densities, by a straightforward computation, we obtain

$$\int \sqrt{f_i(x)g_j(x)}dx = \exp\left(-\left\|a_i(1-\tau_i) + b_i\tau_i - a_j(1-\tau_j') - b_j\tau_j'\right\|^2/8\right)$$

so that by (4.48) and (4.49), we obtain that

$$\int \sqrt{f_i(x)g_j(x)}dx = I\{\tau_i = \tau_i'\} + I\{\tau_i \neq \tau_i'\}e^{-\delta^2/8} \qquad \text{for } i = j$$

and

$$\int \sqrt{f_i(x)g_j(x)}dx \leq e^{-M^2/8} \qquad \text{for } i \neq j.$$

As a result, we obtain

$$\begin{aligned}
\frac{1}{2}\mathfrak{H}^2(f_\tau, f_{\tau'}) &= 1 - \frac{1}{k}\sum_{i=1}^k \int \sqrt{f_i(x)g_i(x)}dx - \frac{1}{k}\sum_{i\neq j}\int \sqrt{f_i(x)g_j(x)}dx \\
&\geq 1 - \frac{1}{k}\sum_{i=1}^k I\{\tau_i = \tau_i'\} - \frac{e^{-\delta^2/8}}{k}\Upsilon(\tau,\tau') - \frac{k^2-k}{k}e^{-M^2/8} \\
&= \frac{1}{k}\Upsilon(\tau,\tau')\left(1 - e^{-\delta^2/8}\right) - (k-1)e^{-M^2/8} \qquad (4.50)
\end{aligned}$$

for every $\tau \neq \tau' \in \{0,1\}^k$. Now let us fix τ, τ' with $\Upsilon(\tau, \tau') = 1$ and bound from above the total variation distance between P_{f_τ} and $P_{f_{\tau'}}$. Without loss of generality, we can assume that $\tau_1 \neq \tau_1'$ and that $\tau_i = \tau_i'$ for $i \geq 2$. Below $D(P_{f_\tau}||P_{f_{\tau'}})$ denotes the Kullback-Leibler divergence between P_{f_τ} and $P_{f_{\tau'}}$. Also $D(f_\tau||f_{\tau'})$ and $\chi^2(f_\tau, f_{\tau'})$ denote the Kullback-Leibler divergence and chi-squared divergence between the densities f_τ and $f_{\tau'}$ respectively. By Pinsker's inequality and the fact that $D(f_\tau||f_{\tau'}) \leq \chi^2(f_\tau, f_{\tau'})$, we obtain

$$\left\|P_{f_\tau} - P_{f_{\tau'}}\right\|_{TV} \leq \sqrt{\frac{1}{2}D(P_{f_\tau}||P_{f_{\tau'}})} = \sqrt{\frac{n}{2}D(f_\tau||f_{\tau'})} \leq \sqrt{\frac{n}{2}\chi^2(f_\tau||f_{\tau'})}.$$

Further

$$\begin{aligned}
\chi^2(f_\tau||f_{\tau'}) &= \int \frac{(f_\tau(x) - f_{\tau'}(x))^2}{f_{\tau'}(x)}dx \\
&= \frac{1}{k^2}\int \frac{(\phi_d(x - a_1(1-\tau_1) - b_1\tau_1) - \phi_d(x - a_1(1-\tau_1') - b_1\tau_1'))^2}{f_{\tau'}(x)}dx \\
&\leq \frac{1}{k}\int \frac{(\phi_d(x - a_1(1-\tau_1) - b_1\tau_1) - \phi_d(x - a_1(1-\tau_1') - b_1\tau_1'))^2}{\phi_d(x - a_1(1-\tau_1') - b_1\tau_1')}dx.
\end{aligned}$$

By a routine calculation, it now follows that

$$\begin{aligned}
\chi^2(f_\tau||f_{\tau'}) &\leq \frac{1}{k}\left\{\exp\left(\|a_1(1-\tau_1) + b_1\tau_1 - a_1(1-\tau_1') - b_1\tau_1'\|^2\right) - 1\right\} \\
&= \frac{1}{k}\left\{\exp\left(\|a_1 - b_1\|^2\right) - 1\right\} = \frac{1}{k}\left(e^{\delta^2} - 1\right).
\end{aligned}$$

We have therefore proved that

$$\left\|P_{f_\tau} - P_{f_{\tau'}}\right\|_{TV} \leq \sqrt{\frac{n}{2k}\left(e^{\delta^2} - 1\right)} \qquad \text{for every } \tau, \tau' \in \{0,1\}^k \text{ with } \Upsilon(\tau, \tau') = 1. \tag{4.51}$$

Combining (4.50) and (4.51), we obtain

$$\mathcal{R}(\mathcal{M}_k) \geq \frac{k}{4}\left(\frac{1}{k}\left(1 - e^{-\delta^2/8}\right) - \frac{(k-1)}{\Upsilon(\tau, \tau')}e^{-M^2/8}\right)\left(1 - \sqrt{\frac{n}{2k}\left(e^{\delta^2} - 1\right)}\right).$$

This inequality holds for every $\delta > 0$ and $M > 0$. So we can let M tend to ∞ to deduce

$$\mathcal{R}(\mathcal{M}_k) \geq \frac{1}{4}\left(1 - e^{-\delta^2/8}\right)\left(1 - \sqrt{\frac{n}{2k}\left(e^{\delta^2} - 1\right)}\right)$$

for every $\delta > 0$. The inequalities $1 - e^{-t} \geq t/2$ and $e^t - 1 \leq 2t$ for $0 \leq t \leq 1$ imply that

$$\mathcal{R}(\mathcal{M}_k) \geq \frac{\delta^2}{64}\left(1 - \sqrt{\frac{n}{k}}\delta\right) \qquad \text{for every } 0 \leq \delta \leq 1.$$

The choice $\delta = \sqrt{k/4n}$ now proves (4.20). □

4.6.5 Proof of Proposition 4.3.5

Proof of Proposition 4.3.5. Note that

$$h^*(x) = \sum_{j=1}^{k} w_j \phi_d(x; \mu_j, \Sigma_j) = \sum_{j=1}^{k} w_j \det(\Sigma_j^{-1/2}) \phi_d\left(\Sigma_j^{-1/2}(x - \mu_j)\right)$$

where $\phi_d(z) := (2\pi)^{-d/2} \exp\left(-\|z\|^2/2\right)$ denotes the standard d-dimensional normal density. It is then easy to see that $X_1, \ldots, X_n$ (where $X_i = Y_i/\sigma_{\min}$) are independent observations having the density f^* where

$$f^*(x) = \sigma_{\min}^d h^*(\sigma_{\min} x) = \sum_{j=1}^{k} w_j \left[\det\left(\sigma_{\min}^{-2}\Sigma_j\right)^{-1/2}\right] \phi_d\left(\left\{\sigma_{\min}^{-2}\Sigma_j\right\}^{-1}\left(x - \sigma_{\min}^{-1}\mu_j\right)\right).$$

This means that f^* is the density of the normal mixture:

$$\sum_{j=1}^{k} w_j N\left(\sigma_{\min}^{-1}\mu_j, \sigma_{\min}^{-2}\Sigma_j\right)$$

where $N(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector μ and covariance matrix Σ. It follows from here that f^* equals f_{G^*} (in the notation (4.1)) where G^* is the distribution of the normal mixture

$$\sum_{j=1}^{k} w_j N\left(\sigma_{\min}^{-1}\mu_j, \sigma_{\min}^{-2}\Sigma_j - I_d\right)$$

where I_d is the $d \times d$ identity matrix.

We can now use Corollary 4.3.2 to bound $\mathfrak{H}^2(\hat{f}_n, f^*)$ (note that $\hat{f}_n$ is an NPMLE based on $X_1, \ldots, X_n$). Specifically we shall use inequality (4.16) with

$$S := \left\{\sigma_{\min}^{-1}\mu_1, \ldots, \sigma_{\min}^{-1}\mu_k\right\}.$$

In order to verify (4.15), observe first that $\bar{G}_n$ in Corollary 4.3.2 is G^* since $X_1, \ldots, X_n$ are i.i.d f_{G^*} and that

$$\mathfrak{d}_S(\theta) = \min_{1 \le i \le k} \left\|\sigma_{\min}^{-1}\mu_i - \theta\right\|$$

As a result, for every $p \ge 1$ and $Z \sim N(0, I_d)$, we have

$$\mu_p(\mathfrak{d}_S) \le \left(\mathbb{E} \max_{1 \le j \le k} \left\|\left(\sigma_{\min}^{-2}\Sigma_j - I_d\right)^{1/2} Z\right\|^p\right)^{1/p} \le \sqrt{\frac{\sigma_{\max}^2}{\sigma_{\min}^2} - 1}\left(\mathbb{E}\|Z\|^p\right)^{1/p} \le C_d \tau \sqrt{p}.$$

Thus (4.15) holds with $K := C_d \max(1, \tau)$ and $\alpha = 2$ and inequality (4.16) then gives

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*) \le C_d \frac{\mathrm{Vol}(S^1)}{n} \left(\max(1, \tau)\right)^d \left(\sqrt{\log n}\right)^{d + (4-d)_+}.$$

As S is a finite set of cardinality k, we have $\text{Vol}(S^1) \leq kC_d$ so that

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*) \leq C_d \left(\frac{k}{n}\right) (\max(1,\tau))^d \left(\sqrt{\log n}\right)^{d+(4-d)_+}.$$

We now use the fact that the Hellinger distance is invariant under scale transformations which implies that $\mathfrak{H}(\hat{f}_n, f^*) = \mathfrak{H}(\hat{h}_n, h^*)$. This proves inequality (4.24). □

4.7 Proofs of Results in Section 4.4

4.7.1 Proof of Theorem 4.4.1

The proof of Theorem 4.4.1 is similar to Jiang and Zhang [63, Proof of Theorem 5]. It uses ingredients that are proved in Section 4.8, Section 4.9 and Section A. More precisely, crucial roles are played by the metric entropy results of Section 4.8 (specifically Corollary 4.8.2) and Theorem 4.9.1 in Section 4.9 which relates the denoising error to Hellinger distance (thereby allowing the application of Theorem 4.3.1). Additionally, Lemma A.0.2, Lemma A.0.4, Lemma A.0.6, Lemma A.0.9 and Lemma A.0.10 from Section A will also be used.

The notation described at the beginning of Section 4.6 will be followed in this section as well.

Proof of Theorem 4.4.1. The goal is to bound

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i - \hat{\theta}_i^*\|^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \left\| X_i + \frac{\nabla \hat{f}_n(X_i)}{\hat{f}_n(X_i)} - X_i - \frac{\nabla f_{\bar{G}_n}(X_i)}{f_{\bar{G}_n}(X_i)} \right\|^2\right)$$

It is convenient to introduce some notation here. Let $\mathbf{X}$ denote the $d \times n$ matrix whose columns are the observed data vectors $X_1, \ldots, X_n$. For a density $f \in \mathcal{M}$, let $T_f(\mathbf{X})$ denote the $d \times n$ matrix whose i^{th} column is given by the $d \times 1$ vector:

$$X_i + \frac{\nabla f(X_i)}{f(X_i)} \qquad \text{for } i = 1, \ldots, n.$$

With this notation, we can clearly rewrite $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ as

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) = \mathbb{E}\left(\frac{1}{n}\left\|T_{\hat{f}_n}(\mathbf{X}) - T_{f_{\bar{G}_n}}(\mathbf{X})\right\|_F^2\right)$$

where $\|\cdot\|_F$ denotes the usual Frobenius norm for matrices.

To bound the above, we first observe that since $\hat{f}_n$ is an NPMLE defined as in (4.2), it follows from the general maximum likelihood theorem (see, for example, Böhning [18, Theorem 2.1]) that

$$\frac{1}{n}\sum_{i=1}^n \frac{\phi_d(X_i - \theta)}{\hat{f}_n(X_i)} \leq 1 \tag{4.52}$$

for every $\theta \in \mathbb{R}^d$. Taking $\theta = X_i$ in the above inequality, we deduce that

$$1 \geq \frac{\phi_d(X_i - \theta)}{n\hat{f}_n(X_i)} = \frac{\phi_d(0)}{n\hat{f}_n(X_i)}$$

so that $\hat{f}_n(X_i) \geq \phi_d(0)/n = (2\pi)^{-d/2}n^{-1}$. Since this is true for each $i = 1, \ldots, n$, this means that

$$\min_{1 \leq i \leq n} \hat{f}_n(X_i) \geq \rho_n := \frac{(2\pi)^{-d/2}}{n}. \tag{4.53}$$

As a result, $\hat{f}_n(X_i) = \max(\hat{f}_n(X_i), \rho_n)$ for each i so that $T_{\hat{f}_n}(\mathbf{X}) = T_{\hat{f}_n}(\mathbf{X}, \rho_n)$ where for $f \in \mathcal{M}$ and $\rho > 0$, we define $T_f(\mathbf{X}, \rho)$ to be the $d \times n$ matrix whose i^{th} column is given by the $d \times 1$ vector:

$$X_i + \frac{\nabla f(X_i)}{\max(f(X_i), \rho)} \qquad \text{for } i = 1, \ldots, n.$$

This gives

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) = \mathbb{E}\left(\frac{1}{n}\left\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X})\right\|_F^2\right).$$

A difficulty in dealing with the expectation on the right hand side above comes from the fact that $\hat{f}_n$ is random. This is handled by covering the random $\hat{f}_n$ by an ϵ-net for a specific ϵ in the following way. First fix a compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$. Note that by Theorem 4.3.1 (specifically inequality (4.12) applied to $\alpha = \beta = 0.5$ and $t = 1$), we deduce that the following inequality holds with probability at least $1 - (2/n)$:

$$\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \leq \tilde{C}_d \epsilon_n(M, S). \tag{4.54}$$

Here $\tilde{C}_d$ is a positive constant depending on d alone and $\epsilon_n(M, S)$ is defined as in (4.10). Let E_n denote the event that (4.54) holds. We now obtain a covering of

$$\{f \in \mathcal{M} : \mathfrak{H}(f, f_{\bar{G}_n}) \leq \tilde{C}_d \epsilon_n(M, S)\} \tag{4.55}$$

under the pseudometric given by

$$\|f - g\|_{S^M, \nabla}^{\rho_n} := \sup_{x \in S^M} \left\| \frac{\nabla f(x)}{\max(f(x), \rho_n)} - \frac{\nabla g(x)}{\max(g(x), \rho_n)} \right\| \tag{4.56}$$

where $S^M := \{x \in \mathbb{R}^d : \mathfrak{d}_S(x) \leq M\}$. We have proved covering number bounds under this pseudometric in Corollary 4.8.2 which will be used in this proof. Let $f_{G_1}, \ldots, f_{G_N}$ denote a maximal subset of (4.55) such that for every $i \neq j$, we have

$$\left\|f_{G_i} - f_{G_j}\right\|_{S^M, \nabla}^{\rho_n} \geq 2\eta^* \tag{4.57}$$

where η^* is defined in terms of

$$\eta^* := \left(\frac{1}{\rho_n} + \sqrt{\frac{1}{\rho_n^2} \log \frac{1}{(2\pi)^d \rho_n^2}}\right) \eta \quad \text{and} \quad \eta := \frac{\rho_n}{n}. \tag{4.58}$$

By the usual relation between packing and covering numbers, the integer N is then bounded from above by $N(\eta^*, \mathcal{M}, \|\cdot\|_{S^M N,\nabla}^{\rho_n})$ which is bounded in Corollary 4.8.2. Specifically, Corollary 4.8.2 (applied to S^M) gives

$$\log N \leq C_d N(a, (S^M)^a)|\log \eta|^2 \leq C_d N(a, S^{M+a})(\log n)^2$$

where

$$a := \sqrt{2\log(2\sqrt{2\pi}n^2)}. \tag{4.59}$$

This further implies (via the use of inequality (A.26) in Lemma A.0.9 to bound $N(a, S^{M+a})$ as $N(a, S^{M+a}) \leq C_d a^{-d}\text{Vol}(S^{M+(3a/2)})$) that

$$\log N \leq C_d(\log n)^2 a^{-d}\text{Vol}(S^{M+(3a/2)}) \leq C_d(\log n)^{2-(d/2)}\text{Vol}(S^{M+(3a/2)}).$$

Using (A.27) in Lemma A.0.9 to bound $\text{Vol}(S^{M+(3a/2)})$ in terms of $\text{Vol}(S^1)$ (and the fact that $a \leq C\sqrt{10\log n} \leq CM$), we obtain

$$\log N \leq C_d \text{Vol}(S^1) M^d (\log n)^{2-(d/2)}. \tag{4.60}$$

Also because $f_{G_1}, \ldots, f_{G_N}$ is a maximal subset of (4.55) satisfying (4.57), we have

$$\max_{1\leq j\leq N} \mathfrak{H}(f_{G_j}, f_{\bar{G}_n}) \leq \tilde{C}_d \epsilon_n(M, S) \tag{4.61}$$

and, on the event E_n,

$$\min_{1\leq j\leq N} \left\| \hat{f}_n - f_{G_j} \right\|_{S^M,\nabla}^{\rho_n} \leq 2\eta^*. \tag{4.62}$$

We are now ready to bound the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$. The strategy is to break down the risk into various terms involving the densities $f_{G_1}, \ldots, f_{G_N}$.

Breakdown of the risk: The risk

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) = \mathbb{E}\left(\frac{1}{n}\left\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X})\right\|_F^2\right)$$

will be broken down via the inequality:

$$\begin{aligned}\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X})\|_F &\leq \|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F + \|T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X})\|_F \\ &\leq (\zeta_{1n} + \zeta_{2n} + \zeta_{3n} + \zeta_{4n}) + \zeta_{5n}\end{aligned} \tag{4.63}$$

where

$$\begin{aligned}
\zeta_{1n} &:= \|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F I(E_n^c) \\
\zeta_{2n} &:= \left(\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F - \max_{1\leq j\leq N} \|T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F\right)_+ I(E_n) \\
\zeta_{3n} &:= \max_{1\leq j\leq N}\left(\|T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F - \mathbb{E}\|T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F\right)_+ \\
\zeta_{4n} &:= \max_{1\leq j\leq N} \mathbb{E}\|T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F \\
\zeta_{5n} &:= \|T_{f_{\bar{G}_n}}(\mathbf{X}) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F
\end{aligned}$$

In conjunction with the elementary inequality $(a_1 + \cdots + a_5)^2 \leq 5(a_1^2 + \cdots + a_5^2)$, inequality (4.63) gives

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq 5 \sum_{i=1}^{5} \frac{\mathbb{E}\zeta_{in}^2}{n}.$$

The proof of Theorem 4.4.1 will be completed below by showing the existence of a positive constant C_d such that, for every $i = 1, \ldots, 5$,

$$\begin{aligned}\mathbb{E}\zeta_{in}^2 &\leq C_d n \epsilon_n^2(M, S) \left(\log n\right)^{\max(d,3)} \\ &= C_d \left(\mathrm{Vol}(S^1) M^d \left(\sqrt{\log n}\right)^{(4-d)_+} + n \left(\log n\right) \inf_{p \geq \frac{d+1}{2 \log n}} \left(\frac{2\mu_p(\mathfrak{d}_S)}{M} \right)^p \right) \left(\log n\right)^{\max(d,3)}.\end{aligned} \tag{4.64}$$

It may be noted that ζ_{4n} is non-random so that the expectation above can be removed for $i = 4$. Every other ζ_{in} is random.

Bounding $\mathbb{E}\zeta_{1n}^2$: We write

$$\begin{aligned}\mathbb{E}\zeta_{1n}^2 &= \mathbb{E}\left(\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\|_F^2 I(E_n^c) \right) \\ &= \sum_{i=1}^{n} \mathbb{E}\left(\left\| \frac{\nabla \hat{f}_n(X_i)}{\max(\hat{f}_n(X_i), \rho_n)} - \frac{\nabla f_{\bar{G}_n}(X_i)}{\max(f_{\bar{G}_n}(X_i), \rho_n)} \right\|^2 I(E_n^c) \right).\end{aligned}$$

Inequality (A.2) in Lemma A.0.2 now gives

$$\left\| \frac{\nabla \hat{f}_n(X_i)}{\max(\hat{f}_n(X_i), \rho_n)} - \frac{\nabla f_{\bar{G}_n}(X_i)}{\max(f_{\bar{G}_n}(X_i), \rho_n)} \right\|^2 \leq 4 \log \frac{(2\pi)^d}{\rho_n^2} \tag{4.65}$$

provided $\rho_n \leq (2\pi)^{-d/2} e^{-1/2}$ which is equivalent to $n \geq \sqrt{e}$ and hence holds for all $n \geq 2$. This gives (note that $\mathbb{P}(E_n^c) \leq 2/n$)

$$\mathbb{E}\zeta_{1n}^2 \leq 4n \left(\log \frac{(2\pi)^d}{\rho_n^2} \right) \mathbb{P}(E_n^c) \leq 8 \left(\log \frac{(2\pi)^d}{\rho_n^2} \right) \leq C_d \log n \leq C_d \mathrm{Vol}(S^1) M^d (\sqrt{\log n})^{(4-d)_+}$$

which proves (4.64) for $i = 1$.

Bounding $\mathbb{E}\zeta_{2n}^2$: For this, we write

$$\begin{aligned}\zeta_{2n}^2 &\leq \min_{1 \leq j \leq N} \left\| T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{G_j}}(\mathbf{X}, \rho_n) \right\|_F^2 I(E_n) \\ &= \min_{1 \leq j \leq N} \sum_{i=1}^{n} \left\| \frac{\nabla \hat{f}_n(X_i)}{\max(\hat{f}_n(X_i), \rho_n)} - \frac{\nabla f_{G_j}(X_i)}{\max(f_{G_j}(X_i), \rho_n)} \right\|^2 I(E_n) \\ &\leq \min_{1 \leq j \leq N} \left(\left\| \hat{f}_n - f_{G_j} \right\|_{S^M, \nabla}^{\rho_n} \right)^2 \left(\sum_{i=1}^{n} I\{X_i \in S^M\} \right) I(E_n) + \left(4 \log \frac{(2\pi)^d}{\rho_n^2} \right) \left(\sum_{i=1}^{n} I\{X_i \notin S^M\} \right) I(E_n).\end{aligned}$$

where we have used the notation (4.56) in the first term above and the inequality (4.65) in the second term. We can simplify the above bound as

$$\zeta_{2n}^2 \le n\left(\min_{1\le j\le N}\left\|\hat{f}_n - f_{G_j}\right\|_{S^M,\nabla}^{\rho_n}\right)^2 I(E_n) + \left(4\log\frac{(2\pi)^d}{\rho_n^2}\right)\left(\sum_{i=1}^n I\{X_i \notin S^M\}\right).$$

Inequality (4.62) and the expression (4.58) for η^* now give

$$\begin{aligned}\mathbb{E}\zeta_{2n}^2 &\le \frac{4}{n}\left(1+\sqrt{\log\frac{1}{(2\pi)^d\rho_n^2}}\right)^2 + \left(4\log\frac{(2\pi)^d}{\rho_n^2}\right)\left(\sum_{i=1}^n \mathbb{P}\{X_i\notin S^M\}\right)\\ &\le C_d\frac{\log n}{n} + C_d(\log n)\left(\sum_{i=1}^n \mathbb{P}\{X_i\notin S^M\}\right).\end{aligned}$$

To control the second term above, we use inequality (A.14) (which is a consequence of Lemma A.0.4). Note that $\mathbb{P}\{X_i \notin S^M\} \le \mathbb{P}\{\mathfrak{d}_S(X_i)\ge M\}$. Inequality (A.14) therefore gives

$$\mathbb{E}\zeta_{2n}^2 \le C_d\frac{\log n}{n} + C_d(\log n)M^{d-2} + C_d(n\log n)\inf_{p\ge\frac{d+1}{2\log n}}\left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p.$$

This proves (4.64) for $i=2$ (note that $(\log n)M^{d-2}\le M^d$ as $M\ge\sqrt{10\log n}$).

Bounding ζ_{3n}^2: Here Lemma A.0.6 and the bound (4.60) will be crucially used. Let us first write $\zeta_{3n} := \max_{1\le j\le N}\zeta_{3n.j}$ where

$$\zeta_{3n.j} := \left(\|T_{f_{G_j}}(\mathbf{X},\rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X},\rho_n)\|_F - \mathbb{E}\|T_{f_{G_j}}(\mathbf{X},\rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X},\rho_n)\|_F\right)_+.$$

Lemma A.0.6 then gives

$$\mathbb{P}\left\{\zeta_{3n.j}\ge x\right\} \le \exp\left(\frac{-x^2}{8L^4(\rho_n)}\right) \qquad \text{for every } 1\le j\le N \text{ and } x>0.$$

where

$$L(\rho_n) = \sqrt{\log\frac{1}{(2\pi)^d\rho_n^2}} = \sqrt{\log n}. \tag{4.66}$$

By the union bound, we have

$$\mathbb{P}\left\{\zeta_{3n}\ge x\right\} \le N\exp\left(\frac{-x^2}{8L^4(\rho_n)}\right) \qquad \text{for every } x>0$$

so that, for every $x_0>0$,

$$\begin{aligned}\mathbb{E}\zeta_{3n}^2 &\le \int_0^\infty \mathbb{P}\left\{\zeta_{3n}\ge\sqrt{x}\right\}dx\\ &\le x_0 + \int_{x_0}^\infty N\exp\left(\frac{-x}{8L^4(\rho_n)}\right)dx = x_0 + 8NL^4(\rho_n)\exp\left(\frac{-x_0}{8L^4(\rho_n)}\right).\end{aligned}$$

Minimizing the above bound over $x_0 > 0$, we deduce that

$$\mathbb{E}\zeta_{3n}^2 \leq 8L^4(\rho_n)\log(eN).$$

The bound (4.60) (along with (4.66)) then gives

$$\mathbb{E}\zeta_{3n}^2 \leq C_d \mathrm{Vol}(S^1) M^d (\sqrt{\log n})^{8-d} \leq C_d \mathrm{Vol}(S^1) M^d (\sqrt{\log n})^{(4-d)_+} (\log n)^3$$

which proves (4.64) for $i = 3$.

Bounding ζ_{4n}^2: To bound the non-random quantity ζ_{4n}^2, we only need to bound

$$\Gamma_j^2 := \mathbb{E}\left\| T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n) \right\|_F^2$$

for each $1 \leq j \leq N$. We can clearly write

$$\begin{aligned}
\Gamma_j^2 &= \sum_{i=1}^n \mathbb{E}\left\| \frac{\nabla f_{G_j}(X_i)}{\max(f_{G_j}(X_i), \rho_n)} - \frac{\nabla f_{\bar{G}_n}(X_i)}{\max(f_{\bar{G}_n}(X_i), \rho_n)} \right\|^2 \\
&= n \int \left\| \frac{\nabla f_{G_j}(x)}{\max(f_{G_j}(x), \rho_n)} - \frac{\nabla f_{\bar{G}_n}(x)}{\max(f_{\bar{G}_n}(x), \rho_n)} \right\|^2 f_{\bar{G}_n}(x) dx.
\end{aligned}$$

The above term can be bounded by a direct application of Theorem 4.9.1 which furnishes a bound in terms of $\mathfrak{H}(f_{G_j}, f_{\bar{G}_n})$. Indeed, because $n \geq 2$, we have $\rho_n \leq (2\pi)^{-d/2} e^{-1/2}$ so that Theorem 4.9.1 applies (with $G = G_j$ and $G_0 = \bar{G}_n$) and we obtain

$$\begin{aligned}
\frac{1}{n}\Gamma_j^2 &\leq C_d \max\left\{ \left(\log \frac{(2\pi)^{-d/2}}{\rho_n} \right)^3, |\log \mathfrak{H}(f_{G_j}, f_{\bar{G}_n})| \right\} \mathfrak{H}^2(f_{G_j}, f_{\bar{G}_n}) \\
&= C_d \max\left\{ (\log n)^3, |\log \mathfrak{H}(f_{G_j}, f_{\bar{G}_n})| \right\} \mathfrak{H}^2(f_{G_j}, f_{\bar{G}_n}).
\end{aligned}$$

We now use the fact that $\mathfrak{H}(f_{G_j}, f_{\bar{G}_n})$ is bounded from above by $\tilde{C}_d \epsilon_n(M, S)$ (see (4.61)). We can then work with two cases. If $\tilde{C}_d \epsilon_n(M, S) \leq e^{-1/2}$, then using the fact that $h \mapsto h^2 |\log h|$ is increasing on $(0, e^{-1/2}]$, we have

$$\frac{1}{n}\Gamma_j^2 \leq C_d \tilde{C}_d^2 \max\left\{ (\log n)^3, \left| \log(\tilde{C}_d \epsilon_n(M, S)) \right| \right\} \epsilon_n^2(M, S).$$

The trivial observation $\epsilon_n(M, S) \geq K_d/n$ for a constant K_d now gives

$$\Gamma_j^2 \leq n C_d (\log n)^3 \epsilon_n^2(M, S). \tag{4.67}$$

On the other hand when $\tilde{C}_d \epsilon_n(M, S) > e^{-1/2}$, then we can simply bound $|\log \mathfrak{H}(f_{G_j}, f_{\bar{G}_n})| \mathfrak{H}^2(f_{G_j}, f_{\bar{G}_n})$ by a constant (the function $h \mapsto h^2 |\log h|$ is bounded on $h \in (0, 2]$) so that the inequality (4.67) still holds. The bound in the right hand side of (4.67) does not depend on j so that it is an upper bound for ζ_{4n}^2 as well. This proves (4.64) for $i = 4$.

Bounding $\mathbb{E}\zeta_{5n}^2$: We write

$$\begin{aligned}
\mathbb{E}\zeta_{5n}^2 &= \mathbb{E}\left\|T_{f_{\bar{G}_n}}(\mathbf{X}) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho_n)\right\|_F^2 \\
&= \sum_{i=1}^n \mathbb{E}\left\|\frac{\nabla f_{\bar{G}_n}(X_i)}{f_{\bar{G}_n}(X_i)} - \frac{\nabla f_{\bar{G}_n}(X_i)}{\max(f_{\bar{G}_n}(X_i), \rho_n)}\right\|^2 \\
&= n\int \left\|\frac{\nabla f_{\bar{G}_n}(x)}{f_{\bar{G}_n}(x)} - \frac{\nabla f_{\bar{G}_n}(x)}{\max(f_{\bar{G}_n}(x), \rho_n)}\right\|^2 f_{\bar{G}_n}(x)dx \\
&= n\int \left(1 - \frac{f_{\bar{G}_n}}{\max\left(f_{\bar{G}_n}, \rho\right)}\right)^2 \frac{\|\nabla f_{\bar{G}_n}\|^2}{f_{\bar{G}_n}} = n\Delta(\bar{G}_n, \rho_n)
\end{aligned}$$

where we define

$$\Delta(G, \rho) := \int \left(1 - \frac{f_G}{\max\left(f_G, \rho\right)}\right)^2 \frac{\|\nabla f_G\|^2}{f_G}$$

for probability measures G on $\mathbb{R}^d$ and $\rho > 0$. We now use Lemma A.0.10 to bound $\Delta(\bar{G}_n, \rho_n)$. Specifically, inequality (A.29) in Lemma A.0.10 applied to the compact set S^M gives

$$\Delta(\bar{G}_n, \rho_n) \leq C_d N\left(\frac{4}{L(\rho_n)}, S^M\right) L^d(\rho_n)\rho_n + d\ \bar{G}_n((S^M)^c). \tag{4.68}$$

The first term above is bounded using Lemma A.0.9 as follows (note that $\rho_n = (2\pi)^{-d/2}/n$ and $L(\rho_n) = \sqrt{\log n}$ as shown in (4.66)):

$$\begin{aligned}
N\left(\frac{4}{L(\rho_n)}, S^M\right) L^d(\rho_n)\rho_n &= N\left(\frac{4}{\sqrt{\log n}}, S^M\right)(\log n)^{d/2}\frac{(2\pi)^{-d/2}}{n} \\
&\leq C_d(4/\sqrt{\log n})^{-d}\mathrm{Vol}((S^M)^{2/\sqrt{\log n}})\frac{(\log n)^{d/2}}{n} \quad \text{(using inequality (A.26))} \\
&\leq \frac{C_d}{n}(\log n)^d \mathrm{Vol}(S^{M+2/\sqrt{\log n}}) \\
&\leq \frac{C_d}{n}(\log n)^d \mathrm{Vol}(S^1)\left(1 + \frac{M}{4} + \frac{1}{2\sqrt{\log n}}\right)^d \quad \text{(using inequality (A.27))} \\
&\leq \frac{C_d}{n}(\log n)^d M^d \mathrm{Vol}(S^1).
\end{aligned}$$

For the second term in (4.68), note that

$$\bar{G}_n((S^M)^c) \leq \int I\{\mathfrak{d}_S(\theta) \geq M\} d\bar{G}_n(\theta) \leq \inf_{p \geq \frac{d+1}{2\log n}} \left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p.$$

We have therefore proved that

$$\mathbb{E}\zeta_{5n}^2 \leq n\Delta(\bar{G}_n, \rho_n) \leq C_d\left\{(\log n)^d M^d \mathrm{Vol}(S^1) + n \inf_{p \geq \frac{d+1}{2\log n}} \left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^p\right\}$$

which evidently implies (4.64). The proof of Theorem 4.4.1 is now complete. □

4.7.2 Proof of Corollary 4.4.3

The idea is to choose M and S appropriately under each of the assumptions on $\bar{G}_n$ and then to appropriately bound $\epsilon_n(M, S)$. The necessary work for this is already done in Corollary 4.3.2 from which Corollary 4.4.3 immediately follows.

4.7.3 Proof of Proposition 4.4.4

The assumption (4.28) implies that the empirical measure $\bar{G}_n$ of $\theta_1, \ldots, \theta_n$ is supported on

$$S := \cup_{j=1}^{k} B(a_j, R) \qquad \text{where } B(a_j, R) := \big\{x \in \mathbb{R}^d : \|x - a_j\| \leq R\big\}.$$

We can therefore apply inequality (4.25) in Corollary 4.4.3 to bound $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$. The conclusion (4.29) then immediately follows from (4.25) because

$$\text{Vol}(S^1) \leq \sum_{j=1}^{k} \text{Vol}(B(a_j, 1 + R)) \leq C_d k (1 + R)^d.$$

4.7.4 Proof of Lemma 4.4.5

The proof of Lemma 4.4.5 uses Assouad's lemma (see, for example, Tsybakov [135, Chapter 2] as well as Lemma A.0.12 (stated and proved in Section A).

Proof of Lemma 4.4.5. Fix k and n with $1 \leq k \leq n$. Also fix $\delta > 0$ and $M \geq 2$. Let $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ be points in $\mathbb{R}^d$ such that

$$\min\left(\min_{i \neq j} \|a_i - a_j\|, \min_{i \neq j} \|b_i - b_j\|, \min_{i \neq j} \|a_i - b_j\|\right) \geq M \tag{4.69}$$

and such that

$$\|a_i - b_i\| = \delta \qquad \text{for every } 1 \leq i \leq k. \tag{4.70}$$

We now define a partition $S_1, \ldots, S_k, S_{k+1}$ of $\{1, \ldots, n\}$ via

$$S_i := \{(i-1)m + 1, \ldots, im\} \qquad \text{for } i = 1, \ldots, k$$

and $S_{k+1} := \{km + 1, \ldots, n\}$ where $m := [n/k]$ (for $x > 0$, we define $[x]$ as usual to be the largest integer that is smaller than or equal to x). Note that the cardinality of S_j equals m for $i = 1, \ldots, k$ and that S_{k+1} will be empty if n is a multiple of k.

Now for every $\tau \in \{0, 1\}^k$, we define n vectors $\theta_1(\tau), \ldots, \theta_n(\tau)$ in $\mathbb{R}^d$ via

$$\theta_i(\tau) := (1 - \tau_j)a_j + \tau_j b_j \qquad \text{provided } i \in S_j \text{ for some } 1 \leq j \leq k$$

and for $i \in S_{k+1}$, we take $\theta_i(\tau) := a_1$.

Let $\Theta(\tau)$ denote the collection of all n-tuples $(\theta_1(\tau), \ldots, \theta_n(\tau))$ as τ ranges over $\{0,1\}^k$. It is easy to see that $\Theta(\tau) \subseteq \Theta_{n,d,k}$ so that

$$\mathcal{R}^*(\Theta_{n,d,k}) \geq \mathcal{R}^*(\Theta(\tau)) := \inf_{\tilde{\theta}_1,\ldots,\tilde{\theta}_n} \sup_{(\theta_1,\ldots,\theta_n)\in\Theta(\tau)} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\tilde{\theta}_i - \hat{\theta}_i^*\right\|^2\right].$$

The elementary inequality $\|a-b\|^2 \geq \|a\|^2/2 - \|b\|^2$ for vectors $a, b \in \mathbb{R}^d$ gives

$$\frac{1}{n}\sum_{i=1}^n \left\|\tilde{\theta}_i - \hat{\theta}_i^*\right\|^2 \geq \frac{1}{2n}\sum_{i=1}^n \left\|\tilde{\theta}_i - \theta_i\right\|^2 - \frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2$$

for every $\theta_1, \ldots, \theta_n$ and estimators $\tilde{\theta}_1 \ldots, \tilde{\theta}_n$. As a result, we deduce that

$$\mathcal{R}^*(\Theta(\tau)) \geq \breve{\mathcal{R}}(\Theta(\tau)) - \sup_{(\theta_1,\ldots,\theta_n)\in\Theta(\tau)} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2\right] \tag{4.71}$$

where

$$\breve{\mathcal{R}}(\Theta(\tau)) := \inf_{\tilde{\theta}_1,\ldots,\tilde{\theta}_n} \sup_{(\theta_1,\ldots,\theta_n)\in\Theta(\tau)} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\tilde{\theta}_i - \theta_i\right\|^2\right].$$

We first bound $\breve{\mathcal{R}}(\Theta(\tau))$ from below via Assouad's lemma. For $\tau, \tau' \in \{0,1\}^k$, let

$$\mathfrak{L}(\tau,\tau') := \frac{1}{n}\sum_{i=1}^n \|\theta_i(\tau) - \theta_i(\tau')\|^2.$$

Also let P_τ denote the joint distribution of the independent random variables $X_1, \ldots, X_n$ with $X_i \sim N(\theta_i(\tau), I_d)$ for $i = 1, \ldots, n$. Assouad's lemma then gives

$$\breve{\mathcal{R}}(\Theta(\tau)) \geq \frac{k}{8}\min_{\tau\neq\tau'} \frac{\mathfrak{L}(\tau,\tau')}{\Upsilon(\tau,\tau')} \min_{\Upsilon(\tau,\tau')=1} (1 - \|P_\tau - P_{\tau'}\|_{TV}) \tag{4.72}$$

where $\Upsilon(\tau,\tau') := \sum_{j=1}^k I\{\tau_j \neq \tau_j'\}$ is the Hamming distance and $\|P_\tau - P_{\tau'}\|_{TV}$ denotes the variation distance between P_τ and $P_{\tau'}$. We now bound the terms appearing in the right hand side of (4.72). For $\tau, \tau' \in \{0,1\}^k$, observe that

$$\mathfrak{L}(\tau,\tau') = \frac{1}{n}\sum_{j=1}^k \sum_{i:i\in S_j} \|a_j - b_j\|^2 I\{\tau_j \neq \tau_j'\} = \frac{1}{n}\sum_{j=1}^k |S_j| \|a_j - b_j\|^2 I\{\tau_j \neq \tau_j'\} = \frac{m\delta^2}{n}\Upsilon(\tau,\tau') \tag{4.73}$$

where $|S_j|$ denotes the cardinality of S_j. We have used above the fact that $|S_j| = m$ for $1 \leq j \leq k$ and (4.70).

To bound the last term in (4.72), we use Pinsker's inequality (below D stands for Kullback-Leibler divergence) to obtain

$$\|P_\tau - P_{\tau'}\|_{TV} \leq \sqrt{\frac{1}{2}D(P_\tau||P_{\tau'})} = \frac{1}{2}\sqrt{\sum_{i=1}^n \|\theta_i(\tau) - \theta_i(\tau')\|^2} = \frac{1}{2}\sqrt{n\mathfrak{L}(\tau,\tau')}.$$

Thus, from (4.73), we deduce that for $\Upsilon(\tau,\tau') = 1$,

$$\|P_\tau - P_{\tau'}\|_{TV} \leq \frac{1}{2}\sqrt{m\delta^2}.$$

Inequality (4.72) thus gives

$$\breve{\mathcal{R}}(\Theta(\tau)) \geq \frac{km\delta^2}{8n}\left(1 - \frac{\sqrt{m\delta^2}}{2}\right). \tag{4.74}$$

To bound the second term in (4.71), we use Lemma A.0.12 which gives that for every $\theta_1, \ldots, \theta_n \in \Theta(\tau)$, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2\right] \leq \frac{k}{2\sqrt{2\pi}} \sum_{j,l:j\neq l} (p_j + p_l)\|c_j - c_l\| \exp\left(-\frac{1}{8}\|c_j - c_l\|^2\right)$$

where $c_1, \ldots, c_{k+1}$ denote the distinct elements from $\theta_1, \ldots, \theta_n$ and $p_j, j = 1, \ldots, k+1$ are nonnegative real numbers summing to one. Now each c_j equals either a_j or b_j and hence, by (4.69), we have $\|c_j - c_l\| \geq M$ for every $j \neq l$. As $x \mapsto xe^{-x^2/8}$ is decreasing for $x > 2$ and $M > 2$, we deduce that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2\right] \leq \frac{k}{2\sqrt{2\pi}} Me^{-M^2/8} \sum_{j,l:j\neq l} (p_j + p_l) \leq \frac{k}{\sqrt{2\pi}} Me^{-M^2/8}. \tag{4.75}$$

We obtain therefore from (4.71), (4.74) and (4.75), that

$$\mathcal{R}^*(\Theta_{n,d,k}) \geq \frac{km\delta^2}{8n}\left(1 - \frac{\sqrt{m\delta^2}}{2}\right) - \frac{k}{\sqrt{2\pi}} Me^{-M^2/8}.$$

The left hand side above does not depend on M so we can let $M \to \infty$ to obtain

$$\mathcal{R}^*(\Theta_{n,d,k}) \geq \frac{km\delta^2}{8n}\left(1 - \frac{\sqrt{m\delta^2}}{2}\right).$$

We now make the choice $\delta := 1/\sqrt{m}$ to obtain $\mathcal{R}^*(\Theta_{n,d,k}) \geq k/(16n)$ which proves Lemma 4.4.5. □

4.8 Main Metric Entropy Results and Proofs

For a compact set $S \subseteq \mathbb{R}^d$, let $\|\cdot\|_S$ and $\|\cdot\|_{S,\nabla}$ denote two pseudonorms given by

$$\|f\|_S := \sup_{x \in S} |f(x)| \quad \text{and} \quad \|f\|_{S,\nabla} := \sup_{x \in S} \|\nabla f(x)\|$$

for densities $f \in \mathcal{M}$. These naturally lead to two pseudometrics on $\mathcal{M}$ and we shall denote the η-covering numbers of $\mathcal{M}$ under these pseudometrics by $N(\eta, \mathcal{M}, \|\cdot\|_S)$ and $N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla})$ respectively. The notion of covering numbers is defined at the beginning of Section 4.6. The following theorem gives upper bounds for $N(\eta, \mathcal{M}, \|\cdot\|_S)$ and $N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla})$. Recall the notation introduced at the beginning of Section 4.6.

Theorem 4.8.1. *There exists a positive constant C_d depending on d alone such that for every compact set $S \subseteq \mathbb{R}^d$ and $0 < \eta \leq \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\sqrt{e}}$, we have*

$$\log N(\eta, \mathcal{M}, \|\cdot\|_S) \leq C_d N(a, S^a) |\log \eta|^2 \tag{4.76}$$

and

$$\log N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla}) \leq C_d N(a, S^a) |\log \eta|^2 \tag{4.77}$$

where a is defined as

$$a := \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\eta}}. \tag{4.78}$$

Theorem 4.8.1 immediately implies a covering number result for $\mathcal{M}$ in terms of another pseudometric that is defined in terms of both $f(x)$ and $\nabla f(x)$. This is given in the next corollary which was used in the proof of Theorem 4.4.1.

Corollary 4.8.2. *For a compact set $S \in \mathbb{R}^d$ and $\rho > 0$, define the pseudometric:*

$$\|f - g\|_{S,\nabla}^{\rho} := \sup_{x \in S} \left\| \frac{\nabla f(x)}{\max(f(x), \rho)} - \frac{\nabla g(x)}{\max(g(x), \rho)} \right\| \tag{4.79}$$

for functions $f : \mathbb{R}^d \to \mathbb{R}$ which are bounded on S and whose derivatives are bounded on S. Let the ϵ-covering number of $\mathcal{M}$ in the pseudometric given by (4.79) be denoted by $N(\epsilon, \mathcal{M}, \|\cdot\|_{S,\nabla}^{\rho})$. Then there exists a positive constant C_d depending on d alone such that for every $\rho > 0$, $0 < \eta \leq \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\sqrt{e}}$ and compact subset $S \subseteq \mathbb{R}^d$, we have

$$\log N(\eta^*, \mathcal{M}, \|\cdot\|_{S,\nabla}^{\rho}) \leq C_d N(a, S^a) |\log \eta|^2 \tag{4.80}$$

where a is defined as in (4.78) and

$$\eta^* := \left(\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} \log \frac{1}{(2\pi)^d \rho^2}} \right) \eta. \tag{4.81}$$

Remark 4.8.3. *When $d = 1$ and $S = [-M, M]$, we have*

$$N(a, S^a) \leq C \max\left\{\frac{M}{\sqrt{|\log \eta|}}, 1\right\}$$

so that inequalities (4.76)*,* (4.77) *and* (4.80) *become*

$$\log N(\eta, \mathcal{M}, \|\cdot\|_{[-M,M]}) \leq C|\log \eta|^2 \max\left\{\frac{M}{\sqrt{|\log \eta|}}, 1\right\}, \tag{4.82}$$

$$\log N(\eta, \mathcal{M}, \|\cdot\|_{[-M,M],\nabla}) \leq C|\log \eta|^2 \max\left\{\frac{M}{\sqrt{|\log \eta|}}, 1\right\}, \tag{4.83}$$

and

$$\log N(\eta^*, \mathcal{M}, \|\cdot\|^{\rho}_{[-M,M],\nabla}) \leq C|\log \eta|^2 \max\left\{\frac{M}{\sqrt{|\log \eta|}}, 1\right\} \tag{4.84}$$

respectively. Inequality (4.82) *has previously appeared in Zhang [147, Lemma 2] (improving an earlier result of Ghosal and Vaart [48]). Inequality* (4.83) *does not seem to have been stated explicitly previously but is implicit in Jiang and Zhang [63, Proof of Proposition* 3*]. Inequality* (4.84) *has previously appeared as Jiang and Zhang [63, Proposition* 3*]. Our contribution therefore lies in generalizing these results to multiple dimensions and further in allowing S to take the form of any compact subset of $\mathbb{R}^d$.*

The rest of this section is devoted to the proofs of Theorem 4.8.1 and Corollary 4.8.2.

4.8.1 Proof of Theorem 4.8.1

Moment Matching Lemma

Recall that for $x \in \mathbb{R}^d$ and $a > 0$, we denote the closed Euclidean ball of radius a centered at x by $B(x, a)$. We also let

$$\mathring{B}(x, a) := \{u \in \mathbb{R}^d : \|u - x\| < a\}$$

denote the open ball of radius a centered at x.

Lemma 4.8.4. *Let G and G' be two arbitrary probability measures on $\mathbb{R}^d$. Fix $a \geq 1$ and $x \in \mathbb{R}^d$. Let A be a subset of $\mathbb{R}^d$ such that*

$$\mathring{B}(x, a) \subseteq A \subseteq B(x, ca)$$

for some $c \geq 1$. Suppose that, for some $m \geq 1$, we have

$$\int_A \theta_j^k dG(\theta) = \int_A \theta_j^k dG'(\theta) \qquad \text{for every } 1 \leq j \leq d \text{ and } 0 \leq k \leq 2m + 1. \tag{4.85}$$

Then

$$|f_G(x) - f_{G'}(x)| \leq \frac{1}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} + \frac{e^{-a^2/2}}{(2\pi)^{d/2}}. \tag{4.86}$$

and

$$\|\nabla f_G(x) - \nabla f_{G'}(x)\| \leq \frac{ca}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} + \frac{a e^{-a^2/2}}{(2\pi)^{d/2}} \tag{4.87}$$

Proof of Lemma 4.8.4. First write

$$f_G(x) - f_{G'}(x) = \int \phi_d(x-\theta) \left(G(d\theta) - G'(d\theta) \right)$$

and

$$\nabla f_G(x) - \nabla f_{G'}(x) = \int (\theta - x) \phi_d(x-\theta) \left(G(d\theta) - G'(d\theta) \right).$$

We split each integral above into two terms by restricting their range first over A and then over A^c, the complement set of A:

$$f_G(x) - f_{G'}(x) = \int_A \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) + \int_{A^c} \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \tag{4.88}$$

$$\nabla f_G(x) - \nabla f_{G'}(x) = \int_A (\theta - x) \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) + \int_{A^c} (\theta - x) \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \tag{4.89}$$

Because $A \supseteq \mathring{B}(x, a)$, it is clear that

$$\sup_{\theta \in A^c} \phi_d(x-\theta) \leq \sup_{\theta : \|x-\theta\| \geq a} \phi_d(x-\theta) \leq (2\pi)^{-d/2} \exp(-a^2/2)$$

$$\sup_{\theta \in A^c} \|\theta - x\| \, \phi_d(x-\theta) \leq \sup_{\theta : \|x-\theta\| \geq a} \|x-\theta\| \, \phi_d(x-\theta) \leq (2\pi)^{-d/2} \sup_{u \geq a} u e^{-u^2/2} = (2\pi)^{-d/2} a e^{-a^2/2}$$

because $a \geq 1$. Therefore the second terms on the right hand side on (4.88) and (4.89) are respectively bounded in absolute value by the final terms in (4.86) and (4.87). It only remains to prove the following pair of inequalities

$$\left| \int_A \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| \leq \frac{1}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} \tag{4.90}$$

$$\left| \int_A (\theta - x) \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| \leq \frac{ca}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} \tag{4.91}$$

For this, we use Taylor expansion and the moment matching condition (4.85). Taylor's formula for e^u is

$$e^u = \sum_{i=0}^{m} \frac{u^i}{i!} + \frac{u^{m+1}}{(m+1)!} e^v$$

for every u where v is some real number lying between 0 and u. Using this for $u = -t^2/2$, we obtain

$$\exp(-t^2/2) = \sum_{i=0}^{m} \frac{(-t^2/2)^i}{i!} + (-1)^{m+1} \frac{(t^2/2)^{m+1}}{(m+1)!} e^v$$

where v lies between 0 and $-t^2/2$. Because $e^v \leq 1$, this gives

$$\left| \exp(-t^2/2) - \sum_{i=0}^{m} \frac{(-t^2/2)^i}{i!} \right| \leq \frac{(t^2/2)^{m+1}}{(m+1)!}.$$

We can therefore write $\phi_d(z) = P_d(z) + R_d(z)$ for every $z \in \mathbb{R}^d$ where $P_d(z)$ is a polynomial of degree $2m$ in z and $R_d(z)$ is a remainder term which satisfies

$$|R_d(z)| \leq \frac{(\|z\|^2/2)^{m+1}}{(2\pi)^{d/2}(m+1)!}.$$

Using this for $z = x - \theta$, we can write

$$\left| \int_A \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| \leq \left| \int_A P_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| + \left| \int_A R_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right|$$

and similarly,

$$\begin{aligned} \left| \int_A (\theta - x) \phi_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| &\leq \left| \int_A (\theta - x) P_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| \\ &\quad + \left| \int_A (\theta - x) R_d(x-\theta) \left(dG(\theta) - dG'(\theta) \right) \right| \end{aligned}$$

The first terms in the above two equations are zero because of condition (4.85) and the fact that $P_d(x-\theta)$ is a polynomial in θ with degree $2m$ (implying that for every j, $(\theta_j - x_j) P_d(x-\theta)$ is a polynomial of degree $2m+1$). Because $A \subseteq B(x, ca)$, we have $\|x - \theta\| \leq ca$ for every $\theta \in A$ so that

$$|R_d(x-\theta)| \leq \frac{(2\pi)^{-d/2}}{(m+1)!} \left(\frac{\|x-\theta\|^2}{2} \right)^{m+1} \leq \frac{(2\pi)^{-d/2}}{(m+1)!} \left(\frac{c^2 a^2}{2} \right)^{m+1}.$$

Stirling's formula $n! \geq \sqrt{2\pi n}(n/e)^n \geq \sqrt{2\pi}(n/e)^n$ applied to $n = m+1$ yields

$$|R_d(x-\theta)| \leq \frac{1}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} \qquad \text{for every } \theta \in A$$

and

$$\|\theta - x\| \, |R_d(x-\theta)| \leq \frac{ca}{(2\pi)^{(d+1)/2}} \left(\frac{c^2 a^2 e}{2(m+1)} \right)^{m+1} \qquad \text{for every } \theta \in A$$

which completes the proof. □

Approximation by mixtures with discrete mixing measures

Given any distribution f_G, what is a bound on ℓ such that we can approximate f_G by another gaussian mixture $f_{G'}$ where G' is a discrete measure with at most ℓ atoms. The following lemma addresses this question where approximation is in terms of the pseudometrics $\sup_{x\in S}|f_G(x)-f_{G'}(x)|$ as well as $\sup_{x\in S}\|\nabla f_G(x)-\nabla f_{G'}(x)\|$.

Recall that for a subset S of $\mathbb{R}^d$, we write $N(\eta, S)$ to mean its η covering number (defined as the smallest number of closed balls of radius η whose union contains S).

Lemma 4.8.5. *Let G be an arbitrary probability measure on $\mathbb{R}^d$ and let S denote an arbitrary compact subset of $\mathbb{R}^d$. Also let $a \geq 1$. Then there exists a discrete probability measure G' that is supported on $S^a := \cup_{x\in S}B(x,a)$ and having at most*

$$\ell := d\left(2\lfloor(13.5)a^2\rfloor + 2\right)N(a, S^a) + 1 \tag{4.92}$$

atoms such that

$$\sup_{x\in S}|f_G(x)-f_{G'}(x)| \leq \left(1+\frac{1}{\sqrt{2\pi}}\right)(2\pi)^{-d/2}e^{-a^2/2} \tag{4.93}$$

and

$$\sup_{x\in S}\|\nabla f_G(x)-\nabla f_{G'}(x)\| \leq \left(a+\frac{3a}{\sqrt{2\pi}}\right)(2\pi)^{-d/2}e^{-a^2/2}. \tag{4.94}$$

Proof of Lemma 4.8.5. Let $\mathring{S}^a := \cup_{x\in S}\mathring{B}(x,a)$ (here $\mathring{B}(x,a)$ denotes the open ball of radius a centered at x) and let $L := N(a, \mathring{S}^a)$ denote the a-covering number of $\mathring{S}^a$. Note that $L \leq N(a, S^a)$. Let $B_1, \ldots, B_L$ denote closed balls of radius a whose union contains $\mathring{S}^a$. Let $E_1, \ldots, E_L$ denote the standard disjointification of the sets $B_1, \ldots, B_L$ i.e., $E_1 := B_1$ and $E_i := B_i \setminus (\cup_{j<i}B_j)$ for $i = 2, \ldots, L$. We can also ensure that $\cup_{i=1}^L E_i = \mathring{S}^a$ by removing the set $\mathring{S}^a \setminus \cup_i E_i$ from each set E_i.

Let $m := \lfloor(13.5)a^2\rfloor$, suppose that a probability measure G' is chosen so that G and G' have the same moments up to order $2m+1$ on each set E_i for $i = 1, \ldots, L$ i.e.,

$$\int_{E_i}\theta_j^k dG(\theta) = \int_{E_i}\theta_j^k dG'(\theta) \qquad \text{for } 1 \leq j \leq d, 0 \leq k \leq 2m+1 \text{ and } 1 \leq i \leq L. \tag{4.95}$$

We shall then prove below that inequalities (4.93) and (4.94) are satisfied. Fix $x \in S$. Because $\mathring{B}(x,a)$ is contained in $\mathring{S}^a$, the sets $E_1, \ldots, E_L$ cover $\mathring{B}(x,a)$ i.e.,

$$\mathring{B}(x,a) \subseteq \cup_{i\in F}E_i$$

where $F := \left\{1 \leq i \leq L : E_i \cap \mathring{B}(x,a) \neq \emptyset\right\}$. Also because the diameter of $E_i \subseteq B_i$ is at most $2a$, we deduce that

$$\mathring{B}(x,a) \subseteq \cup_{i\in F}E_i \subseteq B(x,3a).$$

We now use Lemma 4.8.4 with $A = \cup_{i \in F} E_i$ and $c = 3$ to deduce that

$$|f_G(x) - f_{G'}(x)| \leq \frac{1}{\sqrt{2\pi}} \frac{1}{(2\pi)^{d/2}} \left(\frac{9a^2 e}{2(m+1)} \right)^{m+1} + \frac{e^{-a^2/2}}{(2\pi)^{d/2}}$$
$$\|\nabla f_G(x) - \nabla f_{G'}(x)\| \leq \frac{3a}{\sqrt{2\pi}(2\pi)^{d/2}} \left(\frac{9a^2 e}{2(m+1)} \right)^{m+1} + \frac{ae^{-a^2/2}}{(2\pi)^{d/2}}$$

Because $m := \lfloor 13.5a^2 \rfloor$, we have $m + 1 \geq 13.5a^2$ and consequently,

$$\left(\frac{9a^2 e}{2(m+1)} \right)^{m+1} \leq \left(\frac{e}{3} \right)^{m+1} \leq \exp\left(-\frac{m+1}{12} \right) \leq \exp\left(\frac{-27a^2}{24} \right) \leq e^{-a^2/2}$$

where we have also used that $(e/3)^6 \leq e^{-1/2}$. This proves both inequalities (4.93) and (4.94) .

It therefore remains to prove that a discrete probability G' satisfying (4.95) can be chosen with at most ℓ atoms where ℓ is given by (4.92). This is guaranteed by Caratheodory's theorem as argued below. Let $\mathcal{P}(\mathbb{R}^d)$ denote the collection of all probability measures on $\mathbb{R}^d$ and let

$$T := \left\{ \left(\int \theta_j^k \{\theta \in E_i\} dG(\theta), 1 \leq j \leq d, 0 \leq k \leq 2m+1, 1 \leq i \leq L \right) : G \in \mathcal{P}(\mathbb{R}^d) \right\}.$$

This set T is clearly a convex subset of $\mathbb{R}^p$ for $p := d(2m+2)L$. Moreover, it is easy to see that T is simply the convex hull of

$$C := \left\{ \left(\theta_j^k \{\theta \in E_i\}, 1 \leq j \leq d, 0 \leq k \leq 2m+1, 1 \leq i \leq L \right) : \theta \in S^a \right\}.$$

Therefore, by Caratheodory's theorem, every element of T can be written as a convex combination of at most $p+1$ elements of C. We therefore take G' to be the discrete probability measure supported upon these elements with probabilities given by the weights of this convex combination. Note that the number of atoms of G' is bounded from above by ℓ given in (4.92). It is also easy to see that G' is supported on S^a. This completes the proof. □

Proof of Theorem 4.8.1

Proof. Fix $0 < \eta \leq \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\sqrt{e}}$ and define a as in (4.78). Note that $a \geq 1$. Fix $G \in \mathcal{G}$. According to Lemma 4.8.5, there exists a discrete probability measure G' supported on S^a and having ℓ atoms (with ℓ as in (4.92)) such that:

$$\sup_{x \in S} |f_G(x) - f_{G'}(x)| \leq \left(1 + \frac{1}{\sqrt{2\pi}} \right) (2\pi)^{-d/2} e^{-a^2/2} \tag{4.96}$$

and

$$\sup_{x \in S} \|\nabla f_G(x) - \nabla f_{G'}(x)\| \leq \left(a + \frac{3a}{\sqrt{2\pi}} \right) (2\pi)^{-d/2} e^{-a^2/2}. \tag{4.97}$$

Now let $\alpha > 0$ and let $s_1, \ldots, s_D$ be an $\alpha-$cover of S^a (i.e., $\sup_{s\in S^a} \inf_{1\le i\le D} \|s - s_i\| \le \alpha$) with $D = N(\alpha, S^a)$. Now if $G' = \sum_{i=1}^{\ell} w_i \delta_{a_i}$ (for some probability vector $(w_1, \ldots, w_\ell)$ and atoms $a_1, \ldots, a_\ell \in S^a$), then let $G'' := \sum_{i=1}^{\ell} w_i \delta_{b_i}$ where $b_i \in \{s_1, \ldots, s_D\}$ and $\|a_i - b_i\| \le \alpha$. Then, for every $x \in S$,

$$\begin{aligned}
|f_{G'}(x) - f_{G''}(x)| &= \left| \sum_{i=1}^{\ell} w_i \phi_d(x - a_i) - \sum_{i=1}^{\ell} w_i \phi_d(x - b_i) \right| \\
&\le \sum_{i=1}^{\ell} w_i \left| \phi_d(x - a_i) - \phi_d(x - b_i) \right| \\
&\le \sum_{i=1}^{\ell} w_i \sup_t \|\nabla \phi_d(t)\| \alpha \\
&\le \alpha \sup_t \|\nabla \phi_d(t)\| = \alpha (2\pi)^{-d/2} \sup_t \|t\| e^{-\|t\|^2/2} = \alpha (2\pi)^{-d/2} e^{-1/2}.
\end{aligned}$$

We shall now bound $\|\nabla f_{G'}(x) - \nabla f_{G''}(x)\|$ using similar arguments. By the mean value theorem, there exists u_i on the line segment joining $x - a_i$ and $x - b_i$ such that,

$$\phi_d(x - b_i) = \phi_d(x - a_i) + (a_i - b_i)^\top \nabla \phi_d(u_i)$$

and consequently

$$x - b_i = u_i + \zeta_i \qquad \text{for some } \zeta_i \text{ satisfying } \|\zeta_i\| \le \alpha.$$

Similarly,

$$\begin{aligned}
\|\nabla f_{G'}(x) - \nabla f_{G''}(x)\| &= \sum_{i=1}^{\ell} w_i \left\| \nabla \phi_d(x - a_i) - \nabla \phi_d(x - b_i) \right\| \\
&= \sum_{i=1}^{\ell} w_i \left\| (a_i - x)\phi_d(x - a_i) - (b_i - x)\phi_d(x - b_i) \right\| \\
&= \sum_{i=1}^{\ell} w_i \left\| (b_i - a_i)\phi_d(x - a_i) + (u_i + \zeta_i) \left[(a_i - b_i)^\top \nabla \phi_d(u_i) \right] \right\| \\
&\le \alpha \sup_t \phi_d(t) + \alpha \sup_t (\|t\| + \alpha) \|\nabla \phi_d(t)\| \\
&\le \alpha \sup_t \phi_d(t) + \alpha \sup_t \|t\|^2 \phi_d(t) + \alpha^2 \sup_t \|t\| \phi_d(t) \\
&= \frac{\alpha}{(2\pi)^{d/2}} \left[1 + \frac{2}{e} + \alpha \frac{1}{\sqrt{e}} \right]
\end{aligned}$$

Now if $G''' := \sum_{i=1}^{\ell} w_i' \delta_{b_i}$ for some other probability vector $w' := (w_1', \ldots, w_\ell')$, then clearly

$$\begin{aligned}
|f_{G'}(x) - f_{G''}(x)| &= \left| \sum_{i=1}^{\ell} (w_i - w_i')\phi(x - b_i) \right| \leq (2\pi)^{-d/2} \sum_{i=1}^{\ell} |w_i - w_i'| \\
\|\nabla f_{G'}(x) - \nabla f_{G''}(x)\| &= \left\| \sum_{i=1}^{\ell} (w_i - w_i')\nabla\phi(x - b_i) \right\| \\
&\leq \sum_{i=1}^{\ell} |w_i - w_i'| \left[\sup_t \|t\phi_d(t)\| \right] = (2\pi)^{-d/2} e^{-1/2} \sum_{i=1}^{\ell} |w_i - w_i'|
\end{aligned}$$

Therefore if $\sum_{i=1} |w_i - w_i'| \leq v$, then

$$\sup_{x \in S} |f_G(x) - f_{G'''}(x)| \leq \left(1 + \frac{1}{\sqrt{2\pi}}\right) (2\pi)^{-d/2} e^{-a^2/2} + \alpha(2\pi)^{-d/2} e^{-1/2} + (2\pi)^{-d/2} v$$

and

$$\begin{aligned}
\sup_{x \in S} \|\nabla f_G(x) - \nabla f_{G'''}(x)\| \leq &\left(a + \frac{3a}{\sqrt{2\pi}}\right) (2\pi)^{-d/2} e^{-a^2/2} + \alpha(2\pi)^{-d/2} \left[1 + \frac{2}{e} + \alpha\frac{1}{\sqrt{e}}\right] \\
&+ (2\pi)^{-d/2} e^{-1/2} v.
\end{aligned}$$

By choosing

$$v = \alpha = \frac{(2\pi)^{d/2}}{2\sqrt{2\pi}}\eta \quad \text{and} \quad a = \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\eta}} = \sqrt{2 \log \frac{1}{\alpha}},$$

we obtain

$$\begin{aligned}
\sup_{x \in S} |f_G(x) - f_{G'''}(x)| &< \frac{\alpha}{(2\pi)^{d/2}} \left[2 + \frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{e}}\right] < \eta \\
\sup_{x \in S} \|\nabla f_G(x) - \nabla f_{G'''}(x)\| &\leq \frac{a\alpha}{(2\pi)^{d/2}} \left[2 + \frac{3}{\sqrt{2\pi}} + \frac{3}{e} + \frac{1}{\sqrt{e}}\right] < a\eta
\end{aligned}$$

where we have noted that $a \geq 1$ and $\alpha \leq e^{-1/2}$.

It only remains to count the number of ways of choosing the discrete probability measure G'''. The number of ways of choosing the atoms of G''' is clearly

$$\binom{D}{\ell} \leq \frac{D^\ell}{\ell!} \leq \left(\frac{De}{\ell}\right)^\ell$$

where we used that $\ell! \geq (\ell/e)^\ell$, a fact that follows from Stirling's formula.

The probability vector $w' = (w'_1, \ldots, w'_\ell)$ can be chosen to belong to a v-covering set for all ℓ-dimensional probability vectors under the L^1 norm. This covering number is well known to be at most: $(1 + (2/v))^\ell$. Therefore $N(\eta, \mathcal{M}, d_{S,\alpha_1,\alpha_2})$ is bounded from above by:

$$\left[\frac{De}{\ell}\left(1 + \frac{2}{v}\right)\right]^\ell = A^\ell \qquad \text{where } A := \frac{De}{\ell}\left(1 + \frac{2}{v}\right).$$

We shall bound A below. Below C_d will denote a constant that depends on d alone. Because $v \le e^{-1/2}$,

$$1 + \frac{2}{v} \le \left(\frac{1}{\sqrt{e}} + 2\right)\frac{1}{v} = \frac{C_d}{\eta}.$$

Also note that from the expression for ℓ given in (4.92), we have $\ell \ge N(a, S^a)$ and hence

$$\frac{D}{\ell} \le \frac{N(\alpha, S^a)}{N(a, S^a)} \le N(\alpha, B(0, a)) \le \left(1 + \frac{a}{\alpha}\right)^d \le C_d \left(\frac{1}{\eta}\right)^{3d/2}.$$

where we have used the trivial fact that

$$a = \sqrt{2\log\frac{1}{\alpha}} \le \sqrt{\frac{4}{\alpha}} = C_d \frac{1}{\sqrt{\eta}}. \tag{4.98}$$

We thus have

$$A \le C_d \eta^{-1-3d/2}$$

so that,

$$\log N(\eta, \mathcal{M}, \|\cdot\|_S) \le \ell \log A \le C_d \ell \log\frac{1}{\eta}$$

which along with the expression (4.92) for ℓ proves (4.88). Similarly,

$$\log N(a\eta, \mathcal{M}, \|\cdot\|_{S,\nabla}) \le C_d \ell \log\frac{1}{\eta} \le C_d N(a, S^a)|\log\eta|^2.$$

This implies that

$$\log N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla}) \le C_d N(a, S^a)\left|\log\frac{\eta}{a}\right|^2 \le C_d N(a, S^a)|\log\eta|^2$$

where the last inequality follows from (4.98). This completes the proof of (4.89) and consequently Theorem 4.8.1.

□

4.8.2 Proof of Corollary 4.8.2

Proof of Corollary 4.8.2. Fix $\rho > 0$, $0 < \eta \leq \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\sqrt{e}}$ and compact subset $S \subseteq \mathbb{R}^d$. For $a, b \in \mathbb{R}$, we shall denote the maximum of a and b by $a \vee b$. Note first that for every pair of densities $f_G, f_H \in \mathcal{M}$ and $x \in S$, we have

$$\begin{aligned}\left\|\frac{\nabla f_G(x)}{\rho \vee f_G(x)} - \frac{\nabla f_H(x)}{\rho \vee f_H(x)}\right\| &= \left\|\frac{\nabla f_G(x)}{\rho \vee f_G(x)} - \frac{\nabla f_G(x)}{\rho \vee f_H(x)} + \frac{\nabla f_G(x)}{\rho \vee f_H(x)} - \frac{\nabla f_H(x)}{\rho \vee f_H(x)}\right\| \\ &\leq \frac{\|\nabla f_G(x)\|}{\rho \vee f_G(x)} \frac{|\rho \vee f_G(x) - \rho \vee f_H(x)|}{\rho \vee f_H(x)} + \frac{1}{\rho}\|\nabla f_G(x) - \nabla f_H(x)\|\end{aligned}$$

Using inequality (A.2) (in Lemma A.0.2) and the fact that $t \mapsto \rho \vee t$ is 1-Lipschitz, we deduce from the above that

$$\left\|\frac{\nabla f_G(x)}{\rho \vee f_G(x)} - \frac{\nabla f_H(x)}{\rho \vee f_H(x)}\right\| \leq \sqrt{\frac{1}{\rho^2}\log\frac{1}{(2\pi)^d\rho^2}}\,|f_G(x) - f_H(x)| + \frac{1}{\rho}\|f_G(x) - f_H(x)\|.$$

Because this is true for every $x \in S$, we have

$$\|f_G - f_H\|^{\rho}_{S,\nabla} \leq \sqrt{\frac{1}{\rho^2}\log\frac{1}{(2\pi)^d\rho^2}}\,\|f_G - f_H\|_S + \frac{1}{\rho}\|f_G - f_H\|_{S,\nabla}.$$

We thus have

$$N(\eta^*, \mathcal{T}_\rho, \|\cdot\|_S) \leq N(\eta, \mathcal{M}, \|\cdot\|_S) + N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla})$$

from which (4.80) follows. □

4.9 Bounding Bayes Discrepancy via Hellinger Distance

The purpose of this section is to state and prove the following theorem relating the quantity:

$$\Gamma(G_0, G, \rho) := \left(\int \left\|\frac{\nabla f_G(x)}{\max(f_G(x), \rho)} - \frac{\nabla f_{G_0}(x)}{\max(f_{G_0}(x), \rho)}\right\|^2 f_{G_0}(x)dx\right)^{1/2}. \tag{4.99}$$

for $\rho > 0$ and two probability measures G_0 and G on $\mathbb{R}^d$ in terms of the squared Hellinger distance between f_G and f_{G_0}. This result is crucial for the proof of Theorem 4.4.1.

Theorem 4.9.1. *There exists a universal positive constant C such that for every pair of probability measures G and G_0 on $\mathbb{R}^d$ and $0 < \rho \leq (2\pi)^{-d/2}e^{-1/2}$, we have*

$$\Gamma^2(G_0, G, \rho) \leq Cd\max\left\{\left(\log\frac{(2\pi)^{-d/2}}{\rho}\right)^3, |\log \mathfrak{H}(f_G, f_{G_0})|\right\}\mathfrak{H}^2(f_G, f_{G_0}) \tag{4.100}$$

where $\Gamma(G_0, G, \rho)$ is defined as in (4.99).

The above theorem is a generalization of Jiang and Zhang [63, Theorem 3] to the case when $d \geq 1$. Its proof given below follows Jiang and Zhang [63, Proof of Theorem 3] with appropriate changes to deal with the $d \geq 1$ case. Lemma A.0.2 and Lemma A.0.3 from Section A will be used in this proof.

Proof of Theorem 4.9.1. For real numbers a and b, we denote $\max(a, b)$ by $a \vee b$. For functions $u : \mathbb{R}^d \to \mathbb{R}^d$, we let

$$\|u\|_0 := \left(\int \|u(x)\|^2 f_{G_0}(x) dx \right)^{1/2}$$

so that

$$\begin{aligned}
\Gamma(G_0, G, \rho) &= \left\| \frac{\nabla f_G}{f_G \vee \rho} - \frac{\nabla f_{G_0}}{f_{G_0} \vee \rho} \right\|_0 \\
&= \left\| \frac{\nabla f_G}{f_G \vee \rho} - \frac{2\nabla f_G}{f_G \vee \rho + f_{G_0} \vee \rho} + 2\frac{\nabla f_G - \nabla f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho} + \frac{2\nabla f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho} - \frac{\nabla f_{G_0}}{f_{G_0} \vee \rho} \right\|_0 \\
&\leq 2 \max_{H \in \{G, G_0\}} \left\| \frac{(\nabla f_H)|f_G \vee \rho - f_{G_0} \vee \rho|}{(f_H \vee \rho)(f_G \vee \rho + f_{G_0} \vee \rho)} \right\|_0 + 2 \left\| \frac{\nabla f_G - \nabla f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho} \right\|_0,
\end{aligned}$$

where we have used the triangle inequality for $\|\|_0$ in the last step. Let us represent the two terms on the right hand side above by T_1 and T_2 respectively so that $\Gamma(G_0, G, \rho) \leq T_1 + T_2$. We shall now bound T_1 and T_2 separately. For T_1, we use inequality (A.2) in Lemma A.0.2 (note that we have assumed $0 < \rho \leq (2\pi)^{-d/2} e^{-1/2}$). This inequality allows us to bound T_1 as follows:

$$\begin{aligned}
\frac{1}{4} T_1^2 &= \max_{H \in \{G, G_0\}} \int \frac{\|\nabla f_H\|^2 (f_G \vee \rho - f_{G_0} \vee \rho)^2}{(f_H \vee \rho)^2 (f_G \vee \rho + f_{G_0} \vee \rho)^2} f_{G_0} \\
&\leq \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \frac{(f_G \vee \rho - f_{G_0} \vee \rho)^2}{(f_G \vee \rho + f_{G_0} \vee \rho)^2} f_{G_0} \\
&\leq \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \frac{(f_G - f_{G_0})^2}{(f_G \vee \rho + f_{G_0} \vee \rho)^2} f_{G_0} \\
&= \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \left(\sqrt{f_G} - \sqrt{f_{G_0}} \right)^2 \frac{\left(\sqrt{f_G} + \sqrt{f_{G_0}} \right)^2}{(f_G \vee \rho + f_{G_0} \vee \rho)^2} f_{G_0} \\
&\leq 2 \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \left(\sqrt{f_G} - \sqrt{f_{G_0}} \right)^2 \frac{(f_G + f_{G_0})}{(f_G \vee \rho + f_{G_0} \vee \rho)^2} f_{G_0} \\
&= 2 \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \left(\sqrt{f_G} - \sqrt{f_{G_0}} \right)^2 \left\{ \frac{f_G + f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho} \right\} \left\{ \frac{f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho} \right\} \\
&\leq 2 \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \int \left(\sqrt{f_G} - \sqrt{f_{G_0}} \right)^2 = 2 \left[\log \frac{(2\pi)^{-d}}{\rho^2} \right] \mathfrak{H}^2(f_G, f_{G_0})
\end{aligned}$$

which gives

$$T_1 \leq 2\sqrt{2}\mathfrak{H}(f_G, f_{G_0})\sqrt{\log \frac{(2\pi)^{-d}}{\rho^2}}. \tag{4.101}$$

We shall now deal with T_2. This requires an elaborate argument. Start by writing

$$\begin{aligned}\frac{1}{4}T_2^2 &= \int \frac{\|\nabla f_G - \nabla f_{G_0}\|^2}{\left(f_G \vee \rho + f_{G_0} \vee \rho\right)^2} f_{G_0} \\ &= \int \frac{\|\nabla f_G - \nabla f_{G_0}\|^2}{f_G \vee \rho + f_{G_0} \vee \rho}\left(\frac{f_{G_0}}{f_G \vee \rho + f_{G_0} \vee \rho}\right) \leq \int \frac{\|\nabla f_G - \nabla f_{G_0}\|^2}{f_G \vee \rho + f_{G_0} \vee \rho} = \sum_{i=1}^d \Delta_{i,1}^2 \end{aligned} \tag{4.102}$$

where, for $1 \leq i \leq d$ and $k \geq 0$,

$$\Delta_{i,k}^2 := \int \frac{\left(\partial_i^k (f_G - f_{G_0})\right)^2}{f_G \vee \rho + f_{G_0} \vee \rho} \qquad \text{with } \partial_i^k f := \frac{\partial^k}{\partial x_i^k} f.$$

The next task therefore is to bound $\Delta_{i,1}^2$ from above. Before dealing with $\Delta_{i,1}^2$, let us first note that it is easy to bound $\Delta_{i,0}$ by the Hellinger distance between f_G and f_{G_0}. Indeed, we can write

$$\begin{aligned}\Delta_{i,0}^2 = \int \frac{(f_G - f_{G_0})^2}{f_G \vee \rho + f_{G_0} \vee \rho} &= \int \left(\sqrt{f_G} - \sqrt{f_{G_0}}\right)^2 \frac{(\sqrt{f_G} + \sqrt{f_{G_0}})^2}{f_G \vee \rho + f_{G_0} \vee \rho} \\ &\leq 2\int \left(\sqrt{f_G} - \sqrt{f_{G_0}}\right)^2 \frac{(f_G + f_{G_0})}{f_G \vee \rho + f_{G_0} \vee \rho} \leq 2\mathfrak{H}^2(f_G, f_{G_0}).\end{aligned} \tag{4.103}$$

A simple upper bound for $\Delta_{i,k}^2$ for general $k \geq 1$ can be obtained via Lemma A.0.3. Indeed, noting that (via $f_G \vee \rho + f_{G_0} \vee \rho \geq 2\rho$)

$$\Delta_{i,k}^2 \leq \frac{1}{2\rho}\int \left(\partial_i^k \left(f_G - f_{G_0}\right)\right)^2$$

we can apply Lemma A.0.3 to deduce that

$$\Delta_{i,k}^2 \leq \frac{2(2\pi)^{-d/2}}{\rho}\left\{a^{2k}\mathfrak{H}^2(f_G, f_{G_0}) + \sqrt{\frac{2}{\pi}}a^{2k-1}e^{-a^2}\right\} \qquad \text{for every } a \geq \sqrt{2k-1}. \tag{4.104}$$

The problem with this bound is the presence of ρ in the denominator. This ρ will be, in applications of Theorem 4.9.1, of the order n^{-1} which makes the above bound quite large. The more refined argument below will get rid of the ρ factor in the denominator. This argument involves integration by parts for bounding $\Delta_{i,1}^2$. It will be clear that the use of integration by parts will result in expressions involving $\Delta_{i,k}^2$ for $k \geq 2$. It will then become

necessary to deal with $\Delta_{i,k}^2$ for $k \geq 2$ even though we are only interested in $\Delta_{i,1}^2$. Indeed, integration by parts gives, for $k \geq 1$,

$$\begin{aligned}\Delta_{i,k}^2 = &- \int \left[\partial_i^{k-1}\left(f_G - f_{G_0}\right)\right]\left[\partial_i^{k}\left(f_G - f_{G_0}\right)\right]\partial_i\left(\frac{1}{f_G \vee \rho + f_{G_0} \vee \rho}\right) \\ &- \int \frac{\left[\partial_i^{k-1}\left(f_G - f_{G_0}\right)\right]\left[\partial_i^{k+1}\left(f_G - f_{G_0}\right)\right]}{f_G \vee \rho + f_{G_0} \vee \rho}. \end{aligned} \tag{4.105}$$

Note now that, almost surely

$$\begin{aligned}\left|\partial_i\left(\frac{1}{f_G \vee \rho + f_{G_0} \vee \rho}\right)\right| &\leq \frac{|\partial_i f_G| + |\partial_i f_{G_0}|}{\left(f_G \vee \rho + f_{G_0} \vee \rho\right)^2} \\ &\leq \frac{|\partial_i f_G|/(f_G \vee \rho) + |\partial_i f_{G_0}|/(f_{G_0} \vee \rho)}{f_G \vee \rho + f_{G_0} \vee \rho} \\ &\leq \frac{\|\nabla f_G\|/(f_G \vee \rho) + \|\nabla f_{G_0}\|/(f_{G_0} \vee \rho)}{f_G \vee \rho + f_{G_0} \vee \rho} \\ &\leq \frac{2}{f_G \vee \rho + f_{G_0} \vee \rho}\sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}}\end{aligned}$$

where, in the last inequality, we used (A.2) in Lemma A.0.2. Imputing the above inequality into (4.105), we obtain

$$\Delta_{i,k}^2 \leq 2\sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}}\int \frac{\left|\partial_i^{k-1}\left(f_G - f_{G_0}\right)\right|\left|\partial_i^{k}\left(f_G - f_{G_0}\right)\right|}{f_G \vee \rho + f_{G_0} \vee \rho} + \int \frac{\left|\partial_i^{k-1}\left(f_G - f_{G_0}\right)\right|\left|\partial_i^{k+1}\left(f_G - f_{G_0}\right)\right|}{f_G \vee \rho + f_{G_0} \vee \rho}.$$

Applying the Cauchy-Schwarz inequality to each of the two terms on the right hand side above, we obtain

$$\begin{aligned}\Delta_{i,k}^2 \leq &\, 2\sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}}\sqrt{\int \frac{\left(\partial_i^{k-1}(f_G - f_{G_0})\right)^2}{f_G \vee \rho + f_{G_0} \vee \rho}}\sqrt{\int \frac{\left(\partial_i^{k}(f_G - f_{G_0})\right)^2}{f_G \vee \rho + f_{G_0} \vee \rho}} \\ &+ \sqrt{\int \frac{\left(\partial_i^{k-1}(f_G - f_{G_0})\right)^2}{f_G \vee \rho + f_{G_0} \vee \rho}}\sqrt{\int \frac{\left(\partial_i^{k+1}(f_G - f_{G_0})\right)^2}{f_G \vee \rho + f_{G_0} \vee \rho}}\end{aligned}$$

which can be rewritten as

$$\Delta_{i,k}^2 \leq \Upsilon\Delta_{i,k-1}\Delta_{i,k} + \Delta_{i,k-1}\Delta_{i,k+1} \qquad \text{where } \Upsilon := 2\sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}}. \tag{4.106}$$

The strategy to bound $\Delta_{i,1}$ is now as follows. Divide both sides of (4.106) by $\Delta_{i,k-1}\Delta_{i,k}$ to get

$$\frac{\Delta_{i,k}}{\Delta_{i,k-1}} \leq \Upsilon + \frac{\Delta_{i,k+1}}{\Delta_{i,k}} \qquad \text{for every } k \geq 1. \tag{4.107}$$

Fix an integer $k_0 \geq 1$ and a real number $\beta > 0$. Our bound on $\Delta_{i,1}$ will depend on k_0 and β and the bound will be optimized for k_0 and β at the end.

Suppose first that there exists an integer $1 \leq k \leq k_0$ such that $\Delta_{i,k+1} \leq \beta \Delta_{i,k}$. Then applying (4.107) recursively for $1, \ldots, k$, we obtain

$$\frac{\Delta_{i,1}}{\Delta_{i,0}} \leq k\Upsilon + \beta$$

so that, by (4.103),

$$\Delta_{i,1} \leq (k\Upsilon + \beta)\, \Delta_{i,0} \leq \sqrt{2}\,(k\Upsilon + \beta)\, \mathfrak{H}(f_G, f_{G_0}) \leq \sqrt{2}\,(k_0\Upsilon + \beta)\, \mathfrak{H}(f_G, f_{G_0}). \tag{4.108}$$

Now suppose that $\Delta_{i,k+1} > \beta \Delta_{i,k}$ for every integer $1 \leq k \leq k_0$. In this case, we deduce from (4.107) that

$$\frac{\Delta_{i,k}}{\Delta_{i,k-1}} \leq \Upsilon + \frac{\Delta_{i,k+1}}{\Delta_{i,k}} \leq \left(1 + \frac{\Upsilon}{\beta}\right) \frac{\Delta_{i,k+1}}{\Delta_{i,k}} \qquad \text{for every } k = 0, \ldots, k_0.$$

A recursive application of this inequality implies that

$$\frac{\Delta_{i,1}}{\Delta_{i,0}} \leq \left(1 + \frac{\Upsilon}{\beta}\right)^k \frac{\Delta_{i,k+1}}{\Delta_{i,k}} \qquad \text{for every } k = 0, \ldots, k_0.$$

To obtain a bound for $\Delta_{i,1}/\Delta_{i,0}$ that depends only on Δ_{i,k_0+1} and $\Delta_{i,0}$, one can take the geometric mean of the above inequality for $k = 0, 1, \ldots, k_0$. This gives

$$\frac{\Delta_{i,1}}{\Delta_{i,0}} \leq \left(\prod_{k=0}^{k_0} \left(1 + \frac{\Upsilon}{\beta}\right)^k \frac{\Delta_{i,k+1}}{\Delta_{i,k}}\right)^{1/(k_0+1)} = \left(1 + \frac{\Upsilon}{\beta}\right)^{k_0/2} \Delta_{i,k_0+1}^{1/(k_0+1)} \Delta_{i,0}^{-1/(k_0+1)}$$

which is same as

$$\Delta_{i,1} \leq \left(1 + \frac{\Upsilon}{\beta}\right)^{k_0/2} \Delta_{i,k_0+1}^{1/(k_0+1)} \Delta_{i,0}^{k_0/(k_0+1)}$$

Now using (4.103) and the bound (4.104) (with $k = k_0 + 1$), we obtain

$$\Delta_{i,1} \leq \left(1 + \frac{\Upsilon}{\beta}\right)^{\frac{k_0}{2}} \left(\frac{2(2\pi)^{-d/2}}{\rho}\left[a^{2k_0+2}\mathfrak{H}^2(f_G, f_{G_0}) + \sqrt{\frac{2}{\pi}} a^{2k_0+1} e^{-a^2}\right]\right)^{\frac{1}{2k_0+2}} \left(2\mathfrak{H}^2(f_G, f_{G_0})\right)^{\frac{k_0}{2k_0+2}} \tag{4.109}$$

for every $a \geq \sqrt{2k_0+1}$. The final bound obtained for $\Delta_{i,1}$ is the maximum of the right hand side above and the right hand side of (4.108). This bound will need to be optimized by choosing k_0, β and $a \geq \sqrt{2k_0+1}$ appropriately.

β will be chosen as $\beta = k_0\Upsilon$ so that the bound (4.108) becomes $2\sqrt{2}k_0\Upsilon\mathfrak{H}(f_G, f_{G_0})$ and the term $(1 + \Upsilon/\beta)^{k_0/2}$ appearing in (4.109) is bounded by $\sqrt{e}$. To select k_0, the key is to focus on the term involving ρ in (4.109) which is

$$\left(\frac{(2\pi)^{-d/2}}{\rho}\right)^{1/(2k_0+2)} = \exp\left(\frac{\Upsilon^2}{16(k_0+1)}\right).$$

This suggests taking k_0 to be the smallest integer ≥ 1 such that $k_0 + 1 \geq \Upsilon^2/8$ so that the above term is at most $\sqrt{e}$. Finally a will be taken to be

$$a := \max\left(\sqrt{2k_0+1}, \sqrt{2\left|\log \mathfrak{H}(f_G, f_{G_0})\right|}\right)$$

which will ensure that $e^{-a^2} \leq \mathfrak{H}^2(f_G, f_{G_0})$ and the term involving a in (4.109) can then be bounded by

$$\begin{aligned}\left(a^{2k_0+2}\mathfrak{H}^2(f_G, f_{G_0}) + \sqrt{\frac{2}{\pi}}a^{2k_0+1}e^{-a^2}\right)^{\frac{1}{2k_0+2}} &\leq a\left(1+\sqrt{\frac{2}{\pi}}\right)^{\frac{1}{2k_0+2}} \left(\mathfrak{H}(f_G, f_{G_0})\right)^{\frac{1}{k_0+1}} \\ &\leq \left(1+\sqrt{\frac{2}{\pi}}\right) a\left(\mathfrak{H}(f_G, f_{G_0})\right)^{\frac{1}{k_0+1}} \\ &\leq 2a\left(\mathfrak{H}(f_G, f_{G_0})\right)^{\frac{1}{k_0+1}}.\end{aligned}$$

We have therefore proved that the right hand side in (4.109) is bounded from above by $2\sqrt{2}ea\mathfrak{H}(f_G, f_{G_0})$. Because $\Delta_{i,1}$ is bounded by the maximum of the bounds given by (4.108) and (4.109), we obtain:

$$\Delta_{i,1} \leq 2\sqrt{2}\max\left\{k_0\Upsilon, ea\right\}\mathfrak{H}(f_G, f_{G_0}) \leq 2\sqrt{2}\max\left\{k_0\Upsilon, e\sqrt{2k_0+1}, e\sqrt{2|\log \mathfrak{H}(f_G, f_{G_0})|}\right\}\mathfrak{H}(f_G, f_{G_0}).$$

Now because k_0 is chosen to be the smallest integer ≥ 1 such that $k_0 + 1 \geq \Upsilon^2/8$, we have

$$k_0 \leq 1 + \frac{\Upsilon^2}{8} = \log\frac{e(2\pi)^{-d/2}}{\rho} \leq \frac{3}{2}\log\frac{(2\pi)^{-d/2}}{\rho}$$

because $\rho \leq (2\pi)^{-d/2}e^{-1/2}$. This, along with the expression for Υ, gives

$$\Delta_{i,1} \leq C\max\left\{\left(\log\frac{(2\pi)^{-d/2}}{\rho}\right)^{3/2}, \sqrt{|\log \mathfrak{H}(f_G, f_{G_0})|}\right\}\mathfrak{H}(f_G, f_{G_0})$$

where C is a universal positive constant. Combining with (4.102), we deduce that

$$T_2^2 \leq Cd\max\left\{\left(\log\frac{(2\pi)^{-d/2}}{\rho}\right)^3, |\log \mathfrak{H}(f_G, f_{G_0})|\right\}\mathfrak{H}^2(f_G, f_{G_0}).$$

The proof of Theorem 4.9.1 is now completed by combining the above inequality with the bound (4.101) and the fact that $\Gamma(G_0, G, \rho) \leq T_1 + T_2$ (which implies that $\Gamma^2(G_0, G, \rho) \leq 2T_1^2 + 2T_2^2$). □

Appendix A

Auxiliary results for Chapter 4

This section collects various results which were used in the proofs of the main results of the chapter. We start with the following lemma which is standard and is stated without proof.

Lemma A.0.1. *Suppose* $X|\theta \sim N(\theta, \sigma^2 I_d)$ *and* $\theta \sim G$. *Then*

$$\mathbb{E}\left(\theta|X\right) = X + \sigma^2 \frac{\nabla f_G(X)}{f_G(X)}$$

and

$$\mathbb{E}\left\|X + \sigma^2 \frac{\nabla f_G(X)}{f_G(X)} - \theta\right\|^2 = d\sigma^2 - \sigma^4 \int \left\|\frac{\nabla f_G}{f_G}\right\|^2 f_G.$$

The following lemma generalizes Jiang and Zhang [63, Lemma A.1] to the case $d \geq 1$.

Lemma A.0.2. *Fix a probability measure* G *on* $\mathbb{R}^d$. *For every* $x \in \mathbb{R}^d$, *we have* ($\|\cdot\|$ *denotes the usual Euclidean norm on* $\mathbb{R}^d$)

$$\left(\frac{\|\nabla f_G(x)\|}{f_G(x)}\right)^2 \leq tr\left(I_d + \frac{H f_G(x)}{f_G(x)}\right) \leq \log \frac{(2\pi)^{-d}}{f_G^2(x)} \tag{A.1}$$

where ∇ *and* H *stand for gradient and Hessian respectively and* tr *denotes trace.*

Also for every $x \in \mathbb{R}^d$, *we have*

$$\frac{\|\nabla f_G(x)\|}{\max\left(f_G(x), \rho\right)} \leq \sqrt{\log \frac{(2\pi)^{-d}}{\rho^2}} \qquad 0 < \rho \leq (2\pi)^{-d/2} e^{-1/2} \tag{A.2}$$

and

$$\left(\frac{\|\nabla f_G(x)\|}{f_G(x)}\right)^2 \frac{f_G(x)}{f_G(x) \vee \rho} \leq \log \frac{(2\pi)^{-d}}{\rho^2} \qquad \text{for } 0 < \rho \leq (2\pi)^{-d/2} e^{-1}. \tag{A.3}$$

Proof of Lemma A.0.2. If $\theta \sim G$ and $X|\theta \sim N(\theta, I_d)$, then it is easy to verify that, for every $x \in \mathbb{R}^d$,

$$\frac{\nabla f_G(x)}{f_G(x)} = \mathbb{E}\left(\theta - X | X = x\right)$$

and

$$\frac{Hf_G(x)}{f_G(x)} = -I_d + \mathbb{E}\left((\theta - X)(\theta - X)^T | X = x\right). \tag{A.4}$$

From here, we can deduce that

$$\begin{aligned} I_d + \frac{Hf_G(x)}{f_G(x)} &= \mathbb{E}\left((\theta - X)(\theta - X)^T | X = x\right) \\ &= \left(\mathbb{E}(\theta - X | X = x)\right)\left(\mathbb{E}(\theta - X | X = x)\right)^T + \mathbb{E}\left((\theta - \mathbb{E}(\theta|X=x))(\theta - \mathbb{E}(\theta|X=x))^T | X = x\right) \\ &= \frac{\nabla f_G(x)}{f_G(x)}\frac{(\nabla f_G(x))^T}{f_G(x)} + \mathbb{E}\left((\theta - \mathbb{E}(\theta|X=x))(\theta - \mathbb{E}(\theta|X=x))^T | X = x\right) \end{aligned}$$

and hence

$$I_d + \frac{Hf_G(x)}{f_G(x)} \succeq \frac{\nabla f_G(x)}{f_G(x)}\frac{(\nabla f_G(x))^T}{f_G(x)} \tag{A.5}$$

where $A \succeq B$ means that $A - B$ is non-negative definite.

Also from (A.4) and the convexity of $A \mapsto \exp(tr(A)/2)$ ($tr(A)$ denotes the trace of the $d \times d$ matrix A), we have

$$\begin{aligned} \exp\left(\frac{1}{2}tr\left(I_d + \frac{Hf_G(x)}{f_G(x)}\right)\right) &= \exp\left(\frac{1}{2}tr\left(\mathbb{E}\left((\theta - X)(\theta - X)^T | X = x\right)\right)\right) \\ &\leq \mathbb{E}\left(\exp\left(\frac{1}{2}tr(\theta - X)(\theta - X)^T\right) | X = x\right) \\ &= \mathbb{E}\left(\exp\left(\frac{1}{2}\left\|X - \theta\right\|^2\right) | X = x\right) = \frac{(2\pi)^{-d/2}}{f_G(x)} \end{aligned}$$

so that we have

$$tr\left(I_d + \frac{Hf_G(x)}{f_G(x)}\right) \leq \log\frac{(2\pi)^{-d}}{f_G^2(x)}.$$

Combining with (A.5), we obtain (A.1).

To prove (A.2), note first from (A.1) that

$$\frac{\|\nabla f_G(x)\|}{\max(f_G(x), \rho)} \leq \sqrt{\log\frac{(2\pi)^{-d}}{f_G^2(x)}}\frac{f_G(x)}{\max(f_G(x), \rho)} = \begin{cases} \sqrt{\log\frac{(2\pi)^{-d}}{f_G^2(x)}} \leq \sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}} & \text{if } f_G(x) > \rho \\ \sqrt{\log\frac{(2\pi)^{-d}}{f_G^2(x)}}\frac{f_G(x)}{\rho} & \text{if } f_G(x) \leq \rho \end{cases}$$

The function $v \mapsto v\log\left((2\pi)^{-d}/v\right)$ is non-decreasing on $(0, (2\pi)^{-d}/e]$ and hence when $f_G^2(x) \leq \rho^2 \leq (2\pi)^{-d}/e$, the inequality

$$\sqrt{\log\frac{(2\pi)^{-d}}{f_G^2(x)}}\frac{f_G(x)}{\rho} \leq \sqrt{\log\frac{(2\pi)^{-d}}{\rho^2}}$$

holds and this proves (A.2).

We now turn to (A.3). Whenever $f_G(x) \geq \rho$, note that (A.3) follows directly from (A.2). Thus, (A.3) only needs to be established when $f_G(x) < \rho$. In this case using (A.1),

$$\left(\frac{\|\nabla f_G(x)\|}{f_G(x)}\right)^2 \frac{f_G(x)}{\max\{f_G(x),\rho\}} \leq \left(\frac{f_G(x)}{\rho}\right) \log \frac{(2\pi)^{-d}}{f_G^2(x)} = 2\log\frac{(2\pi)^{-d/2}}{f_G(x)}\frac{f_G(x)}{\rho}$$

From here we note that $v \mapsto v\log\left((2\pi)^{-d}/v^2\right)$ is non-decreasing on $(0,(2\pi)^{-d/2}/e]$. This, along with $f_G(x) < \rho$, immediately implies (A.3). □

For an infinitely differentiable function $u : \mathbb{R}^d \to \mathbb{R}$, $1 \leq i \leq d$ and $k \geq 1$, let $\partial_i^k u : \mathbb{R}^d \to \mathbb{R}$ denote the function

$$(\partial_i^k u)(x) := \frac{\partial^k}{\partial x_i^k} u(x).$$

Lemma A.0.3. *For every pair of probability measures G and G_0 on $\mathbb{R}^d$, $1 \leq i \leq d$ and $k \geq 1$, we have*

$$\int \left\{\partial_i^k (f_G(x) - f_{G_0}(x))\right\}^2 dx \leq 4(2\pi)^{-d/2} \inf_{a \geq \sqrt{2k-1}} \left\{ a^{2k}\mathfrak{H}^2(f_G, f_{G_0}) + \sqrt{\frac{2}{\pi}} a^{2k-1} e^{-a^2} \right\}. \tag{A.6}$$

Proof of Lemma A.0.3. Fix $a \geq \sqrt{2k-1}$ and assume, without loss of generality, that $i = 1$. Let

$$f_{G,1}^*(u, x_2, \ldots, x_d) := \int e^{iux_1} f_G(x) dx_1$$

denote the Fourier transform of f_G treated as a function of x_1. The function $f_{G_0,1}^*$ is defined analogously. For ease of notation, we shall suppress the dependence of $f_{G,1}^*(u, x_2, \ldots, x_d)$ (resp. $f_{G_0}^*(u, x_2, \ldots, x_d)$) on $x_2, \ldots, x_d$ below and write it simply as $f_{G_1}^*(u)$ (resp. $f_{G_0}^*(u)$).

For every $x_2, \ldots, x_d$, we then have (by Plancherel's identity)

$$\begin{aligned}
2\pi \int \left\{\partial_1^k (f_G(x) - f_{G_0}(x))\right\}^2 dx_1 &= \int u^{2k} \left| f_{G,1}^*(u) - f_{G_0,1}^*(u) \right|^2 du \\
&\leq a^{2k} \int \left| f_{G,1}^*(u) - f_{G_0,1}^*(u) \right|^2 du + \int_{|u|>a} u^{2k} \left| f_{G,1}^*(u) - f_{G_0,1}^*(u) \right|^2 du \\
&= (2\pi) a^{2k} \int (f_G(x) - f_{G_0}(x))^2 dx_1 + \int_{|u|>a} u^{2k} \left| f_{G,1}^*(u) - f_{G_0,1}^*(u) \right|^2 du
\end{aligned} \tag{A.7}$$

for every $a > 0$. Also note that for every $u, x_2, \ldots, x_d \in \mathbb{R}$,

$$\begin{aligned} f^*_{G,1}(u) &= \int e^{iux_1} \left(\int \phi_d(x-\theta) dG(\theta) \right) dx_1 \\ &= \int \left(\int e^{iux_1} \phi_d(x-\theta) dx_1 \right) dG(\theta) \\ &= \int (2\pi)^{-d/2} \left[\int e^{iux_1} e^{-(x_1-\theta_1)^2/2} dx_1 \right] \exp\left(-\sum_{j\neq 1} (x_j-\theta_j)^2/2 \right) dG(\theta) \\ &= (2\pi)^{-(d-1)/2} \int e^{iux_1} e^{-u^2/2} \exp\left(-\sum_{j\neq 1} (x_j-\theta_j)^2/2 \right) dG(\theta) \end{aligned}$$

so that

$$\left| f^*_{G,1}(u) \right| \leq (2\pi)^{-(d-1)/2} e^{-u^2/2} \int \exp\left(-\sum_{j\neq 1} (x_j-\theta_j)^2/2 \right) dG(\theta).$$

An analogous bound also holds for $|f^*_{G_0,1}(u)|$. Using these bounds for $f^*_{G,1}(u)$ and $f^*_{G_0,1}(u)$, the second term in (A.7) can be bounded from above as

$$\int_{|u|>a} u^{2k} \left| f^*_{G,1}(u) - f^*_{G_0,1}(u) \right|^2 du \leq 2(2\pi)^{-(d-1)} \int \exp\left(-\sum_{j\neq 1} (x_j-\theta_j)^2 \right) \{dG(\theta) + dG_0(\theta)\} \int_{|u|>a} u^{2k} e^{-u^2} d$$

Thus integrating both sides of (A.7) with respect to $x_2, \ldots, x_d$, we deduce that

$$2\pi \int \left\{ \partial_1^k (f_G(x) - f_{G_0}(x)) \right\}^2 dx \leq (2\pi) a^{2k} \int (f_G - f_{G_0})^2 + 4(2\pi)^{-(d-1)/2} \int_{|u|>a} u^{2k} e^{-u^2} du.$$

which implies that

$$\int \left\{ \partial_1^k (f_G(x) - f_{G_0}(x)) \right\}^2 dx \leq a^{2k} \int (f_G - f_{G_0})^2 + 8(2\pi)^{-(d+1)/2} \int_{u>a} u^{2k} e^{-u^2} du.$$

We now use the integration by parts argument in Jiang and Zhang [63, Page 1675] which gives

$$\int_{u>a} u^{2k} e^{-u^2} du \leq a^{2k-1} e^{-a^2} \qquad \text{provided } a \geq \sqrt{2k-1}.$$

The proof of Lemma A.0.3 is now completed by noting that

$$\int (f_G - f_{G_0})^2 \leq \int \left(\sqrt{f_G} - \sqrt{f_{G_0}} \right)^2 \left(\sqrt{f_G} + \sqrt{f_{G_0}} \right)^2 \leq 4(2\pi)^{-d/2} \mathfrak{H}^2(f_G, f_{G_0})$$

where we have used that every Gaussian mixture density f_G is bounded from above by $(2\pi)^{-d/2}$. □

Lemma A.0.4. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim f_{G_i}$ and $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $g : \mathbb{R}^d \to [0, \infty)$ be a 1-Lipschitz function i.e.,*

$$g(x) - g(y) \leq \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d.$$

Also let $\mu_p(g)$ denote the p^{th} moment of g under the measure $\bar{G}_n$ i.e.,

$$\mu_p(g) := \left(\int_{\mathbb{R}^d} g(\theta)^p d\bar{G}_n(\theta) \right)^{1/p}.$$

There then exists a positive constant C_d depending only on d such that

$$\mathbb{E} \left\{ \prod_{i=1}^n |ag(X_i)|^{I\{g(X_i) \geq M\}} \right\}^\lambda \leq \exp \left\{ C_d a^\lambda M^{\lambda + d - 2} + (aM)^\lambda n \left(\frac{2\mu_p(g)}{M} \right)^p \right\} \tag{A.8}$$

for every $a > 0, M \geq \sqrt{8 \log n}$ and $0 < \lambda \leq \min(1, p)$.

Further, there exists a positive constant C_d depending only on d such that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{P}\left[g(X_i) \geq M \right] \leq C_d \frac{M^{d-2}}{n} + \inf_{p \geq \frac{d+1}{2 \log n}} \left(\frac{2\mu_p(g)}{M} \right)^p \tag{A.9}$$

for any $M \geq \sqrt{8 \log n}$.

Proof of Lemma A.0.4. We write

$$\begin{aligned}
\mathbb{E} \left\{ \prod_{i=1}^n |ag(X_i)|^{I\{g(X_i) \geq M\}} \right\}^\lambda &= \prod_{i=1}^n \mathbb{E} \left| ag(X_i) \right|^{\lambda I\{g(X_i) \geq M\}} \\
&\leq \prod_{i=1}^n \left\{ 1 + a^\lambda \mathbb{E}\left[(g(X_i))^\lambda I\{g(X_i) \geq M\} \right] \right\} \\
&\leq \prod_{i=1}^n \exp\left(a^\lambda \mathbb{E}(g(X_i))^\lambda I\{g(X_i) \geq M\} \right) \\
&= \exp\left(a^\lambda \sum_{i=1}^n \mathbb{E}\left[(g(X_i))^\lambda I\{g(X_i) \geq M\} \right] \right) \\
&= \exp\left(n a^\lambda \int (g(x))^\lambda I\{g(x) \geq M\} f_{\bar{G}_n}(x) dx \right) = \exp\left(n a^\lambda U \right)
\end{aligned}$$

where

$$U := \int (g(x))^\lambda I\{g(x) \geq M\} f_{\bar{G}_n}(x) dx = \mathbb{E}\left[(g(\theta + Z))^\lambda I\{g(\theta + Z) \geq M\} \right]$$

with independent random variables $Z \sim N(0, I_d)$ and $\theta \sim \bar{G}_n$. Because of the 1-Lipschitz property of g, we have $g(\theta + z) \leq g(\theta) + \|z\|$ so that

$$U \leq \mathbb{E}(2\|Z\|)^\lambda I\left\{2\|Z\| \geq M\right\} + \mathbb{E}(2g(\theta))^\lambda I\left\{2g(\theta) \geq M\right\}. \tag{A.10}$$

The first term above will be bounded as

$$\begin{aligned}
\mathbb{E}\left[(2\|Z\|)^\lambda I\{2\|Z\| \geq M\}\right] &= M^\lambda \mathbb{E}\left[\left(\frac{\|Z\|}{M/2}\right)^\lambda I\{\|Z\| \geq M/2\}\right] \\
&\leq M^\lambda \mathbb{E}\left[\left(\frac{\|Z\|}{M/2}\right) I\{\|Z\| \geq M/2\}\right] \qquad \text{since } \lambda \leq 1 \\
&= 2M^{\lambda-1}\frac{1}{(2\pi)^{d/2}}\int_{\|x\|\geq M/2} \|x\| e^{-\|x\|^2/2} dx \\
&\leq C_d M^{\lambda-1}\int_{r \geq M/2} r e^{-r^2/2} r^{d-1} dr \leq C_d M^{\lambda+d-2} e^{-M^2/8}
\end{aligned}$$

where the last inequality follows from Lemma A.0.7. Because $M \geq \sqrt{8\log n}$, we have $e^{-M^2/8} \leq 1/n$ and this gives

$$\mathbb{E}\left[(2\|Z\|)^\lambda I\{2\|Z\| \geq M\}\right] \leq \frac{C_d}{n} M^{\lambda+d-2}. \tag{A.11}$$

For the second term in (A.10), note that (because $\lambda \leq p$)

$$\begin{aligned}
\mathbb{E}\left[(2g(\theta))^\lambda I\{2g(\theta) \geq M\}\right] &= M^\lambda \int_{g(\theta)\geq M/2} \left(\frac{g(\theta)}{M/2}\right)^\lambda G_n(d\theta) \\
&\leq M^\lambda \int \left(\frac{g(\theta)}{M/2}\right)^p G_n(d\theta) = M^\lambda \left(\frac{2\mu_p(g)}{M}\right)^p.
\end{aligned} \tag{A.12}$$

The proof of (A.8) is now completed by putting together inequalities (A.10), (A.11) and (A.12).

For (A.9), we first use an argument similar to the above to write

$$\frac{1}{n}\sum_{i=1}^n \mathbb{P}\left[g(X_i) \geq M\right] = \mathbb{P}\left[g(\theta + Z) \geq M\right]$$

where $\theta \sim \bar{G}_n$ and $Z \sim N(0, I_d)$ are independent. Since g is 1-Lipschitz, $g(\theta+z) \leq g(\theta)+\|z\|$. Consequently,

$$\mathbb{P}\left[g(\theta + Z) \geq M\right] \leq \mathbb{P}\left[2g(\theta) \geq M\right] + \mathbb{P}\left[2\|Z\| \geq M\right]$$

Applying (A.11) and (A.12) with $\lambda = 0$ then concludes the proof of (A.9). □

Remark A.0.5. *We shall apply Lemma A.0.4 to the function*

$$\mathfrak{d}_S(x) := \inf_{u \in S} \|x - u\|$$

for a fixed subset S of $\mathbb{R}^d$. This function is clearly nonnegative and 1-Lipschitz. Inequality (A.8) *in Lemma A.0.4 then gives the inequality*

$$\mathbb{E}\left\{\prod_{i=1}^{n} |a\mathfrak{d}_S(X_i)|^{I\{\mathfrak{d}_S(X_i) \geq M\}}\right\}^{\lambda} \leq \exp\left\{C_d a^{\lambda} M^{\lambda+d-2} + (aM)^{\lambda} n \left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^{p}\right\} \tag{A.13}$$

for all $a > 0, M \geq \sqrt{8 \log n}$ and $0 < \lambda \leq \min(1, p)$.

Further, inequality (A.9) *for $g = \mathfrak{d}_S$ gives*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left[\mathfrak{d}_S(X_i) \geq M\right] \leq C_d \frac{M^{d-2}}{n} + \inf_{p \geq \frac{d+1}{2 \log n}} \left(\frac{2\mu_p(\mathfrak{d}_S)}{M}\right)^{p} \tag{A.14}$$

for all $M \geq \sqrt{8 \log n}$.

These two inequalities (A.13) *and* (A.14) *hold under the same assumptions on $X_1, \ldots, X_n$ as in Lemma A.0.4.*

Lemma A.0.6. *Fix $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$. Suppose $X_1, \ldots, X_n$ are independent random vectors with $X_i \sim N(\theta_i, I_d)$ for $i = 1, \ldots, n$. Let $\mathbf{X}$ denote the $d \times n$ matrix whose columns are $X_1, \ldots, X_n$. For $f \in \mathcal{M}$ and ρ, let $T_f(\mathbf{X}, \rho)$ be defined as in the proof of Theorem 4.4.1 as the $d \times n$ matrix whose i^{th} column is given by the $d \times 1$ vector:*

$$X_i + \frac{\nabla f(X_i)}{\max(f(X_i), \rho)} \qquad \text{for } i = 1, \ldots, n.$$

Then for every $f \in \mathcal{M}$, $0 < \rho \leq (2\pi)^{-d/2} e^{-3/2}$ and $x > 0$, we have

$$\mathbb{P}\left\{\left\|T_f(\mathbf{X}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho)\right\|_F \geq \mathbb{E}\left\|T_f(\mathbf{X}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho)\right\|_F + x\right\} \leq \exp\left(\frac{-x^2}{8L^4(\rho)}\right) \tag{A.15}$$

where

$$L(\rho) := \sqrt{\log \frac{1}{(2\pi)^d \rho^2}}$$

and $\bar{G}_n$ denotes the empirical measure corresponding to $\theta_1, \ldots, \theta_n$.

Proof of Lemma A.0.6. Let

$$F(\mathbf{X}) := \left\|T_f(\mathbf{X}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho)\right\|_F.$$

We shall prove that $F(\mathbf{X})$, as a function of $\mathbf{X}$, is Lipschitz with constant $2L^2(\rho)$ under the Frobenius matrix on $\mathbf{X}$ i.e.,

$$|F(\mathbf{X}) - F(\mathbf{Y})| \leq 2L^2(\rho) \|\mathbf{X} - \mathbf{Y}\|_F . \tag{A.16}$$

Inequality (A.0.6) would then directly follow from the standard concentration inequality for Lipschitz functions of Gaussian random vectors (see, for example, Boucheron, Lugosi, and Massart [19, Theorem 5.6]). To prove (A.16), note first that

$$\begin{aligned} |F(\mathbf{X}) - F(\mathbf{Y})| &= \left| \left\| T_f(\mathbf{X}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{X}, \rho) \right\|_F - \left\| T_f(\mathbf{Y}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{Y}, \rho) \right\|_F \right| \\ &\leq \|T_f(\mathbf{X}, \rho) - T_f(\mathbf{Y}, \rho)\|_F + \left\| T_{f_{\bar{G}_n}}(\mathbf{X}, \rho) - T_{f_{\bar{G}_n}}(\mathbf{Y}, \rho) \right\|_F . \end{aligned}$$

Note now that

$$\|T_f(\mathbf{X}, \rho) - T_f(\mathbf{Y}, \rho)\|_F^2 = \sum_{i=1}^{n} \|t_f(X_i, \rho) - t_f(Y_i, \rho)\|^2 \tag{A.17}$$

where

$$t_f(x, \rho) := x + \frac{\nabla f(x)}{\max(f(x), \rho)}.$$

To bound $\|t_f(X_i, \rho) - t_f(Y_i, \rho)\|$, we compute the Jacobian of the map $x \mapsto t_f(x, \rho)$ as

$$Jt_f(x, \rho) = \begin{cases} I_d + \frac{Hf(x)}{\rho} & \text{if } f(x) < \rho \\ I_d + \frac{Hf(x)}{f(x)} - \left(\frac{\nabla f(x)}{f(x)}\right)\left(\frac{\nabla f(x)}{f(x)}\right)^T & \text{if } f(x) > \rho \end{cases}$$

where ∇ and H denote gradient and Hessian respectively. We shall now argue that

$$0 \preceq Jt_f(x, \rho) \preceq L^2(\rho) I_d \tag{A.18}$$

where $A \preceq B$ means that $B - A$ is a nonnegative definite matrix. Before proving (A.18), let us first note that (A.18) implies

$$\|t_f(x, \rho) - t_f(y, \rho)\| \leq L^2(\rho) \|x - y\|$$

which further implies, via (A.17), that

$$\|T_f(\mathbf{X}, \rho) - T_f(\mathbf{Y}, \rho)\|_F^2 \leq L^2(\rho) \|\mathbf{X} - \mathbf{Y}\|_F^2 .$$

Since this inequality holds for every $f \in \mathcal{M}$, it also holds for $f_{\bar{G}_n}$ which gives (A.16) and completes the proof of Lemma A.0.6.

It remains to prove (A.18). For this, we shall use the above expression for $Jt_f(x, \rho)$ as well as inequality (A.1) from Lemma A.0.2 and inequality (A.5) from the proof of Lemma A.0.2. First when $f(x) > \rho$, note that

$$Jt_f(x, \rho) = I_d + \frac{Hf(x)}{f(x)} - \left(\frac{\nabla f(x)}{f(x)}\right)\left(\frac{\nabla f(x)}{f(x)}\right)^T$$

which is $\succeq 0$ from (A.5) and, by (A.1), we get

$$0 \preceq Jt_f(x,\rho) \preceq I_d + \frac{Hf(x)}{f(x)} \preceq tr\left(I + \frac{Hf(x)}{f(x)}\right) I_d \preceq L^2(f(x))I_d \preceq L^2(\rho)I_d$$

where, in the last inequality, we have used that $L(\cdot)$ is a decreasing function. Here tr denotes trace. This proves (A.18) when $f(x) > \rho$. Now let $f(x) < \rho$. Then

$$Jt_f(x,\rho) = I_d + \frac{Hf(x)}{\rho} = \left(1 - \frac{f(x)}{\rho}\right) I_d + \frac{f(x)}{\rho}\left(I_d + \frac{Hf}{f}\right)$$

which is $\succeq 0$ because $f(x) < \rho$ and because of (A.5). Also, by (A.1),

$$\begin{aligned} Jt_f(x,\rho) &= \left(1 - \frac{f(x)}{\rho}\right) I_d + \frac{f(x)}{\rho}\left(I_d + \frac{Hf}{f}\right) \\ &\preceq \left(1 - \frac{f(x)}{\rho}\right) I_d + \frac{f(x)}{\rho} I_d tr\left(I_d + \frac{Hf}{f}\right) \\ &\preceq \left(1 + \frac{f(x)}{\rho}\left(\log\frac{(2\pi)^{-d}}{f^2(x)} - 1\right)\right) I_d = \left(1 + \frac{f(x)}{\rho}\left(L^2(f(x)) - 1\right)\right) I_d \end{aligned}$$

The right hand side above is $\preceq L^2(\rho)I_d$ because $t \mapsto t(L^2(t) - 1)$ is non-decreasing on $t \in (0, (2\pi)^{-d/2}e^{-3/2}]$ so that when $f(x) < \rho$, we have

$$1 + \frac{f(x)}{\rho}\left(L^2(f(x)) - 1\right) \le L^2(\rho).$$

This proves (A.18) which completes the proof of Lemma A.0.6. □

Lemma A.0.7. *There exists a positive constant A_d depending only on d such that for every $M \ge 1$ and $d \in \{0, 1, 2, \dots\}$, we have*

$$I(d) := \int_{r \ge M} r^d e^{-r^2/2} dr \le A_d M^{d-1} e^{-M^2/2}. \tag{A.19}$$

Proof of Lemma A.0.7. Let $A_0 := 1$, $A_1 := 1$ and define A_d for $d \ge 2$ via the recursion $A_d := 1 + (d-1)A_{d-2}$. Clearly

$$I(0) = \int_{r \ge M} e^{-r^2/2} dr \le \int_{r \ge M} \frac{r}{M} e^{-r^2/2} = M^{-1} e^{-M^2/2}$$

and

$$I(1) = \int_{r \ge M} r e^{-r^2/2} dr = e^{-M^2/2}$$

and thus inequality (A.19) holds for $d = 0$ and $d = 1$. For $d \ge 2$, integration by parts gives

$$I(d) = M^{d-1} e^{-M^2/2} + (d-1) I(d-2).$$

Inequality (A.19) for $d \ge 2$ now easily follows by induction on d. □

Lemma A.0.8. *Let S be a compact subset of $\mathbb{R}^d$. For $\eta, M > 0$, define*

$$v(x) := \begin{cases} \eta & \text{if } x \in S^M \\ \eta \left(\frac{M}{\mathfrak{d}_S(x)}\right)^{d+1} & \text{otherwise} \end{cases} \tag{A.20}$$

Then, for some constant C_d depending only on d,

$$\int v(x)dx \leq C_d \eta \, Vol\left(S^M\right) \tag{A.21}$$

Proof of Lemma A.0.8. We first write

$$\int v(x)dx = \eta \mathrm{Vol}\left(S^M\right) + \eta M^{d+1} \int_{x \notin S^M} \frac{1}{\mathfrak{d}_S(x)^{d+1}} dx \tag{A.22}$$

Let N be the maximal integer such that there exist $u_1, \ldots, u_N \in S$ with

$$\min_{i \neq j} \|u_i - u_j\| \geq M/2. \tag{A.23}$$

The maximality of N implies that $\sup_{u \in S} \min_{1 \leq i \leq N} \|u - u_i\| \leq M/2$. As a result, for every $x \in \mathbb{R}^d$, by triangle inequality, we have

$$\mathfrak{d}_S(x) = \min_{u \in S} \|x - u\| \geq \min_{1 \leq i \leq N} \|x - u_i\| - \frac{M}{2}$$

so that

$$\begin{aligned}
\int_{x \notin S^M} \frac{dx}{(\mathfrak{d}_S(x))^{d+1}} &\leq \int_{x \notin S^M} \left(\frac{1}{\min_{1 \leq i \leq N} \|x - u_i\| - M/2}\right)^{d+1} dx \\
&\leq \sum_{i=1}^N \int_{x \notin S^M} \left(\frac{1}{\|x - u_i\| - M/2}\right)^{d+1} dx \\
&\leq \sum_{i=1}^N \int_{\|x - u_i\| \geq M} \left(\frac{1}{\|x - u_i\| - M/2}\right)^{d+1} dx \\
&= N \int_{\|x\| \geq M} \left(\frac{1}{\|x\| - M/2}\right)^{d+1} dx \\
&= N C_d \int_M^\infty \left(\frac{1}{r - M/2}\right)^{d+1} r^{d-1} dr \\
&= N C_d \int_{M/2}^\infty t^{-d-1} \left(\frac{M}{2} + t\right)^{d-1} dt \leq N C_d \int_{M/2}^\infty t^{-d-1} (2t)^{d-1} dt = \frac{N C_d 2^d}{M}.
\end{aligned} \tag{A.24}$$

Note now that because of (A.23), the balls $B(u_i, M/4), i = 1, \ldots, N$ have disjoint interiors and are all contained in $S^{M/4}$. As a result

$$N \leq \frac{\text{Vol}(S^{M/4})}{\text{Vol}(B(0, M/4))} \leq C_d \frac{\text{Vol}(S^M)}{M^d}. \tag{A.25}$$

The proof of Lemma A.0.8 is completed by putting together inequalities (A.22), (A.24) and (A.25). □

Lemma A.0.9. *There exists a positive constant C_d such that for every compact set $K \subseteq \mathbb{R}^d$ and real numbers $\epsilon > 0$ and $M > 0$, we have*

$$N(\epsilon, K) \leq C_d \epsilon^{-d} \, Vol(K^{\epsilon/2}) \tag{A.26}$$

and

$$\text{Vol}(K^{2M}) \leq C_d \text{Vol}(K^{\epsilon/2}) \left(1 + \frac{M}{\epsilon}\right)^d \tag{A.27}$$

Proof of Lemma A.0.9. Let us first prove (A.26). Let $a_1, \ldots, a_N \in K$ be a maximal set of points such that $\min_{i \neq j} \|a_i - a_j\| \geq \epsilon$. Then clearly $N(\epsilon, K) \leq N$. The balls $B(a_i, \epsilon/2)$ for $i = 1, \ldots, N$ have disjoint interiors and are all contained in $K^{\epsilon/2}$. As a result

$$N(\epsilon, K) \leq N \leq \frac{\text{Vol}(K^{\epsilon/2})}{\text{Vol}(B(0, \epsilon/2))} \tag{A.28}$$

from which (A.26) follows.

To prove (A.27), note that the K is contained in the union of the balls $B(a_i, \epsilon)$ for $i = 1, \ldots, N$. This implies that

$$K^{2M} \subseteq \cup_{i=1}^N B(a_i, \epsilon + 2M)$$

so that

$$\text{Vol}(K^{2M}) \leq N \text{Vol}(B(0, \epsilon + 2M)).$$

Inequality (A.28) then gives

$$\text{Vol}(K^{2M}) \leq \frac{\text{Vol}(K^{\epsilon/2})}{\text{Vol}(B(0, \epsilon/2))} \text{Vol}(B(0, \epsilon + 2M)) \leq C_d \text{Vol}(K^{\epsilon/2}) \left(1 + \frac{M}{\epsilon}\right)^d.$$

□

Lemma A.0.10. *Fix a probability measure G on $\mathbb{R}^d$ and let $0 < \rho \leq (2\pi)^{-d/2}/\sqrt{e}$. Let*

$$L(\rho) := \sqrt{\log \frac{1}{(2\pi)^d \rho^2}}.$$

Then there exists a positive constant C_d such that for every compact set $S \subseteq \mathbb{R}^d$, we have

$$\Delta(G, \rho) := \int \left(1 - \frac{f_G}{\max(f_G, \rho)}\right)^2 \frac{\|\nabla f_G\|^2}{f_G} \leq C_d N\left(\frac{4}{L(\rho)}, S\right) L^d(\rho)\rho + d \; G(S^c). \tag{A.29}$$

Proof of Lemma A.0.10. The proof uses Lemma A.0.11.

Fix a compact set S. Suppose first that G is supported on S so that the second term in (A.29) equals 0.

We consider two further special cases. First assume that S is contained in a ball of radius $a := 4/L(\rho)$. Without loss of generality, we may assume that the ball is centered at the origin. Because G is assumed to be supported on S, we have $\|\theta\| \leq a$ almost surely under G.

For $\theta \sim G$ and $X|\theta \sim N(\theta, I_d)$, we can write

$$\frac{\nabla f_G(x)}{f_G(x)} = \mathbb{E}\left(\theta - X | X = x\right)$$

so that

$$\frac{\|\nabla f_G(x)\|}{f_G(x)} = \|\mathbb{E}\left(\theta - X | X = x\right)\| \leq \mathbb{E}\left(\|\theta - X\| \, | X = x\right) \leq \|x\| + a. \tag{A.30}$$

Note also that

$$(2\pi)^{-d/2} \exp\left(-\frac{1}{2}\left(\|x\| + a\right)^2\right) \leq f_G(x) \leq (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\left(\|x\| - a\right)_+^2\right) \tag{A.31}$$

because $(\|x\| - a)_+ \leq \|x - \theta\| \leq \|x\| + a$ whenever $\|\theta\| \leq a$. This also implies that whenever $f_G(x) \leq \rho$, we have

$$\rho \geq (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\left(\|x\| + a\right)^2\right)$$

which gives

$$\|x\| + a \geq L(\rho) := \sqrt{\log \frac{1}{(2\pi)^d \rho^2}}. \tag{A.32}$$

Putting together (A.30), (A.31) and (A.32), we deduce that

$$\begin{aligned}
\Delta(G, \rho) &\leq \int \{f_G \leq \rho\} \left(\frac{\|\nabla f_G\|}{f_G}\right)^2 f_G \\
&\leq \int_{\{\|x\| + a \geq L(\rho)\}} \left(\|x\| + a\right)^2 (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\left(\|x\| - a\right)_+^2\right) dx.
\end{aligned}$$

Moving to polar coordinates, we deduce

$$\Delta(G, \rho) \leq C_d \int_{(L(\rho) - a)_+}^{\infty} (r + a)^2 \exp\left(-(r - a)_+^2/2\right) r^{d-1} dr.$$

Note now that with $a := 4/L(\rho)$ and $\rho \leq (2\pi)^{-d/2}/\sqrt{e}$, we have $4a \leq L(\rho)$ so that

$$\Delta(G, \rho) \leq C_d \int_{L(\rho) - a}^{\infty} (r + a)^2 \exp\left(-(r - a)^2/2\right) r^{d-1} dr.$$

By a change of variable $r - a \mapsto r$, we obtain

$$\Delta(G, \rho) \leq C_d \int_{L(\rho)-2a}^{\infty} (s + 2a)^2 \exp\left(-s^2/2\right) (s + a)^{d-1} ds.$$

Because $4a \leq L(\rho)$, we have

$$s + a \leq s + 2a \leq s + L(\rho) - 2a \leq 2s$$

whenever $s \geq L(\rho) - 2a$. Thus

$$\Delta(G, \rho) \leq C_d \int_{L(\rho)-2a}^{\infty} s^{d+1} e^{-s^2/2} ds.$$

By Lemma A.0.7, we deduce that

$$\Delta(G, \rho) \leq C_d (L(\rho))^d \exp\left(-\frac{1}{2}\left(L(\rho) - 2a\right)^2\right) \leq C_d (L(\rho))^d e^{2aL(\rho)} e^{-L^2(\rho)/2} = C_d \rho (L(\rho))^d e^{2aL(\rho)}.$$

We now take

$$a := \frac{4}{L(\rho)}$$

which gives

$$\Delta(G, \rho) \leq C_d \rho (L(\rho))^d \tag{A.33}$$

whenever G is supported on a set that is contained in a ball of radius $a = 4/L(\rho)$.

For the rest of the proof, we shall use Lemma A.0.11. Now suppose that G is supported on a general compact set S. Then, for $N := N(a, S)$ (where $a := 4/L(\rho)$), let $E_1, \ldots, E_N$ denote a disjoint covering of S such that each E_i is contained in a ball of radius a. We can then write

$$G := \sum_{j=1}^{N} w_j H_j$$

where $w_j := G(E_j)$ and H_j is the probability measure G conditioned on H_j. The bound (A.35) in Lemma A.0.11 then gives

$$\Delta(G, \rho) \leq \sum_{j=1}^{N} w_j \Delta(H_j, \rho/w_j).$$

Because H_j is supported on a ball of radius at most a, we can use (A.33) on each H_j to deduce that

$$\Delta(G, \rho) \leq C_d \sum_{j=1}^{N} w_j \frac{\rho}{w_j} L^d(\rho/w_j) \leq C_d \rho N(a, S) L^d(\rho). \tag{A.34}$$

To bound $\Delta(G, \rho)$ for an arbitrary probability measure G, we write

$$G = w_1 H_1 + w_2 H_2$$

where $w_1 = G(S) = 1 - w_2$ and H_1 and H_2 are the probability measures obtained by conditioning G on S and S^c respectively. Then clearly H_1 is supported on a compact set S so that the bound (A.34) can be used for $\Delta(H_2, \rho/w_2)$. For $\Delta(H_1, \rho/w_1)$, we use the trivial bound d (see the first part of Lemma A.0.11). This gives (via (A.35))

$$\Delta(G, \rho) \leq C_d G(S) N(a, S) L^d(\rho)\rho + d\ G(S^c) \leq C_d N(a, S) L^d(\rho)\rho + d\ G(S^c)$$

which completes the proof of Lemma A.0.10. □

Lemma A.0.11. *For a probability measure G on $\mathbb{R}^d$ and $\rho > 0$, let*

$$\Delta(G, \rho) := \int \left(1 - \frac{f_G}{\max(f_G, \rho)}\right)^2 \frac{\|\nabla f_G\|^2}{f_G}$$

The following pair of statements are then true.

1. *For every G and $\rho > 0$, we have $\Delta(G, \rho) \leq d$.*
2. *Suppose $G = \sum_{j=1}^m w_j H_j$ for some probability measures $H_1, \ldots, H_m$ and weights $w_1, \ldots, w_m$. Then*

$$\Delta(G, \rho) \leq \sum_{j=1}^m w_j \Delta(H_j, \rho/w_j). \tag{A.35}$$

Proof of Lemma A.0.11. To prove that $\Delta(G, \rho) \leq d$, note that if $\theta \sim G$ and $X|\theta \sim N(\theta, I_d)$, then

$$\frac{\nabla f_G(x)}{f_G(x)} = \mathbb{E}(\theta - X | X = x).$$

As a result

$$\Delta(G, \rho) \leq \int \frac{\|\nabla f_G\|^2}{f_G} = \mathbb{E}\|\mathbb{E}(\theta - X|X)\|^2 \leq \mathbb{E}\|\theta - X\|^2 = d.$$

For proving (A.35), note first that by the convexity of $x \mapsto \|x\|^2$, we have

$$\begin{aligned}
\frac{\|\nabla f_G\|^2}{f_G} &= \frac{\left\|\sum_j w_j \nabla f_{H_j}\right\|^2}{\sum_j w_j f_{H_j}} \\
&= \left\|\sum_j \left(\frac{w_j f_{H_j}}{\sum_j w_j f_{H_j}}\right) \frac{\nabla f_{H_j}}{f_{H_j}}\right\|^2 \left(\sum_j w_j f_{H_j}\right) \\
&\leq \left\{\sum_j \left(\frac{w_j f_{H_j}}{\sum_j w_j f_{H_j}}\right) \frac{\|\nabla f_{H_j}\|^2}{f_{H_j}^2}\right\} \left(\sum_j w_j f_{H_j}\right) = \sum_j w_j \frac{\|\nabla f_{H_j}\|^2}{f_{H_j}}.
\end{aligned}$$

This, along with the trivial inequality (here $a \vee b$ stands for $\max(a, b)$)

$$\left(1 - \frac{f_G}{f_G \vee \rho}\right)^2 \leq \left(1 - \frac{f_{H_j}}{f_{H_j} \vee (\rho/w_j)}\right)^2 \qquad \text{for every } 1 \leq j \leq m$$

yields (A.35). □

Lemma A.0.12. *Suppose $X_1, \ldots, X_n$ are independent observations with $X_i \sim N(\theta_i, I_d)$ for some $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$. Let the Oracle Bayes estimators $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ be defined as in (4.5) where $\bar{G}_n$ is the empirical measure of $\theta_1, \ldots, \theta_n$. Suppose that $\bar{G}_n$ is supported on a set $\{a_1, \ldots, a_k\}$ of cardinality k with $\bar{G}_n\{a_i\} = p_i$ for $i = 1, \ldots, k$ with $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Then*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2\right] \leq \frac{k-1}{2\sqrt{2\pi}} \sum_{j,l:j\neq l} (p_j + p_l) \left\|a_j - a_l\right\| \exp\left(-\frac{1}{8}\left\|a_j - a_l\right\|^2\right). \tag{A.36}$$

Proof of Lemma A.0.12. Note first that $\hat{\theta}_i^*$ has the following expression

$$\hat{\theta}_i^* = \frac{\sum_{j=1}^k a_j p_j \phi_d(X_i - a_j)}{\sum_{j=1}^k p_j \phi_d(X_i - a_j)} \qquad \text{for } i = 1, \ldots, n.$$

The above expression and the fact that $X_i - \theta_i \sim N(0, I_d)$ lets us write

$$\begin{aligned} R := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\hat{\theta}_i^* - \theta_i\right\|^2\right] &= \sum_{l=1}^k p_l \mathbb{E}\left\|\frac{\sum_{j=1}^k a_j p_j \phi_d(a_l + Z - a_j)}{\sum_{j=1}^k p_j \phi_d(a_l + Z - a_j)} - a_l\right\|^2 \\ &= \sum_{l=1}^k p_l \mathbb{E}\left\|\frac{\sum_{j=1}^k (a_j - a_l)\, p_j \phi_d(a_l + Z - a_j)}{\sum_{j=1}^k p_j \phi_d(a_l + Z - a_j)}\right\|^2 \\ &= \sum_{l=1}^k p_l \mathbb{E}\left\|\sum_{j:j\neq l} (a_j - a_l) w_{jl}(Z)\right\|^2 \end{aligned}$$

where $Z \sim N(0, I_d)$ and

$$w_{jl}(Z) := \frac{p_j \phi_d(a_l + Z - a_j)}{\sum_{u=1}^k p_u \phi_d(a_l + Z - a_u)} \qquad \text{for } 1 \leq j, l \leq k.$$

The elementary inequality $\left\|\sum_{i=1}^m \alpha_i\right\|^2 \leq m \sum_{i=1}^m \left\|\alpha_i\right\|^2$ for vectors $\alpha_1, \ldots, \alpha_m \in \mathbb{R}^d$ now lets us write

$$R \leq (k-1) \sum_{l=1}^k p_l \sum_{j:j\neq l} \left\|a_j - a_l\right\|^2 \mathbb{E} w_{jl}^2(Z). \tag{A.37}$$

We now bound $\mathbb{E} w_{jl}^2(Z)$ in the following way. Let

$$U := \left\{z \in \mathbb{R}^d : \left\|a_j - a_l\right\|^2 \geq 2\left\langle Z, a_j - a_l\right\rangle\right\}.$$

When $Z \notin U$, we shall use the trivial upper bound $w_{jl}^2(Z) \leq 1$. When $Z \in U$, we shall use the bound

$$w_{jl}^2(Z) \leq w_{jl}(Z) \leq \frac{p_j \phi_d(a_l + Z - a_j)}{p_l \phi_d(a_l + Z - a_l)} = \frac{p_j \phi_d(a_l + Z - a_j)}{p_l \phi_d(Z)}.$$

This gives

$$\mathbb{E} w_{jl}^2(Z) \leq \mathbb{P}\{Z \notin U\} + \int \frac{p_j \phi_d(a_l + z - a_j)}{p_l \phi_d(z)} I\{\|a_j - a_l\|^2 \geq 2\langle z, a_j - a_l\rangle\} \phi_d(z) dz$$

The change of variable $x = a_l + z - a_j$ in the integral above allows us to write

$$\begin{aligned}\mathbb{E} w_{jl}^2(Z) &\leq \mathbb{P}\left\{\langle Z, a_j - a_l\rangle > \frac{1}{2}\|a_j - a_l\|^2\right\} + \frac{p_j}{p_l}\mathbb{P}\left\{\langle Z, a_j - a_l\rangle \leq -\frac{1}{2}\|a_j - a_l\|^2\right\} \\ &\leq \left(1 + \frac{p_j}{p_l}\right)\left(1 - \Phi\left(\frac{1}{2}\|a_j - a_l\|\right)\right)\end{aligned}$$

where Φ is the standard univariate Gaussian cumulative distribution function. The bound $1 - \Phi(t) \leq \phi(t)/t$ for $t > 0$ now gives

$$\mathbb{E} w_{jl}^2(Z) \leq \frac{1}{2\sqrt{2\pi}}\left(1 + \frac{p_j}{p_l}\right)\frac{1}{\|a_j - a_l\|}\exp\left(-\frac{1}{8}\|a_j - a_l\|^2\right).$$

This bound, when combined with (A.37), yields (A.36) and hence completes the proof of Lemma A.0.12. □

Bibliography

[1] Jayadev Acharya et al. "Sample-optimal density estimation in nearly-linear time". In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2017, pp. 1278–1289.

[2] Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE Trans. Automatic Control* AC-19 (1974). System identification and time-series analysis, pp. 716–723. ISSN: 0018-9286.

[3] S. M Ali and S. D Silvey. "A general class of coefficients of divergence of one distribution from another". In: *Journal of the Royal Statistical Society, Series B* 28 (1966), pp. 131–142.

[4] André Altmann et al. "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10 (2010), pp. 1340–1347.

[5] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *Ann. Statist.* 45.1 (2017), pp. 77–120.

[6] I. Bárány and R. N. Karasev. "Notes about the Carathéodory number". In: *Discrete and Computational Geometry* 48 (2012), pp. 783–792.

[7] A. Barron. "Entropy and the central limit theorem". In: *Annals of Probability* 14 (1986), pp. 336–342.

[8] Andrew Barron, Lucien Birgé, and Pascal Massart. "Risk bounds for model selection via penalization". In: *Probab. Theory Related Fields* 113.3 (1999), pp. 301–413. ISSN: 0178-8051. DOI: 10.1007/s004400050210. URL: http://dx.doi.org/10.1007/s004400050210.

[9] Sumanta Basu et al. "Iterative Random Forests to detect predictive and stable high-order interactions". In: *arXiv preprint arXiv:1706.08457* (2017).

[10] Peter Benner, Volker Mehrmann, and Danny C Sorensen. *Dimension reduction of large-scale systems*. Vol. 45. Springer, 2005.

[11] Richard Berk et al. "Valid post-selection inference". In: *The Annals of Statistics* 41.2 (2013), pp. 802–837.

[12] Aditya Bhaskara, Ananda Suresh, and Morteza Zadimoghaddam. "Sparse solutions to nonnegative linear systems and applications". In: *Artificial Intelligence and Statistics*. 2015, pp. 83–92.

[13] Ella Bingham and Heikki Mannila. "Random projection in dimensionality reduction: applications to image and text data". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2001, pp. 245–250.

[14] L. Birgé and P. Massart. "From model selection to adaptive estimation". In: *A Festschrift for Lucien Le Cam*. Ed. by D. Pollard, E. Torgersen, and G. L. Yang. New York: Springer-Verlag, 1995, pp. 55–87.

[15] Adam Bloniarz et al. "Lasso adjustments of treatment effect estimates in randomized experiments". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7383–7390.

[16] Avrim Blum. "Random projection, margins, kernels, and feature-selection". In: *Subspace, Latent Structure and Feature Selection*. Springer, 2006, pp. 52–68.

[17] Dankmar Böhning. "A review of reliable maximum likelihood algorithms for semiparametric mixture models". In: *J. Statist. Plann. Inference* 47.1-2 (1995). Statistical modelling (Leuven, 1993), pp. 5–28. ISSN: 0378-3758. URL: https://doi.org/10.1016/0378-3758(94)00119-G.

[18] Dankmar Böhning. *Computer-assisted analysis of mixtures and applications*. Vol. 81. Monographs on Statistics and Applied Probability. Meta-analysis, disease mapping and others. Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. xii+260. ISBN: 0-8493-0385-0.

[19] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5. URL: https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[20] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[21] L. D. Brown. "Admissible estimators, recurrent diffusions, and insoluble boundary value problems". In: *Ann. Math. Statist.* 42 (1971), pp. 855–903. ISSN: 0003-4851. URL: https://doi.org/10.1214/aoms/1177693318.

[22] Lawrence D. Brown and Eitan Greenshtein. "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means". In: *Ann. Statist.* 37.4 (2009), pp. 1685–1704. ISSN: 0090-5364. URL: https://doi.org/10.1214/08-AOS630.

[23] Peter Bühlmann. "Statistical significance in high-dimensional linear models". In: *Bernoulli* 19.4 (2013), pp. 1212–1242.

[24] Siu-On Chan et al. "Efficient density estimation via piecewise polynomial approximation". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM. 2014, pp. 604–613.

[25] Siu-On Chan et al. "Learning mixtures of structured distributions over discrete domains". In: *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2013, pp. 1380–1394.

[26] Arindam Chatterjee and Soumendra Nath Lahiri. "Bootstrapping lasso estimators". In: *Journal of the American Statistical Association* 106.494 (2011), pp. 608–625.

[27] Gary K Chen et al. "Convex clustering: an attractive alternative to hierarchical clustering". In: *PLoS computational biology* 11.5 (2015), e1004228.

[28] Eric C Chi and Kenneth Lange. "Splitting methods for convex clustering". In: *Journal of Computational and Graphical Statistics* 24.4 (2015), pp. 994–1013.

[29] Thomas Cover and Joy Thomas. *Elements of Information Theory*. 2nd ed. Wiley, 2006.

[30] I. Csiszar. "A note on Jensen's inequality". In: *Studia Scientarium Mathematicarum Hungarica* 1 (1966), pp. 185–188.

[31] I. Csiszar. "Information-type measures of difference of probability distributions and indirect observations". In: *Studia Scientarium Mathematicarum Hungarica* 2 (1967), pp. 299–318.

[32] I. Csiszar. "On topological properties of $f-$divergences". In: *Studia Scientarium Mathematicarum Hungarica* 2 (1967), pp. 329–339.

[33] I. Csiszár and P. Shields. "Information theory and statistics: a tutorial". In: *Foundations and Trends in Communications and Information Theory* 1 (2004), pp. 417–528.

[34] Constantinos Daskalakis and Gautam Kamath. "Faster and sample near-optimal algorithms for proper learning mixtures of gaussians". In: *Conference on Learning Theory*. 2014, pp. 1183–1213.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 1–38.

[36] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer-Verlag, 2001.

[37] Ruben Dezeure et al. "High-dimensional Inference: Confidence intervals, p-values and R-Software hdi". In: *arXiv preprint arXiv:1408.4026* (2014).

[38] Lee H Dicker and Sihai D Zhao. "High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference". In: *Biometrika* 103.1 (2016), pp. 21–34.

[39] David Donoho and Galen Reeves. "Achieving Bayes MMSE performance in the sparse signal+ Gaussian white noise model when the noise level is unknown". In: *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE. 2013, pp. 101–105.

[40] Bradley Efron. "Tweedie's formula and selection bias". In: *J. Amer. Statist. Assoc.* 106.496 (2011), pp. 1602–1614. ISSN: 0162-1459. URL: https://doi.org/10.1198/jasa.2011.tm11181.

[41] Bradley Efron and Carl Morris. "Empirical Bayes on vector observations: An extension of Stein's method". In: *Biometrika* 59.2 (1972), pp. 335–347.

[42] Bradley Efron and Carl Morris. "Multivariate empirical Bayes and estimation of covariance matrices". In: *The Annals of Statistics* (1976), pp. 22–32.

[43] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Monographs on Applied Probability and Statistics. Chapman & Hall, London-New York, 1981, pp. x+143. ISBN: 0-412-22420-8.

[44] Alexei Fedotov, Peter Harremoës, and Flemming Topsoe. "Refinements of Pinsker's Inequality". In: *IEEE Transactions on Information Theory* 49 (2003), pp. 1491–1498.

[45] Long Feng and Lee H Dicker. "Nonparametric maximum likelihood inference for mixture models via convex optimization". In: *arXiv preprint arXiv:1606.02011* (2016).

[46] Dmitriy Fradkin and David Madigan. "Experiments with random projections for machine learning". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 517–522.

[47] Sara Van de Geer et al. "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3 (2014), pp. 1166–1202.

[48] Subhashis Ghosal and Aad van der Vaart. "Posterior convergence rates of Dirichlet mixtures at smooth densities". In: *Ann. Statist.* 35.2 (2007), pp. 697–723. ISSN: 0090-5364. URL: https://doi.org/10.1214/009053606000001271.

[49] Subhashis Ghosal and Aad W. van der Vaart. "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *Ann. Statist.* 29.5 (2001), pp. 1233–1263. ISSN: 0090-5364. URL: https://doi.org/10.1214/aos/1013203453.

[50] A. L. Gibbs and F. E. Su. "On choosing and bounding probability metrics". In: *International Statistical Review* 70 (2002), pp. 419–435.

[51] Gustavo L. Gilardoni. "On the minimum f-divergence for given total variation". In: *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math* 343 (2006), pp. 763–766.

[52] A. Guntuboyina. "Lower bounds for the minimax risk using f divergences, and applications". In: *IEEE Transactions on Information Theory* 57 (2011), pp. 2386–2399.

[53] Adityanand Guntuboyina. "Minimax Lower Bounds". Submitted. PhD thesis. Yale, 2011.

[54] P. Harremoës. *Convergence to the Poisson distribution in information divergence*. Tech. rep. Preprint Series, No. 2. University of Copenhagen, Copenhagen, Denmark, 2003.

[55] P. Harremoes and I. Vajda. "On pairs of f-divergences and their joint range". In: *IEEE Transactions on Information Theory* 57 (2011), pp. 3230–3235.

[56] Christina Heinze, Brian McWilliams, and Nicolai Meinshausen. "DUAL-LOCO: Distributing Statistical Estimation Using Random Projections". In: *arXiv preprint arXiv:1506.02554* (2015).

[57] Toby Dylan Hocking et al. "Clusterpath an algorithm for clustering using convex fusion penalties". In: *28th international conference on machine learning*. 2011, p. 1.

[58] Paul W Holland. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[59] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[60] Piotr Indyk and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM. 1998, pp. 604–613.

[61] William James and Charles Stein. "Estimation with quadratic loss". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1961. 1961, pp. 361–379.

[62] Adel Javanmard and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869–2909.

[63] Wenhua Jiang and Cun-Hui Zhang. "General maximum likelihood empirical Bayes estimation of normal means". In: *Ann. Statist.* 37.4 (2009), pp. 1647–1684. ISSN: 0090-5364. URL: https://doi.org/10.1214/08-AOS638.

[64] Iain M. Johnstone and Bernard W. Silverman. "Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences". In: *Ann. Statist.* 32.4 (2004), pp. 1594–1649. ISSN: 0090-5364. DOI: 10.1214/009053604000000030. URL: http://dx.doi.org/10.1214/009053604000000030.

[65] Nandakishore Kambhatla and Todd K Leen. "Dimension reduction by local principal component analysis". In: *Neural computation* 9.7 (1997), pp. 1493–1516.

[66] J. H. B. Kemperman. "On the optimum rate of transmitting information". In: *Probability and Information Theory*. Lecture Notes in Mathematics, 89, pages 126–169. Springer-Verlag, 1969.

[67] J. Kiefer and J. Wolfowitz. "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: *Ann. Math. Statist.* 27 (1956), pp. 887–906. ISSN: 0003-4851. URL: https://doi.org/10.1214/aoms/1177728066.

[68] Roger Koenker and Ivan Mizera. "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules". In: *Journal of the American Statistical Association* 109.506 (2014), pp. 674–685.

[69] S. Kullback. "A lower bound for discrimination information in terms of variation". In: *IEEE Transactions on Information Theory* 13 (1967), pp. 126–127.

[70] Nan Laird. "Nonparametric maximum likelihood estimation of a mixed distribution". In: *J. Amer. Statist. Assoc.* 73.364 (1978), pp. 805–811. ISSN: 0003-1291. URL: http://links.jstor.org/sici?sici=0162-1459(197812)73:364<805:NMLEOA>2.0.CO;2-J&origin=MSN.

[71] Danial Lashkari and Polina Golland. "Convex clustering with exemplar-based models". In: *Advances in neural information processing systems*. 2008, pp. 825–832.

[72] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.

[73] Jason D Lee and Jonathan E Taylor. "Exact post model selection inference for marginal screening". In: *Advances in Neural Information Processing Systems*. 2014, pp. 136–144.

[74] Jason D Lee et al. "Exact post-selection inference with the lasso". In: *arXiv preprint arXiv:1311.6238* (2013).

[75] Jerry Li and Ludwig Schmidt. "A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k". In: *arXiv preprint arXiv:1506.01367* (2015).

[76] Ping Li. "Sign Stable Random Projections for Large-Scale Learning". In: *arXiv preprint arXiv:1504.07235* (2015).

[77] Friedrich Liese. "ϕ-divergences, sufficiency, Bayes sufficiency, and deficiency". In: *Kybernetika* 48 (2012), pp. 690–713.

[78] Friedrich Liese and Igor Vajda. "On divergences and informations in statistics and information theory". In: *IEEE Transactions on Information Theory* 52 (2006), pp. 4394–4412.

[79] Winston Lin. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique". In: *The Annals of Applied Statistics* 7.1 (2013), pp. 295–318.

[80] Bruce G Lindsay. "Mixture models: theory, geometry and applications". In: *NSF-CBMS regional conference series in probability and statistics*. JSTOR. 1995, pp. i–163.

[81] Bruce G. Lindsay. "The geometry of mixture likelihoods: a general theory". In: *Ann. Statist.* 11.1 (1983), pp. 86–94. ISSN: 0090-5364. URL: https://doi.org/10.1214/aos/1176346059.

[82] Bruce G. Lindsay. "The geometry of mixture likelihoods. II. The exponential family". In: *Ann. Statist.* 11.3 (1983), pp. 783–792. ISSN: 0090-5364. URL: https://doi.org/10.1214/aos/1176346245.

[83] Bruce G. Lindsay and Mary L. Lesperance. "A review of semiparametric mixture models". In: *J. Statist. Plann. Inference* 47.1-2 (1995). Statistical modelling (Leuven, 1993), pp. 29–39. ISSN: 0378-3758. URL: https://doi.org/10.1016/0378-3758(94)00120-K.

[84] Fredrik Lindsten, Henrik Ohlsson, and Lennart Ljung. *Just relax and come clustering!: A convexification of k-means clustering*. Linköping University Electronic Press, 2011.

[85] Hanzhong Liu and Bin Yu. "Asymptotic properties of Lasso+ mLS and Lasso+ Ridge in sparse high-dimensional linear regression". In: *Electronic Journal of Statistics* 7 (2013), pp. 3124–3169.

[86] Richard Lockhart et al. "A significance test for the lasso". In: *Annals of statistics* 42.2 (2014), p. 413.

[87] Odalric Maillard and Rémi Munos. "Compressed least-squares regression". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1213–1221.

[88] Odalric-Ambrym Maillard and Rémi Munos. "Linear regression with random projections". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2735–2772.

[89] K. Marton. "A measure concentration inequality for contracting Markov chains". In: *Geometric and Functional Analysis* 6 (1996), pp. 556–571.

[90] K. Marton. "A simple proof of the Blowing-Up lemma". In: *IEEE transformations on Information theory* 32 (1986), pp. 445–446.

[91] K. Marton. "Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration". In: *Annals of Probability* 24 (1996), pp. 857–866.

[92] P. Massart. *Concentration inequalities and model selection. Lecture notes in Mathematics*. Vol. 1896. Berlin: Springer, 2007.

[93] Cathy Maugis and Bertrand Michel. "A non asymptotic penalized criterion for Gaussian mixture model selection". In: *ESAIM Probab. Stat.* 15 (2011), pp. 41–68. ISSN: 1292-8100. URL: https://doi.org/10.1051/ps/2009004.

[94] Cathy Maugis and Bertrand Michel. "Data-driven penalty calibration: a case study for Gaussian mixture model selection". In: *ESAIM Probab. Stat.* 15 (2011), pp. 320–339. ISSN: 1292-8100. URL: https://doi.org/10.1051/ps/2010002.

[95] C. Maugis-Rabusseau and B. Michel. "Adaptive density estimation for clustering with Gaussian mixtures". In: *ESAIM Probab. Stat.* 17 (2013), pp. 698–724. ISSN: 1292-8100. URL: https://doi.org/10.1051/ps/2012018.

[96] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[97] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[98] Nicolai Meinshausen. "Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.5 (2015), pp. 923–945.

[99] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473.

[100] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. "P-values for high-dimensional regression". In: *Journal of the American Statistical Association* (2012).

[101] Lucas Mentch and Giles Hooker. "Quantifying uncertainty in random forests via confidence intervals and hypothesis tests". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 841–881.

[102] ApS Mosek. "The MOSEK optimization toolbox for MATLAB manual". In: *Version 7.1 (Revision 28)* (2015), p. 17.

[103] F. Osterreicher and I. Vajda. "Statistical information and discrimination". In: *IEEE Trans. Inform. Theory* 39 (1993), pp. 1036–1039.

[104] Jérôme Paul, Michel Verleysen, Pierre Dupont, et al. "Identification of statistically significant features from random forests". In: *ECML workshop on Solving Complex Machine Learning Problems with Ensemble Methods*. 2013, pp. 69–80.

[105] Saurabh Paul et al. "Random projections for support vector machines". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. 2013, pp. 498–506.

[106] Robert R. Phelps. *Lectures on Choquet's theorem*. Van Nostrand, 1966.

[107] Mert Pilanci and Martin J Wainwright. "Randomized sketches of convex programs with sharp guarantees". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 921–925.

[108] M. S Pinsker. *Information and information stability of random variables and processes*. Moscow: Izv. Akad. Nauk, 1960.

[109] Peter Radchenko and Gourab Mukherjee. "Consistent clustering using an ℓ_1 fusion penalty". In: *arXiv preprint arXiv:1412.0753* (2014).

[110] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*. 2007, pp. 1177–1184.

[111] William M Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.

[112] M. D. Reid and R. C. Williamson. "Information, divergence and risk for binary experiments". In: *Journal of Machine Learning Research* 12 (2011), pp. 731–817.

[113] Mark D. Reid and Robert C. Williamson. "Generalized Pinsker Inequalities". In: *Proceedings of the 22nd Annual Conference on Learning Theory*. 2009.

[114] Herbert Robbins. "A GENERALIZATION OF THE METHOD OF MAXIMUM LIKELIHOOD-ESTIMATING A MIXING DISTRIBUTION". In: *Annals of Mathematical Statistics*. Vol. 21. 2. 1950, pp. 314–315.

[115] Herbert Robbins. "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1956, pp. 157–163. URL: http://projecteuclid.org/euclid.bsmsp/1200501653.

[116] Herbert Robbins. "Asymptotically subminimax solutions of compound statistical decision problems". In: *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1951. 1951, pp. 131–148.

[117] Herbert Robbins. "The empirical Bayes approach to statistical decision problems". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 1–20.

[118] Donald B Rubin. "Causal inference using potential outcomes". In: *Journal of the American Statistical Association* (2011).

[119] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[120] Walter Rudin. *Functional Analysis*. Second. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., 1991.

[121] Peter Schlattmann. *Medical applications of finite mixture models*. Springer, 2009.

[122] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[123] Jerzy Splawa-Neyman, DM Dabrowska, and TP Speed. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9". In: *Statistical Science* 5.4 (1990), pp. 465–472.

[124] Philip Stark. "Inference in infinite-dimensional inverse problems- Discretization and duality". In: *Journal of Geophysical Research* 97.B10 (1992), pp. 14055–14082.

[125] Charles M. Stein. "Estimation of the mean of a multivariate normal distribution". In: *Ann. Statist.* 9.6 (1981), pp. 1135–1151. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(198111)9:6<1135:EOTMOA>2.0.CO;2-5&origin=MSN.

[126] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC bioinformatics* 9.1 (2008), p. 307.

[127] Ananda Theertha Suresh et al. "Near-optimal-sample estimators for spherical gaussian mixtures". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1395–1403.

[128] Kean Ming Tan and Daniela Witten. "Statistical properties of convex clustering". In: *Electronic journal of statistics* 9.2 (2015), p. 2324.

[129] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[130] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[131] Andrei Nikolajevits Tihonov. "Solution of incorrectly formulated problems and the regularization method". In: *Soviet Math* 4 (1963), pp. 1035–1038.

[132] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1985, pp. x+243. ISBN: 0-471-90763-4.

[133] F. Topsøe. "Information theoretical optimization techniques". In: *Kybernetika* 15 (1979), pp. 8–27.

[134] Flemming Topsøe. "Some inequalities for information divergence and related measures of discrimination". In: *IEEE Trans. Inform. Theory* 46 (2000), pp. 1602–1609.

[135] Alexandre Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

[136] Aad Van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag, 1996.

[137] I. Vajda. "Note on discrimination information and variation". In: *IEEE Trans. Inform. Theory* 16 (1970), pp. 771–773.

[138] Santosh S Vempala. *The random projection method*. Vol. 65. American Mathematical Soc., 2005.

[139] Stefan Wager et al. "High-dimensional regression adjustments in randomized experiments". In: *Proceedings of the National Academy of Sciences* 113.45 (2016), pp. 12673–12678.

[140] Binhuan Wang et al. "Sparse convex clustering". In: *arXiv preprint arXiv:1601.04586* (2016).

[141] Larry Wasserman and Kathryn Roeder. "High dimensional variable selection". In: *Annals of statistics* 37.5A (2009), p. 2178.

[142] Michiko Watanabe and Kazunori Yamaguchi. *The EM algorithm and related statistical models*. CRC Press, 2003.

[143] Wing Hung Wong and Xiaotong Shen. "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs". In: *The Annals of Statistics* (1995), pp. 339–362.

[144] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[145] Chong Wu et al. "A New Algorithm and Theory for Penalized Regression-based Clustering". In: *Journal of Machine Learning Research* 17.188 (2016), pp. 1–25.

[146] Bin Yu. "Assouad, Fano, and Le Cam". In: *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*. Ed. by D. Pollard, E. Torgersen, and G. L. Yang. New York: Springer-Verlag, 1997, pp. 423–435.

[147] Cun-Hui Zhang. "Generalized maximum likelihood estimation of normal mixture densities". In: *Statistica Sinica* 19.3 (2009), p. 1297.

[148] Cun-Hui Zhang and Stephanie S Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217–242.

[149] Lijun Zhang et al. "Random Projections for Classification: A Recovery Approach". In: *Information Theory, IEEE Transactions on* 60.11 (2014), pp. 7300–7316.

[150] Lijun Zhang et al. "Recovering the optimal solution by dual random projection". In: *arXiv preprint arXiv:1211.3046* (2012).

[151] Shuheng Zhou, Larry Wasserman, and John D Lafferty. "Compressed regression". In: *Advances in Neural Information Processing Systems*. 2008, pp. 1713–1720.

[152] Changbo Zhu et al. "Convex optimization procedure for clustering: Theoretical revisit". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1619–1627.