Estimating Probability Distributions and their Properties

Shashank Singh

August 2019
CMU-ML-19-114

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Dr. Barnabás Póczos (Chair)
Dr. Ryan Tibshirani
Dr. Larry Wasserman
Dr. Bharath Sriperumbudur (Pennsylvania State University)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2019 Shashank Singh

This research was supported by the National Science Foundation award numbers DGE1252522 and DGE1745016, a grant from the Richard King Mellon Foundation, a grant from JP Morgan Chase Bank, and a grant from the Lockheed Martin Corporation.

Keywords: Nonparametric Statistics, Density Estimation, Density Functional Estimation, Entropy Estimation, Information Estimation, Divergence Estimation, Smoothness Estimation, Kozachenko-Leonenko Estimator, Wasserstein Distance, Integral Probability Metric, IPM, Holder Space, Sobolev Space, Besov Space, Nonparanormal Model, Gaussian Copula Model, Generative Adversarial Network, Implicit Generative Model, Explicit Generative Model, Statistical Minimax Theory



CARNEGIE MELLON UNIVERSITY

DOCTORAL THESIS

Estimating Probability Distributions and their Properties

Author:
Shashank SINGH

Advisor:
Dr. Barnabás PÓCZOS

Thesis Committee:
Dr. Barnabás Póczos (Chair)
Dr. Ryan Tibshirani
Dr. Larry Wasserman
Dr. Bharath Sriperumbudur (Pennsylvania State University)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Machine Learning Department
& Department of Statistics & Data Science

August 23, 2019

Declaration of Authorship

I, Shashank SINGH, declare that this thesis titled, "Estimating Probability Distributions and their Properties" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

This thesis studies several theoretical problems in nonparametric statistics and machine learning, mostly in the areas of *nonparametric density functional estimation* (estimating an integral functional of the population distribution from which the data are drawn) and *nonparametric density estimation* (estimating the entire population distribution from which the data are drawn). A consistent theme is that, although nonparametric density estimation is traditionally thought to be intractable in high-dimensions, several equally (or more) useful tasks are relatively more tractable, even with similar or weaker assumptions on the distribution.

Our work on density functional estimation focuses on several types of integral functionals, such as information theoretic quantities (entropies, mutual informations, and divergences), measures of smoothness, and measures of (dis)similarity between distributions, which play important roles as subroutines elsewhere in statistics, machine learning, and signal processing. For each of these quantities, under a variety of nonparametric models, we provide some combination of (a) new estimators, (b) upper bounds on convergence rates of these new estimators, (c) new upper bounds on the convergence rates of established estimators, (d) concentration bounds or asymptotic distributions for estimators, or (e) lower bounds on the minimax risk of estimation. We briefly discuss some applications of these density functional estimators to hypothesis testing problems such as two-sample (homogeneity) or (conditional) independence testing.

For density estimation, whereas the majority of prior work has focused on estimation under $\mathcal{L}^2$ or other $\mathcal{L}^p$ losses, we consider minimax convergence rates under several new losses, including the whole spectrum of Wasserstein distances and a large class of metrics called integral probability metrics (IPMs) that includes, for example, $\mathcal{L}^p$, total variation, Kolmogorov-Smirnov, earth-mover, Sobolev, Besov, and some RKHS distances. These losses open several new possibilities for nonparametric density estimation in certain cases; some examples include

- convergence rates with no or reduced dependence on dimension
- density-free distribution estimation, for data lying in general (e.g., non-Euclidean) metric spaces, or for data whose distribution may not be absolutely continuous with respect to Lebesgue measure
- convergence rates depending only on intrinsic dimension of data

Our main results here are the derivation of minimax convergence rates. However, we also briefly discuss several consequences of our results. For example, we show that IPMs have close connections with generative adversarial networks (GANs), and we leverage our results to prove the first finite-sample guarantees for GANs, in an idealized model of GANs as density estimators. These results may help explain why these tools appear to perform well at problems that are intractable from traditional perspectives of nonparametric statistics. We also briefly discuss consequences for estimation of certain density functionals, Monte Carlo integration of smooth functions, and distributionally robust optimization.

Acknowledgements

Despite the obligatory claim two pages ago that "this thesis is entirely my own work", this thesis is in fact the result of a collaboration between a great many individuals, who have contributed to my intellectual development or to my overall well-being over the years; without any of them this thesis might not exist in its present form. I would like to thank, among others and in no particular order,

- **Joel Bezaire**, **Sarah Ice**, **Cindy Crenshaw**, **Debbie Davies**, **Warren Davidson**, **Eric Appelt**, **Bill Rodriguez**, and especially **Adnan Rubai** for introducing me to the world of creative mathematics and science.
- **Diane Sorrel** and **Ann Wheeler** for guiding me beyond mathematics, to the world of literature, where I encountered many ideas that have helped shape my perspectives on life.
- **Justin Khim**, **Dan Dugmore**, **Reed Jones**, **Andrew Tiller**, **John Davidson**, and **Mark Arildsen** for teaching me to explore, from USN to the present; any critical thinking skills I possess were developed in the battlefield of our perpetual banter, discussion, debate, and argument.
- My stalwart and patient student collaborators at CMU, including **Manzil Zaheer**, **Chun-Liang Li**, and **Simon Du**, and especially **Ananya Uppal**, **Sabrina Rashid**, and **Yang Yang** for extended collaborations across several papers each.
- My cohorts in both Statistics and ML Departments for encouragement, discussion, and fun times along the way; especially to **Bryan Hooi** and **Yotam Hechtlinger** for project collaborations through which I learned a lot.
- **Bill Hrusa** for encouraging me to apply to graduate school, for putting up with me through three semesters of functional analysis and calculus of variations, and for agreeing to be on my undergraduate thesis committee. Of the many great professors by whom I was taught during my undergrad years, none were as careful as educators or as supportive of students.
- **Tai Sing Lee** for giving me my first exposure to the science of data analysis and thereby setting me on this path of studying machine learning.
- **Ziv Bar-Joseph** for allowing me to dip my toes in the field of mathematical biology, and for encouraging publishing in the best journals, even in the face of multiple rejections.
- **Saket Navlakha** for his collaborative spirit and for several thought-provoking discussions about algorithms in nature.
- **Diane Stidle** for ensuring that I always had a comfortable office, for ensuring that I felt like a full member of the ML department even when my office was elsewhere, and for being an amazingly knowledgeable and helpful reference through every stage of the PhD process.
- **Naman Tyagi** for his hospitality and guidance while I was an intern at Google (Mountain View, Summer 2015), and **Ashish Khetan** and **Zohar Karnin** for theirs while I was at Amazon (New York, Fall 2018).

- The **National Science Foundation (NSF)** for generously funding three years of my PhD, as well as the **Richard King Mellon Foundation** for funding a fourth. Beyond financial assistance, these fellowships were great confidence boosts along the way.

- **Alexandre Tsybakov** for his advice on working in the field of nonparametric statistics, for authoring my first and favorite statistics textbook, *Introduction to Nonparametric Estimation*, and for being willing to sign my copy thereof, when he visited CMU.

- **Erik Thiessen** and **Anna Fisher** for keeping an open mind towards what I hope has been a useful collaboration for all involved.

- **Larry Wasserman**, **Ryan Tibshirani**, and **Bharath Sriperumbudur** for agreeing to be on my thesis committee, for their many helpful suggestions, and for helping to shape my understanding of the practice and theory of nonparametric statistics.

- **Jian Ma** for engaging me in several interesting collaborative problems, and for supporting me (with both time and resources) as if I was his own advisee. Also, the members of Jian's lab for both helpful discussions and technical assistance.

- My advisor, **Barnabás Poczós**, for assigning me the perfect first problem to start my research, for allowing me incredible latitude to direct my own work and allocate my time, for being patient through my long periods without productivity, for helping me see the big picture of my work, and for always reminding me that curiosity is the ultimate driving force behind research. If I had to squeeze the university motto anywhere into this thesis, it would be to describe my advisor; his heart is in the work.

- My grandfather, **Krishan S. Chopra**, for surrounding me with wisdom long before I could recognize it, for teaching me that differing perspectives can always be reconciled, and for teaching me how to love my craft.

- My father, **Pradumna P. Singh**, for encouraging me to look beyond the horizon of what seems easily possible, for helping me transform my passions into a career, and for always having my back.

- My mother, **Sudha P. Singh**, for giving me the foundations of learning, reading, and reasoning, and for teaching me that any problem can be solved with calm, focused effort.

- My wife, **Jaeah Kim**, for keeping me young, healthy, adventurous, open-minded, and sane, for tolerating my often obsessive work habits, for sharing her brainspace, and for asking all the big questions.

Finally, innumerable causes that were entirely beyond the control of myself or anyone else conspired create this thesis. Post-hoc inferences aside, I can only conclude that I have been inexplicably lucky to have had the privilege of authoring this work.

Contents

Declaration of Authorship ii

Abstract iii

Acknowledgements iv

1 Introduction 1
1.1 Organization of this thesis . . . 2

2 Plug-in Kernel Density Estimates 9
2.1 Introduction . . . 9
2.1.1 Motivations and Goals . . . 9
2.1.2 Related Work . . . 10
2.2 Density Functional Estimator . . . 11
2.2.1 Notation . . . 11
2.2.2 Problem Statement . . . 12
2.2.3 Estimator . . . 12
2.3 Main Results . . . 12
2.4 Bias Bound . . . 14
2.4.1 Proof of Bias Bound . . . 14
2.5 Variance Bound . . . 15
2.5.1 Proof of Variance Bound . . . 16
2.6 Extension to Conditional Density Functionals . . . 16
2.6.1 Problem Statement, Assumptions, and Estimator . . . 17
2.6.2 Proof of Bounds for Conditional Density Functionals . . . 17
2.6.3 Application to Rényi-α Conditional Mutual Information . . . 18
2.7 Experimental Results . . . 19

3 Bias Corrected Fixed-k Nearest Neighbor Estimates 21
3.1 Introduction . . . 21
3.1.1 Some history . . . 21
3.1.2 Some intuition for BCFk estimators . . . 22
3.2 Introduction . . . 23
3.3 Problem statement and notation . . . 25
3.4 Estimator . . . 26
3.4.1 k-NN density estimation and plug-in functional estimators . . . 26
3.4.2 Fixed-k functional estimators . . . 26
3.5 Related work . . . 27
3.5.1 Estimating information theoretic functionals . . . 27
3.5.2 Prior analysis of fixed-k estimators . . . 28
3.6 Discussion of assumptions . . . 29
3.7 Preliminary lemmas . . . 30
3.8 Main results . . . 31

3.9 Conclusions and discussion . . . 33
3.10 A More General Setting . . . 33
3.11 Proofs of Lemmas . . . 34
3.11.1 Applications of Lemma 4 . . . 37
3.12 Proof of Bias Bound . . . 38
3.13 Proof of Variance Bound . . . 40

4 Nonparanormal Information Estimation **42**
4.1 Introduction . . . 42
4.2 Problem statement and notation . . . 42
4.3 Related Work and Our Contributions . . . 45
4.3.1 The Nonparanormal . . . 45
4.3.2 Information Estimation . . . 45
4.4 Nonparanormal Information Estimators . . . 47
4.4.1 Estimating Σ by Gaussianization . . . 47
4.4.2 Estimating Σ via Rank Correlation . . . 47
4.4.3 Regularization and estimating I . . . 48
4.5 Upper Bounds on the Error of $\widehat{I}_{\rho,z}$ . . . 49
4.6 Lower Bounds in terms of Σ . . . 50
4.7 Empirical Results . . . 51
4.8 Estimating Entropy . . . 53
4.9 Conclusions and Future Work . . . 54
4.10 Lemmas . . . 55
4.11 Proofs of Main Results . . . 56
4.12 Upper bounds on the MSE of $\widehat{I}_{\rho}$ . . . 56
4.12.1 Lower bound for rank-based estimators in terms of Σ . . . 58
4.13 Details of Experimental Methods . . . 59
4.13.1 Computational Considerations . . . 59
4.14 Additional Experimental Results . . . 59
4.14.1 Effects of Other Marginal Transformations . . . 59
4.15 Specific Assumptions for Estimating $H(X)$ . . . 60
4.16 Lower bounding the eigenvalues of a bandable matrix . . . 60

5 Fourier-weighted Quadratic Functionals **62**
5.1 Introduction . . . 62
5.1.1 Motivating Examples . . . 64
5.2 Notation . . . 65
5.3 Formal Problem Statement . . . 66
5.4 Related Work . . . 66
5.4.1 Prior work on special cases . . . 66
5.4.2 Estimation of related functionals . . . 68
5.4.3 Applications . . . 69
5.5 Upper Bounds . . . 70
5.5.1 Proposed Estimators . . . 70
5.5.2 Bounding the risk of $\widehat{S}_Z$ . . . 71
5.5.3 Discussion of Upper Bounds . . . 72
5.6 Lower Bounds . . . 73
5.7 Special Cases . . . 73
5.7.1 Sobolev . . . 74
5.7.2 Gaussian RKHS . . . 74
5.7.3 Exponential RKHS . . . 75

5.7.4 Logarithmic decay . . . 76
5.7.5 Sinc RKHS . . . 76
5.8 Discussion . . . 76
5.9 Proofs . . . 78
5.9.1 Proof of Proposition 24 . . . 78
5.9.2 Proof of Proposition 25 . . . 79
General Proof Setup . . . 79
Variance Bounds in the Fourier Basis . . . 81
5.9.3 Proof of Theorem 28 . . . 83

6 Wasserstein Convergence of the Empirical Measure 88
6.1 Introduction . . . 88
6.2 Background . . . 88
6.3 Notation and Problem Setting . . . 90
6.3.1 Problem Setting . . . 90
6.3.2 Definitions for Stating our Results . . . 91
6.4 Related Work . . . 93
6.4.1 Upper Bounds . . . 94
6.4.2 Lower Bounds . . . 96
6.5 Example Applications . . . 96
6.6 Conclusion . . . 100
6.6.1 Future Work . . . 100
6.7 Preliminary Lemmas and Proof Sketch of Theorem 38 . . . 101
6.8 Proof Sketch of Theorem 38 . . . 104
6.9 Proofs of Lemmas . . . 106
6.10 Proof of Lower Bound . . . 110
6.11 Proofs of Minimax Lower Bound in terms of Moment Bounds . . . 113

7 Distribution Estimation under Adversarial Losses 115
7.1 Introduction . . . 115
7.2 Background . . . 116
7.2.1 Adversarial Losses . . . 116
7.3 Problem Statement and Notation . . . 118
7.3.1 Notation . . . 118
7.4 Related Work . . . 119
7.5 Upper Bounds for Orthogonal Series Estimators . . . 121
7.6 Minimax Lower Bound . . . 122
7.7 Examples . . . 123
7.8 Consequences for Generative Adversarial Neural Networks (GANs) . . . 124
7.9 Minimax Comparison of Explicit and Implicit Generative Models . . . 125
7.9.1 A Minimax Framework for Implicit Generative Models . . . 125
7.9.2 Comparison of Explicit and Implicit Generative Models . . . 126
7.10 Conclusions . . . 128
7.11 Further Related Work . . . 128
7.11.1 Other statistical analyses of GANs . . . 129
7.12 Proof of Upper Bound . . . 130
7.13 Proof of Lower Bound . . . 133
7.14 Proofs and Further Discussion of Applications in Section 7.7 . . . 135
7.14.1 Wavelet Basis . . . 138
7.15 Proofs and Applications of Explicit & Implicit Generative Modeling Results (Section 7.9) . . . 139

7.15.1 Proofs of Theorem 59 and Converse . . . 139
7.15.2 Applications . . . 141
7.16 Experimental Results . . . 143
7.17 Future Work . . . 143

8 Open Questions, Preliminary Results, and Future Work 145
8.1 Distribution estimation under Besov IPM losses . . . 145
8.1.1 Summary of Results for Besov IPMs . . . 146
8.2 Some further implications of convergence rates for density estimation under IPMs . . . 149
8.2.1 Monte Carlo Integration . . . 149
8.2.2 Distributionally Robust Optimization . . . 150
8.3 Asymptotic Distributions for BCF-k Estimators . . . 151

A Other Projects 153
A.1 Distributed Gradient Descent and Bacterial Foraging . . . 153
A.2 Sequence-based Prediction of Enhancer-Promoter Interactions . . . 154
A.3 Reconstruction Risk of Convolutional Sparse Dictionary Learning . . . 155
A.4 A Hidden Markov Model for Eye-Tracking Data Analysis . . . 156
A.5 DARC: Differentiable Architecture Compression . . . 158

B A Condensed Summary of Results on Density Functional Estimation 159

Bibliography 162

List of Figures

2.1 A data point $x^1 \in C_{(1,2,\emptyset,\emptyset)} \subseteq [0,1]^2$ (using the notation of Singh and Póczos (2014b)), along with its three reflected copies.The sum of the integrals over $\mathcal{X}$ of (the absolute values of) the four kernels (with shaded support) is $\|K\|_1^2$. 13
2.2 Two possible graphs of dependence between variables X, Y, and Z. The left graph corresponds to $I(X;Y|Z) = 0$, whereas the right graph corresponds to $I(X;Y|Z) > 0$. These can thus be distinguished using an estimate of conditional mutual information. 19
2.3 Log-log plot of squared error (averaged over 100 IID trials) of our Rényi-0.8 estimator for various sample sizes n, alongside our theoretical bound. Error bars indicate standard deviation of estimator over 100 trials. 19

3.1 Illustration of kNN density estimation at a point (red) with $k = 3$, $n = 10$, $D = 2$. 23
3.2 Illustrations of kissing numbers $N_{1,2} = 6$ and $N_{1,3} = 12$, which bound the number of points for which any fixed point can be the nearest neighbor. The existence of such a constant, together with the Efron-Stein inequality, form the basis for bounds on the variance of BCFk estimators. 40

4.1 Surface and contour plots of three example nonparanormal densities. Figure taken from Liu, Lafferty, and Wasserman (2009). 44
4.2 Plots of $\log_{10}$(MSE) plotted over (a) log-sample-size $\log_{10}(n)$, (b) fraction α of dimensions with non-Gaussian marginals, (c) fraction β of outlier samples in each dimension, and (d) covariance $\Sigma_{1,2} = \mathrm{Cov}(X_1, X_2)$. Note that the x-axis in (d) is decreasing. 52
4.3 Cartoon phase diagram showing when each kind of estimator (Gaussian, nonparanormal, fully nonparametric) estimator can be useful. Nonparanormal estimators can help fill the large gap in the setting of moderately high-dimensional data with non-Gaussian marginal distributions. 54
4.4 Semi-log plot of mean squared error of various estimators over the fraction of non-Gaussian marginals $\alpha \in [0,1]$, for various marginal transforms T. 60

7.1 Examples of probability distributions P and Q and corresponding discriminator functions f^*. In (a), P and Q are single Dirac masses at $+1$ and -1, respectively, and $\mathcal{F}$ is the 1-Lipschitz class, so that $d_{\mathcal{F}}$ is the Wasserstein metric. In (b), P and Q are standard Gaussian and standard Laplace distributions, respectively, and $\mathcal{F}$ is a ball in an RKHS with a Gaussian kernel, so that $d_{\mathcal{F}}$ is the Gaussian Maximum Mean Discrepancy (MMD). 117

7.2 Simple synthetic experiments to showcase the tightness of our bound on convergence rates under adversarial losses in the Sobolev case. . . . 143

8.1 Emission spectrum of a metal halide lamp, shown as an example of a density function with very inhomogeneous smoothness. Any linear density estimator, depending on its tuning, will either over-fit the data in smooth regions (e.g., around 500nm or 700nm) or under-fit the data in spiky regions (e.g., around 550nm or 600nm). On the other hand, a nonlinear wavelet-thresholding estimator can adaptively allocate a finer representation to the spikier portions of the density. *Image credits: Philips Lighting* (`https://commons.wikimedia.org/wiki/File:MHL.png`), "MHL", `https://creativecommons.org/licenses/by-sa/2.5/nl/deed.en` 146
8.2 The first five levels of the Haar wavelet basis, the simplest of wavelet basis. The level (or scale) is indexed by $j \in \mathbb{N}$, while the shift (or offset) is indexed by $k \in [2^j]$. Note that our results actually require using slightly smoother wavelets. 147
8.3 Minimax convergence rates as functions of discriminator smoothness σ_d and distribution function smoothness σ_g, for (a) general and (b) linear estimators, in the case dimension $D = 4$, and Besov parameters $p_d = 1.2$, $p_g = 2$. Color shows exponent α of minimax convergence rate $n^{-\alpha}$, ignoring polylogarithmic factors. 149

A.1 Terrain model for bacterial food search. Obstacles are placed are regular intervals, and the food source is at the center of the region; contours display the diffusion of the food source gradient. The bacterial swarm, in the bottom right area, is depicted as a set of black points, each corresponding to an individual cell. 154
A.2 Schematic of the deep learning model, SPEID, that we designed to predict enhancer-promoter interactions from DNA sequence. Key steps involving rectification, batch normalization, and dropout are annotated. Note that the final output step is essentially a logistic regression in SPEID which provides a probability to indicate whether the input enhancer element and promoter element interact. 155
A.3 Illustration of how, in a convolutional dictionary model, a long, rich signal ("True Data", black) can be decomposed into a sum of convolutions of long, sparse signals ("Encoding", red/blue) with short, simple signals ("Patterns", orange/green). 156
A.4 An example trial of the standard TrackIt task (endogenous condition), on a 4×4 grid with 4 distractor objects. The target object here is the grey triangle, as indicated before the trial. Videos of example TrackIt trials can be found at `https://github.com/sss1/eyetracking/tree/master/videos`. 156
A.5 (a) Graphical model schematic of HMM. The initial state (object) $S(1)$ is sampled uniformly at random. At each time point t, we observe a gaze data point $X(t)$, distributed according to a Gaussian centered around the state $S(t)$. At the next time point $t+1$, a new state $S(t+1)$ is sampled according to a distribution depending on $S(t)$, and the process repeats. (b) Example conditional distribution of $E(t)$ given $S(t) =$ "Blue Moon". 157

A.6 Example of a "cell" (here, a mixture of a full convolution, a depthwise separable (DS) convolution, and a shift operation), used as the basic network building block in DARC. A cost-weighted ℓ_1 penalty is placed on α during training. After training, operations with corresponding $\alpha_j = 0$ are removed from the network. 158

List of Tables

3.1 Functionals with known bias-corrected k-NN estimators, their bias corrections, and references. All expectations are over $X \sim P$. $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx$ is the gamma function, and $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$ is the digamma function. $\alpha \in \mathbb{R}\backslash\{1\}$ is a free parameter. *For KL divergence, bias corrections for p and q cancel. 25

5.1 Minimax convergence rates for different combinations of a_z and b_z. Results are given up to $\log n$ factors, except the case when both a_z and b_z are logarithmic, which is given up to $\log\log n$ factors. Note that, in this last case, only the upper bound is known. A value of ∞ indicates that the estimand itself may be ∞ and consistent estimation is impossible. 77

B.1 Table of density functionals studied in this thesis. 'CI' indicates the existence of a concentration inequality around the estimator's mean. 'CLT' indicates the existence of a central limit theorem (under additional assumptions). 'Minimax' indicates that the convergence rate matches known minimax lower bounds (up to polylogarithmic factors), for all values for s and d. 's-Adaptive' (resp., 'd-Adaptive') indicates that the estimator does not require knowledge of the true smoothness s (resp., the true support dimension d) of the density. Results in green are novel contributions of this thesis. 'Intrinsic d' indicates that d denotes the *intrinsic* dimension of the support of the density (which is often much smaller than the *ambient* data dimension) 161

List of Abbreviations

BCFk	**B**ias-**C**orrected **F**ixed-k
CSDL	**C**onvolutional **S**parse **D**ictionary **L**earning
DRO	**D**istributionally **R**obust **O**ptimization
GAN	**G**enerative **A**dversarial **N**etwork
HMM	**H**idden **M**arkov **M**odel
IID	**I**ndependent and **I**dentically **D**istributed
IPM	**I**ntegral **P**robability **M**etric
KDE	**K**ernel **D**ensity **E**stimate
KL	**K**ozachenko-**L**eonenko (Chapter 3) or **K**ullback-**L**eibler (elsewhere)
k**-NN**	k-**N**earest **N**eighbor
MSE	**M**ean **S**quared **E**rror

Chapter 1

Introduction

This thesis is broadly concerned with a theoretical perspective on the following question:

> Given some data, when and how precisely can we estimate the data's underlying population distribution, or some property thereof, under minimal prior assumptions on the population?

Since this problem is, more or less, the concern of all of statistics, I'll now be more specific. Suppose we observe n IID samples $X_{1:n} = X_1, ..., X_n \overset{IID}{\sim} P$ from an unknown probability distribution P lying in a nonparametric class $\mathcal{P}$ of distributions. Within the minimax estimation framework, this thesis addresses special cases of several basic statistical problems (listed here in decreasing order of my focus):

1. **Distribution Functional Estimation:** Given a (known, nonlinear) functional $F : \mathcal{P} \to \mathbb{R}$, we want to estimate its value $F(P)$ at the unknown distribution P. That is, we want to compute a function $\widehat{F} : \mathcal{X}^n \to \mathbb{R}$ that has small worst-case $\mathcal{L}^2$ risk:
$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\left(F(P) - \widehat{F}(X_{1:n}) \right)^2 \right].$$

2. **Distribution Estimation:** Given a loss function $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty)$, we want to estimate the entire distribution P. That is, we want to compute a function $\widehat{P} : \mathcal{X}^n \to \mathcal{P}$ that has small worst-case risk under ℓ:
$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\ell \left(P, \widehat{P}(X_{1:n}) \right) \right].$$

3. **Implicit Distribution Estimation (a.k.a., Sampling):** Given a loss function $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ and a "latent" random variable Z with a known distribution on a space $\mathcal{Z}$, we want to learn a transformation f such that the distribution of $f(Z)$ is close to P. That is, we want to compute a function $\widehat{f} : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ such that, if $P_{\widehat{f}(X_{1:n}, Z) | X_{1:n}} \in \mathcal{P}$ is the conditional distribution of $\widehat{f}(X_{1:n}, Z)$ given $X_{1:n}$, then
$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X \sim P} \left[\ell \left(P, P_{\widehat{f}(X_{1:n}, Z) | X_{1:n}} \right) \right].$$

4. **Hypothesis Testing:** Given a partition $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ into two disjoint subsets, we would like to determine whether $P \in \mathcal{P}_0$ or $P \in \mathcal{P}_1$, under a constraint on the Type 1 error probability. That is, given a level $\alpha \in (0, 1)$, we would like to compute a test statistic $\widehat{T} : \mathcal{X}^n \to \{\mathcal{P}_0, \mathcal{P}_1\}$ that has high power
$$\inf_{P \in \mathcal{P}} \Pr_{X_{1:n} \overset{IID}{\sim} P} \left[\widehat{T}(X_{1:n}) = \mathcal{P}_1 \right]$$

whenever $P \in \mathcal{P}_1$, subject to the type-1 error bound

$$\sup_{P \in \mathcal{P}_0} \Pr_{X_{1:n} \overset{IID}{\sim} P} \left[\widehat{T}(X_{1:n}) = \mathcal{P}_1 \right] \leq \alpha.$$

In each of the above problems, certain parameters need to be specified to give a well-defined statistical problem. The first is the hypothesis class $\mathcal{P}$ of distributions under consideration. Second, each problem has its own specific parameters that need to be fixed: the functional F, the loss ℓ, the latent variable Z, or the null hypothesis $\mathcal{P}_0$.

In most of this thesis, the class $\mathcal{P}$ of distributions will be a bounded set in some nonparametric function space, such as a Hölder, Sobolev, Besov, or reproducing kernel Hilbert space. A variety of choices of parameters F, ℓ, etc. are considered, based largely on a variety of downstream applications.

1.1 Organization of this thesis

The majority of this document was produced by collecting (read: copying and pasting large portions of) several papers into a single document, and most of the chapters are organized in the manners of the original papers, with the addition of a short introductory section at the beginning of each chapter. The resulting thesis has the benefit that each chapter can be read in relative isolation. The remainder the current chapter seeks to bind these into a somewhat coherent whole.

We now give a brief overview of each chapter, including notes on the papers through which the contents of the chapter were originally disseminated.

Chapter 2 This chapter presents my first work, conducted in 2013-2014, on density functional estimation, based on Barnabás' suggestion that I generalize a convergence rate and concentration inequality (due to Liu, Wasserman, and Lafferty (2012)) for estimating the entropy of a 2-dimensional random variable under 2^{nd}-order Hölder assumptions. I first generalized this rate to estimation of Rényi-α divergences in arbitrary dimensions, under general Hölder conditions, a task which primarily relied on careful exposition in terms of general-order multivariate Taylor series. I then further generalized this to a broad class of smooth functionals, including, as examples, $\mathcal{L}^p$ distances, many entropy, mutual information, and divergence measures, and their conditional analogues. These results were published in two papers:

- Shashank Singh and Barnabás Póczos (2014b). "Generalized Exponential Concentration Inequality for Rényi Divergence Estimation". In: *Proceedings of The 31st International Conference on Machine Learning*, pp. 333–341
- Shashank Singh and Barnabás Póczos (2014a). "Exponential concentration of a density functional estimator". In: *Advances in Neural Information Processing Systems*, pp. 3032–3040

At the time, not much was known in the machine learning community about estimation of integral functionals of densities, and the approach in these papers is somewhat elementary: essentially a kernel density estimate of the data is plugged into the functional of interest. For many of the information theoretic functionals, this approach requires assuming that the density was lower bounded away from 0, enforcing in turn that its support was bounded. The

primary challenge then involves correcting for boundary bias along the edge of the density's support; this is performed using a mirroring trick, which relies on strong assumptions on the behavior of the density near the boundary.

Of the two papers listed above, the results in the latter paper largely supersede those in the former paper (albeit in somewhat less detail), and so this chapter presents only material from the latter paper. I am obliged to note that this work formed the bulk of my undergraduate honors thesis, in part for which I was granted an MS degree in Mathematical Sciences, and I am including it in this thesis only because it is thematically quite revelant. Hopefully, this thesis contains enough novel content to justify the PhD degree even with the exclusion of this chapter.

Chapter 3 This chapter presents work, conducted from 2015-2016, on a family of k-nearest neighbor entropy and divergence estimators, based on classic work of Kozachenko and Leonenko (1987). While quite popular in practice, these estimators had, at the time very limited theoretical guarantees, and their convergence rate was unknown (except in a very particular case studied by (Tsybakov and Meulen, 1996)). For these estimators, I derived the first general convergence rates for smooth densities in arbitrary dimensions. These results were released in a technical report, on the special case of entropy estimation, and a 2016 NeurIPS paper on general smooth functionals; Chapter 3 presents the latter paper:

- Shashank Singh and Barnabás Póczos (2016a). "Analysis of k-Nearest Neighbor Distances with Application to Entropy Estimation". In: *arXiv preprint arXiv:1603.08578*
- Shashank Singh and Barnabás Póczos (2016b). "Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functional Estimators". In: *Advances in Neural Information Processing Systems*, pp. 1217–1225

Interestingly (and as far as I know, purely coincidentally), two other papers (Berrett, Samworth, and Yuan (2019) and Gao, Oh, and Viswanath (2017a)) appeared later that year providing closely related results – this was somewhat remarkable, given the nearly 30 years that has passed since the proposal of the estimator by Kozachenko and Leonenko (1987). In retrospect, my analysis of these estimators is rather delicate, requiring both bounded support and quite specific conditions on the density near the boundary of its support. Berrett, Samworth, and Yuan (2019), although specific to the case of entropy estimation, gave a more nuanced analysis, allowing certain distributions with unbounded support and showing not only a central limit theorem but also asymptotic efficiency. Gao, Oh, and Viswanath (2017a) showed slower rates (due to a less careful analysis of boundary bias), but also extended their results to the widely used mutual information estimator of Kraskov, Stögbauer, and Grassberger (2004), providing what, as far as I know, continue to be the only rates for that estimator. Thus, all three papers provided essentially distinct results for these estimators. Finally, we note that quite a bit of progress has since been made based on these three papers; state-of-the-art results for these estimators can be found in Berrett and Samworth (2019) and Jiao, Gao, and Han (2018).

Chapter 4 In the previous chapters, we strove to make minimal assumptions on the distribution of the data, focusing on Hölder- or Sobolev-type smoothness assumptions. Unfortunately, minimax convergence rates under these weak assumptions scale very poorly with the data dimension D; minimax lower bounds imply that the number of samples required to guarantee an MSE of at most $\epsilon > 0$ scales, for some constant $c > 0$, as ϵ^{-cD}. Quite simply, these spaces are too large to estimate their parameters except in very low dimensions.

This chapter presents an attempt to scale the estimation of information-theoretic quantities to higher dimensions, by using a smaller semiparametric model, the Gaussian copula (a.k.a., "nonparanormal") model. We focus on multivariate mutual information, in terms of which most other quantities can be expressed. In particular (since there are a number of distinct generalizations of mutual information to more than two variables), we consider estimating the difference between the sum of marginal entropies and the joint entropy:

$$I(X) := \mathop{\mathbb{E}}_{X \sim p}\left[\log\left(\frac{p(X)}{\prod_{j=1}^{D} p_j(X_j)}\right)\right] = \sum_{j=1}^{D} H(X_j) - H(X).$$

To do this, we propose 3 distinct estimators, based on different estimates of the latent covariance structure of the data. For two of these estimators, we prove error bounds of order D^2/n, far better (for large D) than rates achievable by nonparametric estimators. We also experimentally show, in a number of simulations, that the proposed estimators can do well in moderately high-dimensional settings, in which most fully nonparametric estimates are uninformative. We also show that these estimators are relatively robust, compared to optimal estimators for the perfectly Gaussian case, which fail dramatically when the data deviate even moderately from Gaussian. These results were published in ICML 2017:

- Shashank Singh and Barnabás Póczos (2017). "Nonparanormal Information Estimation". In: *International Conference on Machine Learning*, pp. 3210–3219

Chapter 5 Many results in nonparametric statistics rely on the assumption that the data distribution exhibits some degree of smoothness, typically in some sense of having sufficiently small derivatives. This assumption is essentially never verified in practice, and this caused me to become interested in estimating smoothness parameters of a data density. Motivated by this idea, Chapter 5 presents some work on estimating (semi-)inner products, (semi-)norms, and (pseudo-)metrics between densities lying in Hilbert spaces, such as Sobolev-Hilbert and certain reproducing kernel Hilbert spaces. Salient examples include (squared) $\mathcal{L}^2$ norms of the derivatives of the density; e.g., in 1 dimension, for integer s,

$$\|p\|_{\mathcal{H}^s}^2 = \|p^{(s)}\|_2^2 = \int \left(\frac{d^s}{dx^s} p(x)\right)^2 dx.$$

These and more general quantities can be re-expressed in terms of the characteristic function of the density p. We initially published work on estimating Sobolev-Hilbert quantities via empirical characteristic functions in NeurIPS in 2016:

- Shashank Singh, Simon S Du, and Barnabás Póczos (2016). "Efficient Nonparametric Smoothness Estimation". In: *Advances in Neural Information Processing Systems*, pp. 1010–1018

At NeurIPS, Bharath Sriperumbudur pointed out that the same methods could be used to estimate other quadratic functionals of probability distributions, such as RKHS quantities. This observation led to our writing a journal paper (in submission), which comprises the majority of Chapter 5):

- Shashank Singh, Bharath K Sriperumbudur, and Barnabás Póczos (2018a). "Minimax Estimation of Quadratic Fourier Functionals". In: *arXiv preprint arXiv:1803.11451*

Notably (for me), this work was the first of mine to include novel minimax lower bounds, which are relatively complicated to prove for many density functional estimation problems (as compared to full density estimation).

Chapter 6 This is the first of two chapters that switch from the problem of estimating $\mathbb{R}$-valued functionals of distributions to the problem of estimating entire distributions. This chapter considers the case where loss is measured using Wasserstein distances, which enables analysis in an extreme breadth of settings (essentially, arbitrary metric spaces), using the very general tools of covering/packing numbers. The chapter focuses on (a) upper bounding the rate of convergence of the empirical distribution (in terms of covering numbers of balls in the sample space and moment assumptions on the distribution) and (b) lower bounding the minimax rate of distribution estimation (in terms of, essentially, the same quantities). For the most part, we find that the empirical distribution converges at the minimax optimal rate. The following preprint has been submitted to a statistics journal:

- Shashank Singh and Barnabás Póczos (2018). "Minimax Distribution Estimation in Wasserstein Distance". In: *arXiv preprint arXiv:1802.08855*

Chapter 7 This chapter studies nonparametric density estimation under yet another family of loss functions, integral probability metrics (IPMs). That is, given a sample space $\mathcal{X} \subseteq \mathbb{R}^D$, suppose we observe n IID samples $X_1, ..., X_n \overset{IID}{\sim} p$ from a probability density p over $\mathcal{X}$ that is unknown but assumed to lie in a regularity class $\mathcal{P}$. We seek an estimator $\widehat{p} : \mathcal{X}^n \to \mathcal{P}$ of p, with the goal of minimizing a loss

$$d_{\mathcal{F}}(p, \widehat{p}(X_1, ..., X_n)) := \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim p} [f(X)] - \mathop{\mathbb{E}}_{X \sim \widehat{p}(X_1, ..., X_n)} [f(X)] \right|, \qquad (*)$$

where $\mathcal{F}$, called the *discriminator class*, is some class of functions on $\mathcal{X}$.

Metrics of the form $(*)$ are called *integral probability metrics* (IPMs)[1], and can capture a wide variety of metrics on probability distributions by choosing $\mathcal{F}$ appropriately (Müller, 1997). This work studied the case where both $\mathcal{F}$ and

[1] While the name IPM seems to have caught on as the most widely used (Müller, 1997; Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet, 2012; Bottou, Arjovsky, Lopez-Paz, and Oquab, 2018; Zellinger, Moser, Grubinger, Lughofer, Natschläger, and Saminger-Platz, 2019), many other names have been used for these quantities, including *adversarial loss* (Dong and Yang, 2019), *MMD* (Dziugaite, Roy, and Ghahramani, 2015), and *$\mathcal{F}$-distance* or *neural net distance* (Arora, Ge, Liang, Ma, and Zhang, 2017).

$\mathcal{P}$ belong to a family of ellipsoids that includes, as examples, $\mathcal{L}^p$, Sobolev, and RKHS balls. We have two main motivations for studying this problem:

(a) This problem unifies nonparametric density estimation and the central problem of empirical process theory, namely bounding quantities of the form $d_{\mathcal{F}}(P, \widehat{P})$ when $\widehat{P}$ is the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ of the data (Pollard, 1990). Whereas empirical process theory typically avoids restricting $\mathcal{P}$ and fixes the estimator $\widehat{P} = P_n$, focusing on the discriminator class $\mathcal{F}$, nonparametric density estimation typically fixes the loss to be an $\mathcal{L}^p$ distance, and seeks a good estimator $\widehat{P}$ for a given distribution class $\mathcal{P}$. In contrast, we study how constraints on $\mathcal{F}$ and $\mathcal{P}$ *jointly* determine convergence rates of a number of estimates $\widehat{P}$ of P. This perspective allows us to unify, generalize, and extend several classical and recent results in distribution estimation.

(b) This problem is a theoretical framework for analyzing generative adversarial networks (GANs). Specifically, given a GAN whose discriminator and generator networks encode functions in $\mathcal{F}$ and $\mathcal{P}$, respectively, recent work (Liu, Bousquet, and Chaudhuri, 2017; Liang, 2017) showed that a GAN can be seen as a distribution estimate[2]

$$\widehat{P} = \operatorname*{argmin}_{Q\in\mathcal{P}} \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{X\sim Q}[f(X)] - \mathop{\mathbb{E}}_{X\sim\widetilde{P}_n}[f(X)] \right| = \operatorname*{argmin}_{Q\in\mathcal{P}} d_{\mathcal{F}}\left(Q, \widetilde{P}_n\right), \tag{1.1}$$

i.e., an estimate which directly minimizes empirical IPM risk with respect to a (regularized) empirical distribution $\widetilde{P}_n$. While, in the original GAN model (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio, 2014), $\widetilde{P}_n$ was the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ of the data, Liang (2017) showed that, under smoothness assumptions on the population distribution, performance is improved by replacing P_n with a regularized version $\widetilde{P}_n$, equivalent to the instance noise trick that has become standard in GAN training (Sønderby, Caballero, Theis, Shi, and Huszár, 2016; Mescheder, Geiger, and Nowozin, 2018). We show, in particular, that when $\widetilde{P}_n$ is a kernel-smoothed estimate, a GAN based on sufficiently large fully-connected neural networks with ReLU activations learns Sobolev probability distributions at the minimax optimal rate.

This work was published in NeurIPS 2018:

- Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos (2018). "Nonparametric density estimation under adversarial losses". In: *Advances in Neural Information Processing Systems*, pp. 10225–10236

Chapter 8 This chapter is intended to serve as a conclusion, but is really an assortment of short discussions on topics that are closely related to, but not formally part of, this thesis. The topics include a few extensions of the completed work, a number of downstream applications or consequences of this work, and some preliminary thoughts on future work. Most of this chapter is speculative, with the exception of one recent paper (under review) that extends the ideas of Chapter 7 to Besov IPMs and distributions:

[2] We assume a good optimization algorithm for computing (1.1), although this is also an active area of research.

- Ananya Uppal, Shashank Singh, and Barnaás Póczos (2019). "Nonparametric Density Estimation under Besov IPM Losses". In: *arXiv preprint arXiv:1902.03511*

Appendix A This appendix briefly overviews some collaborative projects I have worked on that are unrelated to the themes of this thesis. This includes a few simulation and applied machine learning papers in computational biology:

- Shashank Singh, Sabrina Rashid, Zhicheng Long, Saket Navlakha, Hanna Salman, Zoltán N Oltvai, and Ziv Bar-Joseph (2016). "Distributed Gradient Descent in Bacterial Food Search". In: *arXiv preprint arXiv:1604.03052*
- Sabrina Rashid, Shashank Singh, Saket Navlakha, and Ziv Bar-Joseph (2019). "A bacterial based distributed gradient descent model for mass scale evacuations". In: *Swarm and Evolutionary Computation* 46, pp. 97–103
- Sabrina Rashid, Zhicheng Long, Shashank Singh, Maryam Kohram, Harsh Vashistha, Saket Navlakha, Hanna Salman, Zoltán N Oltvai, and Ziv Bar-Joseph (2019). "Adjustment in tumbling rates improves bacterial chemotaxis on obstacle-laden terrains". In: *Proceedings of the National Academy of Sciences*, p. 201816315
- Yang Yang, Ruochi Zhang, Shashank Singh, and Jian Ma (2017). "Exploiting sequence-based features for predicting enhancer–promoter interactions". In: *Bioinformatics* 33.14, pp. i252–i260
- Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma (2019). "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks". In: *Quantitative Biology*, pp. 1–16

a theoretical paper on convolutional dictionary learning that was intended for but never quite made it to application in computational biology:

- Shashank Singh, Barnabás Póczos, and Jian Ma (2018). "Minimax reconstruction risk of convolutional sparse dictionary learning". In: *International Conference on Artificial Intelligence and Statistics*, pp. 1327–1336

some applied work in psychology on eye-tracking analysis methodology:

- Jaeah Kim, Shashank Singh, Anna Vande Velde, Erik D. Thiessen, and Anna V. Fisher (2018). "A Hidden Markov Model for Analyzing Eye-Tracking of Moving Objects". In: *Proceedings of the 2018 Annual Conference of the Cognitive Science Society (CogSci)*
- Jaeah Kim, Shashank Singh, Erik Thiessen, and Anna Fisher (2019). *A Hidden Markov Model for Analyzing Eye-Tracking of Moving Objects*. DOI: 10.31234/osf.io/mqpnf. URL: psyarxiv.com/mqpnf

and an applied paper on deep neural network compression that sprang out of an internship at Amazon:

- Shashank Singh, Ashish Khetan, and Zohar Karnin (2019). "DARC: Differentiable ARchitecture Compression". In: *arXiv preprint arXiv:1905.08170*

Appendix B The main contents of this short appendix are a list of the types of assumptions commonly made in the study of density functional estimation, and a table that

associates these assumptions with proposed estimators and their known convergence rates, along with a few other notes on what is known about these estimators (central limit theorems, adaptivity results, etc.). This table was originally constructed to give an overview of the field for my thesis proposal. Due the active nature of this area, not all of the most recent work is included, but it was fairly comprehensive as of Summer, 2018.

Chapter 2

Plug-in Kernel Density Estimates

2.1 Introduction

Many important quantities in machine learning and statistics can be viewed as integral functionals of one of more continuous probability densities; that is, quantities of the form

$$F(p_1, \cdots, p_k) = \int_{\mathcal{X}_1 \times \cdots \times X_k} f(p_1(x_1), \ldots, p_k(x_k))\, d(x_1, \ldots, x_k),$$

where $p_1, \cdots, p_k$ are probability densities of random variables taking values in $\mathcal{X}_1, \cdots, X_k$, respectively, and $f : \mathbb{R}^k \to \mathbb{R}$ is some measurable function. For simplicity, we refer to such integral functionals of densities as 'density functionals'. In this chapter, we study the problem of estimating density functionals. In our framework, we assume that the underlying distributions are not given explicitly. Only samples of n independent and identically distributed (i.i.d.) points from each of the unknown, continuous, nonparametric distributions $p_1, \cdots, p_k$ are given.

2.1.1 Motivations and Goals

One density functional of interest is Conditional Mutual Information (CMI), a measure of conditional dependence of random variables, which comes in several varieties including Rényi-α and Tsallis-α CMI (of which Shannon CMI is the $\alpha \to 1$ limit case). Estimating conditional dependence in a consistent manner is a crucial problem in machine learning and statistics; for many applications, it is important to determine how the relationship between two variables changes when we observe additional variables. For example, upon observing a third variable, two correlated variables may become independent, and, conversely, two independent variables may become dependent. Hence, CMI estimators can be used in many scientific areas to detect confounding variables and help distinguish causation from correlation (Pearl, 1998; Montgomery, 2005). Conditional dependencies are also central to Bayesian network learning (Koller and Friedman, 2009; Zhang, Peters, Janzing, and Scholkopf, 2011), where CMI estimation can be used to verify compatibility of a particular Bayes net with observed data under a local Markov assumption.

Other important density functionals are divergences between probability distributions, including Rényi-α (Rényi, 1970) and Tsallis-α (Villmann and Haase, 2010) divergences (of which Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) is the $\alpha \to 1$ limit case), and L_p divergence. Divergence estimators can be used to extend machine learning algorithms for regression, classification, and clustering from the standard setting where inputs are finite-dimensional feature vectors to settings where inputs are sets or distributions (Póczos, Xiong, Sutherland, and

Schneider, 2012; Oliva, Poczos, and Schneider, 2013). Entropy and mutual information (MI) can be estimated as special cases of divergences. Entropy estimators are used in goodness-of-fit testing (Goria, Leonenko, Mergel, and Novi Inverardi, 2005), parameter estimation in semi-parametric models (Wolsztynski, Thierry, and Pronzato, 2005a), and texture classification (Hero, Ma, Michel, and Gorman, 2002b), and MI estimators are used in feature selection (Peng, Long, and Ding, 2005), clustering (Aghagolzadeh, Soltanian-Zadeh, Araabi, and Aghagolzadeh, 2007), optimal experimental design (Lewi, Butera, and Paninski, 2007), and boosting and facial expression recognition (Shan, Gong, and Mcowan, 2005). Both entropy and mutual information estimators are used in independent component and subspace analysis (Learned-Miller and Fisher, 2003; Szabó, Póczos, and Lőrincz, 2007) and image registration (Hero, Ma, Michel, and Gorman, 2002b). Further applications of divergence estimation are in Leonenko, Pronzato, and Savani (2008).

Despite the practical utility of density functional estimators, little is known about their statistical performance, especially for functionals of more than one density. In particular, few density functional estimators have known convergence rates, and, to the best of our knowledge, no finite sample exponential concentration bounds have been derived for general density functional estimators. An important consequence of this exponential bound is that, using a union bound, we can guarantee accuracy of multiple estimates simultaneously. For example, Liu, Wasserman, and Lafferty (2012) shows how this can be applied to optimally analyze forest density estimation algorithms. Because the CMI of variables X and Y given a third variable Z is zero if and only X and Y are conditionally independent given Z, by estimating CMI with a confidence interval, we can test for conditional independence with bounded type I error probabilty.

Our main contribution is to derive convergence rates and an exponential concentration inequality for a particular, consistent, nonparametric estimator for large class of density functionals, including conditional density functionals. We also apply our concentration inequality to the important case of Rényi-α CMI.

2.1.2 Related Work

Although lower bounds are not known for estimation of general density functionals (of arbitrarily many densities), Birgé and Massart (1995) lower bounded the convergence rate for estimators of functionals of a single density (e.g., entropy functionals) by $O\left(n^{-4\beta/(4\beta+d)}\right)$, when the data are d-dimensional and the underlying density is assumed to lie in a β-Hölder class. Krishnamurthy, Kandasamy, Poczos, and Wasserman (2014) extended this lower bound to the two-density cases of L_2, Rényi-α, and Tsallis-α divergences and gave plug-in estimators which achieve this rate. These estimators enjoy the parametric rate of $O\left(n^{-1/2}\right)$ when $\beta > d/4$, and work by optimally estimating the density and then applying a correction to the plug-in estimate. In contrast, our estimator undersmooths the density, and converges at a slower rate of $O\left(n^{-\beta/(\beta+d)}\right)$ when $\beta < d$ (and the parametric rate $O\left(n^{-1/2}\right)$ when $\beta \geq d$), but obeys an exponential concentration inequality, which is not known for the estimators of Krishnamurthy, Kandasamy, Poczos, and Wasserman (2014).

Another exception for f-divergences is provided by Nguyen, Wainwright, and Jordan. (2010), using empirical risk minimization. This approach involves solving an ∞-dimensional convex minimization problem which be reduced to an n-dimensional problem for certain function classes defined by reproducing kernel Hilbert spaces (n is the sample size). When n is large, these optimization problems can still

be very demanding. They studied the estimator's convergence rate, but did not derive concentration bounds.

A number of papers have studied k-nearest-neighbors estimators, primarily for Rényi-α density functionals including entropy (Leonenko, Pronzato, and Savani, 2008), divergence (Wang, Kulkarni, and Verdú, 2009) and conditional divergence and MI (Poczos and Schneider, 2012). These estimators work directly, without the intermediate density estimation step, and generally have proofs of consistency, but their convergence rates and dependence on k, α, and the dimension are unknown. One recent exception is a k-nearest-neighbors based estimator that converges at the parametric rate when $\beta > d$, using an optimally weighted ensemble of weak estimators (Sricharan, Wei, and Hero, 2013; Moon and Hero, 2014b). These estimators appear to perform well in higher dimensions, but rates for these estimators require that $k \to \infty$ as $n \to \infty$, causing computational difficulties for large samples.

Although the literature on dependence measures is huge, few estimators have been generalized to the conditional case (Fukumizu, Gretton, Sun, and Schoelkopf, 2008; Reddi and Poczos, 2013). There is some work on testing conditional dependence (Su and White, 2008; Bouezmarni, Rombouts, and Taamouti, 2009), but, unlike CMI estimation, these tests are intended to simply accept or reject the hypothesis that variables are conditionally independent, rather than to measure conditional dependence. Our exponential concentration inequality also suggests a new test for conditional independence.

This chapter continues a line of work begun by Liu, Wasserman, and Lafferty (2012) and continued by Singh and Póczos (2014b). Liu, Wasserman, and Lafferty (2012) proved an exponential concentration inequality for an estimator of Shannon entropy and MI in the 2-dimensional case. Singh and Póczos (2014b) used similar techniques to derive an exponential concentration inequality for an estimator of Rényi-α divergence in d dimensions, for a larger family of densities. Both used plug-in estimators based on a mirrored kernel density estimator (KDE) on $[0,1]^d$. Our work generalizes these results to a much larger class of density functionals, as well as to conditional density functionals (see Section 6). In particular, we use a plug-in estimator for general density functionals based on the same mirrored KDE, and also use some lemmas regarding this KDE proven by Singh and Póczos (2014b). By considering the more general density functional case, we are also able to significantly simplify the proofs of the convergence rate and exponential concentration inequality.

Organization

In Section 2, we establish the theoretical context of our work, including notation, the precise problem statement, and our estimator. In Section 3, we outline our main theoretical results and state some consequences. Sections 4 and 5 give precise statements and proofs of the results in Section 3. Finally, in Section 6, we extend our results to conditional density functionals, and state the consequences in the particular case of Rényi-α CMI.

2.2 Density Functional Estimator

2.2.1 Notation

For an integer k, $[k] = \{1, \cdots, k\}$ denotes the set of positive integers at most k. Using the notation of multi-indices common in multivariable calculus, $\mathbb{N}^d$ denotes the set

of d-tuples of non-negative integers, which we denote with a vector symbol $\vec{\cdot}$, and, for $\vec{i} \in \mathbb{N}^d$,

$$D^{\vec{i}} := \frac{\partial^{|\vec{i}|}}{\partial^{i_1} x_1 \cdots \partial^{i_d} x_d} \quad \text{and} \quad |\vec{i}| = \sum_{k=1}^{d} i_k.$$

For fixed $\beta, L > 0$, $r \geq 1$, and a positive integer d, we will work with densities in the following bounded subset of a β-Hölder space:

$$C^{\beta}_{L,r}([0,1]^d) := \left\{ p : [0,1]^d \to \mathbb{R} \,\middle|\, \sup_{\substack{x \neq y \in D \\ |\vec{i}|=\ell}} \frac{|D^{\vec{i}} p(x) - D^{\vec{i}} p(y)|}{\|x - y\|^{(\beta-\ell)}} \leq L \right\}, \tag{2.1}$$

where $\ell = \lfloor \beta \rfloor$ is the greatest integer *strictly* less than β, and $\| \cdot \|_r : \mathbb{R}^d \to \mathbb{R}$ is the usual r-norm. To correct for boundary bias, we will require the densities to be nearly constant near the boundary of $[0,1]^d$, in that their derivatives vanish at the boundary. Hence, we work with densities in

$$\Sigma(\beta, L, r, d) := \left\{ p \in C^{\beta}_{L,r}([0,1]^d) \,\middle|\, \max_{1 \leq |\vec{i}| \leq \ell} |D^{\vec{i}} p(x)| \to 0 \text{ as } \operatorname{dist}(x, \partial[0,1]^d) \to 0 \right\}, \tag{2.2}$$

where $\partial[0,1]^d = \{x \in [0,1]^d : x_j \in \{0,1\} \text{ for some } j \in [d]\}$.

2.2.2 Problem Statement

For each $i \in [k]$ let X_i be a d_i-dimensional random vector taking values in $\mathcal{X}_i := [0,1]^{d_i}$, distributed according to a density $p_i : \mathcal{X} \to \mathbb{R}$. For an appropriately smooth function $f : \mathbb{R}^k \to \mathbb{R}$, we are interested in a using random sample of n i.i.d. points from the distribution of each X_i to estimate

$$F(p_1, \ldots, p_k) := \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_k} f(p_1(x_1), \ldots, p_k(x_k))\, d(x_1, \ldots, x_k). \tag{2.3}$$

2.2.3 Estimator

For a fixed bandwidth h, we first estimate each density p_i using the mirrored kernel density estimator (KDE) $\widehat{p}_i$ (described formally in Singh and Póczos (2014b)), which reflects each sample over nearby boundaries to mitigate boundary bias before performing kernel density estimation (as illustrated in Figure 2.1). We then use a plug-in estimate of $F(p_1, \ldots, p_k)$.

$$F(\widehat{p}_1, \ldots, \widehat{p}_k) := \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_k} f(\widehat{p}_1(x_1), \ldots, \widehat{p}_k(x_k))\, d(x_1, \ldots, x_k).$$

Our main results generalize those of Singh and Póczos (2014b) to a broader class of density functionals.

2.3 Main Results

In this section, we outline our main theoretical results, proven in Sections 4 and 5, and also discuss some important corollaries.

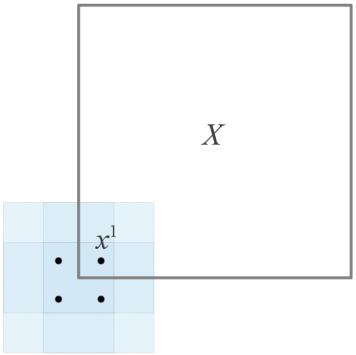


FIGURE 2.1: A data point $x^1 \in C_{(1,2,\emptyset,\emptyset)} \subseteq [0,1]^2$ (using the notation of Singh and Póczos (2014b)), along with its three reflected copies.The sum of the integrals over $\mathcal{X}$ of (the absolute values of) the four kernels (with shaded support) is $\|K\|_1^2$.

We decompose the estimatator's error into bias and a variance-like terms via the triangle inequality:

$$|F(\widehat{p}_1,\ldots,\widehat{p}_k) - F(p_1,\ldots,p_k)| \leq \underbrace{|F(\widehat{p}_1,\ldots,\widehat{p}_k) - \mathbb{E}\,F(\widehat{p}_1,\ldots,\widehat{p}_k)|}_{\text{variance-like term}} + \underbrace{|\,\mathbb{E}\,F(\widehat{p}_1,\ldots,\widehat{p}_k) - F(p_1,\ldots,p_k)|}_{\text{bias term}}.$$

We will prove the "variance" bound

$$\mathbb{P}\left(|F(\widehat{p}_1,\ldots,\widehat{p}_k) - \mathbb{E}\,F(\widehat{p}_1,\ldots,\widehat{p}_k)| > \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 n}{C_V^2}\right) \tag{2.4}$$

for all $\varepsilon > 0$ and the bias bound

$$|\,\mathbb{E}\,F(\widehat{p}_1,\ldots,\widehat{p}_k) - F(p_1,\ldots,p_k)| \leq C_B\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right), \tag{2.5}$$

where $d := \max_i d_i$, and C_V and C_B are constant in the sample size n and bandwidth h for exact values. To the best of our knowledge, this is the first time an exponential inequality like (2.4) has been established for general density functional estimation. This variance bound does not depend on h and the bias bound is minimized by $h \asymp n^{-\frac{1}{\beta+d}}$, we have the convergence rate

$$|\,\mathbb{E}\,F(\widehat{p}_1,\ldots,\widehat{p}_k) - F(p_1,\ldots,p_k)| \in O\left(n^{-\frac{\beta}{\beta+d}}\right).$$

It is interesting to note that, in optimizing the bandwidth for our density functional estimate, we use a smaller bandwidth than is optimal for minimizing the bias of the KDE. Intuitively, this reflects the fact that the plug-in estimator, as an integral functional, performs some additional smoothing.

We can use our exponential concentration bound to obtain a bound on the true variance of $F(\widehat{p}_1,\ldots,\widehat{p}_k)$. If $G : [0,\infty) \to \mathbb{R}$ denotes the cumulative distribution

function of the squared deviation of $F(\widehat{p}_1, \ldots, \widehat{p}_k)$ from its mean, then

$$1 - G(\varepsilon) = \mathbb{P}\left(\left(F(\widehat{p}_1, \ldots, \widehat{p}_k) - \mathbb{E}\, F(\widehat{p}_1, \ldots, \widehat{p}_k)\right)^2 > \varepsilon \right) \leq 2 \exp\left(-\frac{2\varepsilon n}{C_V^2} \right).$$

Thus,

$$\begin{aligned} \mathbb{V}[F(\widehat{p}_1, \ldots, \widehat{p}_k)] &= \mathbb{E}\left[\left(F(\widehat{p}_1, \ldots, \widehat{p}_k) - \mathbb{E}\, F(\widehat{p}_1, \ldots, \widehat{p}_k)\right)^2 \right] \\ &= \int_0^\infty 1 - G(\varepsilon)\, d\varepsilon \leq 2 \int_0^\infty \exp\left(-\frac{2\varepsilon n}{C_V^2} \right) = C_V^2 n^{-1}. \end{aligned}$$

We then have a mean squared error of

$$\mathbb{E}\left[\left(F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(p_1, \ldots, p_k)\right)^2 \right] \in O\left(n^{-1} + n^{-\frac{2\beta}{\beta+d}} \right),$$

which is in $O(n^{-1})$ if $\beta \geq d$ and $O\left(n^{-\frac{2\beta}{\beta+d}}\right)$ otherwise.

It should be noted that the constants in both the bias bound and the variance bound depend exponentially on the dimension d. Lower bounds in terms of d are unknown for estimating most density functionals of interest, and an important open problem is whether this dependence can be made asymptotically better than exponential.

2.4 Bias Bound

In this section, we precisely state and prove the bound on the bias of our density functional estimator, as introduced in Section 3.

Assume each $p_i \in \Sigma(\beta, L, r, d)$ (for $i \in [k]$), assume $f : \mathbb{R}^k \to \mathbb{R}$ is twice continuously differentiable, with first and second derivatives all bounded in magnitude by some $C_f \in \mathbb{R}$, [1] and assume the kernel $K : \mathbb{R} \to \mathbb{R}$ has bounded support $[-1, 1]$ and satisfies

$$\int_{-1}^{1} K(u)\, du = 1 \quad \text{and} \quad \int_{-1}^{1} u^j K(u)\, du = 0 \quad \text{for all } j \in \{1, \ldots, \ell\}.$$

Then, there exists a constant $C_B \in \mathbb{R}$ such that

$$\left| \mathbb{E}\, F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(p_1, \ldots, p_k) \right| \leq C_B \left(h^\beta + h^{2\beta} + \frac{1}{nh^d} \right).$$

2.4.1 Proof of Bias Bound

By Taylor's Theorem, $\forall x = (x_1, \ldots, x_k) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$, for some $\xi \in \mathbb{R}^k$ on the line segment between $\widehat{p}(x) := (\widehat{p}_1(x_1), \ldots, \widehat{p}_k(x_k))$ and $p(x) := (p_1(x_1), \ldots, p_k(x_k))$,

[1] If $p_1(\mathcal{X}_1) \times \cdots \times p_k(\mathcal{X}_k)$ is known to lie within some cube $[\kappa_1, \kappa_2]^k$, then it suffices for f to be twice continuously differentiable on $[\kappa_1, \kappa_2]^k$ (and the boundedness condition follows immediately). This will be important for our application to Rényi-α Conditional Mutual Information.

letting H_f denote the Hessian of f

$$|\mathbb{E} f(\widehat{p}(x)) - f(p(x))| = \left|\mathbb{E}(\nabla f)(p(x)) \cdot (\widehat{p}(x) - p(x)) + \frac{1}{2}(\widehat{p}(x) - p(x))^T H_f(\xi)(\widehat{p}(x) - p(x))\right|$$
$$\leq C_f \left(\sum_{i=1}^{k} |B_{p_i}(x_i)| + \sum_{i<j\leq k} |B_{p_i}(x_i) B_{p_j}(x_j)| + \sum_{i=1}^{k} \mathbb{E}[\widehat{p}_i(x_i) - p_i(x_i)]^2 \right)$$

where we used that $\widehat{p}_i$ and $\widehat{p}_j$ are independent for $i \neq j$. Applying Hölder's Inequality,

$$|\mathbb{E} F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(p_1, \ldots, p_k)| \leq \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_k} |\mathbb{E} f(\widehat{p}(x)) - f(p(x))| \, dx$$
$$\leq C_f \left(\sum_{i=1}^{k} \int_{\mathcal{X}_i} |B_{p_i}(x_i)| + \mathbb{E}[\widehat{p}_i(x_i) - p_i(x_i)]^2 \, dx_i + \sum_{i<j\leq k} \int_{\mathcal{X}_i} |B_{p_i}(x_i)| \, dx_i \int_{\mathcal{X}_j} |B_{p_j}(x_j)| \, dx_j \right)$$
$$\leq C_f \left(\sum_{i=1}^{k} \sqrt{\int_{\mathcal{X}_i} B_{p_i}^2(x_i) \, dx_i} + \int_{\mathcal{X}_i} \mathbb{E}[\widehat{p}_i(x_i) - p_i(x_i)]^2 \, dx_i \right.$$
$$\left. + \sum_{i<j\leq k} \sqrt{\int_{\mathcal{X}_i} B_{p_i}^2(x_i) \, dx_i \int_{\mathcal{X}_j} B_{p_j}^2(x_j) \, dx_j} \right).$$

We now make use of the so-called Bias Lemma proven by (Singh and Póczos, 2014b), which bounds the integrated squared bias of the mirrored KDE $\widehat{p}$ on $[0,1]^d$ for an arbitrary $p \in \Sigma(\beta, L, r, d)$. Writing the bias of $\widehat{p}$ at $x \in [0,1]^d$ as $B_p(x) = \mathbb{E}\,\widehat{p}(x) - p(x)$, (Singh and Póczos, 2014b) showed that there exists $C > 0$ constant in n and h such that

$$\int_{[0,1]^d} B_p^2(x) \, dx \leq C h^{2\beta}. \tag{2.6}$$

Applying the Bias Lemma and certain standard results in kernel density estimation (see, for example, Propositions 1.1 and 1.2 of (Tsybakov, 2008)) gives

$$|\mathbb{E} F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(p_1, \ldots, p_k)| \leq C \left(k^2 h^\beta + k h^{2\beta} \right) + \frac{\|K\|_1^d}{n h^d} \leq C_B \left(h^\beta + h^{2\beta} + \frac{1}{n h^d} \right),$$

where $\|K\|_1$ denotes the 1-norm of the kernel. □

2.5 Variance Bound

In this section, we precisely state and prove the exponential concentration inequality for our density functional estimator, as introduced in Section 3. Assume that f is Lipschitz continuous with constant C_f in the 1-norm on $p_1(\mathcal{X}_1) \times \cdots \times p_k(\mathcal{X}_k)$ (i.e.,

$$|f(x) - f(y)| \leq C_f \sum_{k=1}^{\infty} |x_i - y_i|, \quad \forall x, y \in p_1(\mathcal{X}_1) \times \cdots \times p_k(\mathcal{X}_k)). \tag{2.7}$$

and assume the kernel $K \in L_1(\mathbb{R})$ (i.e., it has finite 1-norm). Then, there exists a constant $C_V \in \mathbb{R}$ such that $\forall \varepsilon > 0$,

$$\mathbb{P}\left(|F(\widehat{p}_1, \ldots, \widehat{p}_k) - \mathbb{E}\, F(\widehat{p}_1, \ldots, \widehat{p}_k)|\right) \leq 2 \exp\left(-\frac{2\varepsilon^2 n}{C_V^2} \right).$$

Note that, while we require no assumptions on the densities here, in certain specific applications, such us for some Rényi-α quantities, where $f = \log$, assumptions such as lower bounds on the density may be needed to ensure f is Lipschitz on its domain.

2.5.1 Proof of Variance Bound

Consider i.i.d. samples $(x_1^1, \ldots, x_k^n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ drawn according to the product distribution $p = p_1 \times \cdots p_k$. In anticipation of using McDiarmid's Inequality (McDiarmid, 1989), let $\widehat{p}'_j$ denote the j^{th} mirrored KDE when the sample x_j^i is replaced by new sample $(x_j^i)'$. Then, applying the Lipschitz condition (2.7) on f,

$$|F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(\widehat{p}_1, \ldots, \widehat{p}'_j, \ldots, \widehat{p}_k)| \leq C_f \int_{\mathcal{X}_j} |p_j(x) - p'_j(x)|\, dx,$$

since most terms of the sum in (2.7) are zero. Expanding the definition of the kernel density estimates $\widehat{p}_j$ and $\widehat{p}'_j$ and noting that most terms of the mirrored KDEs $\widehat{p}_j$ and $\widehat{p}'_j$ are identical gives

$$|F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(\widehat{p}_1, \ldots, \widehat{p}'_j, \ldots, \widehat{p}_k)| = \frac{C_f}{nh^{d_j}} \int_{\mathcal{X}_j} \left| K_{d_j}\left(\frac{x - x_j^i}{h}\right) - K_{d_j}\left(\frac{x - (x_j^i)'}{h}\right) \right| dx$$

where K_{d_j} denotes the d_j-dimensional mirrored product kernel based on K. Performing a change of variables to remove h and applying the triangle inequality followed by the bound on the integral of the mirrored kernel proven in (Singh and Póczos, 2014b),

$$\begin{aligned} |F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(\widehat{p}_1, \ldots, \widehat{p}'_j, \ldots, \widehat{p}_k)| &\leq \frac{C_f}{n} \int_{h\mathcal{X}_j} \left| K_{d_j}(x - x_j^i) - K_{d_j}(x - (x_j^i)') \right| dx \\ &\leq \frac{2C_f}{n} \int_{[-1,1]^{d_j}} |K_{d_j}(x)|\, dx \leq \frac{2C_f}{n} \|K\|_1^{d_j} = \frac{C_V}{n}, \end{aligned} \tag{2.8}$$

for $C_V = 2C_f \max_j \|K\|_1^{d_j}$. Since $F(\widehat{p}_1, \ldots, \widehat{p}_k)$ depends on kn independent variables, McDiarmid's Inequality then gives, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|F(\widehat{p}_1, \ldots, \widehat{p}_k) - F(p_1, \ldots, p_k)| > \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{knC_V^2/n^2}\right) = 2\exp\left(-\frac{2\varepsilon^2 n}{kC_V^2}\right). \quad \square$$

2.6 Extension to Conditional Density Functionals

Our convergence result and concentration bound can be fairly easily adapted to to KDE-based plug-in estimators for many functionals of interest, including Rényi-α and Tsallis-α entropy, divergence, and MI, and L_p norms and distances, which have either the same or analytically similar forms as as the functional (2.3). As long as the density of the variable being conditioned on is lower bounded on its domain, our results also extend to conditional density functionals of the form [2]

$$F(P) = \int_{\mathcal{Z}} P(z) f\left(\int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_k} g\left(\frac{P(x_1, z)}{P(z)}, \frac{P(x_2, z)}{P(z)}, \ldots, \frac{P(x_k, z)}{P(z)}\right) d(x_1, \ldots, x_k) \right) dz \tag{2.9}$$

[2] We abuse notation slightly and also use P to denote all of its marginal densities.

including, for example, Rényi-α conditional entropy, divergence, and mutual information, where f is the function $x \mapsto \frac{1}{1-\alpha}\log(x)$. The proof of this extension for general k is essentially the same as for the case $k = 1$, and so, for notational simplicity, we demonstrate the latter.

2.6.1 Problem Statement, Assumptions, and Estimator

For given dimensions $d_x, d_z \geq 1$, consider random vectors X and Z distributed on unit cubes $\mathcal{X} := [0,1]^{d_x}$ and $\mathcal{Z} := [0,1]^{d_z}$ according to a joint density $P : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. We use a random sample of $2n$ i.i.d. points from P to estimate a conditional density functional $F(P)$, where F has the form (2.9).

Suppose that P is in the Hölder class $\Sigma(\beta, L, r, d_x+d_z)$, noting that this implies an analogous condition on each marginal of P, and suppose that P bounded below and above, i.e., $0 < \kappa_1 := \inf_{x\in\mathcal{X},z\in\mathcal{Z}} P(z)$ and $\infty > \kappa_2 := \inf_{x\in\mathcal{X},z\in\mathcal{Z}} P(x,z)$. Suppose also that f and g are continuously differentiable, with

$$C_f := \sup_{x\in[c_g,C_g]} |f(x)| \quad \text{and} \quad C_{f'} := \sup_{x\in[c_g,C_g]} |f'(x)|, \tag{2.10}$$

where

$$c_g := \inf g\left(\left[0, \frac{\kappa_2}{\kappa_1}\right]\right) \quad \text{and} \quad C_g := \sup g\left(\left[0, \frac{\kappa_2}{\kappa_1}\right]\right).$$

After estimating the densities $P(z)$ and $P(x,z)$ by their mirrored KDEs, using n independent data samples for each, we clip the estimates of $P(x,z)$ and $P(z)$ below by κ_1 and above by κ_2 and denote the resulting density estimates by $\widehat{P}$. Our estimate $F(\widehat{P})$ for $F(P)$ is simply the result of plugging $\widehat{P}$ into equation (2.9).

2.6.2 Proof of Bounds for Conditional Density Functionals

We bound the error of $F(\widehat{P})$ in terms of the error of estimating the corresponding unconditional density functional using our previous estimator, and then apply our previous results.

Suppose P_1 is either the true density P or a plug-in estimate of P computed as described above, and P_2 is a plug-in estimate of P computed in the same manner but using a different data sample. Applying the triangle inequality twice,

$$\begin{aligned}
|F(P_1) - F(P_2)| \leq& \int_{\mathcal{Z}} \left| P_1(z) f\left(\int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) dx\right) - P_2(z) f\left(\int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) dx\right)\right| \\
&+ \left| P_2(z) f\left(\int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) dx\right) - P_2(z) f\left(\int_{\mathcal{X}} g\left(\frac{P_2(x,z)}{P_2(z)}\right) dx\right)\right| dz \\
\leq& \int_{\mathcal{Z}} |P_1(z) - P_2(z)| \left| f\left(\int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) dx\right)\right| \\
&+ P_2(z) \left| f\left(\int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) dx\right) - f\left(\int_{\mathcal{X}} g\left(\frac{P_2(x,z)}{P_2(z)}\right) dx\right)\right| dz
\end{aligned}$$

Applying the Mean Value Theorem and the bounds in (2.10) gives

$$\begin{aligned}
|F(P_1) - F(P_2)| \leq& \int_{\mathcal{Z}} C_f |P_1(z) - P_2(z)| + \kappa_2 C_{f'} \left| \int_{\mathcal{X}} g\left(\frac{P_1(x,z)}{P_1(z)}\right) - g\left(\frac{P_2(x,z)}{P_2(z)}\right) dx \right| dz \\
=& \int_{\mathcal{Z}} C_f |P_1(z) - P_2(z)| + \kappa_2 C_{f'} \left| G_{P_1(z)}(P_1(\cdot, z)) - G_{P_2(z)}(P_2(\cdot, z)) \right| dz,
\end{aligned}$$

where G_z is the density functional

$$G_{P(z)}(Q) = \int_{\mathcal{X}} g\left(\frac{Q(x)}{P(z)}\right) dx.$$

Note that, since the data are split to estimate $P(z)$ and $P(x,z)$, $G_{\widehat{P}(z)}(\widehat{P}(\cdot,z))$ depends on each data point through only one of these KDEs. In the case that P_1 is the true density P, taking the expectation and using Fubini's Theorem gives

$$\begin{aligned}
\mathbb{E}\,|F(P) - F(\widehat{P})| \leq & \int_{\mathcal{Z}} C_f\,\mathbb{E}\,|P(z) - \widehat{P}(z)| + \kappa_2 C_{f'}\,\mathbb{E}\left|G_{P(z)}(P(\cdot,z)) - G_{\widehat{P}(z)}(\widehat{P}(\cdot,z))\right| dz, \\
\leq & C_f \sqrt{\int_{\mathcal{Z}} \mathbb{E}(P(z) - \widehat{P}(z))^2 dz} + 2\kappa_2 C_{f'} C_B \left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right) \\
\leq & \left(2\kappa_2 C_{f'} C_B + C_f C\right)\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right)
\end{aligned}$$

applying Hölder's Inequality and our bias bound (2.5), followed by the bias lemma (2.6). This extends our bias bound to conditional density functionals. For the variance bound, consider the case where P_1 and P_2 are each mirrored KDE estimates of P, but with one data point resampled (as in the proof of the variance bound, setting up to use McDiarmid's Inequality). By the same sequence of steps used to show (2.8),

$$\int_{\mathcal{Z}} |P_1(z) - P_2(z)|\, dz \leq \frac{2\|K\|_1^{d_z}}{n},$$

and

$$\int_{\mathcal{Z}} \left|G_{P(z)}(P(\cdot,z)) - G_{\widehat{P}(z)}(\widehat{P}(\cdot,z))\right| dz \leq \frac{C_V}{n}.$$

(by casing on whether the resampled data point was used to estimate $P(x,z)$ or $P(z)$), for an appropriate C_V depending on $\sup_{x\in[\kappa_1/\kappa_2,\kappa_2/\kappa_1]} |g'(x)|$. Then, by McDiarmid's Inequality,

$$\mathbb{P}\left(|F(\widehat{p}_1,\ldots,\widehat{p}_k) - F(p_1,\ldots,p_k)| > \varepsilon\right) = 2\exp\left(-\frac{\varepsilon^2 n}{4C_V^2}\right). \quad \square$$

2.6.3 Application to Rényi-α Conditional Mutual Information

As a concrete example, our estimator can be applied to estimate Rényi-α Conditional Mutual Information (CMI). As an application, one might use an estimate of this sort to test for conditional independence, i.e., for distinguishing between the two graphical models presented in Figure 2.2. Consider random vectors X, Y, and Z on $\mathcal{X} = [0,1]^{d_x}$, $\mathcal{Y} = [0,1]^{d_y}$, $\mathcal{Z} = [0,1]^{d_z}$, respectively. $\alpha \in (0,1)\cup(1,\infty)$, the Rényi-$\alpha$ CMI of X and Y given Z is

$$I(X;Y|Z) = \frac{1}{1-\alpha}\int_{\mathcal{Z}} P(z)\log\int_{\mathcal{X}\times\mathcal{Y}}\left(\frac{P(x,y,z)}{P(z)}\right)^{\alpha}\left(\frac{P(x,z)P(y,z)}{P(z)^2}\right)^{1-\alpha} d(x,y)\,dz. \tag{2.11}$$

In this case, the estimator which plugs mirrored KDEs for $P(x,y,z)$, $P(x,z)$, $P(y,z)$, and $P(z)$ into (2.11) obeys the concentration inequality (2.4) with $C_V = \kappa^*\|K\|_1^{d_x+d_y+d_z}$, where κ^* depends only on α, κ_1, and κ_2.

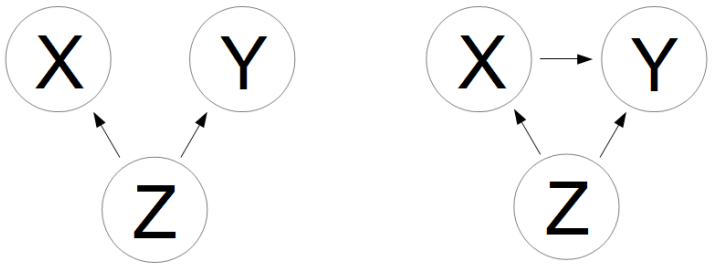


FIGURE 2.2: Two possible graphs of dependence between variables X, Y, and Z. The left graph corresponds to $I(X;Y|Z) = 0$, whereas the right graph corresponds to $I(X;Y|Z) > 0$. These can thus be distinguished using an estimate of conditional mutual information.

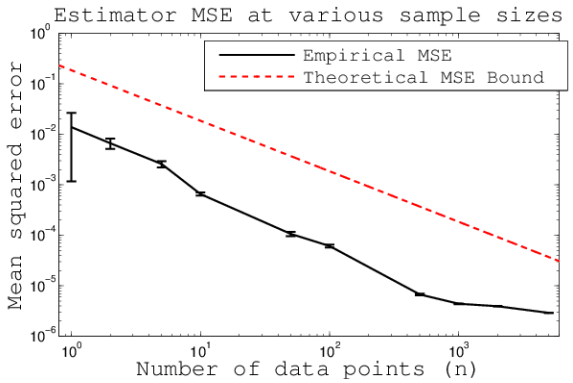


FIGURE 2.3: Log-log plot of squared error (averaged over 100 IID trials) of our Rényi-0.8 estimator for various sample sizes n, alongside our theoretical bound. Error bars indicate standard deviation of estimator over 100 trials.

2.7 Experimental Results

This sections provides a very simple validation experiment, in which we used our estimator to estimate the Rényi-α divergence between two normal distributions, restricted to the unit cube $[0,1]^3$, with different means and identical, isotropic covariances. In particular,

$$\vec{\mu_1} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \vec{\mu_2} = \begin{bmatrix} 0.7 \\ 0.7 \\ 0.7 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}.$$

For each $n \in \{1, 2, 5, 10, 50, 100, 500, 1000, 2000, 5000\}$, n data points were sampled according to each distribution and constrained (via rejection sampling) to lie within $[0,1]^3$. Our estimator was computed from these samples, for $\alpha = 0.8$,using the Epanechnikov Kernel

$$K(u) = \frac{3}{4}(1-u^2)1_{[-1,1]},$$

with bandwidth $h = 0.25$. The true α-divergence was approximated by numerical integration. Bias and variance of our estimator were then computed in the usual manner based on 100 trials. Figure 2.3 shows the error and variance of our estima-

tor for each n. We compare our estimator's empirical error to an approximation of our theoretical bound (also shown in Figure 2.3). Since the distributions used are infinitely differentiable, $\beta = \infty$, and so the estimator's MSE should converge at the rate $O(n^{-1})$. An appropriate constant multiple was computed from our bounds.

Chapter 3

Bias Corrected Fixed-k Nearest Neighbor Estimates

3.1 Introduction

This chapter focuses on a family of estimators for nonlinear expectation functionals, based on k-nearest neighbor (k-NN) statistics, that are popular in practice due to excellent empirical performance, but whose statistical properties have remained elusive for decades.

3.1.1 Some history

These estimators, which we call "bias-corrected fixed-k" (BCFk) estimators were broadly inspired by the 1-NN entropy estimator of Kozachenko and Leonenko (1987). This method for estimating entropy has since been generalized extensively, by Goria, Leonenko, Mergel, and Novi Inverardi (2005) to use $k > 1$ nearest neighbors, by Wang, Kulkarni, and Verdú (2009) to estimate KL divergence, by Leonenko, Pronzato, and Savani (2008) (with corrections in Leonenko and Pronzato (2010)) to estimate Rényi entropies, by Póczos and Schneider (2011) to estimate Rényi and Tsallis divergences, and by Poczos and Schneider (2012) to estimate conditional entropies and divergences; see Poczos, Xiong, and Schneider (2011) for a survey of these estimators and discussion of their asymptotic consistency.

Excepting the analysis of Tsybakov and Meulen (1996) for a truncated variant of the Kozachenko-Leonenko estimator in the 1-dimensional case, the convergence rates of these estimators were unknown until recently. In contrast, beginning in 2016 (almost 30 years after the seminal paper of Kozachenko and Leonenko (1987)) there has been a flurry of work studying this problem. In particular, in 2016, our NIPS paper Singh and Póczos (2016b), as well as the thesis Berrett, Samworth, and Yuan (2019) of Thomas Berrett in Richard Samworth's group at Cambridge, and work (Gao, Oh, and Viswanath, 2017a) by Weihao Gao and others at UIUC independently but simultaneously provided the first general upper bounds on the convergence rates of the original Kozachenko-Leonenko estimator (and of the generalization to $k > 1$ by Goria, Leonenko, Mergel, and Novi Inverardi (2005)).

Of these papers, Berrett, Samworth, and Yuan (2019) provides the most nuanced analysis of the original Kozachenko-Leonenko estimator (for general k), proving asymptotic normality and even computing the asymptotic variance, while making the weakest tail assumptions of the true probability distribution.

Gao, Oh, and Viswanath (2017a) require the density to have bounded support, and the upper bound of Gao, Oh, and Viswanath (2017a) is somewhat loose, due to a somewhat loose analysis of boundary bias. Intriguingly, they are able to extend their results to the mutual information estimator proposed by Kraskov, Stögbauer,

and Grassberger (2004), which is undoubtedly the most widely used nonparametric estimator of mutual information. This estimator is based on a variant of the Kozachenko-Leonenko entropy estimator, but, rather than simply using three different entropy estimators (via the identity $I(X,Y) = H(X) + H(Y) + H(X,Y)$), the KSG estimator uses a clever coupling of the choices of k, which causes a substantial cancellation in the biases of these three estimators. Unfortunately, the variation does not lend itself to easily adapting our analysis. The bounds of Gao, Oh, and Viswanath (2017a) are the first theoretical results on this estimator and offer some innovative insights into the performance advantage of the KSG estimator, although I again believe their the rate of their error bound is loose given their smoothness assumptions.

As described in the remainder of this chapter, Singh and Póczos (2016b) also require the density to have bounded support, but prove general results, in that they apply not only to the Kozachenko-Leonenko estimator, but in fact to any estimator based on the same k-NN approach, as described in the next section. In particular, this includes the estimators of Wang, Kulkarni, and Verdú (2009) for KL divergence and of Leonenko, Pronzato, and Savani (2008).[1]

It is worth noting that Gao, Oh, and Viswanath (2017b) recently studied estimation of a general simple integral functional F based on applying the estimator of Leonenko, Pronzato, and Savani (2008) to estimate terms of the F's Taylor expansion.

Leaving the finite-sample setting, (Bulinski and Dimitrov, 2018) recently proved asymptotic unbiasedness and $\mathcal{L}_2$ consistency of the Kozachenko-Leonenko estimator, even without making the Hölder or Sobolev smoothness assumptions typically made when analyzing these estimators.

Finally, a very recent result due to Jiao, Gao, and Han (2018), namely that the asymptotic convergence *rate* for estimating certain functionals, including entropy, depends crucially on whether we assume the density to be lower bounded away from zero (as this determines the smoothness of the functional over the class of distributions). For such functionals, if the density is allowed to take arbitrarily small positive values, then, rather than the rate of $O\left(n^{-\min\left\{\frac{8s}{4s+D},1\right\}}\right)$ MSE rate that is optimal for estimation of smooth functionals, or even the $O\left(n^{-\min\left\{\frac{2s}{D},1\right\}}\right)$ rate that we here derive for k-NN functionals in smooth settings, the minimax rate becomes, up to logarithmic factors, a slower rate of $O\left(n^{-\min\left\{\frac{2s}{s+D},1\right\}}\right)$. This result implies that, under these weaker assumptions, BCFk estimators (or at least Kozachenko-Leonenko-type entropy estimators) are minimax optimal, providing some theoretical explanation for their strong empirical performance.

3.1.2 Some intuition for BCFk estimators

These estimators are generally based on the following line of reasoning:

Let $\epsilon_k(x)$ denote the distance from x to its k-nearest neighbor in the sample $X_1, ..., X_n$ (i.e., $\epsilon_k = \inf\{\epsilon \geq 0 : \sum_{i=1}^n 1_{\{X_i \in B_\epsilon(x)\}} \geq k\}$). Then,

$$p(x) \approx \frac{P\left(B_{\epsilon_k(x)}(x)\right)}{\mu\left(B_{\epsilon_k(x)}(x)\right)} \approx \frac{k/n}{\mu\left(B_{\epsilon_k(x)}(x)\right)}.$$

[1] Note that early results along these lines, containing the key analysis ideas but specific to entropy estimation, are available in an unpublished technical report (Singh and Póczos, 2016a).

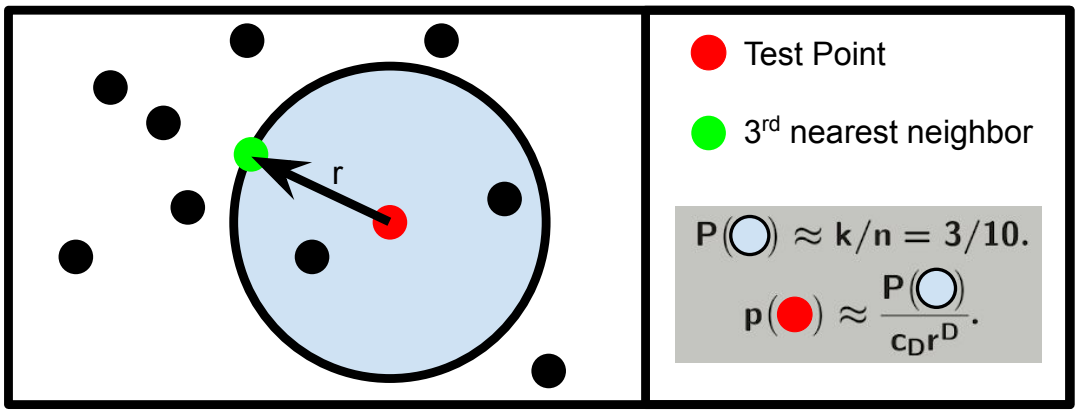


FIGURE 3.1: Illustration of kNN density estimation at a point (red) with $k = 3$, $n = 10$, $D = 2$.

An illustration of this estimate is provided in Figure 3.1.

In fact, it can be shown that, for Lebesgue-almost-all $x \in \mathcal{X}$, $\mathbb{E}\left[\frac{k}{n\epsilon_k^D(x)}\right] \to p(x)$ as $n \to \infty$ (and, moreover, this convergence is uniform over x when p is somewhat smooth). However, if k is fixed as $n \to \infty$, it is *not* necessarily the case that $\mathbb{V}\left[\frac{k}{n\epsilon_k^D(x)}\right] \to 0$ as $n \to \infty$; specifically, $\frac{k/n}{\epsilon_k^D(X)}$ converges, in distribution, to an Erlang distribution with shape parameter k and rate parameter depending on p. In particular, one can often analytically compute the expectation of the function f under this Erlang distribution, typically giving in a simple expression in terms of the estimand $F(p)$. BCFk solve this expression for $F(p)$, which provides a formula for asymptotic bias correction, resulting in an asymptotically unbiased estimate using a fixed value of fixed k.

It should be noted that the estimators we study are distinct from those recently studied by Sricharan, Raich, and Hero (2011), Sricharan, Wei, and Hero (2013), Sricharan, Raich, and Hero III (2012), Moon and Hero (2014b), Moon and Hero (2014a), and Moon (2016), which are also based on k-NN statistics. These estimators are based on plugging consistent k-NN density estimates into the desired functional, and hence, to be consistent, these estimates require $k \to \infty$ as $n \to \infty$. This increases the bias of these estimators (and their computational overhead), and they thus tend to converge more slowly than the bias-corrected estimators we study. However, Moon and Hero (2014b) showed that the convergence rate can be accelerated somewhat by using an ensemble of estimates (with different parameters k).[2] See Kevin Moon's Thesis (Moon, 2016) for an extensive discussion of these estimators.

3.2 Introduction

Estimating entropies and divergences of probability distributions in a consistent manner is of importance in a number of problems in machine learning. Entropy estimators have applications in goodness-of-fit testing (Goria, Leonenko, Mergel, and Novi Inverardi, 2005), parameter estimation in semi-parametric models (Wolsztynski, Thierry, and Pronzato, 2005a), studying fractal random walks (Alemany and Zanette, 1994), and texture classification (Hero, Ma, Michel, and Gorman, 2002a; Hero, Ma, Michel, and Gorman, 2002b). Divergence estimators have been used to

[2] In fact, Berrett, Samworth, and Yuan (2019) recently showed how this ensemble approach can also reduce bias of bias-corrected k-NN estimators, especially in higher dimensions.

generalize machine learning algorithms for regression, classification, and clustering from inputs in $\mathbb{R}^D$ to sets and distributions (Póczos, Xiong, Sutherland, and Schneider, 2012; Oliva, Poczos, and Schneider, 2013).

Divergences also include mutual informations as a special case; mutual information estimators have applications in feature selection (Peng, Long, and Ding, 2005), clustering (Aghagolzadeh, Soltanian-Zadeh, Araabi, and Aghagolzadeh, 2007), causality detection (Hlaváckova-Schindler, Palušb, Vejmelkab, and Bhattacharya, 2007), optimal experimental design (Lewi, Butera, and Paninski, 2007; Póczos and Lőrincz, 2009), fMRI data analysis (Chai, Walther, Beck, and Fei-Fei, 2009), prediction of protein structures (Adami, 2004), and boosting and facial expression recognition Shan, Gong, and Mcowan, 2005. Both entropy estimators and mutual information estimators have been used for independent component and subspace analysis (Learned-Miller and Fisher, 2003; Szabó, Póczos, and Lőrincz, 2007; Póczos and Lőrincz, 2005; Hulle, 2008), as well as for image registration (Hero, Ma, Michel, and Gorman, 2002a; Hero, Ma, Michel, and Gorman, 2002b). Further applications can be found in (Leonenko, Pronzato, and Savani, 2008).

This paper considers the more general problem of estimating functionals of the form

$$F(P) := \mathop{\mathbb{E}}_{X\sim P}\left[f(p(X))\right], \tag{3.1}$$

using n IID samples from P, where P is an unknown probability measure with smooth density function p and f is a known smooth function. We are interested in analyzing a class of nonparametric estimators based on k-nearest neighbor (k-NN) distance statistics. Rather than plugging a consistent estimator of p into (3.1), which requires $k \to \infty$ as $n \to \infty$, these estimators derive a bias correction for the plug-in estimator with *fixed* k; hence, we refer to this type of estimator as a fixed-k estimator. Compared to plug-in estimators, fixed-k estimators are faster to compute. As we show, fixed-k estimators can also exhibit superior rates of convergence.

As shown in Table 3.1, several authors have derived bias corrections necessary for fixed-k estimators of entropies and divergences, including, most famously, the Shannon entropy estimator of (Kozachenko and Leonenko, 1987). [3] The estimators in Table 3.1 estimators are known to be weakly consistent, [4] but, except for Shannon entropy, no finite sample bounds are known. The **main goal of this paper** is to provide finite-sample analysis of these estimators, via unified analysis of the estimator after bias correction. Specifically, we show conditions under which, for β-Hölder continuous ($\beta \in (0, 2]$) densities on D dimensional space, the bias of fixed-k estimators decays as $O\left(n^{-\beta/D}\right)$ and the variance decays as $O\left(n^{-1}\right)$, giving a mean squared error of $O\left(n^{-2\beta/D} + n^{-1}\right)$. Hence, the estimators converge at the parametric $O(n^{-1})$ rate when $\beta \geq D/2$, and at the slower rate $O(n^{-2\beta/D})$ otherwise. A modification of the estimators would be necessary to leverage additional smoothness for $\beta > 2$, but we do not pursue this here. Along the way, we prove a finite-sample version of the useful fact (Leonenko, Pronzato, and Savani, 2008) that (normalized) k-NN distances have an Erlang asymptotic distribution, which may be of independent interest.

Here, we present our results for distributions P supported on the unit cube in $\mathbb{R}^D$ because this significantly simplifies the statements of our results, but, as we discuss in the supplement, our results generalize fairly naturally, for example to distributions supported on smooth compact manifolds. In this context, it is worth noting

[3] MATLAB code for these estimators is in the ITE toolbox https://bitbucket.org/szzoli/ite/ (Szabó, 2014).

[4] Several of these proofs contain errors regarding the use of integral convergence theorems when their conditions do not hold, as described in (Poczos and Schneider, 2012).

Functional Name	Functional Form	Bias Correction	Ref.
Shannon Entropy	$\mathbb{E}\left[\log p(X)\right]$	Add. const.: $\psi(n) - \psi(k) + \log(k/n)$	Kozachenko and Leonenko (1987)Goria, Leonenko, Mergel, and Novi Inverardi (2005)
Rényi-α Entropy	$\mathbb{E}\left[p^{\alpha-1}(X)\right]$	Mult. const.: $\frac{\Gamma(k)}{\Gamma(k+1-\alpha)}$	Leonenko, Pronzato, and Savani (2008) and Leonenko and Pronzato (2010)
KL Divergence	$\mathbb{E}\left[\log \frac{p(X)}{q(X)}\right]$	None*	Wang, Kulkarni, and Verdú (2009)
Rényi-α Divergence	$\mathbb{E}\left[\left(\frac{p(X)}{q(X)}\right)^{\alpha-1}\right]$	Mult. const.: $\frac{\Gamma^2(k)}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$	Poczos and Schneider (2012)

TABLE 3.1: Functionals with known bias-corrected k-NN estimators, their bias corrections, and references. All expectations are over $X \sim P$. $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx$ is the gamma function, and $\psi(x) = \frac{d}{dx}\log\left(\Gamma(x)\right)$ is the digamma function. $\alpha \in \mathbb{R}\backslash\{1\}$ is a free parameter.
*For KL divergence, bias corrections for p and q cancel.

that our results scale with the *intrinsic* dimension of the manifold. As we discuss later, we believe deriving finite sample rates for distributions with *unbounded* support may require a truncated modification of the estimators we study (as in (Tsybakov and Meulen, 1996)), but we do not pursue this here.

3.3 Problem statement and notation

Let $\mathcal{X} := [0,1]^D$ denote the unit cube in $\mathbb{R}^D$, and let μ denote the Lebesgue measure. Suppose P is an unknown μ-absolutely continuous Borel probability measure supported on $\mathcal{X}$, and let $p : \mathcal{X} \to [0,\infty)$ denote the density of P. Consider a (known) differentiable function $f : (0,\infty) \to \mathbb{R}$. Given n samples $X_1, ..., X_n$ drawn IID from P, we are interested in estimating the functional

$$F(P) := \mathop{\mathbb{E}}_{X\sim P}\left[f(p(X))\right].$$

Somewhat more generally (as in divergence estimation), we may have a function $f : (0,\infty)^2 \to \mathbb{R}$ of two variables and a second unknown probability measure Q, with density q and n IID samples $Y_1, ..., Y_n$. Then, we are interested in estimating

$$F(P,Q) := \mathop{\mathbb{E}}_{X\sim P}\left[f(p(X), q(X))\right].$$

Fix $r \in [1,\infty]$ and a positive integer k. We will work with distances induced by the r-norm

$$\|x\|_r := \left(\sum_{i=1}^D x_i^r\right)^{1/r} \quad \text{and define} \quad c_{D,r} := \frac{\left(2\Gamma(1+1/r)\right)^D}{\Gamma(1+D/r)} = \mu(B(0,1)),$$

where $B(x,\varepsilon) := \{y \in \mathbb{R}^D : \|x-y\|_r < \varepsilon\}$ denotes the open radius-ε ball centered at x. Our estimators use k-nearest neighbor (k-NN) distances:

Definition 1. (k-NN distance): Given n IID samples $X_1, ..., X_n$ from P, for $x \in \mathbb{R}^D$, we define the *k-NN distance* $\varepsilon_k(x)$ by $\varepsilon_k(x) = \|x - X_i\|_r$, where X_i is the k^{th}-nearest element (in $\|\cdot\|_r$) of the set $\{X_1, ..., X_n\}$ to x. For divergence estimation, given n samples $Y_1, ..., Y_n$ from Q, then we similarly define $\delta_k(x)$ by $\delta_k(x) = \|x - Y_i\|_r$, where Y_i is the k^{th}-nearest element of $\{Y_1, ..., Y_n\}$ to x.

μ-absolute continuity of P precludes the existence of atoms (i.e., $\forall x \in \mathbb{R}^D$, $P(\{x\}) = \mu(\{x\}) = 0$). Hence, each $\varepsilon_k(x) > 0$ a.s. We will require this to study quantities such as $\log \varepsilon_k(x)$ and $1/\varepsilon_k(x)$.

3.4 Estimator

3.4.1 k-NN density estimation and plug-in functional estimators

The k-NN density estimator

$$\widehat{p}_k(x) = \frac{k/n}{\mu(B(x, \varepsilon_k(x)))} = \frac{k/n}{c_D \varepsilon_k^D(x)}$$

is well-studied nonparametric density estimator (Loftsgaarden and Quesenberry, 1965), motivated by noting that, for small $\varepsilon > 0$,

$$p(x) \approx \frac{P(B(x, \varepsilon))}{\mu(B(x, \varepsilon))},$$

and that, $P(B(x, \varepsilon_k(x))) \approx k/n$. One can show that, for $x \in \mathbb{R}^D$ at which p is continuous, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then $\widehat{p}_k(x) \to p(x)$ in probability ((Loftsgaarden and Quesenberry, 1965), Theorem 3.1). Thus, a natural approach for estimating $F(P)$ is the plug-in estimator

$$\widehat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f\left(\widehat{p}_k(X_i)\right). \tag{3.2}$$

Since $\widehat{p}_k \to p$ in probability pointwise as $k, n \to \infty$ and f is smooth, one can show $\widehat{F}_{PI}$ is consistent, and in fact derive finite sample convergence rates (depending on how $k \to \infty$). For example, (Sricharan, Raich, and Hero, 2011) show a convergence rate of $O\left(n^{-\min\left\{\frac{2\beta}{\beta+D}, 1\right\}}\right)$ for β-Hölder continuous densities (after sample splitting and boundary correction) by setting $k \asymp n^{\frac{\beta}{\beta+d}}$.

Unfortunately, while necessary to ensure $\mathbb{V}\left[\widehat{p}_k(x)\right] \to 0$, the requirement $k \to \infty$ is computationally burdensome. Furthermore, increasing k can increase the bias of $\widehat{p}_k$ due to over-smoothing (see (3.5) below), suggesting that this may be sub-optimal for estimating $F(P)$. Indeed, similar work based on kernel density estimation (Singh and Póczos, 2014a) suggests that, for plug-in functional estimation (as compared to density estimation), *under-smoothing* may be preferable, since the empirical mean results in additional smoothing.

3.4.2 Fixed-k functional estimators

An alternative approach is to fix k as $n \to \infty$. Since $\widehat{F}_{PI}$ is itself an empirical mean, unlike $\mathbb{V}\left[\widehat{p}_k(x)\right]$, $\mathbb{V}\left[\widehat{F}_{PI}\right] \to 0$ as $n \to \infty$. The more critical complication of fixing k is

bias. Since f is typically non-linear, the non-vanishing variance of $\widehat{p}_k$ translates into asymptotic bias. A solution adopted by several papers is to derive a bias correction function $\mathcal{B}$ (depending only on known factors) such that

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[\mathcal{B}\left(f\left(\frac{k/n}{\mu(B(x,\varepsilon_k(x))}\right)\right)\right] = \mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[f\left(\frac{P(B(x,\varepsilon_k(x)))}{\mu(B(x,\varepsilon_k(x))}\right)\right]. \tag{3.3}$$

For continuous p, the quantity

$$p_{\varepsilon_k(x)}(x) := \frac{P(B(x,\varepsilon_k(x)))}{\mu(B(x,\varepsilon_k(x))} \tag{3.4}$$

is a consistent estimate of $p(x)$ with k fixed, but it is not computable, since P is unknown. The bias correction $\mathcal{B}$ gives us an asymptotically unbiased estimator

$$\widehat{F}_{\mathcal{B}}(P) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{B}\left(f\left(\widehat{p}_k(X_i)\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{B}\left(f\left(\frac{k/n}{\mu(B(X_i,\varepsilon_k(X_i))}\right)\right).$$

that uses k/n in place of $P(B(x,\varepsilon_k(x)))$. This estimate extends naturally to divergences:

$$\widehat{F}_{\mathcal{B}}(P,Q) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{B}\left(f\left(\widehat{p}_k(X_i),\widehat{q}_k(X_i)\right)\right).$$

As an example, if $f = \log$ (as in Shannon entropy), then it can be shown that, for any continuous p,

$$\mathbb{E}\left[\log P(B(x,\varepsilon_k(x)))\right] = \psi(k) - \psi(n).$$

Hence, for $B_{n,k} := \psi(k) - \psi(n) + \log(n) - \log(k)$,

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[f\left(\frac{k/n}{\mu(B(x,\varepsilon_k(x))}\right)\right] + B_{n,k} = \mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[f\left(\frac{P(B(x,\varepsilon_k(x)))}{\mu(B(x,\varepsilon_k(x))}\right)\right].$$

giving the estimator of (Kozachenko and Leonenko, 1987). Other examples of functionals for which the bias correction is known are given in Table 3.1.

In general, deriving an appropriate bias correction can be quite a difficult problem specific to the functional of interest, and it is not our goal presently to study this problem; rather, we are interested in bounding the error of $\widehat{F}_{\mathcal{B}}(P)$, *assuming the bias correction is known*. Hence, our results apply to all of the estimators in Table 3.1, as well as any estimators of this form that may be derived in the future.

3.5 Related work

3.5.1 Estimating information theoretic functionals

Recently, there has been much work on analyzing estimators for entropy, mutual information, divergences, and other functionals of densities. Besides bias-corrected fixed-k estimators, most of this work has taken one of three approaches. One series of papers (Liu, Wasserman, and Lafferty, 2012; Singh and Póczos, 2014a; Singh and Póczos, 2014b) studied a boundary-corrected plug-in approach based on undersmoothed kernel density estimation. This approach has strong finite sample guarantees, but requires prior knowledge of the support of the density, and can have a slow rate of convergence. A second approach (Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014)

uses von Mises expansion to partially correct the bias of optimally smoothed density estimates. This is statistically more efficient, but can require computationally demanding numerical integration over the support of the density. A final line of work (Moon and Hero, 2014b; Moon and Hero, 2014a; Sricharan, Raich, and Hero, 2011; Sricharan, Wei, and Hero, 2013) studied plug-in estimators based on consistent, boundary corrected k-NN density estimates (i.e., with $k \to \infty$ as $n \to \infty$). (Nguyen, Wainwright, and Jordan., 2010) study a divergence estimator based on convex risk minimization, but this relies of the context of an RKHS, making results are difficult to compare.

Rates of Convergence: For densities over $\mathbb{R}^D$ satisfying a Hölder smoothness condition parametrized by $\beta \in (0, \infty)$, the minimax mean squared error rate for estimating functionals of the form $\int f(p(x))\,dx$ has been known since (Birgé and Massart, 1995) to be $O\left(n^{-\min\left\{\frac{8\beta}{4\beta+D},1\right\}}\right)$. (Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014) recently derived identical minimax rates for divergence estimation.

Most of the above estimators have been shown to converge at the rate $O\left(n^{-\min\left\{\frac{2\beta}{\beta+D},1\right\}}\right)$. Only the von Mises approach (Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014) is known to achieve the minimax rate for general β and D, but due to its computational demand ($O(2^D n^3)$), [5] the authors suggest using other statistically less efficient estimators for moderate sample size. Here, we show that, for $\beta \in (0, 2]$, bias-corrected fixed-k estimators converge at the relatively fast rate $O\left(n^{-\min\left\{\frac{2\beta}{D},1\right\}}\right)$. For $\beta > 2$, modifications are needed for the estimator to leverage the additional smoothness of the density. Notably, this rate is *adaptive*; that is, it does not require selecting a smoothing parameter depending on the unknown β; our results (Theorem 5) imply the above rate is achieved for *any* fixed choice of k. On the other hand, since no empirical error metric is available for cross-validation, parameter selection is an obstacle for competing estimators.

3.5.2 Prior analysis of fixed-k estimators

As of writing this paper, the only finite-sample results for $\widehat{F}_{\mathcal{B}}(P)$ were those of (Biau and Devroye, 2015a) for the Kozachenko-Leonenko (KL) [6] Shannon entropy estimator. (Kozachenko and Leonenko, 1987) Theorem 7.1 of (Biau and Devroye, 2015a) shows that, if the density p has compact support, then the variance of the KL estimator decays as $O(n^{-1})$. They also claim (Theorem 7.2) to bound the bias of the KL estimator by $O(n^{-\beta})$, under the assumptions that p is β-Hölder continuous ($\beta \in (0, 1]$), bounded away from 0, and supported on the interval $[0, 1]$. However, in their proof, (Biau and Devroye, 2015a) neglect to bound the additional bias incurred near the boundaries of $[0, 1]$, where the density cannot simultaneously be bounded away from 0 and continuous. In fact, because the KL estimator does not attempt to correct for boundary bias, it is not clear that the bias should decay as $O(n^{-\beta})$ under these conditions; we require additional conditions at the boundary of $\mathcal{X}$.

(Tsybakov and Meulen, 1996) studied a closely related entropy estimator for which they prove $\sqrt{n}$-consistency. Their estimator is identical to the KL estimator, except that it truncates k-NN distances at $\sqrt{n}$, replacing $\varepsilon_k(x)$ with $\min\{\varepsilon_k(x), \sqrt{n}\}$. This sort of truncation may be necessary for certain fixed-k estimators to satisfy

[5] Fixed-k estimators can be computed in $O\left(Dn^2\right)$ time, or $O\left(2^D n \log n\right)$ via k-d trees for small D.

[6] Not to be confused with Kullback-Leibler (KL) divergence, for which we also analyze an estimator.

finite-sample bounds for densities of *unbounded* support, though consistency can be shown regardless.

Finally, two very recent papers (Gao, Oh, and Viswanath, 2017a; Berrett, Samworth, and Yuan, 2019) have analyzed the KL estimator. In this case, (Gao, Oh, and Viswanath, 2017a) generalize the results of (Biau and Devroye, 2015a) to $D > 1$, and (Berrett, Samworth, and Yuan, 2019) weaken the regularity and boundary assumptions required by our bias bound, while deriving the same rate of convergence. Moreover, they show that, if k increases with n at the rate $k \asymp \log^5 n$, the KL estimator is asymptotically efficient (i.e., asymptotically normal, with optimal asymptotic variance). As explained in Section 3.9, together with our results this elucidates the role of k in the KL estimator: fixing k optimizes the convergence rate of the estimator, but increasing k slowly can further improve error by constant factors.

3.6 Discussion of assumptions

The lack of finite-sample results for fixed-k estimators is due to several technical challenges. Here, we discuss some of these challenges, motivating the assumptions we make to overcome them.

First, these estimators are sensitive to regions of low probability (i.e., $p(x)$ small), for two reasons:

1. Many functions f of interest (e.g., $f = \log$ or $f(z) = z^\alpha$, $\alpha < 0$) have singularities at 0.
2. The k-NN estimate $\widehat{p}_k(x)$ of $p(x)$ is highly biased when $p(x)$ is small. For example, for p β-Hölder continuous ($\beta \in (0, 2]$), one has ((Mack and Rosenblatt, 1979), Theorem 2)

$$\text{Bias}(\widehat{p}_k(x)) \asymp \left(\frac{k}{np(x)}\right)^{\beta/D}. \tag{3.5}$$

For these reasons, it is common in analysis of k-NN estimators to assume the following (Biau and Devroye, 2015a; Poczos and Schneider, 2012):

(A1) p is bounded away from zero on its support. That is, $p_* := \inf_{x \in \mathcal{X}} p(x) > 0$.

Second, unlike many functional estimators (see e.g., (Pál, Póczos, and Szepesvári, 2010; Sricharan, Raich, and Hero III, 2012; Singh and Póczos, 2014a)), the fixed-k estimators we consider do not attempt correct for boundary bias (i.e., bias incurred due to discontinuity of p on the boundary $\partial\mathcal{X}$ of $\mathcal{X}$). [7] The boundary bias of the density estimate $\widehat{p}_k(x)$ does vanish at x in the interior $\mathcal{X}^\circ$ of $\mathcal{X}$ as $n \to \infty$, but additional assumptions are needed to obtain finite-sample rates. Either of the following assumptions would suffice:

(A2) p is continuous not only on $\mathcal{X}^\circ$ but also on $\partial\mathcal{X}$ (i.e., $p(x) \to 0$ as $\text{dist}(x, \partial\mathcal{X}) \to 0$).

(A3) p is supported on all of $\mathbb{R}^D$. That is, the support of p has no boundary. This is the approach of (Tsybakov and Meulen, 1996), but we reiterate that, to handle an unbounded domain, they require truncating $\varepsilon_k(x)$.

Unfortunately, both assumptions **(A2)** and **(A3)** are inconsistent with **(A1)**. Our approach is to assume **(A2)** and replace assumption **(A1)** with a much milder assumption that p is *locally lower bounded* on its support in the following sense:

(A4) There exist $\rho > 0$ and a function $p_* : \mathcal{X} \to (0, \infty)$ such that, for all $x \in \mathcal{X}, r \in (0, \rho]$, $p_*(x) \leq \frac{P(B(x,r))}{\mu(B(x,r))}$.

[7] This complication was omitted in the bias bound (Theorem 7.2) of (Biau and Devroye, 2015a) for entropy estimation.

We show in Lemma 2 that assumption **(A4)** is in fact very mild; in a metric measure space of positive dimension D, as long as p is continuous on $\mathcal{X}$, such a p_* exists for *any* desired $\rho > 0$. For simplicity, we will use $\rho = \sqrt{D} = \operatorname{diam}(\mathcal{X})$.

As hinted by (3.5) and the fact that $F(P)$ is an expectation, our bounds will contain terms of the form

$$\mathop{\mathbb{E}}_{X\sim P}\left[\frac{1}{(p_*(X))^{\beta/D}}\right] = \int_{\mathcal{X}} \frac{p(x)}{(p_*(x))^{\beta/D}}\, d\mu(x)$$

(with an additional $f'(p_*(x))$ factor if f has a singularity at zero). Hence, our key assumption is that these quantities are finite. This depends primarily on *how quickly* p approaches zero near $\partial\mathcal{X}$. For many functionals, Lemma 6 gives a simple sufficient condition.

3.7 Preliminary lemmas

Here, we present some lemmas, both as a means of summarizing our proof techniques and also because they may be of independent interest for proving finite-sample bounds for other k-NN methods. Due to space constraints, all proofs are given in the appendix. Our first lemma states that, if p is continuous, then it is locally lower bounded as described in the previous section.

Lemma 2. (Existence of Local Bounds) *If p is continuous on $\mathcal{X}$ and strictly positive on the interior $\mathcal{X}^\circ$ of $\mathcal{X}$, then, for $\rho := \sqrt{D} = \operatorname{diam}(\mathcal{X})$, there exists a continuous function $p_* : \mathcal{X}^\circ \to (0,\infty)$ and a constant $p^* \in (0,\infty)$ such that*

$$0 < p_*(x) \le \frac{P(B(x,r))}{\mu(B(x,r))} \le p^* < \infty, \quad \forall x \in \mathcal{X}, r \in (0,\rho].$$

We now use these local lower and upper bounds to prove that k-NN distances concentrate around a term of order $(k/(np(x)))^{1/D}$. Related lemmas, also based on multiplicative Chernoff bounds, are used by (Kpotufe and Luxburg, 2011; Chaudhuri, Dasgupta, Kpotufe, and Luxburg, 2014) and (Chaudhuri and Dasgupta, 2014; Kontorovich and Weiss, 2015) to prove finite-sample bounds on k-NN methods for cluster tree pruning and classification, respectively. For cluster tree pruning, the relevant inequalities bound the error of the k-NN density estimate, and, for classification, they lower bound the probability of nearby samples of the same class. Unlike in cluster tree pruning, we are not using a consistent density estimate, and, unlike in classification, our estimator is a function of k-NN distances themselves (rather than their ordering). Thus, our statement is somewhat different, bounding the k-NN distances themselves:

Lemma 3. (Concentration of k-NN Distances) *Suppose p is continuous on $\mathcal{X}$ and strictly positive on $\mathcal{X}^\circ$. Let p_* and p^* be as in Lemma 2. Then, for any $x \in \mathcal{X}^\circ$,*

1. *if* $r > \left(\frac{k}{p_*(x)n}\right)^{1/D}$, *then* $\mathbb{P}\left[\varepsilon_k(x) > r\right] \le e^{-p_*(x)r^D n}\left(e\frac{p_*(x)r^D n}{k}\right)^k$.
2. *if* $r \in \left[0, \left(\frac{k}{p^* n}\right)^{1/D}\right)$, *then* $\mathbb{P}\left[\varepsilon_k(x) < r\right] \le e^{-p_*(x)r^D n}\left(\frac{ep^* r^D n}{k}\right)^{kp_*(x)/p^*}$.

It is worth noting an asymmetry in the above bounds: counter-intuitively, the lower bound depends on p_*. This asymmetry is related to the large bias of k-NN density estimators when p is small (as in (3.5)).

The next lemma uses Lemma 3 to bound expectations of monotone functions of the ratio $\widehat{p}_k/p_*$. As suggested by the form of integrals (3.6) and (3.7), this is essentially a finite-sample statement of the fact that (appropriately normalized) k-NN distances have Erlang asymptotic distributions; this asymptotic statement is key to consistency proofs of (Leonenko, Pronzato, and Savani, 2008) and (Poczos and Schneider, 2012) for α-entropy and divergence estimators.

Lemma 4. *Let p be continuous on $\mathcal{X}$ and strictly positive on $\mathcal{X}^\circ$. Define p_* and p^* as in Lemma 2. Suppose $f : (0,\infty) \to \mathbb{R}$ is continuously differentiable and $f' > 0$. Then, we have the upper bound* [8]

$$\sup_{x\in\mathcal{X}^\circ} \mathbb{E}\left[f_+\left(\frac{p_*(x)}{\widehat{p}_k(x)}\right)\right] \le f_+(1) + e\sqrt{k}\int_k^\infty \frac{e^{-y}y^k}{\Gamma(k+1)} f_+\left(\frac{y}{k}\right)\,dy, \tag{3.6}$$

and, for all $x \in \mathcal{X}^\circ$, for $\kappa(x) := kp_*(x)/p^*$, the lower bound

$$\mathbb{E}\left[f_-\left(\frac{p_*(x)}{\widehat{p}_k(x)}\right)\right] \le f_-(1) + e\sqrt{\frac{k}{\kappa(x)}}\int_0^{\kappa(x)} \frac{e^{-y}y^{\kappa(x)}}{\Gamma(\kappa(x)+1)} f_-\left(\frac{y}{k}\right)\,dy \tag{3.7}$$

Note that plugging the function $z \mapsto f\left(\left(\frac{kz}{c_{D,r}np_*(x)}\right)^{\frac{1}{D}}\right)$ into Lemma 4 gives bounds on $\mathbb{E}\left[f(\varepsilon_k(x))\right]$. As one might guess from Lemma 3 and the assumption that f is smooth, this bound is roughly of the order $\asymp \left(\frac{k}{np(x)}\right)^{\frac{1}{D}}$. For example, for any $\alpha > 0$, a simple calculation from (3.6) gives

$$\mathbb{E}\left[\varepsilon_k^\alpha(x)\right] \le \left(1+\frac{\alpha}{D}\right)\left(\frac{k}{c_{D,r}np_*(x)}\right)^{\frac{\alpha}{D}}. \tag{3.8}$$

(3.8) is used for our bias bound, and more direct applications of Lemma 4 are used in variance bound.

3.8 Main results

Here, we present our main results on the bias and variance of $\widehat{F}_{\mathcal{B}}(P)$. Again, due to space constraints, all proofs are given in the appendix. We begin with bounding the bias:

Theorem 5. (Bias Bound) *Suppose that, for some $\beta \in (0,2]$, p is β-Hölder continuous with constant $L > 0$ on $\mathcal{X}$, and p is strictly positive on $\mathcal{X}^\circ$. Let p_* and p^* be as in Lemma 2. Let $f : (0,\infty) \to \mathbb{R}$ be differentiable, and define $M_{f,p} : \mathcal{X} \to [0,\infty)$ by*

$$M_{f,p}(x) := \sup_{z\in[p_*(x),p^*]} \left|\frac{d}{dz}f(z)\right|$$

Assume

$$C_f := \mathop{\mathbb{E}}_{X\sim p}\left[\frac{M_{f,p}(X)}{(p_*(X))^{\frac{\beta}{D}}}\right] < \infty. \quad \textit{Then,} \quad \left|\mathbb{E}\,\widehat{F}_{\mathcal{B}}(P) - F(P)\right| \le C_f L\left(\frac{k}{n}\right)^{\frac{\beta}{D}}.$$

[8] $f_+(x) = \max\{0, f(x)\}$ and $f_-(x) = -\min\{0, f(x)\}$ denote the positive and negative parts of f. Recall that $\mathbb{E}\left[f(X)\right] = \mathbb{E}\left[f_+(X)\right] - \mathbb{E}\left[f_-(X)\right]$.

The statement for divergences is similar, assuming that q is also β-Hölder continuous with constant L and strictly positive on $\mathcal{X}^\circ$. Specifically, we get the same bound if we replace $M_{f,o}$ with

$$M_{f,p}(x) := \sup_{(w,z)\in[p_*(x),p^*]\times[q_*(x),q^*]} \left|\frac{\partial}{\partial w} f(w,z)\right|$$

and define $M_{f,q}$ similarly (i.e., with $\frac{\partial}{\partial z}$) and we assume that

$$C_f := \mathop{\mathbb{E}}_{X\sim p}\left[\frac{M_{f,p}(X)}{(p_*(X))^{\frac{\beta}{D}}}\right] + \mathop{\mathbb{E}}_{X\sim p}\left[\frac{M_{f,q}(X)}{(q_*(X))^{\frac{\beta}{D}}}\right] < \infty.$$

As an example of the applicability of Theorem 5, consider estimating the Shannon entropy. Then, $f(z) = \log(x)$, and so we need $C_f = \int_{\mathcal{X}} (p_*(x))^{-\beta/D}\, d\mu(x) < \infty$.

The assumption $C_f < \infty$ is not immediately transparent. For the functionals in Table 3.1, C_f has the form $\int_{\mathcal{X}} (p(x))^{-c}\, dx$, for some $c > 0$, and hence $C_f < \infty$ intuitively means $p(x)$ cannot approach zero too quickly as $\operatorname{dist}(x, \partial\mathcal{X}) \to 0$. The following lemma gives a formal sufficient condition:

Lemma 6. (Boundary Condition) *Let $c > 0$. Suppose there exist $b_\partial \in (0, \frac{1}{c})$, $c_\partial, \rho_\partial > 0$ such that, for all $x \in \mathcal{X}$ with $\varepsilon(x) := \operatorname{dist}(x, \partial\mathcal{X}) < \rho_\partial$, $p(x) \geq c_\partial \varepsilon^{b_\partial}(x)$. Then, $\int_{\mathcal{X}} (p_*(x))^{-c}\, d\mu(x) < \infty$.*

In the supplement, we give examples showing that this condition is fairly general, satisfied by densities proportional to x^{b_∂} near $\partial\mathcal{X}$ (i.e., those with at least b_∂ nonzero one-sided derivatives on the boundary).

We now bound the variance. The main obstacle here is that the fixed-k estimator is an empirical mean of *dependent* terms (functions of k-NN distances). We generalize the approach used by (Biau and Devroye, 2015a) to bound the variance of the KL estimator of Shannon entropy. The key insight is the geometric fact that, in $(\mathbb{R}^D, \|\cdot\|_p)$, there exists a constant $N_{k,D}$ (independent of n) such that any sample X_i can be amongst the k-nearest neighbors of at most $N_{k,D}$ other samples. Hence, at most $N_{k,D}+1$ of the terms in (3.2) can change when a single X_i is added, suggesting a variance bound via the Efron-Stein inequality (Efron and Stein, 1981), which bounds the variance of a function of random variables in terms of its expected change when its arguments are resampled. (Evans, 2008) originally used this approach to prove a general Law of Large Numbers (LLN) for nearest-neighbors statistics. Unfortunately, this LLN relies on bounded kurtosis assumptions that are difficult to justify for the $\log$ or negative power statistics we study.

Theorem 7. (Variance Bound) *Suppose $\mathcal{B} \circ f$ is continuously differentiable and strictly monotone. Assume $C_{f,p} := \mathbb{E}_{X\sim P}\left[\mathcal{B}^2(f(p_*(X)))\right] < \infty$, and $C_f := \int_0^\infty e^{-y} y^k f(y) < \infty$. Then, for*

$$C_V := 2\,(1 + N_{k,D})\,(3 + 4k)\,(C_{f,p} + C_f)\,, \quad \textit{we have} \quad \mathbb{V}\left[\widehat{F}_{\mathcal{B}}(P)\right] \leq \frac{C_V}{n}.$$

As an example, if $f = \log$ (as in Shannon entropy), then, since $\mathcal{B}$ is an additive constant, we simply require $\int_{\mathcal{X}} p(x) \log^2(p_*(x)) < \infty$. In general, $N_{k,D}$ is of the order $k2^{cD}$, for some $c > 0$. Our bound is likely quite loose in k; in practice, $\mathbb{V}\left[\widehat{F}_{\mathcal{B}}(P)\right]$ typically decreases somewhat with k.

3.9 Conclusions and discussion

In this paper, we gave finite-sample bias and variance error bounds for a class of fixed-k estimators of functionals of probability density functions, including the entropy and divergence estimators in Table 3.1. The bias and variance bounds in turn imply a bound on the mean squared error (MSE) of the bias-corrected estimator via the usual decomposition into squared bias and variance:

Corollary 8. (MSE Bound) *Under the conditions of Theorems 5 and 7,*

$$\mathbb{E}\left[\left(\widehat{F}_{\mathcal{B}}(P) - F(P)\right)^2\right] \leq C_f^2 L^2 \left(\frac{k}{n}\right)^{2\beta/D} + \frac{C_V}{n}. \tag{3.9}$$

Choosing k: Contrary to the name, fixing k is not *required* for "fixed-k" estimators. (Pérez-Cruz, 2009) empirically studied the effect of changing k with n and found that fixing $k = 1$ gave best results for estimating $F(P)$. However, there has been no theoretical justification for fixing k. Assuming tightness of our bias bound in k, we provide this in a worst-case sense: since our bias bound is nondecreasing in k and our variance bound is no larger than the minimax MSE rate for these estimation problems, reducing variance (i.e., increasing k) does not improve the (worst-case) convergence rate. On the other hand, (Berrett, Samworth, and Yuan, 2019) recently showed that slowly increasing k can improves the asymptotic variance of the estimator, with the rate $k \asymp \log^5 n$ leading to asymptotic efficiency. In view of these results, we suggest that increasing k can improve error by constant factors, but cannot improve the convergence rate.

Finally, we note that (Pérez-Cruz, 2009) found increasing k quickly (e.g., $k = n/2$) was *best* for certain hypothesis tests based on these estimators. Intuitively, this is because, in testing problems, bias is less problematic than variance (e.g., an asymptotically biased estimator can still lead to a consistent test).

3.10 A More General Setting

In the main paper, for the sake of clarity, we discussed only the setting of distributions on the D-dimensional unit cube $[0,1]^D$. For sake of generality, we prove our results in the significantly more general setting of a set equipped with a metric, a base measure, a probability density, and an appropriate definition of dimension. This setting subsumes Euclidean spaces, in which k-NN methods are usually analyzed, but also includes, for instance, Riemannian manifolds.

Definition 1. (Metric Measure Space): A quadruple $(\mathbb{X}, d, \Sigma, \mu)$ is called a *metric measure space* if $(\mathbb{X}, d)$ is a complete metric space, $(\mathbb{X}, \Sigma, \mu)$ is a σ-finite measure space, and Σ contains the Borel σ-algebra induced by d.

Definition 2. (Scaling Dimension): A metric measure space $(\mathbb{X}, d, \Sigma, \mu)$ has *scaling dimension* $D \in [0, \infty)$ if there exist constants $\mu_*, \mu^* > 0$ such that, $\forall r > 0$, $x \in \mathbb{X}$, $\mu_* \leq \frac{\mu(B(x,r))}{r^D} \leq \mu^*$. [9]

Remark 3. The above definition of dimension coincides with D in $\mathbb{R}^D$, where, under the L^p metric and Lebesgue measure,

$$\mu_* = \mu^* = \frac{(2\Gamma(1+1/p))^D}{\Gamma(1+D/p)}$$

[9] $B(x,r) := \{y \in \mathbb{X} : d(x,y) < r\}$ denotes the open ball of radius r centered at x.

is the usual volume of the unit ball. However, it is considerably more general than the vector-space definition of dimension. It includes, for example, the case that $\mathbb{X}$ is a smooth Riemannian manifold, with the standard metric and measure induced by the Riemann metric. In this case, our results scale with the *intrinsic* dimension of data, rather than the dimension of a space in which the data are embedded. Often, $\mu_* = \mu^*$, but leaving these distinct allows, for example, manifolds with boundary. The scaling dimension is slightly more restrictive than the well-studied doubling dimension of a measure, (Luukkainen and Saksman, 1998) which enforces only an upper bound on the rate of growth.

3.11 Proofs of Lemmas

Lemma 2. *Consider a metric measure space $(\mathbb{X}, d, \Sigma, \mu)$ of scaling dimension D, and a μ-absolutely continuous probability measure P, with density function $p : \mathbb{X} \to [0, \infty)$ supported on*

$$\mathcal{X} := \{x \in \mathbb{X} : p(x) > 0\}.$$

If p is continuous on $\mathcal{X}$, then, for any $\rho > 0$, there exists a function $p_* : \mathcal{X} \to (0, \infty)$ such that

$$0 < p_*(x) \leq \inf_{r \in (0,\rho]} \frac{P(B(x,r))}{\mu(B(x,r))}, \quad \forall x \in \mathcal{X},$$

and, if p is bounded above by $p^* := \sup_{x \in \mathcal{X}} p(x) < \infty$, then

$$\sup_{r \in (0,\rho]} \frac{P(B(x,r))}{\mu(B(x,r))} \leq p^* < \infty, \quad \forall r \in (0, \rho],$$

Proof: Let $x \in \mathcal{X}$. Since p is continuous and strictly positive at x, there exists $\varepsilon \in (0, \rho]$ such that and, for all $y \in B(x, \varepsilon)$, $p(y) \geq p(x)/2 > 0$. Define

$$p_*(x) := \frac{p(x)}{2} \frac{\mu_*}{\mu^*} \left(\frac{\varepsilon}{\rho}\right)^D.$$

Then, for any $r \in (0, \rho]$, since P is a non-negative measure, and μ has scaling dimension D,

$$\begin{aligned} P(B(x,r)) \geq P(B(x, \varepsilon r/\rho)) &\geq \mu(B(x, \varepsilon r/\rho)) \min_{y \in B(x, \varepsilon r/\rho)} p(y) \\ &\geq \mu(B(x, \varepsilon r/\rho)) \frac{p(x)}{2} \\ &\geq \frac{p(x)}{2} \mu_* \left(\frac{\varepsilon r}{\rho}\right)^D = p_*(x) \mu^* r^D \geq p_*(x) \mu(B(x,r)). \end{aligned}$$

Also, trivially, $\forall r \in (0, \rho]$,

$$P(B(x,r)) \leq \mu(B(x,r)) \max_{y \in B(x, r\rho/\varepsilon)} p(y) \leq p^*(x) \mu(B(x,r)).$$

■

Lemma 3. *Consider a metric measure space $(\mathbb{X}, d, \Sigma, \mu)$ of scaling dimension D, and a μ-absolutely continuous probability measure P, with continuous density function $p : \mathbb{X} \to [0, \infty)$ supported on*

$$\mathcal{X} := \{x \in \mathbb{X} : p(x) > 0\}.$$

For $x \in \mathcal{X}$, if $r > \left(\frac{k}{p_*(x)n}\right)^{1/D}$, then

$$\mathbb{P}\left[\varepsilon_k(x) > r\right] \leq e^{-p_*(x)r^D n}\left(e\frac{p_*(x)r^D n}{k}\right)^k.$$

and, if $r \in \left[0, \left(\frac{k}{p^* n}\right)^{1/D}\right)$, then

$$\mathbb{P}\left[\varepsilon_k(x) \leq r\right] \leq e^{-p_*(x)r^D n}\left(\frac{ep^* r^D n}{k}\right)^{kp_*(x)/p^*}.$$

Proof: Notice that, for all $x \in \mathcal{X}$ and $r > 0$,

$$\sum_{i=1}^n 1_{\{X_i \in B(x,r)\}} \sim \text{Binomial}\left(n, P(B(x,r))\right),$$

and hence that many standard concentration inequalities apply. Since we are interested in small r (and hence small $P(B(x,r))$), we prefer bounds on relative error, and hence apply multiplicative Chernoff bounds. If $r > (k/(p_*(x)n))^{1/D}$, then, by definition of p_*, $P(B(x,r)) < k/n$, and so, applying the multiplicative Chernoff bound with $\delta := \frac{p_*(x)r^D n - k}{p_*(x)r^D n} > 0$ gives

$$\begin{aligned}
\mathbb{P}\left[\varepsilon_k(x) > r\right] &= \mathbb{P}\left[\sum_{i=1}^n 1_{\{X_i \in B(x,r)\}} < k\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^n 1_{\{X_i \in B(x,r)\}} < (1-\delta)nP(B(x,r))\right] \\
&\leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{nP(B(x,r))} \\
&= e^{-p_*(x)r^D n}\left(\frac{ep_*(x)r^D n}{k}\right)^k.
\end{aligned}$$

Similarly, if $r < (k/(p^* n))^{1/D}$, then, applying the multiplicative Chernoff bound with $\delta := \frac{k - p^* r^D n}{p^* r^D n} > 0$,

$$\begin{aligned}
\mathbb{P}\left[\varepsilon_k(x) < r\right] &= \mathbb{P}\left[\sum_{i=1}^n 1_{\{X_i \in B(x,r)\}} \geq k\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^n 1_{\{X_i \in B(x,r)\}} \geq (1+\delta)nP(B(x,r))\right] \\
&\leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{nP(B(x,r))} \\
&\leq e^{-p_*(x)r^D n}\left(\frac{ep^* r^D n}{k}\right)^{kp_*(x)/p^*}
\end{aligned}$$

■

The bound we prove below is written in a somewhat different form from the version of Lemma 4 in the main paper. This form follows somewhat more intuitively from Lemma 3, but does not make obvious the connection to the asymptotic Erlang distribution. To derive the form in the paper, one simply integrates the integral below by parts, plugs in the function $x \mapsto f\left(p_*(x) \Big/ \frac{k/n}{c_D \varepsilon_k^D(x)}\right)$, and applies the bound $(e/k)^k \leq \frac{e}{\sqrt{k}\Gamma(k)}$.

Lemma 4. *Consider the setting of Lemma 3 and assume $\mathcal{X}$ is compact with diameter $\rho := \sup_{x,y\in\mathcal{X}} d(x,y)$. Suppose $f : (0,\rho) \to \mathbb{R}$ is continuously differentiable, with $f' > 0$. Then, for any $x \in \mathcal{X}$, we have the upper bound*

$$\mathbb{E}\left[f_+(\varepsilon_k(x))\right] \leq f_+\left(\left(\frac{k}{p_*(x)n}\right)^{\frac{1}{D}}\right) + \frac{(e/k)^k}{D(np_*(x))^{\frac{1}{D}}} \int_k^{np_*(x)\rho^D} e^{-y} y^{\frac{Dk+1-D}{D}} f'\left(\left(\frac{y}{np_*(x)}\right)^{\frac{1}{D}}\right) dy \tag{3.10}$$

and the lower bound

$$\mathbb{E}\left[f_-(\varepsilon_k(x))\right] \leq f_-\left(\left(\frac{k}{p^*n}\right)^{\frac{1}{D}}\right) + \frac{(e/\kappa(x))^{\kappa(x)}}{D\left(np_*(x)\right)^{\frac{1}{D}}} \int_0^{\kappa(x)} e^{-y} y^{\frac{D\kappa(x)+1-D}{D}} f'\left(\left(\frac{y}{np_*(x)}\right)^{\frac{1}{D}}\right) dy, \tag{3.11}$$

where $f_+(x) = \max\{0, f(x)\}$ and $f_-(x) = -\min\{0, f(x)\}$ denote the positive and negative parts of f, respectively, and $\kappa(x) := kp_*(x)/p^*$.

Proof: For notational simplicity, we prove the statement for $g(x) = f\left(np_*(x)x^D\right)$; the main result follows by substituting f back in.

Define

$$\varepsilon_0^+ = f_+\left(\left(\frac{k}{p_*(x)n}\right)^{\frac{1}{D}}\right) \quad \text{and} \quad \varepsilon_0^- = f_-\left(\left(\frac{k}{p^*n}\right)^{\frac{1}{D}}\right).$$

Writing the expectation in terms of the survival function,

$$\begin{aligned}
\mathbb{E}\left[f_+(\varepsilon_k(x))\right] &= \int_0^\infty \mathbb{P}\left[f(\varepsilon_k(x)) > \varepsilon\right] d\varepsilon \\
&= \int_0^{\varepsilon_0^+} \mathbb{P}\left[f(\varepsilon_k(x)) > \varepsilon\right] d\varepsilon + \int_{\varepsilon_0^+}^{f_+(\rho)} \mathbb{P}\left[f(\varepsilon_k(x)) > \varepsilon\right] d\varepsilon, \\
&\leq \varepsilon_0^+ + \int_{\varepsilon_0^+}^{f_+(\rho)} \mathbb{P}\left[f(\varepsilon_k(x)) > \varepsilon\right] d\varepsilon,
\end{aligned} \tag{3.12}$$

since f is non-decreasing and $\mathbb{P}\left[\varepsilon_k(x) > \rho\right] = 0$. By construction of ε_0^+, for all $\varepsilon > \varepsilon_0^+$, $f^{-1}(\varepsilon) > (k/(p_*(x)n))^{1/D}$. Hence, applying Lemma 3 followed by the change of variables $y = np_*(x)\left(f^{-1}(\varepsilon)\right)^D$ gives [10]

$$\begin{aligned}
\int_{\varepsilon_0^+}^{f_+(\rho)} \mathbb{P}\left[\varepsilon_k(x) > f^{-1}(\varepsilon)\right] d\varepsilon &\leq \int_{\varepsilon_0^+}^{f_+(\rho)} e^{-np_*(x)\left(f^{-1}(\varepsilon)\right)^D} \left(\frac{enp_*(x)\left(f^{-1}(\varepsilon)\right)^D}{k}\right)^k d\varepsilon \\
&= \frac{(e/k)^k}{D(np_*(x))^{\frac{1}{D}}} \int_k^{np_*(x)\rho^D} e^{-y} y^{\frac{kD+1-D}{D}} f'\left(\left(\frac{y}{np_*(x)}\right)^{\frac{1}{D}}\right) dy,
\end{aligned}$$

[10] f need not be surjective, but the generalized inverse $f^{-1} : [-\infty, \infty] \to [0, \infty]$ defined by $f^{-1}(\varepsilon) := \inf\{x \in (0,\infty) : f(x) \geq \varepsilon\}$ suffices here.

Together with (3.12), this gives the upper bound (3.10). Similar steps give

$$\mathbb{E}\left[f(\varepsilon_k(x))\right] \leq \varepsilon_0^- + \int_{\varepsilon_0^-}^{f_-(0)} \mathbb{P}\left[f(\varepsilon_k(x)) < -\varepsilon\right] d\varepsilon. \tag{3.13}$$

Applying Lemma 3 followed the change of variables $y = np_*(x)\left(f^{-1}(-\varepsilon)\right)^D$ gives

$$\int_{\varepsilon_0^-}^{f_-(\rho)} \mathbb{P}\left[\varepsilon_k(x) < f^{-1}(-\varepsilon)\right] d\varepsilon \leq \frac{(e/\kappa(x))^{\kappa(x)}}{D\left(np_*(x)\right)^{\frac{1}{D}}} \int_0^{\kappa(x)} e^{-y} y^{\frac{D\kappa(x)+1-D}{D}} f'\left(\left(\frac{y}{np_*(x)}\right)^{\frac{1}{D}}\right) dy$$

Together with inequality (3.13), this gives the result (3.11). ■

3.11.1 Applications of Lemma 4

When $f(x) = \log(x)$, (3.10) gives

$$\mathbb{E}\left[\log_+(\varepsilon_k(x))\right] \leq \frac{1}{D}\log_+\left(\frac{k}{p_*(x)n}\right) + \left(\frac{e}{k}\right)^k \frac{\Gamma(k,k)}{D} \leq \frac{1}{D}\left(\log_+\left(\frac{k}{p_*(x)n}\right) + 1\right)$$

and (3.11) gives [11]

$$\mathbb{E}\left[\log_-(\varepsilon_k(x))\right] \leq \frac{1}{D}\left(\log_-\left(\frac{k}{p^*n}\right) + \left(\frac{e}{\kappa(x)}\right)^{\kappa(x)} \gamma(\kappa(x),\kappa(x))\right) \tag{3.14}$$

$$\leq \frac{1}{D}\left(\log_-\left(\frac{k}{p^*n}\right) + \frac{1}{\kappa(x)}\right). \tag{3.15}$$

For $\alpha > 0$, $f(x) = x^\alpha$, (3.10) gives

$$\begin{aligned}\mathbb{E}\left[\varepsilon_k^\alpha(x)\right] &\leq \left(\frac{k}{p_*(x)n}\right)^{\frac{\alpha}{D}} + \left(\frac{e}{k}\right)^k \frac{\alpha\Gamma\left(k+\alpha/D,k\right)}{D(np_*(x))^{\alpha/D}} \\ &\leq C_2\left(\frac{k}{p_*(x)n}\right)^{\frac{\alpha}{D}},\end{aligned} \tag{3.16}$$

where $C_2 = 1 + \frac{\alpha}{D}$. For any $\alpha \in [-D\kappa(x), 0]$, when $f(x) = -x^\alpha$, (3.11) gives

$$\mathbb{E}\left[\varepsilon_k^\alpha(x)\right] \leq \left(\frac{k}{p^*n}\right)^{\frac{\alpha}{D}} + \left(\frac{e}{\kappa(x)}\right)^{\kappa(x)} \frac{\alpha\gamma\left(\kappa(x)+\alpha/D,\kappa(x)\right)}{D(np_*(x))^{\alpha/D}} \tag{3.17}$$

$$\leq C_3\left(\frac{k}{p^*n}\right)^{\frac{\alpha}{D}}, \tag{3.18}$$

where $C_3 = 1 + \frac{\alpha}{D\kappa(x)+\alpha}$.

[11] $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}\,dt$ and $\gamma(s,x) := \int_0^x t^{s-1}e^{-t}\,dt$ denote the upper and lower incomplete Gamma functions respectively. We used the bounds $\Gamma(s,x), x\gamma(s,x) \leq x^s e^{-x}$.

3.12 Proof of Bias Bound

Theorem 5. *Consider the setting of Lemma 3. Suppose Suppose p is β-Hölder continuous, for some $\beta \in (0, 2]$. Let $f : (0, \infty) \to \mathbb{R}$ be differentiable, and define $M_f : \mathcal{X} \to [0, \infty)$ by*

$$M_f(x) := \sup_{z \in \left[\frac{p_*(x)}{\mu^*}, \frac{p^*}{\mu_*}\right]} \|\nabla f(z)\|$$

(assuming this quantity is finite for almost all $x \in \mathcal{X}$). Suppose that

$$C_M := \mathop{\mathbb{E}}_{X \sim p}\left[\frac{M_f(X)}{(p_*(X))^{\frac{\beta}{D}}}\right] < \infty.$$

Then, for $C_B := C_M L$,

$$\left|\mathop{\mathbb{E}}_{X, X_1, \ldots, X_n \sim P}\left[f(p_{\varepsilon_k(X)}(X))\right] - F(p)\right| \leq C_B \left(\frac{k}{n}\right)^{\frac{\beta}{D}}.$$

Proof: By construction of p_* and p^*,

$$p_*(x) \leq p_\varepsilon(x) = \frac{P(B(x, \varepsilon))}{\mu(B(x, \varepsilon))} \leq p^*.$$

Also, by the Lebesgue differentiation theorem (Lebesgue, 1910), for μ-almost all $x \in \mathcal{X}$,

$$p_*(x) \leq p(x) \leq p^*.$$

For all $x \in \mathcal{X}$, applying the mean value theorem followed by inequality (3.16),

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim p}\left[\left|f(p(x)) - f(p_{\varepsilon_k(x)}(x))\right|\right] &\leq \mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim p}\left[\|\nabla f(\xi(x))\| \left|p(x) - p_{\varepsilon_k(x)}(x)\right|\right] \\
&\leq M_f(x) \mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim p}\left[\left|p(x) - p_{\varepsilon_k(x)}(x)\right|\right] \\
&\leq \frac{M_f(x) L D}{D + \beta} \mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim P}\left[\varepsilon_k^\beta(x)\right] \\
&\leq \frac{C_2 M_f(x) L D}{D + \beta}\left(\frac{k}{p_*(x) n}\right)^{\frac{\beta}{D}}
\end{aligned}$$

Hence,

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim p}\left[F(p) - \widehat{F}(p)\right]\right| &= \left|\mathop{\mathbb{E}}_{X \sim p}\left[\mathop{\mathbb{E}}_{X_1, \ldots, X_n \sim p}\left[f(p(X)) - f(p_{\varepsilon_k(X)}(X))\right]\right]\right| \\
&\leq \frac{C_2 L D}{D + \beta} \mathop{\mathbb{E}}_{X \sim p}\left[\frac{M_f(X)}{(p_*(X))^{\frac{\beta}{D}}}\right]\left(\frac{k}{n}\right)^{\frac{\beta}{D}} = \frac{C_2 C_M L D}{D + \beta}\left(\frac{k}{n}\right)^{\frac{\beta}{D}}.
\end{aligned}$$

■

Lemma 6. *Let $c > 0$. Suppose there exist $b_\partial \in (0, \frac{1}{c})$, $c_\partial, \rho_\partial > 0$ such that for all $x \in \mathcal{X}$ with $\varepsilon(x) := \operatorname{dist}(x, \partial\mathcal{X}) < \rho_\partial$, $p(x) \geq c_\partial \varepsilon^{b_\partial}(x)$. Then,*

$$\int_{\mathcal{X}} (p_*(x))^{-c}\, d\mu(x) < \infty.$$

Proof: Let $\mathcal{X}_\partial := \{x \in \mathcal{X} : \mathrm{dist}(x, \partial\mathcal{X}) < \rho_\partial\}$ denote the region within ρ_∂ of $\partial\mathcal{X}$. Since p_* is continuous and strictly positive on the compact set $\mathcal{X}\backslash\mathcal{X}_\partial$, it has a positive lower bound $\ell := \inf_{x\in\mathcal{X}\backslash\mathcal{X}_\partial}$ on this set, and it suffices to show

$$\int_{\mathcal{X}\backslash\mathcal{X}_\partial} (p_*(x))^{-c}\, d\mu(x) < \infty.$$

For all $x \in \mathcal{X}_\partial$,

$$p_*(x) \geq \frac{\min\{\ell, c_\partial \varepsilon^{b_\partial}(x)\}}{\mu(B(x, \sqrt{D}))}.$$

Hence,

$$\int_{\mathcal{X}\backslash\mathcal{X}_\partial} (p_*(x))^{-c}\, d\mu(x) \leq \int_{\mathcal{X}\backslash\mathcal{X}_\partial} \ell^{-c}\, d\mu(x) + \int_{\mathcal{X}\backslash\mathcal{X}_\partial} c_\partial^{-c}\varepsilon^{-b_\partial/c}(x)\, d\mu(x).$$

The first integral is trivially bounded by ℓ^{-c}. Since $\partial\mathcal{X}$ is the union of $2D$ "squares" of dimension $D-1$, the second integral can be reduced to the sum of $2D$ integrals of dimension 1, giving the bound

$$2Dc_\partial^{-c} \int_0^{\rho_\partial} x^{-b_\partial/c}(x)\, dx.$$

Since $b_\partial/c < 1$, the integral is finite. ∎

For concreteness, we give an illustrative example of how Lemma 6 is useful.

Example: Consider the one-dimensional density $p(x) = (\alpha+1)x^\alpha$ on $(0,1)$. Though the lower bound p_* provided by Lemma 2 is somewhat loose in this case, notice that, for $x < r \in (0,1)$,

$$\frac{P(B(x,r))}{\mu(B(x,r))} \geq \frac{(x+r)^{\alpha+1}}{2r} \geq \frac{(x(1+1/\alpha))^{\alpha+1}}{2x/\alpha} = \frac{\alpha(1+1/\alpha)^{\alpha+1}}{2} x^\alpha,$$

and, for $r < x \in (0,1)$,

$$\frac{P(B(x,r))}{\mu(B(x,r))} = \frac{(x+r)^{\alpha+1} - (x-r)^{\alpha+1}}{2r} \geq \frac{2rx^\alpha}{2r} = x^\alpha.$$

In either case, for $C_\alpha := \min\left\{1, \alpha(1+1/\alpha)^{\alpha+1}/2\right\}$, we have

$$p_*(x) := C_\alpha x^\sigma \leq \frac{P(B(x,r))}{\mu(B(x,r))}.$$

Thus, we have a local lower bound p_* of the form in Lemma, satisfying the conditions of Lemma 6 with $b_\partial = \alpha$.

Now consider more general densities p on $(0,1)$. If $p(0) = 0$ and p is right-differentiable at 0 with $\lim_{h\to 0} \frac{p(h)}{h} > 0$ (i.e., the one-sided Taylor expansion of p at 0 has a non-zero first-order coefficient), then, near 0, p is proportional to x. This intuition can be formalized to show that the example above extends to quite general distributions.

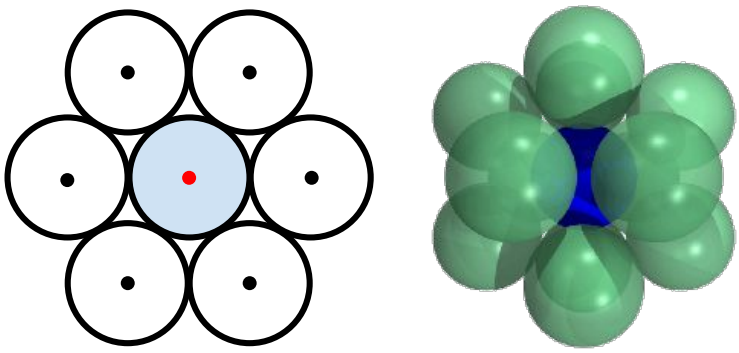


FIGURE 3.2: Illustrations of kissing numbers $N_{1,2} = 6$ and $N_{1,3} = 12$, which bound the number of points for which any fixed point can be the nearest neighbor. The existence of such a constant, together with the Efron-Stein inequality, form the basis for bounds on the variance of BCFk estimators.

3.13 Proof of Variance Bound

Theorem 7. (Variance Bound) *Suppose that $\mathcal{B} \circ f$ is continuously differentiable and strictly monotone. Let $N_{k,D}$ denote the maximum number of points of which a fixed point x can be the k-nearest neighbor. Note that this constant depending only on the geometry of the sample space, as an example, when $k = 1$, $N_{k,D}$ is precisely the kissing number in $\mathbb{R}^D$, illustrated in Figure 3.2. Assume that $C_{f,p} := \mathbb{E}_{X\sim P}\left[\mathcal{B}^2(f(p_*(X)))\right] < \infty$, and that $C_f := \int_0^\infty e^{-y} y^k f(y) < \infty$. Then, for*

$$C_V := 2\left(1 + N_{k,D}\right)(3 + 4k)\left(C_{f,p} + C_f\right), \quad \textit{we have} \quad \mathbb{V}\left[\widehat{F}_{\mathcal{B}}(P)\right] \leq \frac{C_V}{n}.$$

Proof: For convenience, define

$$H_i := \mathcal{B}\left(f\left(\frac{k/n}{\mu\left(B(X_i, \varepsilon_k(X_i))\right)}\right)\right).$$

By the Efron-Stein inequality (Efron and Stein, 1981) and the fact that the $\widehat{F}_{\mathcal{B}}(P)$ is symmetric in $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbb{V}\left[\widehat{F}_{\mathcal{B}}(P)\right] &\leq \frac{n}{2}\,\mathbb{E}\left[\left(\widehat{F}_{\mathcal{B}}(P) - F'_{\mathcal{B}}(P)\right)^2\right] \\
&\leq n\,\mathbb{E}\left[\left(\widehat{F}_{\mathcal{B}}(P) - F_{2:n}\right)^2 + \left(\widehat{F}'_{\mathcal{B}}(P) - F_{2:n}\right)^2\right] \\
&= 2n\,\mathbb{E}\left[\left(\widehat{F}_{\mathcal{B}}(P) - F_{2:n}\right)^2\right],
\end{aligned}$$

where $\widehat{F}'_{\mathcal{B}}(P)$ denotes the estimator after X_1 is resampled, and $F_{2:n} := \frac{1}{n}\sum_{i=2}^n H_i$. Then,

$$n(\widehat{F}_n(P) - F_{2:n}) = H_1 + \sum_{i=2}^n 1_{E_i}\left(H_i - H'_i\right),$$

where 1_{E_i} is the indicator function of the event $E_i = \{\varepsilon_k(X_i) \neq \varepsilon'_k(X_i)\}$. By Cauchy-Schwarz followed by the definition of $N_{k,D}$,

$$\begin{aligned} n^2(\widehat{F}_n(P) - \widehat{F}_{n-1}(P))^2 &= \left(1 + \sum_{i=2}^{n} 1_{E_i}\right)\left(H_1^2 + \sum_{i=2}^{n} 1_{E_i}\left(H_i - H'_i\right)^2\right) \\ &= (1 + N_{k,D})\left(H_1^2 + \sum_{i=2}^{n} 1_{E_i}\left(H_i - H'_i\right)^2\right) \\ &\leq (1 + N_{k,D})\left(H_1^2 + 2\sum_{i=2}^{n} 1_{E_i}\left(H_i^2 + H_i'^2\right)\right). \end{aligned}$$

Taking expectations, since the terms in the summation are identically distributed, we need to bound

$$\mathbb{E}\left[H_1^2\right], \tag{3.19}$$

$$(n-1)\,\mathbb{E}\left[1_{E_2} H_2^2\right] \tag{3.20}$$

and $$(n-1)\,\mathbb{E}\left[1_{E_2} H_2'^2\right]. \tag{3.21}$$

Bounding (3.19)**:** Note that

$$\mathbb{E}\left[H_1^2\right] = \mathbb{E}\left[\mathcal{B}^2\left(f\left(\widehat{p}_k(X_1)\right)\right)\right] = \mathbb{E}\left[\mathcal{B}^2\left(g\left(\frac{p_*(x)}{\widehat{p}_k(x)}\right)\right)\right]$$

for $g(y) = f\left(p_*(x)/y\right)$. Applying the upper bound in Lemma 4, if $\mathcal{B}^2 \circ g$ is increasing,

$$\mathbb{E}\left[H_1^2\right] \leq \mathcal{B}^2(g(1)) + \frac{e\sqrt{k}}{\Gamma(k+1)} C_\uparrow = \mathcal{B}^2(f(p_*(x))) + \frac{e\sqrt{k}}{\Gamma(k+1)} C_\uparrow.$$

If $\mathcal{B}^2 \circ g$ is decreasing, we instead use the lower bound in Lemma 4, giving a similar result. If $\mathcal{B}^2 \circ g$ is not monotone (i.e., if $\mathcal{B} \circ g$ takes both negative and positive values), then, since $\mathcal{B} \circ f$ *is* monotone (by assumption), we can apply the above steps to $(\mathcal{B} \circ g)_-$ and $(\mathcal{B} \circ g)_+$, which *are* monotone, and add the resulting bounds.

Bounding (3.20)**:** Since $\{\varepsilon_k(X_2) \neq \varepsilon'_k(X_2)\}$ is precisely the event that X_1 is amongst the k-NN of X_2, $\mathbb{P}\left[\varepsilon_k(X_i) \neq \varepsilon'_k(X_i)\right] = k/(n-1)$. Thus, since E_2 is independent of $\varepsilon_k(X_2)$ and

$$(n-1)\,\mathbb{E}\left[1_{E_2} H_2^2\right] = (n-1)\,\mathbb{E}\left[1_{E_2}\right]\mathbb{E}\left[H_2^2\right] = k\,\mathbb{E}\left[H_2^2\right] = k\,\mathbb{E}\left[H_1^2\right],$$

and we can use the bound for (3.19).

Bounding (3.21)**:** Since E_2 is independent of $\varepsilon_{k+1}(X_2)$ and

$$\begin{aligned} (n-1)\,\mathbb{E}\left[1_{E_2} H_2'^2\right] &= (n-1)\,\mathbb{E}\left[1_{E_2}\mathcal{B}^2\left(f\left(\widehat{p}_{k+1}(X_2)\right)\right)\right] \\ &= (n-1)\,\mathbb{E}\left[1_{E_2}\right]\mathbb{E}\left[\mathcal{B}^2\left(f\left(\widehat{p}_{k+1}(X_2)\right)\right)\right] = k\,\mathbb{E}\left[\mathcal{B}^2\left(f\left(\widehat{p}_{k+1}(X_2)\right)\right)\right]. \end{aligned}$$

Hence, we can again use the same bound as for (3.19), except with $k+1$ instead of k.

Combining these three terms gives the final result.

■

Chapter 4

Nonparanormal Information Estimation

4.1 Introduction

In the previous chapters, while estimating information theoretic quantities, we've striven to make minimal assumptions on the distribution of the data, focusing on Hölder- or Sobolev-type smoothness assumptions. Unfortunately, minimax convergence rates under these weak assumptions scale very poorly with the dimension; the number of samples required to guarantee an MSE of at most $\epsilon > 0$ scales, for some constant $c > 0$, as ϵ^{-cD}. Quite simply, these spaces are too large to estimate their parameters except in very low dimensions. Empirically, it has been found that information estimators for this setting fail to converge at realistic sample sizes in all but very low dimensions. Moreover, most nonparametric estimators are sensitive to tuning of bandwidth parameters, which is problematic for information estimation, since empirical error estimates are typically not available to enable cross-validation.

At another extreme, for Gaussian data, Cai, Liang, and Zhou (2015) have shown that consistent, parameter-free information estimation is tractable even in the high-dimensional case where D increases quickly with n (specifically, as long as $D/n \to 0$). However, optimal estimators for the Gaussian setting rely strongly on the assumption of joint Gaussianity, and their performance can degrade quickly when the data deviate from Gaussian. Especially in high dimensions, it is unlikely that data are jointly Gaussian, making these estimators brittle in practice.

Given these factors, though the nonparametric and Gaussian cases are fairly well understood in theory, there remains a lack of practical information estimators for the common case where data are neither exactly Gaussian nor very low-dimensional. The **main goal of this chapter** is to fill the gap between these two extreme settings by studying information estimation in a semiparametric compromise between the two, known as the "nonparanormal" (a.k.a. "Gaussian copula") model (see Definition 9 below). The nonparanormal model, analogous to the additive model popular in regression (Friedman and Stuetzle, 1981), limits complexity of interactions among variables but makes minimal assumptions on the marginal distribution of each variable. The result scales better with dimension than nonparametric models, while being more robust than Gaussian models.

4.2 Problem statement and notation

There are a number of distinct generalizations of mutual information to more than two variables. The definition we consider is simply the difference between the sum of marginal entropies and the joint entropy:

Definition 8. (Multivariate mutual information) Let $X_1, \ldots, X_D$ be $\mathbb{R}$-valued random variables with a joint probability density $p : \mathbb{R}^D \to [0, \infty)$ and marginal densities $p_1, ..., p_D : \mathbb{R} \to [0, \infty)$. The *multivariate mutual information* $I(X)$ *of* $X = (X_1, \ldots, X_D)$ is defined by

$$\begin{aligned} I(X) &:= \mathop{\mathbb{E}}_{X \sim p}\left[\log\left(\frac{p(X)}{\prod_{j=1}^{D} p_j(X_j)}\right)\right] \\ &= \sum_{j=1}^{D} H(X_j) - H(X), \end{aligned} \tag{4.1}$$

where $H(X) = -\mathbb{E}_{X \sim p}[\log p(X)]$ denotes entropy of X.

This notion of multivariate mutual information, originally due to Watanabe (1960) (who called it "total correlation") measures total dependency, or redundancy, within a set of D random variables. It has also been called the "multivariate constraint" (Garner, 1962) and "multi-information" (Studený and Vejnarová, 1998). Many related information theoretic quantities can be expressed in terms of $I(X)$, and can thus be estimated using estimators of $I(X)$. Examples include pairwise mutual information $I(X, Y) = I((X, Y)) - I(X) - I(Y)$, which measures dependence between (potentially multivariate) random variables X and Y, conditional mutual information

$$I(X|Z) = I((X, Z)) - \sum_{j=1}^{D} I((X_j, Z)),$$

which is useful for characterizing how much dependence within X can be explained by a latent variable Z (Studený and Vejnarová, 1998), and transfer entropy (a.k.a. "directed information") $T_{X \to Y}$, which measures predictive power of one time series X on the future of another time series Y. $I(X)$ is also related to entropy via Eq. (4.1), but, unlike the above quantities, this relationship depends on the marginal distributions of X, and hence involves some additional considerations, as discussed in Section 4.8.

We now define the class of nonparanormal distributions, from which we assume our data are drawn.

Definition 9. (Nonparanormal distribution, a.k.a. Gaussian copula model) A random vector $X = (X_1, \ldots, X_D)^T$ is said to have a *nonparanormal distribution* (denoted $X \sim \mathcal{NPN}(\Sigma; f)$) if there exist functions $\{f_j\}_{j=1}^{D}$ such that each $f_j : \mathbb{R} \to \mathbb{R}$ is a diffeomorphism [1] and $f(X) \sim \mathcal{N}(0, \Sigma)$, for some (strictly) positive definite $\Sigma \in \mathbb{R}^{D \times D}$ with 1's on the diagonal (i.e., each $\sigma_j = \Sigma_{j,j} = 1$). [2] Σ is called the *latent covariance* of X and f is called the *marginal transformation* of X.

The nonparanormal family relaxes many constraints of the Gaussian family. As illustrated in a few example in Figure 4.1, nonparanormal distributions can be multimodal, skewed, or heavy-tailed, can encode noisy nonlinear dependencies, and need not be supported on all of $\mathbb{R}^D$. Minimal assumptions are made on the marginal distributions; any desired continuously differentiable marginal cumulative distribution function (CDF) F_i of variable X_i corresponds to marginal transformation

[1] A diffeomorphism is a continuously differentiable bijection $g : \mathbb{R} \to R \subseteq \mathbb{R}$ such that g^{-1} is continuously differentiable.

[2] Setting $\mathbb{E}[f(X)] = 0$ and each $\sigma_j = 1$ ensures model identifiability, but does not reduce the model space, since these parameters can be absorbed into the marginal transformation f.

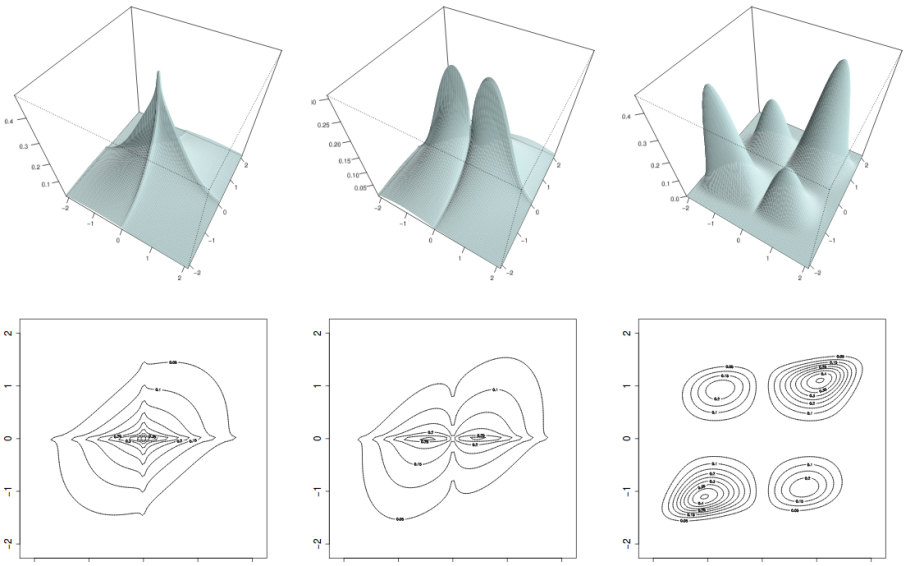


FIGURE 4.1: Surface and contour plots of three example nonparanormal densities. Figure taken from Liu, Lafferty, and Wasserman (2009).

$f_i(x) = \Phi^{-1}(F_i(x))$ (where Φ is the standard normal CDF). As examples, for a Gaussian variable Z, the 2-dimensional case, $X_1 \sim \mathcal{N}(0,1)$, and $X_2 = T(X_1 + Z)$ is nonparanormal when $T(x) = x^3$, $T = \tanh$, $T = \Phi$, or any other diffeomorphism. On the other hand, the limits of the Gaussian copula appear, for example, when $T(x) = x^2$, which is not bijective; then, if $\mathbb{E}[Z] = 0$, the Gaussian copula approximation of (X_1, X_2) models X_1 and X_2 as independent.

We are now ready to formally state our problem:

Formal Problem Statement: *Given n i.i.d. samples $X_1, ..., X_n \sim \mathcal{NPN}(\Sigma; f)$, where Σ and f are both unknown, we would like to estimate $I(X)$.*

Other notation: D denotes the dimension of the data (i.e., $\Sigma \in \mathbb{R}^{D \times D}$ and $f : \mathbb{R}^D \to \mathbb{R}^D$). For a positive integer k, $[k] = \{1, ..., k\}$ denotes the set of positive integers less than k (inclusive). For consistency, where possible, we use $i \in [n]$ to index samples and $j \in [D]$ to index dimensions (so that, e.g., $X_{i,j}$ denotes the j^{th} dimension of the i^{th} sample). Given a data matrix $X \in \mathbb{R}^{n \times D}$, our estimators depend on the empirical rank matrix

$$R \in [n]^{n \times D} \quad \text{with} \quad R_{i,j} := \sum_{k=1}^{n} 1_{\{X_{i,j} \geq X_{k,j}\}}. \tag{4.2}$$

For a square matrix $A \in \mathbb{R}^{k \times k}$, $|A|$ denotes the determinant of A, A^T denotes the transpose of A, and

$$\|A\|_2 := \max_{\substack{x \in \mathbb{R}^k \\ \|x\|_2 = 1}} \|Ax\|_2 \quad \text{and} \quad \|A\|_F := \sqrt{\sum_{i,j \in [k]} A_{i,j}^2}$$

denote the spectral and Frobenius norms of A, respectively. When A is symmetric, $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_D(A)$ are its eigenvalues.

4.3 Related Work and Our Contributions

4.3.1 The Nonparanormal

Nonparanormal models have been used for modeling dependencies among high-dimensional data in a number of fields, such as graphical modeling of gene expression data (Liu, Han, Yuan, Lafferty, and Wasserman, 2012), of neural data (Berkes, Wood, and Pillow, 2009), and of financial time series (Malevergne and Sornette, 2003; Wilson and Ghahramani, 2010; Hernández-Lobato, Lloyd, and Hernández-Lobato, 2013), extreme value analysis in hydrology (Renard and Lang, 2007; Aghakouchak, 2014), and informative data compression (Rey and Roth, 2012). Besides being more robust generalizations of Gaussians, nonparanormal distributions are also theoretically motivated in certain contexts. For example, the output Z of a neuron is often modeled by feeding a weighted linear combination $Y = \sum_{k=1}^{N} w_k X_k$ of inputs into a nonlinear transformation $Z = f(Y)$. When the components of X are independent, the central limit theorem suggests Y is approximately normally distributed, and hence Z is approximately nonparanormally distributed (Szabó, Póczos, Szirtes, and Lőrincz, 2007). It is also useful to note that the nonparanormal assumption can also be tested statistically in practice (Malevergne and Sornette, 2003).

With one recent exception (Ince, Giordano, Kayser, Rousselet, Gross, and Schyns, 2017), previous information estimators for the nonparanormal case (Calsaverini and Vicente, 2009; Ma and Sun, 2011; Elidan, 2013), rely on fully nonparametric information estimators as subroutines, and hence suffer strongly from the curse of dimensionality. Very recently, Ince, Giordano, Kayser, Rousselet, Gross, and Schyns (2017) proposed what we believe is the first mutual information estimator tailored specifically to the nonparanormal case; their estimator is equivalent to one of the estimators (I_G, described in Section 4.4.1) we study. However, they focused on its applications to neuroimaging data analysis, and did not study its performance theoretically or empirically.

4.3.2 Information Estimation

Our motivation for studying the nonparanormal family comes from trying to bridge two recent approaches to information estimation. The first has studied fully nonparametric entropy estimation, assuming only that data are drawn from a smooth probability density p; smoothness is typically quantified by a Hölder or Sobolev exponent $s \in (0, \infty)$, roughly corresponding to the continuous differentiability of s. In this setting, the minimax optimal MSE rate has been shown by Birgé and Massart (1995) to be $O\left(\max\left\{n^{-1}, n^{-\frac{8s}{4s+D}}\right\}\right)$. This rate slows exponentially with the dimension D, and, while many estimators have been proposed

(Pál, Póczos, and Szepesvári, 2010; Sricharan, Raich, and Hero, 2011; Sricharan, Wei, and Hero, 2013; Singh and Póczos, 2014b; Singh and Póczos, 2014a; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014; Moon and Hero, 2014b; Moon and Hero, 2014a; Singh and Póczos, 2016b; Moon, Sricharan, and Hero III, 2017) for this setting, their practical use is limited to a few dimensions[3].

The second area is in the setting where data are assumed to be drawn from a truly Gaussian distribution. Here the high-dimensional case is far more optimistic. While

[3] "Few" depends on s and n, but Kandasamy, Krishnamurthy, Poczos, and Wasserman (2015) suggest nonparametric estimators should only be used with D at most 4-6. Rey and Roth (2012) tried using several nonparametric information estimators on the *Communities and Crime* UCI data set ($n = 2195, D = 10$), but found all too unstable to be useful.

this case had been studied previously (Ahmed and Gokhale, 1989; Misra, Singh, and Demchuk, 2005; Srivastava and Gupta, 2008), Cai, Liang, and Zhou (2015) recently provided a precise finite-sample analysis based on deriving the exact probability law of the log-determinant $\log|\widehat{\Sigma}|$ of the scatter matrix $\widehat{\Sigma}$. From this, they derived a deterministic bias correction, giving an estimator for which they prove an MSE upper bound of $-2\log\left(1-\frac{D}{n}\right)$ and a high-dimensional central limit theorem for the case $D \to \infty$ as $n \to \infty$ (but $D < n$).

Cai, Liang, and Zhou (2015) also prove a minimax lower bound of $2D/n$ on MSE, with several interesting consequences. First, consistent information estimation is possible only if $D/n \to 0$. Second, since, for small x, $-\log(1-x) \approx x$, this lower bound essentially matches the above upper bound when D/n is small. Third, they show this lower bound holds even when restricted to diagonal covariance matrices. Since the upper bound for the general case and the lower bound for the diagonal case essentially match, it follows that Gaussian information estimation is not made easier by structural assumptions such as Σ being bandable, sparse, or Toeplitz, as is common in, for example, stationary Gaussian process models (Cai and Yuan, 2012).

This $2D/n$ lower bound extends to our more general nonparanormal setting. However, we provide a minimax lower bound suggesting that the nonparanormal setting is strictly harder, in that optimal rates depend on Σ. Our results imply nonparanormal information estimation *does* become easier if Σ is assumed to be bandable or Toeplitz.

A closely related point is that known convergence rates for the fully nonparametric case require the density p to be bounded away from 0 or have particular tail behavior, due to singularity of the logarithm near 0 and resulting sensitivity of Shannon information-theoretic functionals to regions of low but non-zero probability. In contrast, Cai, Liang, and Zhou (2015) need no lower-bound-type assumptions in the Gaussian case. In the nonparanormal case, we show *some* such condition is needed to prove a uniform rate, but a weaker condition, a positive lower bound on $\lambda_D(\Sigma)$, suffices.

The **main contributions** of this paper are the following:

1. We propose three estimators, $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$,[4] for the mutual information of a nonparanormal distribution.
2. We prove upper bounds, of order $O(D^2/(\lambda_D^2(\Sigma)n))$ on the mean squared error of $\widehat{I}_\rho$, providing the first upper bounds for a nonparanormal information estimator. This bound suggests nonparanormal estimators scale far better with D than nonparametric estimators.
3. We prove a minimax lower bound suggesting that, unlike the Gaussian case, difficulty of nonparanormal information estimation depends on the true Σ.
4. We give simulations comparing our proposed estimators to Gaussian and nonparametric estimators. Besides confirming and augmenting our theoretical predictions, these help characterize the settings in which each nonparanormal estimator works best.
5. We present entropy estimators based on $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$. Though nonparanormal entropy estimation requires somewhat different assumptions from mutual information estimation, we show that entropy can also be estimated at the rate $O(D^2/(\lambda_D^2(\Sigma)n))$.

[4] Ince, Giordano, Kayser, Rousselet, Gross, and Schyns (2017) proposed $\widehat{I}_G$ for use in neuroimaging data analysis. To the best of our knowledge, $\widehat{I}_\rho$ and $\widehat{I}_\tau$ are novel.

4.4 Nonparanormal Information Estimators

In this section, we present three different estimators, I_G, I_ρ, and I_τ, for the mutual information of a nonparanormal distribution. We begin with a lemma providing common motivation for all three estimators.

Since mutual information is invariant to diffeomorphisms of individual variables, it is easy to see that the mutual information of a nonparanormal random variable is the same as that of the latent Gaussian random variable. Specifically:

Lemma 10. (Nonparanormal mutual information): *Suppose* $X \sim \mathcal{NPN}(\Sigma; f)$. *Then,*

$$I(X) = -\frac{1}{2}\log|\Sigma|. \tag{4.3}$$

Lemma 10 shows that mutual information of a nonparanormal random variable depends only the latent covariance Σ; the marginal transformations are nuisance parameters, allowing us to avoid difficult nonparametric estimation; the estimators we propose all plug different estimates of Σ into Eq. (4.3), after a regularization step described in Section 4.4.3.

4.4.1 Estimating Σ by Gaussianization

The first estimator $\widehat{\Sigma}_G$ of Σ proceeds in two steps. First, the data are transformed to have approximately standard normal marginal distributions, a process Szabó, Póczos, Szirtes, and Lőrincz (2007) referred to as "Gaussianization". By the nonparanormal assumption, the Gaussianized data are approximately jointly Gaussian. Then, the latent covariance matrix is estimated by the empirical covariance of the Gaussianized data.

More specifically, letting Φ^{-1} denote the quantile function of the standard normal distribution and recalling the rank matrix R defined in (4.2), the Gaussianized data

$$\widetilde{X}_{i,j} := \Phi^{-1}\left(\frac{R_{i,j}}{n+1}\right) \quad (\text{for } i \in [n], j \in [D])$$

are obtained by transforming the empirical CDF of the each dimension to approximate Φ. Then, we estimate Σ by the empirical covariance $\widehat{\Sigma}_G := \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^T$.

4.4.2 Estimating Σ via Rank Correlation

The second estimator actually has two variants, I_ρ and I_τ, respectively based on relating the latent covariance to two classic rank-based dependence measures, Spearman's ρ and Kendall's τ. For two random variables X and Y with CDFs $F_X, F_Y : \mathbb{R} \to [0,1]$, ρ and τ are defined by

$$\begin{aligned} &\rho(X,Y) := \mathrm{Corr}(F_X(X), F_Y(Y)) \\ \text{and} \quad &\tau(X,Y) := \mathrm{Corr}(\mathrm{sign}(X-X'), \mathrm{sign}(Y-Y')), \end{aligned}$$

respectively, where

$$\mathrm{Corr}(X,Y) = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

denotes the standard Pearson correlation operator and (X', Y') is an IID copy of (X, Y). ρ and τ generalize to the D-dimensional setting in the form of rank correlation matrices $\rho, \tau \in [-1, 1]^{D \times D}$ with $\rho_{j,k} = \rho(X_j, X_k)$ and $\tau_{j,k} = \tau(X_j, X_k)$ for each $j, k \in [D]$.

I_ρ and I_τ are based on a classical result relating the correlation and rank-correlation of a bivariate Gaussian:

Theorem 11. (Kruskal, 1958): *Suppose (X, Y) has a Gaussian joint distribution with covariance Σ. Then,*

$$\mathrm{Corr}(X, Y) = 2 \sin\left(\frac{\pi}{6}\rho(X, Y)\right) = \sin\left(\frac{\pi}{2}\tau(X, Y)\right).$$

ρ and τ are often preferred over Pearson correlation for their relative robustness to outliers and applicability to non-numerical ordinal data. While these are strengths here as well, the main reason for their relevance is that they are invariant to marginal transformations (i.e., for diffeomorphisms $f, g : \mathbb{R} \to \mathbb{R}$, $\rho(f(X), g(Y)) = \pm\rho(X, Y)$ and $\tau(f(X), g(Y)) = \pm\tau(X, Y)$). As a consequence, the identity provided in Theorem 11 extends unchanged to the case $(X, Y) \sim \mathcal{NPN}(\Sigma; f)$. This suggests an estimate for Σ based on estimating ρ or τ and plugging this element-wise into the transform $x \mapsto 2 \sin\left(\frac{\pi}{6}x\right)$ or $x \mapsto \sin\left(\frac{\pi}{2}x\right)$, respectively. Specifically, Σ_ρ is defined by

$$\widehat{\Sigma}_\rho := 2 \sin\left(\frac{\pi}{6}\widehat{\rho}\right), \quad \text{where} \quad \widehat{\rho} = \widehat{\mathrm{Corr}}(R)$$

is the empirical correlation of the rank matrix R, and sine is applied element-wise. Similarly, $\widehat{\Sigma}_\tau := \sin\left(\frac{\pi}{2}\widehat{\tau}\right)$, where

$$\widehat{\tau}_{j,k} := \frac{1}{\binom{n}{2}} \sum_{i \neq \ell \in [n]} \mathrm{sign}(X_{i,j} - X_{\ell,j}) \, \mathrm{sign}(X_{i,k} - X_{\ell,k}).$$

4.4.3 Regularization and estimating I

Unfortunately, unlike usual empirical correlation matrices, none of $\widehat{\Sigma}_G$, $\widehat{\Sigma}_\rho$, or $\widehat{\Sigma}_\tau$ is almost surely strictly positive definite. As a result, directly plugging into the mutual information functional (4.3) may give ∞ or be undefined.

To correct for this, we propose a regularization step, in which we project each estimated latent covariance matrix onto the (closed) cone $\mathcal{S}(z)$ of symmetric matrices with minimum eigenvalue $z > 0$. Specifically, for any $z > 0$, let

$$\mathcal{S}(z) := \left\{A \in \mathbb{R}^{D \times D} : A = A^T, \lambda_D(A) \geq z\right\}.$$

For any symmetric matrix $A \in \mathbb{R}^{D \times D}$ with eigendecomposition $\widehat{\Sigma} = Q \Lambda Q^{-1}$ (i.e., $QQ^T = Q^T Q = I_D$ and Λ is diagonal), the projection A_z of A onto $\mathcal{S}(z)$ is defined as $A_z := Q \Lambda_z Q^{-1}$, where Λ_z is the diagonal matrix with j^{th} nonzero entry $(\Lambda_z)_{j,j} = \max\{z, \Lambda_{j,j}\}$. We call this a "projection" because $A_z = \mathrm{argmin}_{B \in \mathcal{S}(z)} \|A - B\|_F$ (see, e.g., Henrion and Malick (2012)).

Applying this regularization to $\widehat{\Sigma}_G$, $\widehat{\Sigma}_\rho$, or $\widehat{\Sigma}_\tau$ gives a strictly positive definite estimate $\widehat{\Sigma}_{G,z}$, $\widehat{\Sigma}_{\rho,z}$, or $\widehat{\Sigma}_{\tau,z}$, respectively, of Σ. We can then estimate I by plugging

this into Equation (4.3), giving our three estimators:

$$\widehat{I}_{G,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{G,z}\right|, \qquad \widehat{I}_{\rho,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{\rho,z}\right|$$
$$\text{and} \quad \widehat{I}_{\tau,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{\tau,z}\right|.$$

4.5 Upper Bounds on the Error of $\widehat{I}_{\rho,z}$

Here, we provide finite-sample upper bounds on the error of the estimator $\widehat{I}_\rho$ based on Spearman's ρ. Proofs are given in the Appendix.

We first bound the bias of $\widehat{I}_\rho$:

Proposition 12. *Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. Then, there exists a constant $C > 0$ such that, for any $z > 0$, the bias of $\widehat{I}_{\rho,z}$ is at most*

$$\left|\mathbb{E}\left[\widehat{I}_{\rho,z}\right] - I\right| \leq C\left(\frac{D}{z\sqrt{n}} + \log\frac{|\Sigma_z|}{|\Sigma|}\right),$$

where Σ_z is the projection of Σ onto $\mathcal{S}(z)$.

The first term of the bias stems from nonlinearity of the log-determinant function in Equation 4.3, which we analyze via Taylor expansion. The second term,

$$\log\frac{|\Sigma_z|}{|\Sigma|} = \sum_{\lambda_j(\Sigma)<z} \log\left(\frac{z}{\lambda_j(\Sigma)}\right),$$

is due to the regularization step and is actually exact, but is difficult to simplify or bound without more assumptions on the spectrum of Σ and choice of z, which we discuss later. We now turn to bounding the variance of $\widehat{I}_{\rho,z}$. We first provide an exponential concentration inequality for $\widehat{I}_{\rho,z}$ around its expectation, based on McDiarmid's inequality:

Proposition 13. *Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. Then, for any $z, \varepsilon > 0$,*

$$\mathbb{P}\left[\left|\widehat{I}_{\rho,z} - \mathbb{E}\left[\widehat{I}_{\rho,z}\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{nz^2\varepsilon^2}{18\pi^2 D^2}\right).$$

Such exponential concentration bounds are useful when one wants to simultaneously bound the error of multiple uses of an estimator, and hence we present it separately as it may be independently useful. However, for the purpose of understanding convergence rates, we are more interested in the variance bound that follows as an easy corollary:

Corollary 14. *Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. Then, for any $z > 0$, the variance of $\widehat{I}_{\rho,z}$ is at most*

$$\mathbb{V}\left[\widehat{I}_{\rho,z}\right] \leq \frac{36\pi^2 D^2}{z^2 n}.$$

Given these bias and variance bounds, a bound on the MSE of $\widehat{I}_{\rho,z}$ follows via the usual bias-variance decomposition:

Theorem 15. *Suppose $X \sim \mathcal{NPN}(\Sigma; f)$. Then, there exists a constant C such that*

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq C\left(\frac{D^2}{z^2 n} + \log^2 \frac{|\Sigma_z|}{|\Sigma|}\right). \tag{4.4}$$

A natural question is now how to optimally select the regularization parameter z. While the bound (4.4) is clearly convex in z, it depends crucially on the unknown spectrum of Σ, and, in particular, on the smallest eigenvalues of Σ. As a result, it is difficult to choose z optimally in general, but we can do so for certain common subclasses of covariance matrices. For example, if Σ is Toeplitz or bandable (i.e., for some $c \in (0,1)$, all $|\Sigma_{i,j}| \leq c^{|i-j|}$), then the smallest eigenvalue of Σ can be bounded below (Cai and Yuan, 2012). When Σ is bandable, as we show in the Appendix, this bound can be independent of D. In these cases, the following somewhat simpler MSE bound can be used:

Corollary 16. *Suppose $X \sim \mathcal{NPN}(\Sigma; f)$, and suppose $z \leq \lambda_D(\Sigma)$. Then, there exists a constant $C > 0$ such that*

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq \frac{CD^2}{z^2 n}.$$

4.6 Lower Bounds in terms of Σ

If $X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma)$ are Gaussian, for the plug-in estimator

$$\widehat{I} = -\tfrac{1}{2}\log\left|\widehat{\Sigma}\right| \quad \text{(where} \quad \widehat{\Sigma} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i X_i^T$$

is the empirical covariance matrix), Cai, Liang, and Zhou (2015) showed that the distribution of $\widehat{I} - I$ is independent of the true correlation matrix Σ. This follows from the "stability" of Gaussians (i.e., that nonsingular linear transformations of Gaussian random variables are Gaussian). In particular,

$$\widehat{I} - I = \log|\widehat{\Sigma}| - \log|\Sigma| = \log|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}|,$$

and $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ has the same distribution as $\log\widehat{\Sigma}$ does in the special case that $\Sigma = I_D$ is the identity. This property is both somewhat surprising, given that $I \to \infty$ as $|\Sigma| \to 0$, and useful, leading to a tight analysis of the error of $\widehat{I}$ and confidence intervals that do not depend on Σ.

It would be convenient if any nonparanormal information estimators satisfied this property. Unfortunately, the main result of this section is a negative one, showing that this property is unlikely to hold without additional assumptions:

Proposition 17. *Consider the 2-dimensional case*

$$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma), \quad \text{with} \quad \Sigma = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}, \tag{4.5}$$

and let $\sigma_* \in (0,1)$. Suppose an estimator $\widehat{I} = \widehat{I}(R)$ of $I_\sigma = -\frac{1}{2}\log(1-\sigma^2)$ is a function of the empirical rank matrix $R \in \mathbb{N}^{n\times 2}$ of X. Then, there exists a constant $C > 0$, depending

only n, such that the worst-case MSE of $\widehat{I}$ over $\sigma \in (0, \sigma_*)$ satisfies

$$\sup_{\sigma \in (0,\sigma^*)} \mathbb{E}\left[\left(\widehat{I}(R) - I_\sigma\right)^2\right] \geq \frac{1}{64}\left(C - \log(1-\sigma_*^2)\right)^2$$

Clearly, this lower bound tends to ∞ as $\sigma \to 1$. As written, this result lower bounds the error of *rank-based estimators* in the Gaussian case when $\sigma \approx 1$. However, to the best of our knowledge, all methods for estimating Σ in the nonparanormal case are functions of R, and prior work (Hoff, 2007) has shown that the rank matrix R is a generalized sufficient statistic for Σ (and hence for I) in the nonparanormal model. Thus, it is reasonable to think of lower bounds for rank-based estimators in the Gaussian case as lower bounds for any estimator in the nonparanormal case.

The proof of this result is based on the simple observation that the rank matrix can take only finitely many values. Hence, as $\sigma \to 1$, R tends to be perfectly correlated, providing little information about σ, whereas the dependence of the estimand I_σ on σ increases sharply. This is intuition is formalized in the Appendix using Le Cam's lemma for lower bounds in two-point parameter estimation problems.

4.7 Empirical Results

We compare 5 mutual information estimators:

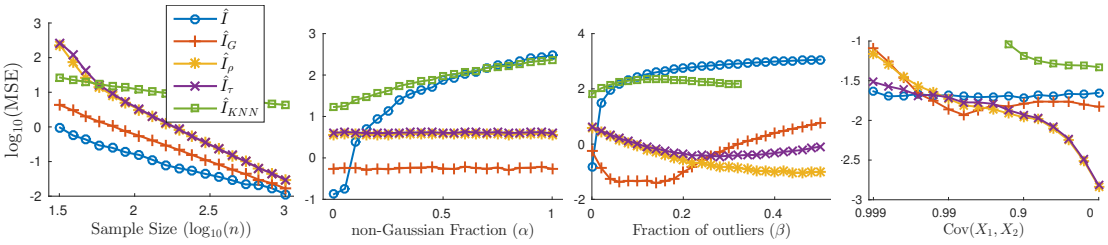
- $\widehat{I}$: Gaussian plug-in estimator with bias-correction (see Cai, Liang, and Zhou (2015)).
- $\widehat{I}_G$: Nonparanormal estimator using Gaussianization.
- $\widehat{I}_\rho$: Nonparanormal estimator using Spearman's ρ.
- $\widehat{I}_\tau$: Nonparanormal estimator using Kendall's τ.
- $\widehat{I}_{k\text{NN}}$: Nonparametric estimator using k-nearest neighbor (kNN) statistics.

For I_ρ and I_τ, we used a regularization constant $z = 10^{-3}$. We did not regularize for I_G. Although this implies $\mathbb{P}[I_G = \infty] > 0$, this is extremely unlikely for even moderate values of n and never occurred during our experiments, which all use $n \geq 32$. We thus omit denoting dependence on z. For $I_{k\text{NN}}$, except as noted in Experiment 3, $k = 2$, based on recent analysis (Singh and Póczos, 2016a) suggesting that small values of k are best for estimation.

Sufficient details to reproduce experiments are given in the Appendix, and MATLAB source code is available on GitHub[5]. We report MSE based on 1000 i.i.d. trials of each condition. 95% confidence intervals were consistently smaller than plot markers and hence omitted to avoid cluttering plots. Except as specified otherwise, each experiment had the following basic structure: In each trial, a correlation matrix Σ was drawn by normalizing a random covariance matrix from a Wishart distribution, and data $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ drawn. All 5 estimators were computed from $X_1, ..., X_n$ and squared error from true mutual information (computed from Σ) was recorded. Unless specified otherwise, $n = 100$ and $D = 25$.

Since our nonparanormal information estimators are functions of ranks of the data, neither the true mutual information nor our non-paranormal estimators depend on the marginal transformations. Thus, except in Experiment 2, where we

[5] https://github.com/sss1/nonparanormal-information



(A) Experiment 1 (B) Experiment 2 (C) Experiment 3 (D) Experiment 4

FIGURE 4.2: Plots of $\log_{10}(\mathrm{MSE})$ plotted over (a) log-sample-size $\log_{10}(n)$, (b) fraction α of dimensions with non-Gaussian marginals, (c) fraction β of outlier samples in each dimension, and (d) covariance $\Sigma_{1,2} = \mathrm{Cov}(X_1, X_2)$. Note that the x-axis in (d) is decreasing.

show the effects of transforming marginals, and Experiment 3, where we add outliers to the data, we perform all experiments on truly Gaussian data, with the understanding that this setting favors the Gaussian estimator.

All experimental results are displayed in Figure 4.2.

Experiment 1 (Dependence on n): We first show nonparanormal estimators have "parametric" $O(n^{-1})$ dependence on n, unlike $\widehat{I}_{k\mathrm{NN}}$, which converges far more slowly. For large n, MSEs of $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$ are close to that of $\widehat{I}$.

Experiment 2 (Non-Gaussian Marginals): Next, we show nonparanormal estimators are robust to non-Gaussianity of the marginals, unlike $\widehat{I}$. We applied a nonlinear transformation f to a fraction $\alpha \in [0, 1]$ of dimensions of Gaussian data. That is, we drew $Z_1, ..., Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ and then used data $X_1, ..., X_n$, where

$$X_{i,j} = \begin{cases} T(Z_{i,j}) & \text{if } j < \alpha D \\ Z_{i,j} & \text{if } j \geq \alpha D \end{cases}, \quad \forall i \in [n], j \in [D],$$

for a diffeomorphism T. Here, we use $T(z) = e^z$. The Appendix shows similar results for several other T. $\widehat{I}$ performs poorly even when α is quite small. Poor performance of $\widehat{I}_{k\mathrm{NN}}$ may be due to discontinuity of the density at $x = 0$.

Experiment 3 (Outliers): We now show that nonparanormal estimators are far more robust to the presence of outliers than $\widehat{I}$ or $\widehat{I}_{k\mathrm{NN}}$. To do this, we added outliers to the data according to the method of Liu, Han, Yuan, Lafferty, and Wasserman (2012). After drawing Gaussian data, we independently select $\lfloor \beta n \rfloor$ samples in each dimension, and replace each i.i.d. uniformly at random from $\{-5, +5\}$. Performance of $\widehat{I}$ degrades rapidly even for small β. $\widehat{I}_{k\mathrm{NN}}$ can fail for atomic distributions, $\widehat{I}_{k\mathrm{NN}} = \infty$ whenever at least k samples are identical. This mitigate this, we increased k to 20 and ignored trials where $\widehat{I}_{k\mathrm{NN}} = \infty$, but $\widehat{I}_{k\mathrm{NN}}$ ceased to give any finite estimates when β was sufficiently large.

For small values of β, nonparanormal estimators surprisingly improve. We hypothesize this is due to convexity of the mutual information functional Eq. (4.3) in Σ. By Jensen's inequality, estimators which plug-in an approximately unbiased estimate $\widehat{\Sigma}$ of Σ are biased towards overestimating I. Adding random (uncorrelated) noise reduces estimated dependence, moving the estimate closer to the true value. If this nonlinearity is indeed a major source of bias, it may be possible to derive a von Mises-type bias correction (see Kandasamy, Krishnamurthy, Poczos, and Wasserman (2015)) accounting for higher-order terms in the Taylor expansion of the log-determinant.

Experiment 4 (Dependence on Σ): Here, we verify our results in Section 4.6 showing that MSE of rank-based estimators approaches ∞ as $|\Sigma| \to 0$, while MSE of $\widehat{I}$ is independent of Σ. Here, we set $D = 2$ and Σ as in Eq. (4.5), varying $\sigma \in [0, 1]$. Indeed, the MSE of $\widehat{I}$ does not change, while the MSEs of $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$ all increase as $\sigma \to 1$. This increase seems mild in practice, with performance worse than of $\widehat{I}$ only when $\sigma > 0.99$. $\widehat{I}_\tau$ appears to perform far better than $\widehat{I}_G$ and $\widehat{I}_\rho$ in this regime. Performance of $I_{k\text{NN}}$ degrades far more quickly as $\sigma \to 1$. This phenomenon is explored by Gao, Ver Steeg, and Galstyan (2015), who lower bound error of $I_{k\text{NN}}$ in the presence of strong dependencies, and proposed a correction to improve performance in this case.

It is also interesting that errors of $\widehat{I}_\rho$ and $\widehat{I}_\tau$ drop as $\sigma \to 0$. This is likely because, for small σ, the main source of error is the variance of $\widehat{\rho}$ and $\widehat{\tau}$ (as $-\log(1-\sigma^2) \approx \sigma^2$ when $\sigma \approx 0$). When $n \to \infty$ and D is fixed, both $2\sin(\pi\widehat{\rho}/6)$ and $\sin(\pi\widehat{\tau}/2)$ are asymptotically normal estimates of σ, with asymptotic variances proportional to $(1-\sigma^2)^2$ (Klaassen and Wellner, 1997). By the delta method, since $\frac{dI}{d\sigma} = \frac{\sigma}{1-\sigma^2}$, $\widehat{I}_\rho$ and $\widehat{I}_\tau$ are asymptotically normal estimates of I, with asymptotic variances proportional to σ^2 and hence vanishing as $\sigma \to 0$.

4.8 Estimating Entropy

Thus far, we have discussed estimation of mutual information $I(X)$. Mutual information is convenient because it is invariant under marginal transformation, and hence $I(X) = I(f(X))$ depends only on Σ. While the entropy $H(X)$ does depend on the marginal transform f, fortunately, by Eq. (4.1), $H(X)$ differs from $I(X)$ only by a sum of univariate entropies. Univariate nonparametric estimation of entropy in has been studied extensively, and there exist several estimators (e.g., based on sample spacings (Beirlant, Dudewicz, Györfi, and Meulen, 1997), kernel density estimates (Moon, Sricharan, Greenewald, and Hero, 2016) or k-nearest neighbor methods (Singh and Póczos, 2016a)) that can estimate $H(X_j)$ at the rate $\asymp n^{-1}$ in MSE under relatively mild conditions on the marginal density p_j. While the precise assumptions vary with the choice of estimator, they are mainly (a) that p_j be lower bounded on its support or have particular (e.g., exponential) tail behavior, and (b) that p_j be smooth, typically quantified by a Hölder or Sobolev condition. Details of these assumptions are in the Appendix.

Under these conditions, since there exist estimators $\widehat{H}_1, ..., \widehat{H}_D$ and a constant $C > 0$ such that

$$\mathbb{E}[(\widehat{H}_j - H(X_j))^2] \leq C/n, \quad \forall j \in [D]. \tag{4.6}$$

Combining these estimators with an estimator, say $\widehat{I}_{\rho,z}$, of mutual information gives an estimator of entropy:

$$\widehat{H}_{\rho,z} := \textstyle\sum_{j=1}^{D} \widehat{H}_j - \widehat{I}_{\rho,z}.$$

If we assume $z = \lambda_D^{-1}(\Sigma)$ is bounded below by a positive constant, combining inequality (4.6) with Corollary 16 gives

$$\mathbb{E}\left[\left(\widehat{H}_{\rho,z} - H(X)\right)^2\right] \leq \frac{CD^2}{n},$$

where C differs from in (4.6) but is independent of n and D.

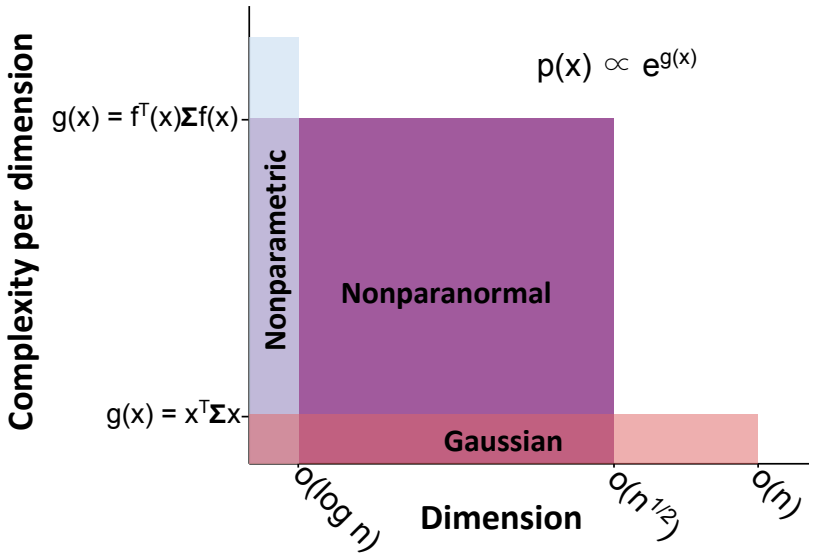


FIGURE 4.3: Cartoon phase diagram showing when each kind of estimator (Gaussian, nonparanormal, fully nonparametric) estimator can be useful. Nonparanormal estimators can help fill the large gap in the setting of moderately high-dimensional data with non-Gaussian marginal distributions.

4.9 Conclusions and Future Work

This paper suggests nonparanormal information estimation as a practical compromise between the intractable nonparametric case and the limited Gaussian case. We proposed three estimators for this problem and provided the first upper bounds for nonparanormal information estimation. We also gave lower bounds showing how dependence on Σ differs from the Gaussian case and demonstrated empirically that nonparanormal estimators are more robust than Gaussian estimators, even in dimensions too high for nonparametric estimators.

Collectively, these results suggest that, by scaling to moderate or high dimensionality without relying on Gaussianity, nonparanormal information estimators may be effective tools with a number of machine learning applications. While the best choice of information estimator inevitably depends on context, as an off-the-shelf guide for practitioners, the estimators we suggest, in order of preference, are:

- fully nonparametric if $D < 6, n > \max\{100, 10^D\}$.
- $\widehat{I}_\rho$ if D^2/n is small and data may have outliers.
- $\widehat{I}_\tau$ if D^2/n is small and dependencies may be strong.
- $\widehat{I}_G$ otherwise.
- $\widehat{I}$ only given strong belief that data are nearly Gaussian.

Figure 4.3 shows, in a simplified phase diagram, when each kind of estimator can be useful.

There are many natural open questions in this line of work. First, in the nonparanormal model, we focused on estimating mutual information $I(X)$, which does not depend on marginal transforms f, and entropy, which decomposes into $I(X)$ and 1-dimensional entropies. In both cases, additional structure imposed by the nonparanormal model allows estimation in higher dimensions than fully nonparametric models. Can nonparanormal assumptions lead to higher dimensional estimators for the many other useful nonlinear functionals of densities (e.g., L_p norms/distances

and more general (e.g., Rényi or Tsallis) entropies, mutual informations, and divergences) that do not decompose?

Second, there is a gap between our upper bound rate of $\|\Sigma^{-1}\|_2^2 D^2/n$ and the only known lower bound of $2D/n$ (from the Gaussian case), though we also showed that bounds for rank-based estimators depend on Σ. Is quadratic dependence on D optimal? How much do rates improve under structural assumptions on Σ? Upper bounds should be derived for other estimators, such as $\widehat{I}_G$ and $\widehat{I}_\tau$. The $2D/n$ lower bound proof of Cai, Liang, and Zhou (2015) for the Gaussian case, based on the Cramer-Rao inequality (Bos, 2007), is unlikely to tighten in the nonparanormal case, since Fisher information is invariant to diffeomorphisms of the data. Hence, a new approach is needed if the lower bound in the nonparanormal case is to be raised.

Finally, our work applies to estimating the log-determinant $\log|\Sigma|$ of the latent correlation in a nonparanormal model. Besides information estimation, the work of Cai, Liang, and Zhou (2015) on estimating $\log|\Sigma|$ in the Gaussian model was motivated by the role of $\log|\Sigma|$ in other multivariate statistical tools, such as quadratic discriminant analysis (QDA) and MANOVA (Anderson, 1984). Can our estimators lead to more robust nonparanormal versions of these tools?

4.10 Lemmas

Our proofs rely on the following lemmas.

Lemma 18. (Convexity of the inverse operator norm): *The function $A \mapsto \|A^{-1}\|_2$ is convex over $A \succ 0$.*

Proof: For $A, B \succ 0$, let $C := \tau A + (1-\tau)B$. Then,

$$\begin{aligned}
\|\widehat{C}^{-1}\|_2 &= \frac{1}{\inf_{x\in\mathbb{R}^D} x^T C x} \\
&= \frac{1}{\inf_{x\in\mathbb{R}^D} \tau x^T A x + (1-\tau) x^T B x} \\
&\leq \frac{1}{\tau \inf_{x\in\mathbb{R}^D} x^T A x + (1-\tau)\inf_{x\in\mathbb{R}^D} x^T B x} \\
&\leq \tau \frac{1}{\inf_{x\in\mathbb{R}^D} x^T A x} + (1-\tau)\frac{1}{\inf_{x\in\mathbb{R}^D} x^T B x} \\
&= \tau \left\|A^{-1}\right\|_2 + (1-\tau)\left\|B^{-1}\right\|_2
\end{aligned}$$

via convexity of the function $x \mapsto 1/x$ on $(0,\infty)$. ∎

Lemma 19. (Mean-Value Bound on the Log-Determinant): *Matrix derivative of log-determinant. Suppose $A, B \succ 0$. Then, for $\lambda := \min\{\lambda_D(A), \lambda_D(B)\}$,*

$$|\log|A| - \log|B|| \leq \frac{1}{\lambda}\|A - B\|_F.$$

Proof: *Proof:* First recall that the log-determinant is continuously differentiable over the strict positive definite cone, with $\nabla_X \log|X| = X^{-1}$ for any $X \succ 0$. Hence, by the matrix-valued version of the mean value theorem,

$$\log|A| - \log|B| = tr(C^{-1}(A-B)),$$

where $C = \tau A + (1-\tau)B$ for some $\tau \in (0,1)$. Since for positive definite matrices, the inner product can be bounded by the product of the operator and Frobenius norms, and clearly $C \succ 0$, we have

$$|\log|A| - \log|B|| = \|C^{-1}\|_2 \|A - B\|_F.$$

Finally, it follows by Lemma 18 that

$$|\log|A| - \log|B|| \leq \frac{1}{\lambda}\|A - B\|_F.$$

■

4.11 Proofs of Main Results

Here, we give proofs of our main theoretical results, beginning with upper bounds on the MSE of $\widehat{I}_\rho$ and proceeding to minimax lower bounds in terms of Σ.

4.12 Upper bounds on the MSE of $\widehat{I}_\rho$

Proposition 20.

$$\left|\mathbb{E}\left[\log|\widehat{\Sigma}_z|\right] - \log|\Sigma|\right| \leq C\left(\|\Sigma\|_2^2 \frac{D}{z^2 n} + \left(\sum_{\lambda_j(\Sigma) < z} \log\left(\frac{z}{\lambda_j(\Sigma)}\right)\right)^2\right).$$

Proof: By the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}\left[\log|\widehat{\Sigma}_z|\right] - \log|\Sigma|\right| &\leq \left|\mathbb{E}\left[\log|\widehat{\Sigma}_z|\right] - \log|\Sigma_z|\right| \\ &\quad + |\log|\Sigma_z| - \log|\Sigma||\end{aligned}$$

For the first term, applying the matrix mean value theorem (Lemma 19) and the inequality $\|A\|_F \leq \sqrt{D}\|A\|_2$

$$\begin{aligned}\left|\mathbb{E}\left[\log\left|\widehat{\Sigma}_z\right|\right] - \log|\Sigma_z|\right| &\leq \mathbb{E}\left[\left|\log\left|\widehat{\Sigma}_z\right| - \log|\Sigma_z|\right|\right] \\ &\leq \frac{1}{z}\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_F\right] \\ &\leq \frac{\sqrt{D}}{z}\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \\ &\leq \frac{C_{MZ}\|\Sigma\|_2 D}{z\sqrt{n}},\end{aligned}$$

where we used Theorem 1 of Mitra and Zhang (2014), which gives a constant C_{MZ} such that

$$\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \leq C_{MZ}\|\Sigma\|_2\sqrt{\frac{D}{n}}.$$

Via the bound $\|\Sigma\|_2 \leq \sqrt{D}\|\Sigma\|_\infty$, this reduces to

$$\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \leq C_{MZ}\frac{D}{\sqrt{n}}.$$

■

Proposition 21. *The variance of the nonparanormal information estimator $\widehat{I}_\rho$ based on Spearman's ρ with regularization parameter z can be bounded as*

$$\mathbb{V}\left[\widehat{I}_{\rho,z}\right] \leq \frac{36\pi^2 D^2}{z^2 n}.$$

Proof: By the Efron-Stein inequality (Efron and Stein, 1981), since $X_1, \ldots, X_n$ are independent and identically distributed,

$$\begin{aligned}\mathbb{V}\left[\widehat{I}\right] &\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(i)}|\right)^2\right] \\ &= \frac{n}{2}\,\mathbb{E}\left[\left(\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right)^2\right],\end{aligned}$$

where $\widehat{\Sigma}_z^{(1)}$ is our estimator after independently re-sampling the first sample X_1. Applying the multivariate mean-value theorem (Lemma 19), we have

$$\left|\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right| \leq \frac{1}{z}\|\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\|_F.$$

$\|\widehat{\Sigma}_\tau^{-1}\|_2 \leq \frac{1}{z}$. Since $\mathcal{S}(z)$ is convex and the Frobenius norm is supported by an inner product, the operation of projecting onto $\mathcal{S}(z)$ is a contraction. In particular, $\left\|\left(\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\right)\right\|_F \leq \left\|\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)\right\|_F$ Applying the mean value theorem to the function $x \mapsto 2\sin\left(\frac{\pi}{6}x\right)$,

$$\left\|\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)\right\|_F^2 = \sum_{j,k=1}^{D}\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)_{j,k}^2 \tag{4.7}$$

$$\leq \frac{\pi^2}{9}\sum_{j,k=1}^{D}\left(\widehat{\rho}_{j,k} - \widehat{\rho}_{j,k}^{(1)}\right)^2 \tag{4.8}$$

$$= \frac{\pi^2}{9}\left\|\widehat{\rho} - \widehat{\rho}^{(1)}\right\|_F^2. \tag{4.9}$$

From the formula

$$\widehat{\rho}_{j,k} = 1 - \frac{6\sum_{i=1}^{n} d_{i,j,k}^2}{n(n^2-1)},$$

(where $d_{i,j,k}$ denotes the difference in ranks of $X_{i,j}$ and $X_{i,k}$ in $X_{1,j}, ..., X_{n,j}$ and $X_{1,k}, ..., Y_{n,k}$, respectively), one can see, since $|d_{1,j,k} - d'_{1,j,k}| \leq n$ and, for $i \neq 1$, $|d_{i,j,k} - d'_{i,j,k}| \leq 1$, that

$$\left|\widehat{\rho}_{j,k} - \widehat{\rho}_{j,k}^{(1)}\right| \leq \frac{18}{n},$$

and hence that

$$\|\widehat{\rho} - \widehat{\rho}^{(1)}\|_F \leq \frac{18D}{n}. \tag{4.10}$$

It follows from inequality (4.9) that

$$\|\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\|_F \leq \frac{6\pi D}{n}.$$

Altogether, this gives

$$\left|\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right| \leq \frac{6\pi D}{zn}.$$

Then, McDiarmid's Inequality gives, for all $\varepsilon > 0$,

$$\mathbb{P}\left[\left|\widehat{I} - \mathbb{E}\left[\widehat{I}\right]\right| > \varepsilon\right] = 2\exp\left(-\frac{nz^2\varepsilon^2}{18\pi^2 D^2}\right).$$

This translates to a variance bound of

$$\mathbb{V}\left[\widehat{I}\right] \leq \frac{36\pi^2 D^2}{z^2 n}.$$

■

4.12.1 Lower bound for rank-based estimators in terms of Σ

One (perhaps surprising) result of Cai, Liang, and Zhou (2015) is that, as long as $D/n \to 0$, the convergence rate of the estimator is independent of the true correlation structure Σ. Here, we show that this desirable property does not hold in the nonparanormal case.

Proposition 22. *Consider the 2-dimensional case*

$$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma), \quad \textit{with} \quad \Sigma = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}, \tag{4.11}$$

and let $\sigma_* \in (0,1)$. Suppose an estimator $\widehat{I} = \widehat{I}(R)$ of $I_\sigma = -\frac{1}{2}\log(1-\sigma^2)$ is a function of the empirical rank matrix $R \in \mathbb{N}^{n\times 2}$ of X (as defined in (4.2)). Then, there exists a constant $C > 0$, depending only n, such that the worst-case MSE of $\widehat{I}$ over $\sigma \in (0, \sigma_*)$ satisfies

$$\sup_{\sigma\in(0,\sigma^*]} \mathbb{E}\left[\left(\widehat{I}(R) - I_\sigma\right)^2\right] \geq \frac{1}{64}\left(C - \log(1-\sigma_*^2)\right)^2$$
$$\to \infty \quad \textit{as} \quad \sigma_* \to 1.$$

Proof: Note that the rank matrix R can take only finitely many values. Let $\mathcal{R}$ be the set of all $(n!)^D$ possible rank matrices and let $\mathcal{R}_1 \subseteq \mathcal{R}$ be the set of $n!$ rank matrices that are perfectly correlated. Then, as $\sigma \to 1$, $\mathbb{P}[R \in \mathcal{R}_1] \to 1$, so, in particular, we can pick σ_0 (depending only on n) such that, for all $\sigma \geq \sigma_0$, $\mathbb{P}[R \in \mathcal{R}_1] \geq \frac{1}{2}$. Since the data are i.i.d., all rank matrices in $\mathcal{R}_1$ have equal probability. It follows that

$$D_{TV}(\mathbb{P}_0||\mathbb{P}_1) = \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_1 \leq \frac{1}{2},$$

where D_{TV} denotes total variation distance. Finally, by Le Cam's Lemma (see, e.g., Section 2.3 of Tsybakov (2008)),

$$\inf_{\widehat{I}} \sup_{\sigma\in\{\sigma_0,\sigma_1\}} \mathbb{E}\left[\left(\widehat{I} - I_\sigma\right)^2\right] \geq \frac{(I_{\sigma_*} - I_{\sigma_0})^2}{8}\left(1 - D_{TV}(P_{\sigma_0}, P_{\sigma_1})\right)$$
$$\geq \frac{(\log(1-\sigma_0^2) - \log(1-\sigma_*^2))^2}{64}$$

■

4.13 Details of Experimental Methods

Here, we present details needed to reproduce our numerical simulations. Note that MATLAB source code for these experiments is available on GitHub[6], including a single runnable script that performs all experiments and generates all figures presented in this paper. Specific details needed to reproduce experiments are given in the Appendix,

In short, experiments report empirical mean squared errors based on 100 i.i.d. trials of each condition. We initially computed 95% confidence intervals, but these intervals were consistently smaller than marker sizes, so we omitted them to avoid cluttering plots. Except as specified otherwise, each experiment followed the same basic structure, as follows: In each trial, a random correlation matrix $\Sigma \in [-1,1]^{D\times D}$ was drawn by normalizing a covariance matrix from a Wishart distribution $W(I_D, D)$ with identity scale matrix and D degrees of freedom. Data $X_1, ..., X_n$ were then drawn i.i.d. from $\mathcal{N}(0, \Sigma)$. All estimators were applied to the same data. Unless specified otherwise, $n = 100$ and $D = 25$.

4.13.1 Computational Considerations

In general, the running time of all the nonparanormal estimators considered is $O(Dn \log n + D^2 n + D^3)$ (i.e., $O(Dn \log n)$ to rank or Gaussianize the variables in each dimension, $D^2 n$ to compute the covariance matrix, and $O(D^3)$ to compute the log-determinant). All log-determinants $\log |\Sigma|$ were computed by summing the logarithms of the diagonal of the Cholesky decomposition of Σ, as this is widely considered to be a fast and numerically stable approach. Note however that faster ($O(D)$-time) randomized algorithms (Han, Malioutov, and Shin, 2015) have been proposed to approximate the log-determinant).

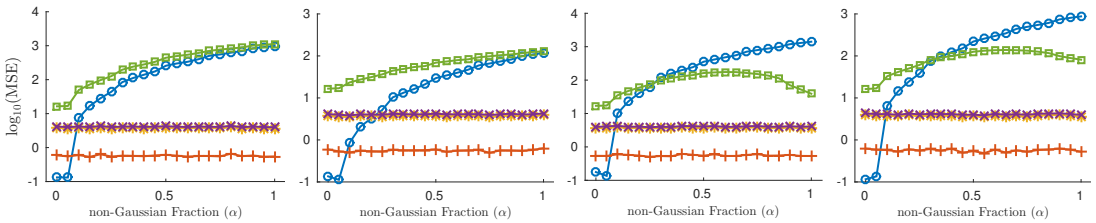
4.14 Additional Experimental Results

Here, we present variants on the experiments presented in the main paper, which support but are not necessary for illustrating our conclusions.

4.14.1 Effects of Other Marginal Transformations

In Section 4.7, we showed that the Gaussian estimator $\widehat{I}$ is highly sensitive to failure of the Gaussian assumption for even a small fraction of marginals. Figure 4.2(b), illustrates this for the transformation $x \mapsto \exp(x)$, but we show here that this is not specific to the exponential transformation. As shown in Figures 4.4 nearly identical results hold when the marginal transformation f is the hyperbolic tangent function $x \mapsto \tanh(x)$, the cubic function $x \mapsto x^3$, sigmoid function $x \mapsto \frac{1}{1+e^{-x}}$, or standard normal CDF.

[6] https://github.com/sss1/nonparanormal-information



(A) $T(x) = x^3$ (B) $T(x) = \tanh(x)$ (C) $T(x) = \frac{1}{1+e^{-x}}$ (D) $T(x) = \Phi(x)$

FIGURE 4.4: Semi-log plot of mean squared error of various estimators over the fraction of non-Gaussian marginals $\alpha \in [0, 1]$, for various marginal transforms T.

4.15 Specific Assumptions for Estimating $H(X)$

As shown in the main paper, to estimate the entropy of a nonparanormal distribution at the rate $O(D^2/n)$, it suffices to the univariate entropy of each variable X_j at the rate $O(1/n)$. To do this, additional assumptions are required on the marginal densities p_j. Here, we give detailed sufficient conditions for this.

Letting $S_j \subseteq \mathbb{R}$ denote the support of p_j, the two key assumptions can be roughly classified as follows:

(a) $\frac{1}{2}$-order smoothness[7]; e.g., a Hölder condition:

$$\sup_{x \neq y \in S_j} \frac{|p_j(x) - p_j(y)|}{|x - y|^{1/2}} < L,$$

or a (slightly weaker) Sobolev condition:

$$\int_{S_j} p_j^2(x)\, dx < \infty \quad \text{and} \quad \int_{S_j} \left(|\xi|^{1/2} |\mathcal{F}[p_j](\xi)|\right)^2 d\xi < L,$$

(where $\mathcal{F}[p_j](\xi)$ denotes the Fourier transform of p_j evaluated at ξ) for some constant $L > 0$.

(b) absolute bounds $p_j(x) \in [\kappa_1, \kappa_2]$ for all $x \in S_j$ or (a_j, b_j)-exponential tail bounds

$$\frac{f(x)}{\exp(-a_j x^{b_j})} \in [\kappa_1, \kappa_2] \quad \text{for all } x \in S_j$$

for some $\kappa_1, \kappa_2 \in (0, \infty)$.

Under these assumptions, there are a variety of nonparametric univariate entropy estimators that have been shown to converge at the rate $O(1/n)$ (Beirlant, Dudewicz, Györfi, and Meulen, 1997; Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015; Singh and Póczos, 2016a; Moon, Sricharan, Greenewald, and Hero, 2016).

4.16 Lower bounding the eigenvalues of a bandable matrix

Recall that, for $c \in (0, 1)$, a matrix $\Sigma \in \mathbb{R}^{D \times D}$ is called *c-bandable* if there exists a constant $c \in (0, 1)$ such that, for all $i, j \in D$, $|\Sigma_{i,j}| \leq c^{|i-j|}$.

[7] This is stronger than the $\frac{1}{4}$-order smoothness mandated by the minimax rate for entropy estimation (Birgé and Massart, 1995), but appears necessary for most practical entropy estimators. See Section 4 of Kandasamy, Krishnamurthy, Poczos, and Wasserman (2015) for further details.

Here, we show simple bounds on the eigenvalues of a bandable correlation matrix Σ. While this result is fairly straightforward, a brief search the literature turned up no comparable results. Bickel and Levina (2008), who originally introduced the class of bandable covariance matrices, separately assumed the existence of lower and upper bounds on the eigenvalues to prove their results. In the context of information estimation, this results of particular interest because, when $c < 1/3$ it implies a dimension-free positive lower bound on the minimum eigenvalue of Σ, hence complementing our upper bound in Theorem 15.

Proposition 23. *Suppose a symmetric matrix $\Sigma \in \mathbb{R}^{D\times D}$ is c-bandable and has identical diagonal entries $\Sigma_{j,j} = 1$. Then, the eigenvalues $\lambda_1(\Sigma), ..., \lambda_D(\Sigma)$ of Σ can be bounded as*

$$\frac{1-3c}{1-c} \leq \lambda_1(\Sigma), ..., \lambda_D(\Sigma) \leq \frac{1+c}{1-c}.$$

In particular, when $c < 1/3$, we have

$$0 < \frac{1-3c}{1-c} \leq \lambda_D(\Sigma).$$

Proof: The proof is based on the Gershgorin circle theorem (Gershgorin, 1931; Varga, 2009). In the case of a real symmetric matrix Σ, this states that the eigenvalues of Σ lie within a union of intervals

$$\{\lambda_1(\Sigma), ..., \lambda_D(\Sigma)\} \subseteq \bigcup_{j=1}^{D} \left[\Sigma_{j,j} - R_j, \Sigma_{j,j} + R_j\right], \tag{4.12}$$

where $R_j := \sum_{k\neq j} |\Sigma_{j,k}|$ is the sum of the absolute values of the non-diagonal entries of the j^{th} row of Σ. In our case, since the diagonal entries of Σ are all $\Sigma_{j,j} = 1$, we simply have to bound

$$\max_{j\in[D]} R_j \leq \sum_{k\neq j} c^{|k-j|}.$$

This geometric sum is maximized when $j = \lceil D/2 \rceil$, giving

$$R_j \leq 2\sum_{\delta=1}^{\lfloor D/2 \rfloor} c^{\delta} = 2c\frac{1-c^{\lfloor D/2 \rfloor}}{1-c} \leq \frac{2c}{1-c}.$$

Finally, the inclusion (4.12) gives

$$\lambda_D(\Sigma) \geq 1 - \frac{2c}{1-c} = \frac{1-3c}{1-c} > 0$$

when $c < 1/3$. $1 + \frac{2c}{1-c} = \frac{1+c}{1-c}$. ∎

Chapter 5

Fourier-weighted Quadratic Functionals

5.1 Introduction

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^D$ endowed with the Borel σ-algebra and let $\mathcal{P}$ denote the family of all Borel probability measures on $\mathcal{X}$. For each $P \in \mathcal{P}$, let $\phi_P : \mathbb{R}^D \to \mathcal{C}$ denote the characteristic function of P given by

$$\phi_P(z) = \mathop{\mathbb{E}}_{X \sim P}\left[\overline{\psi_z(X)}\right] \quad \text{for all } z \in \mathbb{R}^D, \quad \text{where} \quad \psi_z(x) = \exp\left(i\langle z, x\rangle\right) \tag{5.1}$$

denotes the i^{th} Fourier basis element, in which $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbb{R}^D$.

For any family $a = \{a_z\}_{z \in \mathcal{Z}} \subseteq \mathbb{R}$ of real-valued coefficients indexed by a countable set $\mathcal{Z}$, define a set of probability measures

$$\mathcal{H}_a := \left\{ P \in \mathcal{P} : \sum_{z \in \mathcal{Z}} \frac{|\phi_P(z)|^2}{a_z^2} < \infty \right\}.$$

Now fix two unknown probability measures $P, Q \in \mathcal{H}_a$. We study estimation of the semi-inner product[1]

$$\langle P, Q \rangle_a = \sum_{z \in \mathcal{Z}} \frac{\phi_P(z)\overline{\phi_Q(z)}}{a_z^2}, \tag{5.2}$$

as well as the squared seminorm $\|P\|_a^2 := \langle P, P \rangle_a$ and squared pseudometric $\|P - Q\|_a^2$, using n i.i.d. samples $X_1, ..., X_n \overset{i.i.d.}{\sim} P$ and $Y_1, ..., Y_n \overset{i.i.d.}{\sim} Q$ from each distribution. Specifically, we assume that P and Q lie in a smaller subspace $\mathcal{H}_b \subseteq \mathcal{H}_a$ parameterized by a $\mathcal{Z}$-indexed real family $b = \{b_z\}_{z \in \mathcal{Z}}$. In this setting, we study the minimax $\mathcal{L}^2$ error $M(a, b)$ of estimating $\langle P, Q \rangle_a$, over P and Q lying in a (unit) ellipsoid with respect to $\|\cdot\|_b$; that is, the quantity

$$M(a,b) := \inf_{\widehat{S}} \sup_{\|P\|_b, \|Q\|_b \leq 1} \mathop{\mathbb{E}}_{\substack{X_1,...,X_n \sim P \\ Y_1,...,Y_n \sim Q}} \left[\left| \widehat{S}(X_1, ..., X_n, Y_1, ..., Y_n) - \langle P, Q \rangle_a \right|^2 \right], \tag{5.3}$$

where the infimum is taken over all estimators $\widehat{S}$ (i.e., all complex-valued functions $\widehat{S} : \mathcal{X}^{2n} \to \mathcal{C}$ of the data).

[1] For a complex number $\xi = a + bi \in \mathcal{C}$, $\overline{\xi} = a - bi \in \mathcal{C}$ denotes the complex conjugate of ξ. A semi-inner product has all properties of an inner product, except that $\langle P, P \rangle = 0$ does not imply $P = 0$.

We study how the rate of the minimax error $M(a,b)$ is primarily governed by the rates at which a_z and b_z decay to 0 as $\|z\| \to \infty$.[2] This has been studied extensively in the Sobolev (or polynomial-decay) case, where, for some $t > s \geq 0$, $a_z = \|z\|^{-s}$ and $b_z = \|z\|^{-t}$, corresponding to estimation of s-order Sobolev semi-inner products under t-order Sobolev smoothness assumptions on the Lebesgue density functions p and q of P and Q (as described in Example 3 below) (Bickel and Ritov, 1988; Donoho and Nussbaum, 1990; Laurent and Massart, 2000; Singh, Du, and Póczos, 2016). In this case, the rate of $M(a,b)$ has been identified (by Bickel and Ritov (1988), Donoho and Nussbaum (1990), and Singh, Du, and Póczos (2016), in increasing generality) as [3]

$$M(a,b) \asymp \max\left\{n^{-1}, n^{-\frac{8(t-s)}{4t+D}}\right\}, \tag{5.4}$$

so that the "parametric" rate n^{-1} dominates when $t \geq 2s + D/4$, and the slower rate $n^{-\frac{8(t-s)}{4t+D}}$ dominates otherwise. Laurent and Massart (2000) additionally showed that, for $t < 2s + D/4$, $M(a,b)$ increases by a factor of $(\log n)^{\frac{4(t-s)}{4t+D}}$ in the "adaptive" case, when the tail index t is not assumed to be known to the estimator.

However, the behavior of $M(a,b)$ for other (non-polynomial) decay rates of a and b has not been studied, despite the fact that, as discussed in Section 5.1.1, other rates of decay of a and b, such as Gaussian or exponential decay, correspond to inner products and assumptions commonly considered in nonparametric statistics. The goal of this paper is therefore to understand the behavior of $M(a,b)$ for general sequences a and b.

Although our results apply more generally, to simply summarize our results, consider the case where a and b are "radial"; i.e. a_z and b_z are both functions of some norm $\|z\|$. Under mild assumptions, we show that the minimax convergence rate is then a function of the quantities

$$A_{\zeta_n} = \sum_{\|z\| \leq \zeta_n} a_z^{-2} \quad \text{and} \quad B_{\zeta_n} = \sum_{\|z\| \leq \zeta_n} b_z^{-2},$$

which can be thought of as measures of the "strengths" of $\|\cdot\|_a$ and $\|\cdot\|_b$, for a particular choice of a "smoothing" (or "truncation") parameter $\zeta_n \in (0,\infty)$. Specifically, we show

$$M(a,b) \asymp \max\left\{\left(\frac{A_{\zeta_n}}{B_{\zeta_n}}\right)^2, \frac{1}{n}\right\}, \quad \text{where} \quad \zeta_n^D n^2 = B_{\zeta_n}^2. \tag{5.5}$$

While (5.5) is difficult to simplify or express in a closed form in general, it is quite simple to compute given the forms of a and b. In this sense, (5.5) might be considered as an analogue of the Le Cam equation (Yang and Barron, 1999) (which gives a similar implicit formula for the minimax rate of nonparametric density estimation in terms of covering numbers) for estimating inner products and related quantities. It is easy to check that, in the Sobolev case (where $a_z = \|z\|^{-s}$ and $b_z = \|z\|^{-t}$ decay polynomially), (5.5) recovers the previously known rate (5.4). Moreover, our assumptions are also satisfied by other rates of interest, such as exponential (where $a_z = e^{-s\|z\|_1}$ and $b_z = e^{-t\|z\|_1}$) and Gaussian (where $a_z = e^{-s\|z\|_2^2}$ and $b_z = e^{-t\|z\|_2^2}$) rates, for which we are the first to identify minimax rates. As in the Sobolev case, the rates here exhibit the so-called "elbow" phenomenon, where the convergence rates is "parametric" (i.e., of order $\asymp 1/n$) when t is sufficiently large relative to s,

[2] By equivalence of finite-dimensional norms, the choice of norm here affects only constant factors.

[3] Here and elsewhere, $\asymp$ denotes equality up to constant factors.

and slower otherwise. However, for rapidly decaying b such as in the exponential case, the location of this elbow no longer depends directly on the dimension D; the parametric rate is achieved as soon as $t \geq 2s$.

We note that, in all of the above cases, the minimax rate (5.5) is achieved by a simple bilinear estimator:

$$\widehat{S}_{\zeta_n} := \sum_{\|z\| \leq \zeta_n} \frac{\widehat{\phi}_P(z)\overline{\widehat{\phi}_Q(z)}}{a_z^2},$$

where

$$\widehat{\phi}_P(z) := \frac{1}{n}\sum_{i=1}^{n} \psi_z(X_i) \quad \text{and} \quad \widehat{\phi}_Q(z) := \frac{1}{n}\sum_{i=1}^{n} \psi_z(Y_i)$$

are linear estimates of $\phi_P(z)$ and $\phi_Q(z)$, and $\zeta_n \geq 0$ is a tuning parameter.

We also show that, in many cases, a rate-optimal ζ_n can be chosen adaptively (i.e., without knowledge of the space $\mathcal{H}_b$ in which P and Q lie).

5.1.1 Motivating Examples

Here, we briefly present some examples of products $\langle \cdot, \cdot \rangle_a$ and spaces $\mathcal{H}_a$ of the form (5.2) that are commonly encountered in statistical theory and functional analysis. In the following examples, the base measure on $\mathcal{X}$ is taken to be the Lebesgue measure μ, and "probability densities" are with respect to μ. Also, for any integrable function $f \in \mathcal{L}^1(\mathcal{X})$, we use $\widetilde{f}_z = \int_{\mathcal{X}} f\psi_z \, d\mu$ to denote the z^{th} Fourier coefficient of f (where ψ_z is the z^{th} Fourier basis element as in (5.1)).

The simplest example is the standard $\mathcal{L}^2$ inner product:

Example 1. In the "unweighted" case where $a_z = 1$ for all $z \in \mathcal{Z}$, $\mathcal{H}_a$ includes the usual space $\mathcal{L}^2(\mathcal{X})$ of square-integrable probability densities on $\mathcal{X}$, and, for P and Q with square-integrable densities $p, q \in \mathcal{L}^2(\mathcal{X})$, we have

$$\langle p, q \rangle_a = \int_{\mathcal{X}} p(x)q(x) \, dx.$$

Typically, however, we are interested in weight sequences such that $a_z \to 0$ as $\|z\| \to \infty$ and $\mathcal{H}_a$ will be strictly smaller than $\mathcal{L}^2(\mathcal{X})$ to ensure that $\langle \cdot, \cdot \rangle_a$ is finite-valued; this corresponds intuitively to requiring additional smoothness of functions in $\mathcal{H}$. Here are two examples widely used in statistics:

Example 2. If $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) with a symmetric, translation-invariant kernel $K(x, y) = \kappa(x - y)$ (where $\kappa \in \mathcal{L}^2(\mathcal{X})$), one can show via Bochner's theorem (see, e.g., Theorem 6.6 of (Wendland, 2004)) that the semi-inner product induced by the kernel can be written in the form

$$\langle f, g \rangle_{\mathcal{H}_K} := \sum_{z \in \mathcal{Z}} \widetilde{\kappa}_z^{-2} \widetilde{f}_z \overline{\widetilde{g}_z}.$$

Hence, setting each $a_z = \langle \kappa, \psi_z \rangle = \widetilde{\kappa}_z$, $\mathcal{H}_a$ contains any distributions P and Q on $\mathcal{X}$ with densities $p, q \in \mathcal{H}_K = \{p \in \mathcal{L}^2 : \langle p, p \rangle_{\mathcal{H}_K} < \infty\}$, and we have $\langle P, Q \rangle_a = \langle p, q \rangle_{\mathcal{H}_K}$.

Example 3. For $s \in \mathbb{N}$, $\mathcal{H}^s$ is the s-order Sobolev space

$$\mathcal{H}^s := \left\{ f \in \mathcal{L}^2(\mathcal{X}) : f \text{ is } s\text{-times weakly differentiable with } f^{(s)} \in \mathcal{L}^2(\mathcal{X}) \right\},$$

endowed with the semi-inner product of the form

$$\langle p, q\rangle_{\mathcal{H}^s} := \left\langle p^{(s)}, q^{(s)} \right\rangle_{\mathcal{L}^2(\mathcal{X})} = \sum_{z\in\mathcal{Z}} |z|^{2s} \widetilde{f}_z \overline{\widetilde{g}_z} \tag{5.6}$$

where the last equality follows from Parseval's identity. Indeed, (5.6) is commonly used to generalize $\langle f, g\rangle_{\mathcal{H}^s}$, for example, to non-integer values of s. Thus, setting $a_z = |z|^{-s}$, $\mathcal{H}_a$ contains any distributions $P, Q \in \mathcal{P}$ with densities $p, q \in \mathcal{H}^s$, and, moreover, we have $\langle P, Q\rangle_a = \langle p, q\rangle_{\mathcal{H}^s}$. Note that, when $s \geq D/2$, one can show via Bochner's theorem that $\mathcal{H}^s$ is in fact also an RKHS, with symmetric, translation-invariant kernel defined as above by $\kappa(x) = \sum_{z\in\mathcal{Z}} z^{-s}\psi_z$.

Paper Organization

The remainder of this paper is organized as follows: In Section 5.2, we provide notation needed to formally state our estimation problem, given in Section 5.3. Section 5.4 reviews related work on estimation of functionals of probability densities, as well as some applications of this work. Sections 5.5 and 5.6 present our main theoretical results, with upper bounds in Sections 5.5 and minimax lower bounds in Section 5.6; proofs of all results are given in Appendix 5.9. Section 5.7 expands upon these general results in a number of important special cases. Finally, we conclude in Section 5.8 with a discussion of broader consequences and avenues for future work.

5.2 Notation

We assume the sample space $\mathcal{X} \subseteq \mathbb{R}^D$ is a compact subset of $\mathbb{R}^D$, and we use μ to denote the usual Lebesgue measure on $\mathcal{X}$. We use $\{\psi_z\}_{z\in\mathbb{Z}^D}$ to denote the standard orthonormal Fourier basis of $\mathcal{L}_2(\mathcal{X})$, indexed by D-tuples of integer frequencies $z \in \mathbb{Z}^D$. For any function $f \in \mathcal{L}_2(\mathcal{X})$ and $z \in \mathbb{Z}^D$, we use

$$\widetilde{f}_z := \int_{\mathcal{X}} f(x)\overline{\psi_z(x)}\, d\mu(x)$$

to denote the z^{th} Fourier coefficient of f (i.e., the projection of f onto ψ_z), and for any probability distribution $P \in \mathcal{P}$, we use the same notation

$$\phi_P(z) := \mathop{\mathbb{E}}_{X\sim P}\left[\overline{\psi_z(X)}\right] = \int_{\mathcal{X}} \overline{\psi_z(x)} dP(x)$$

to denote the characteristic function of P.

We will occasionally use the notation $\|z\|$ for indices $z \in \mathcal{Z}$. Due to equivalence of finite dimensional norms, the exact choice of norm affects only constant factors; for concreteness, one may take the Euclidean norm.

For certain applications, it is convenient to consider only a subset $\mathcal{Z} \subseteq \mathbb{Z}^D$ of indices of interest (for example, Sobolev seminorms are indexed only over $\mathcal{Z} = \{z \in \mathbb{Z}^D : z_1, ..., z_D \neq 0\}$). The subset $\mathcal{Z}$ may be considered arbitrary but fixed in our work.

Given two $(0,\infty)$-valued sequences[4] $a = \{a_z\}_{z\in\mathcal{Z}}$ and $b = \{b_z\}_{z\in\mathcal{Z}}$, we are interested in products of the form

$$\langle f, g\rangle_a := \sum_{z\in\mathcal{Z}} \frac{\widetilde{f}_z\overline{\widetilde{g}_z}}{a_z^2},$$

and their induced (semi)norms $\|f\|_a = \sqrt{\langle f, f\rangle_a}$ over spaces of the form[5]

$$\mathcal{H}_a = \{f \in \mathcal{L}_2(\mathcal{X}) : \|f\|_a < \infty\}$$

(and similarly when replacing a by b). Typically, we will have $a_z, b_z \to 0$ and $\frac{b_z}{a_z} \to 0$ whenever $\|z\| \to \infty$, implying the inclusion $\mathcal{H}_b \subseteq \mathcal{H}_a \subseteq \mathcal{L}^2(\mathcal{X})$.

5.3 Formal Problem Statement

Suppose we observe n i.i.d. samples $X_1, ..., X_n \overset{i.i.d.}{\sim} P$ and n i.i.d. samples $Y_1, ..., Y_n \overset{i.i.d.}{\sim} Q$, where P and Q are (unknown) distributions lying in the (known) space $\mathcal{H}_a$. We are interested in the problem of estimating the inner product (5.2), along with the closely related (squared) seminorm and pseudometric given by

$$\|P\|_a^2 := \langle P, P\rangle_a \quad \text{and} \quad \|P - Q\|_a^2 := \|P\|_a^2 + \|Q\|_a^2 - 2\langle P, Q\rangle_a. \tag{5.7}$$

We assume P and Q lie in a (known) smaller space $\mathcal{H}_b \subseteq \mathcal{H}_a$, and we are specifically interested in identifying, up to constant factors, the minimax mean squared (i.e., $\mathcal{L}^2$) error $M(a,b)$ of estimating $\langle P, Q\rangle_a$ over P and Q lying in a unit ellipsoid with respect to $\|\cdot\|_b$; that is, the quantity

$$M(a,b) := \inf_{\widehat{S}} \sup_{\|P\|_b, \|Q\|_b \le 1} \mathop{\mathbb{E}}_{\substack{X_1,...,X_n\sim p,\\ Y_1,...,Y_n\sim q}} \left[\left|\widehat{S} - \langle p, q\rangle_a\right|^2\right], \tag{5.8}$$

where the infimum is taken over all estimators (i.e., all functions $\widehat{S} : \mathbb{R}^{2n} \to \mathcal{C}$ of the data $X_1, ..., X_n, Y_1, ..., Y_n$).

5.4 Related Work

This section reviews previous studies on special cases of the problem we study, as well as work on estimating related functionals of probability distributions, and a few potential applications of this work in statistics and machine learning.

5.4.1 Prior work on special cases

While there has been substantial work on estimating unweighted $\mathcal{L}_2$ norms and distances of densities (Schweder, 1975; Anderson, Hall, and Titterington, 1994; Giné and Nickl, 2008), to the best of our knowledge, most work on the more general problem of estimating *weighted* inner products or norms has been on estimating Sobolev quantities (see Example 3 in Section 5.1) by Bickel and Ritov (1988), Donoho and

[4] A more proper mathematical term for a and b would be *net*.

[5] Specifically, we are interested in probability densities, which lie in the simplex $\mathcal{P} := \{f \in \mathcal{L}_1(\mathcal{X}) : f \ge 0, \int_{\mathcal{X}} f\, d\mu = 1\}$, so that we should write, e.g., $p, q \in \mathcal{H} \cap \mathcal{P}$. Henceforth, "density" refers to any function lying in $\mathcal{P}$.

Nussbaum (1990), and Singh, Du, and Póczos (2016). Bickel and Ritov (1988) considered the case of integer-order Sobolev norms, which have the form

$$\|f\|_{\mathcal{H}^s}^2 = \|f^{(s)}\|_{\mathcal{L}^2(\mathcal{X})}^2 = \int \left(f^{(s)}(x)\right)^2 dx, \tag{5.9}$$

for which they upper bounded the error of an estimator based on plugging a kernel density estimate into (5.9) and then applying an analytic bias correction. They also derived matching minimax lower bounds for this problem.[6] Singh, Du, and Póczos (2016) proved rate-matching upper bounds on the error of a much simpler inner product estimator (generalizing an estimator proposed by Donoho and Nussbaum (1990)), which applies for arbitrary $s \in \mathbb{R}$. Our upper and lower bounds are strict generalizations of these results. Specifically, relative to this previous work on the Sobolev case, our work makes advances in three directions:

1. We consider estimating a broader class of inner product functionals $\langle p, q \rangle_z$, for arbitrary sequences $\{a_z\}_{z \in \mathcal{Z}}$. The Sobolev case corresponds to $a_z = \|z\|^{-s}$ for some $s > 0$.

2. We consider a broader range of assumptions on the true data densities, of the form $\|p\|_b, \|q\|_b < \infty$, for arbitrary sequences $\{b_z\}_{z \in \mathcal{Z}}$. The Sobolev case corresponds to $b_z = \|z\|^{-t}$ for some $t > 0$.

3. We prove lower bounds that match our upper bounds, thereby identifying minimax rates. For many cases, such as Gaussian or exponential RKHS inner products or densities, these results are the first concerning minimax rates, and, even in the Sobolev case, our lower bounds address some previously open cases (namely, non-integer s and t, and $D > 1$.

The closely related work of Fan (1991) also generalized the estimator of Donoho and Nussbaum (1990), and proved (both upper and lower) bounds on $M(a, b)$ for somewhat more general sequences, and also considered norms with exponent $p \neq 2$ (i.e., norms not generated by an inner product, such as those underlying a broad class of Besov spaces). However, his analysis placed several restrictions on the rates of a and b; for example, it requires

$$\sup_{Z \subseteq \mathcal{Z}} \frac{|Z| \sup_{z \in Z} a_z^{-2}}{\sum_{z \in Z} a_z^{-2}} < \infty \quad \text{and} \quad \sup_{Z \subseteq \mathcal{Z}} \frac{|Z| \sup_{z \in Z} b_z^{-2}}{\sum_{z \in Z} b_z^{-2}} < \infty.$$

This holds when a and b decay polynomially, but fails in many of the cases we consider, such as exponential decay. The estimation of norms with $p \neq 2$ and a and b decaying non-polynomially, therefore, remains an important unstudied case, which we leave for future work.

Finally, we note that, except Singh, Du, and Póczos (2016), all the above works have considered only $D = 1$ (i.e., when the sample space $\mathcal{X} \subseteq \mathbb{R}$), despite the fact that D can play an important role in the convergence rates of the estimators. The results in this paper hold for arbitrary $D \geq 1$.

[6] Bickel and Ritov (1988) actually make Hölder assumptions on their densities (essentially, an $\mathcal{L}_\infty$ bound on the derivatives of the density), rather than our slightly milder Sobolev assumption (essentially, an $\mathcal{L}_2$ bound on the derivative). However, as we note in Section 5.8, these assumptions are closely related such that the results are comparable up to constant factors.

5.4.2 Estimation of related functionals

There has been quite a large amount of recent work (Nguyen, Wainwright, and Jordan, 2010; Liu, Wasserman, and Lafferty, 2012; Moon and Hero, 2014b; Singh and Póczos, 2014b; Singh and Póczos, 2014a; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014; Moon and Hero, 2014a; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2015; Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015; Gao, Steeg, and Galstyan, 2015a; Gao, Steeg, and Galstyan, 2015b; Mukherjee, Tchetgen, and Robins, 2015; Mukherjee, Tchetgen, and Robins, 2016; Moon, Sricharan, Greenewald, and Hero, 2016; Singh and Póczos, 2016b; Berrett, Samworth, and Yuan, 2019; Gao, Oh, and Viswanath, 2017b; Gao, Kannan, Oh, and Viswanath, 2017; Jiao, Gao, and Han, 2018; Han, Jiao, Weissman, and Wu, 2017; Noshad, Moon, Sekeh, and Hero, 2017; Wisler, Moon, and Berisha, 2017; Singh and Póczos, 2017; Noshad and Hero III, 2018; Bulinski and Dimitrov, 2018; Bulinski and Kozhevin, 2018; Sekeh, Oselio, and Hero, 2018; Berrett and Samworth, 2019; Rubenstein, Bousquet, Djolonga, Riquelme, and Tolstikhin, 2019; Sekeh and Hero, 2019; Ba and Lo, 2019; Goldfeld, Greenewald, Polyanskiy, and Weed, 2019) on practical estimation of nonlinear integral functionals of probability densities, of the form

$$F(p) = \int_{\mathcal{X}} \varphi(p(x))\, dx, \tag{5.10}$$

where $\phi : [0, \infty) \to \mathbb{R}$ is nonlinear but smooth. Whereas minimax optimal estimators have been long established, their computational complexity typically scales as poorly as $O(n^3)$ (Birgé and Massart, 1995; Laurent, 1996; Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015). Hence, this recent work has focused on analyzing more computationally efficient (but less statistically efficient) estimators, as well as on estimating information-theoretic quantities such as variants of entropy, mutual information, and divergence, for which ϕ can be locally non-smooth (e.g., $\phi = \log$), and can hence follow somewhat different minimax rates.

As discussed in detail by Laurent (1996), under Sobolev smoothness assumptions on p, estimation of quadratic functionals (such as those considered in this paper) is key to constructing minimax rate-optimal estimators for general functionals of the form (5.10). The reason for this is that minimax rate-optimal estimators of $F(p)$ can often be constructed by approximating a second-order Taylor (a.k.a., von Mises (Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015)) expansion of F around a density estimate $\widehat{p}$ of p that is itself minimax rate-optimal (with respect to integrated mean squared error). Informally, if we expand $F(p)$ as

$$F(p) = F(\widehat{p}) + \langle \nabla F(\widehat{p}), p - \widehat{p} \rangle_{\mathcal{L}^2} + \left\langle p - \widehat{p}, (\nabla^2 F(\widehat{p}))p - \widehat{p} \right\rangle_{\mathcal{L}^2} + O\left(\|p - q\|_{\mathcal{L}^2}^3 \right), \tag{5.11}$$

where $\nabla F(\widehat{p})$ and $\nabla^2 F(\widehat{p})$ are the first and second order Frechet derivatives of F at $\widehat{p}$.

In the expansion (5.11), the first term is a simple plug-in estimate, and the second term is linear in p, and can therefore be estimated easily by an empirical mean. The remaining term is precisely a quadratic functional of the density, of the type we seek to estimate in this paper. Indeed, to the best of our knowledge, this is the approach taken by *all* estimators that are known to achieve minimax rates (Birgé and Massart, 1995; Laurent, 1996; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014; Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015; Mukherjee, Tchetgen, and Robins, 2015; Mukherjee, Tchetgen, and Robins, 2016) for general functionals of the form (5.10).

Interestingly, the estimators studied in the recent papers above are all based on either kernel density estimators (Singh and Póczos, 2014b; Singh and Póczos, 2014a; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2014; Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2015; Kandasamy, Krishnamurthy, Poczos, and Wasserman, 2015; Moon, Sricharan, Greenewald, and Hero, 2016; Mukherjee, Tchetgen, and Robins, 2015; Mukherjee, Tchetgen, and Robins, 2016) or k-nearest neighbor methods (Moon and Hero, 2014b; Moon and Hero, 2014a; Singh and Póczos, 2016b; Berrett, Samworth, and Yuan, 2019; Gao, Oh, and Viswanath, 2017b). This contrasts with our approach, which is more comparable to orthogonal series density estimation; given the relative efficiency of computing orthogonal series estimates (e.g., via the fast Fourier transform), it may be desirable to try to adapt our estimators to these classes of functionals.

When moving beyond Sobolev assumptions, only estimation of very specific functionals has been studied. For example, under RKHS assumptions, only estimation of maximum mean discrepancy (MMD)(Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Ramdas, Reddi, Póczos, Singh, and Wasserman, 2015; Tolstikhin, Sriperumbudur, and Schölkopf, 2016), has received much attention. Hence, our work significantly expands our understanding of minimax functional estimation in this setting. More generally, our work begins to provide a framework for a unified understanding of functional estimation across different types of smoothness assumptions.

Along a different line, there has also been some work on estimating $\mathcal{L}^p$ norms for *regression functions*, under similar Sobolev smoothness assumptions (Lepski, Nemirovski, and Spokoiny, 1999). However, the problem of norm estimation for regression functions turns out to have quite different statistical properties and requires significantly different estimators and analysis, compared to norm estimation for density functions. Generally, the problem for densities is statistically easier in terms of having a faster convergence rate under a comparable smoothness assumption; this is most obvious when $p = 1$, since the $\mathcal{L}^1$ norm of a density is always 1, while the $\mathcal{L}^1$ norm of a regression function is less trivial to estimate. However, this is true more generally as well. For example, Lepski, Nemirovski, and Spokoiny (1999) showed that, under s-order Sobolev assumptions, the minimax rate for estimating the $\mathcal{L}^2$ norm of a 1-dimensional regression function (up to $\log$ factors) is $\asymp n^{-\frac{4s}{4s+1}}$, whereas the corresponding rate for estimating the $\mathcal{L}^2$ norm of a density function is $\asymp n^{-\min\left\{\frac{8s}{4s+1},1\right\}}$, which is parametric when $s \geq 1/4$. To the best of our knowledge, there has been no work on the natural question of estimating Sobolev or other more general quadratic functionals of regression functions.

5.4.3 Applications

Finally, although this paper focuses on estimation of general inner products from the perspective of statistical theory, we mention a few of the many applications that motivate the study of this problem.

Estimates of quadratic functionals can be directly used for nonparametric goodness-of-fit, independence, and two-sample testing (Anderson, Hall, and Titterington, 1994; Dumbgen, 1998; Ingster and Suslina, 2012; Goria, Leonenko, Mergel, and Novi Inverardi, 2005; Pardo, 2005; Chwialkowski, Ramdas, Sejdinovic, and Gretton, 2015). They can also by used to construct confidence sets for a variety of nonparametric objects (Li, 1989; Baraud, 2004; Genovese and Wasserman, 2005), as well as for parameter estimation in semi-parametric models (Wolsztynski, Thierry, and Pronzato, 2005b).

In machine learning, Sobolev-weighted distances can also be used in transfer learning (Du, Koushik, Singh, and Póczos, 2017) and transduction learning (Quadrianto, Petterson, and Smola, 2009) to measure relatedness between source and target domains, helping to identify when transfer can benefit learning. Semi-inner products can be used as kernels over probability distributions, enabling generalization of a wide variety of statistical learning methods from finite-dimensional vectorial inputs to nonparametric distributional inputs (Sutherland, 2016). This *distributional learning* approach has been applied to many diverse problems, including image classification (Póczos, Xiong, and Schneider, 2011; Póczos, Xiong, Sutherland, and Schneider, 2012), galaxy mass estimation (Ntampaka, Trac, Sutherland, Battaglia, Póczos, and Schneider, 2015), ecological inference (Flaxman, Wang, and Smola, 2015; Flaxman, Sutherland, Wang, and Teh, 2016), aerosol prediction in climate science (Szabó, Gretton, Póczos, and Sriperumbudur, 2015), and causal inference (Lopez-Paz, Muandet, Schölkopf, and Tolstikhin, 2015). Finally, it has recently been shown that the losses minimized in certain implicit generative models can be approximated by Sobolev and related distances (Liang, 2017). Further applications of these quantities can be found in (Principe, 2010).

5.5 Upper Bounds

In this section, we provide upper bounds on minimax risk. Specifically, we propose estimators for semi-inner products, semi-norms, and pseudo-metrics, and bound the risk of the semi-inner product estimator; identical bounds (up to constant factors) follow easily for semi-norms and pseudo-metrics.

5.5.1 Proposed Estimators

Our proposed estimator $\widehat{S}_Z$ of $\langle P, Q\rangle_a$ consists of simply plugging estimates of ϕ_P and $\widetilde{Q}$ into a truncated version of the summation in Equation (5.2). Specifically, since

$$\phi_P(z) = \mathop{\mathbb{E}}_{X\sim P}\left[\overline{\psi_z(X)}\right],$$

we estimate each $\phi_P(z)$ by $\widehat{\phi}_P(z) := \frac{1}{n}\sum_{i=1}^n \psi_z(X_i)$ and each $\phi_Q(z)$ by $\widehat{\phi}_Q(z) := \frac{1}{n}\sum_{i=1}^n \psi_z(Y_i)$. Then, for some finite set $Z \subseteq \mathcal{Z}$ (a tuning parameter to be chosen later) our estimator $\widehat{S}_Z$ for the product (5.2) is

$$\widehat{S}_Z := \sum_{z\in Z} \frac{\widehat{\phi}_P(z)\overline{\widehat{\phi}_Q(z)}}{a_z^2}. \tag{5.12}$$

To estimate the squared semi-norm $\|P\|_a^2$ from a single sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, we use

$$\widehat{N}_Z := \sum_{z\in Z} \frac{\widehat{\phi}_P(z)\overline{\widehat{\phi}_P(z)'}}{a_z^2}. \tag{5.13}$$

where $\widehat{\phi}_P(z)$ is estimated using the first half $X_1, \ldots, X_{\lfloor n/2\rfloor}$ of the sample, $\widehat{\phi}_P(z)'$ is estimated using the second half $X_{\lfloor n/2\rfloor+1}, \ldots, X_n$ of the sample. While it is not clear that sample splitting is optimal in practice, it allows us to directly apply convergence results for the semi-inner product, which assume the samples from the two densities are independent.

To estimate the squared pseudo-metric $\|P-Q\|_a^2$ from two samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Q$, we combine the above inner product and norm estimators according to the formula (5.7), giving

$$\widehat{\rho}_Z = \widehat{N}_Z + \widehat{M}_Z - 2\widehat{S}_Z,$$

where $\widehat{M}_Z$ denotes the analogue of the norm estimator (5.13) applied to $Y_1, \ldots, Y_n$.

5.5.2 Bounding the risk of $\widehat{S}_Z$

Here, we state upper bounds on the bias, variance, and mean squared error of the semi-inner product estimator $\widehat{S}_Z$, beginning with an easy bound on the bias of $\widehat{S}_Z$ (proven in Appendix 5.9.1):

Proposition 24 (Upper bound on bias of $\widehat{S}_Z$)**.** *Suppose $P, Q \in \mathcal{H}_b$. Then,*

$$\left|\mathcal{B}\left[\widehat{S}_Z\right]\right| \leq \|P\|_b \|Q\|_b \sup_{z \in \mathcal{Z} \setminus Z} \frac{b_z^2}{a_z^2}, \tag{5.14}$$

where $\mathcal{B}\left[\widehat{S}_Z\right] := \mathbb{E}\left[\widehat{S}_Z\right] - \langle P, Q \rangle_a$ denotes the bias of $\widehat{S}_Z$.

Note that for the above bound to be non-trivial, we require $b_z \to 0$ faster than a_z as $\|z\| \to \infty$ which ensures that $\sup_{z \in \mathcal{Z} \setminus Z} \frac{b_z}{a_z} < \infty$. While (5.14) does not explicitly depend on the sample size n, in practice, the parameter set Z will be chosen to grow with n, and hence the supremum over $\mathcal{Z} \setminus Z$ will decrease monotonically with n. Next, we provide a bound on the variance of $\widehat{S}_Z$, whose proof, given in Appendix 5.9.2, is more involved.

Proposition 25 (Upper bound on variance of $\widehat{S}_Z$)**.** *Suppose $P, Q \in \mathcal{H}_b$. Then,*

$$\mathbb{V}[\widehat{S}_Z] \leq \frac{2\|P\|_2 \|Q\|_2}{n^2} \sum_{z \in Z} \frac{1}{a_z^4} + \frac{\|Q\|_b^2 \|P\|_b + \|P\|_b^2 \|Q\|_b}{n} R_{a,b,Z} + \frac{2\|P\|_a^2 \|Q\|_a^2}{n} \tag{5.15}$$

where $\mathbb{V}$ denotes the variance operator and

$$R_{a,b,Z} := \left(\sum_{z \in Z} \frac{b_z^4}{a_z^8}\right)^{1/4} \left(\sum_{z \in Z} \left(\frac{b_z}{a_z^2}\right)^8\right)^{1/8} \left(\sum_{z \in Z} b_z^8\right)^{1/8}. \tag{5.16}$$

Having bounded the bias and variance of the estimator $\widehat{S}_Z$, we now turn to the mean squared error (MSE). Via the usual decomposition of MSE into (squared) bias and variance, Propositions 24 and 25 together immediately imply the following bound:

Theorem 26 (Upper bound on MSE of $\widehat{S}_Z$)**.** *Suppose $P, Q \in \mathcal{H}_b$. Then,*

$$\begin{aligned} \mathrm{MSE}\left[\widehat{S}_Z\right] \leq \|P\|_b^2 \|Q\|_b^2 \sup_{z \in \mathcal{Z} \setminus Z} \frac{b_z^4}{a_z^4} + \frac{2\|P\|_2 \|Q\|_2}{n^2} \sum_{z \in Z} \frac{1}{a_z^4} \\ + \frac{\|Q\|_b^2 \|P\|_b + \|P\|_b^2 \|Q\|_b}{n} R_{a,b,Z} + \frac{2\|P\|_a^2 \|Q\|_a^2}{n}, \end{aligned} \tag{5.17}$$

where $R_{a,b,Z}$ is as defined in (5.16).

Corollary 27 (Norm estimation)**.** *In the particular case of norm estimation (i.e., when $Q = P$), this simplifies to:*

$$\mathrm{MSE}\left[\widehat{S}_Z\right] \leq \|P\|_b^4 \sup_{z\in\mathcal{Z}\setminus Z} \frac{b_z^4}{a_z^4} + \frac{2\|P\|_2^2}{n^2}\sum_{z\in Z}\frac{1}{a_z^4} + \frac{2\|P\|_b^3}{n}R_{a,b,Z} + \frac{2\|P\|_a^4}{n}. \tag{5.18}$$

5.5.3 Discussion of Upper Bounds

Two things might stand out that distinguish the above variance bound from many other nonparametric variance bounds: First, the rate depends on the smoothness of $P, Q \in \mathcal{H}_b$. Smoothness assumptions in nonparametric statistics are usually needed only to bound the bias of estimators (Tsybakov, 2008). The reason the smoothness appears in this variance bound is that the estimand in Equation (5.2) includes products of the Fourier coefficients of P and Q. Hence, the estimates $\widehat{\phi}_P(z)$ of $\phi_P(z)$ are scaled by $\widehat{\phi}_Q(z)$, and vice versa, and as a result, the decay rates of $\phi_P(z)$ and $\phi_Q(z)$ affect the variance of the tails of $\widehat{S}_Z$. One consequence of this is that the convergence rates exhibit a phase transition, with a parametric convergence rate when the tails of ϕ_P and $\widetilde{Q}$ are sufficiently light, and a slower rate otherwise.

Second, the bounds are specific to the Fourier basis (as opposed to, say, any uniformly bounded basis, e.g., one with $\sup_{z\in\mathcal{Z},x\in\mathcal{X}}|\psi_z(x)| \leq 1$). The reason for this is that, when expanded, the variance includes terms of the form $\mathbb{E}_{X\sim P}[\phi_y(X)\psi_z(X)]$, for some $y \neq z \in \mathcal{Z}$. In general, these covariance-like terms are difficult to bound tightly; for example, the uniform boundedness assumption above would only give a bound of the form $\mathbb{E}_{X\sim P}[|\phi_y(X)\psi_z(X)|] \leq \min\{\phi_P(y), \phi_P(z)\}$. For the Fourier basis, however, the recurrence relation $\phi_y\psi_z = \phi_{y+z}$ allows us to bound $\mathbb{E}_{X\sim P}[\phi_y(X)\psi_z(X)] = \mathbb{E}_{X\sim P}[\phi_{y+z}(X)] = \phi_P(y+z)$ in terms of assumptions on the decay rates of the coefficients of P. It turns out that $\phi_P(y+z)$ decays significantly faster than $\min\{\phi_P(y), \phi_P(z)\}$, and this tighter bound is needed to prove optimal convergence rates.

More broadly, this suggests that convergence rates for estimating inner products in terms of weights in a particular basis may depend on algebraic properties of that basis. For example, another common basis, the Haar wavelet basis, satisfies a different recurrence relation: $\phi_y\psi_z \in \{0, \phi_y, \psi_z\}$, depending on whether (and how) the supports of ϕ_y and ψ_z are nested or disjoint. We leave investigation of this and other bases for future work.

Clearly, $\sup_{Z\subseteq\mathcal{Z}} R_{a,b,Z} < \infty$ if and only if b_z^4/a_z^8 is summable (i.e., $\sum_{z\in\mathcal{Z}} b_z^4/a_z^8 < \infty$). Thus, assuming $|\mathcal{Z}| = \infty$, this already identifies the precise condition required for the minimax rate to be parametric. When it is the case that

$$\frac{R_{a,b,Z}}{n} \in O\left(\sup_{z\in\mathcal{Z}\setminus Z}\frac{b_z^4}{a_z^4} + \frac{1}{n^2}\sum_{z\in Z}\frac{1}{a_z^4}\right),$$

the third term in (5.17) will be dominated by the first and third terms, and so the upper bound simplifies to order

$$\mathrm{MSE}\left[\widehat{S}_{Z^*}\right] \lesssim \frac{1}{n} + \min_{Z\subseteq\mathcal{Z}}\left[\sup_{z\in\mathcal{Z}\setminus Z}\frac{b_z^4}{a_z^4} + \frac{1}{n^2}\sum_{z\in Z}\frac{1}{a_z^4}\right]. \tag{5.19}$$

This happens for every choice of a_z and b_z we consider in this paper, including the Sobolev (polynomial decay) case and the RKHS case. However, simplifying the

bound further requires some knowledge of the form of a and/or b, and we develop this in several cases in Section 5.7. In Section 5.8, we also consider some heuristics for approximately simplifying (5.19) in certain settings.

5.6 Lower Bounds

In this section, we provide a lower bound on the minimax risk of the estimation problems described in Section 5.3. Specifically, we use a standard information theoretic framework to lower bound the minimax risk for semi-norm estimation; bounds of the same rate follow easily for inner products and pseudo-metrics. In a wide range of cases, our lower bound matches the MSE upper bound (Theorem 26) presented in the previous section.

Theorem 28 (Lower Bound on Minimax MSE)**.** *Suppose $\mathcal{X}$ has finite base measure $\mu(\mathcal{X}) = 1$ and suppose the basis $\{\psi_z\}_{z\in\mathcal{Z}}$ contains the constant function $\phi_0 = 1$ and is uniformly bounded (i.e., $\sup_{z\in\mathcal{Z}} \|\psi_z\|_\infty < \infty$). Define $Z_{\zeta_n} := \{z \in \mathbb{Z}^D : \|z\|_\infty \le \zeta_n\}$, $A_{\zeta_n} := \sum_{z\in Z_{\zeta_n}} a_z^{-2}$ and $B_{\zeta_n} := \sum_{z\in Z_{\zeta_n}} b_z^{-2}$. If $B_{\zeta_n} \in \Omega\left(\zeta_n^{2D}\right)$, then we have the minimax lower bound*

$$\inf_{\widehat{S}} \sup_{\|h\|_b\le 1} \mathbb{E}\left[\left(\widehat{S} - \|h\|_a^2\right)^2\right] \in \Omega\left(\max\left\{\frac{A_{\zeta_n}^2}{B_{\zeta_n}^2}, n^{-1}\right\}\right),$$

where ζ_n is chosen to satisfy $B_{\zeta_n}^2 \asymp \zeta_n^D n^2$. Also, if $B_{\zeta_n} \in o\left(\zeta_n^{2D}\right)$, then we have the (looser) minimax lower bound

$$\inf_{\widehat{S}} \sup_{\|h\|_b\le 1} \mathbb{E}\left[\left(\widehat{S} - \|h\|_a^2\right)^2\right] \in \Omega\left(\max\left\{\frac{A_{n^{2/3D}}^2}{n^{4/3}}, n^{-1}\right\}\right).$$

Remark 29. The uniform boundedness assumption permits the Fourier basis, our main case of interest, but also allows other bases (see, e.g., the "generalized Fourier bases" used in Corollary 2.2 of Liang (2017)).

Remark 30. The condition that $B_{\zeta_n} \in \Omega\left(\zeta_n^{2D}\right)$ is needed to ensure that the "worst-case" densities we construct in the proof of Theorem 28 are indeed valid probability densities (specifically, that they are non-negative). Hence, this condition would no longer be necessary if we proved results in the simpler Gaussian sequence model, as in many previous works on this problem (e.g., (Cai, 1999; Cai and Low, 2005)). However, when $B_{\zeta_n} \in o\left(\zeta_n^{2D}\right)$, density estimation, and hence the related problem of norm estimation, become asymptotically easier than the analogous problems under the Gaussian sequence model.

Remark 31. Intuitively, the ratio A_{ζ_n}/B_{ζ_n} measures the relative strengths of the norms $\|\cdot\|_a$ and $\|\cdot\|_b$. As expected, consistent estimation is possible if and only if $\|\cdot\|_b$ is a stronger norm than $\|\cdot\|_a$.

5.7 Special Cases

In this section, we develop our lower and upper results for several special cases of interest. The results of this section are summarized in Table 5.1.

Notation: Here, for simplicity, we assume that the estimator $\widehat{S}_Z$ uses a choice of Z that is symmetric across dimensions; in particular, $Z = \prod_{j=1}^D \{\phi_{-\zeta_n}, ..., \phi_0, ..., \phi_{\zeta_n}\}$

(for some $\zeta_n \in \mathbb{N}$ depending on n) is the Cartesian product of D sets of the first $2\zeta_n + 1$ integers. Throughout this section, we use $\lesssim$ and $\gtrsim$ to denote inequality up to $\log n$ factors. Although we do not explicitly discuss estimation of $\mathcal{L}^2$ norms, it appears as a special case of the Sobolev case with $s = 0$.

5.7.1 Sobolev

For some $s, t \geq 0$, $a_z = \|z\|^{-s}$ and $b_z = \|z\|^{-t}$.

Upper Bound: By Proposition 24, $\mathcal{B}[\widehat{S}_Z] \lesssim \zeta_n^{2(s-t)}$, and, by Proposition 25,

$$\mathbb{V}[\widehat{S}_Z] \lesssim \frac{\zeta_n^{4s+D}}{n^2} + \frac{\zeta_n^{4s-3t+D/2}}{n} + \frac{1}{n}$$

Thus,

$$\mathrm{MSE}[\widehat{S}_Z] \lesssim \zeta_n^{4(s-t)} + \frac{\zeta_n^{4s+D}}{n^2} + \frac{\zeta_n^{4s-3t+D/2}}{n} + \frac{1}{n}.$$

One can check that $\zeta_n^{4(s-t)} + \frac{\zeta_n^{4s+D}}{n^2}$ is minimized when $\zeta_n \asymp n^{\frac{2}{4t+D}}$, and that, for this choice of ζ_n, the $\frac{\zeta_n^{4s-3t+D/2}}{n}$ term is of lower order, giving the convergence rate

$$\mathrm{MSE}[\widehat{S}_Z] \asymp n^{\frac{8(s-t)}{4t+D}}.$$

Lower Bound: Note that $A_{\zeta_n} \asymp \zeta_n^{2s+D}$ and $B_{\zeta_n} \asymp \zeta_n^{2t+D}$. Solving $B_{\zeta_n}^2 = \zeta_n^D n^2$ gives $\zeta_n \asymp n^{\frac{2}{4t+D}}$. Thus, Theorem 28 gives a minimax lower bound of

$$\inf_{\widehat{S}} \sup_{\|p\|_b, \|q\|_b \leq 1} \mathbb{E}\left[\left(\widehat{S} - \langle p, q\rangle_b\right)^2\right] \gtrsim \frac{A_{\zeta_n}^2}{B_{\zeta_n}^2} = \zeta_n^{4(s-t)} = n^{\frac{8(s-t)}{4t+D}},$$

matching the upper bound. Note that the rate is parametric ($\asymp n^{-1}$) when $t \geq 2s + D/4$, and slower otherwise.

5.7.2 Gaussian RKHS

For some $t \geq s \geq 0$, $a_z = e^{-s\|z\|_2^2}$ and $b_z = e^{-t\|z\|_2^2}$.

Upper Bound: By Proposition 24, $\mathcal{B}[\widehat{S}_Z] \lesssim e^{2(s-t)\zeta_n^2}$. If we use the upper bound

$$\sum_{z \in Z} e^{\theta\|z\|_2^2} \leq C_\theta \zeta_n^D e^{\theta\zeta_n^2},$$

for any $\theta > 0$ and some $C_\theta > 0$, then Proposition 25 gives

$$\mathbb{V}[\widehat{S}_Z] = \frac{\zeta_n^D e^{4s\zeta_n^2}}{n^2} + \frac{\zeta_n^{89D/20} e^{(4s-3t)\zeta_n^2}}{n} + \frac{1}{n}.$$

Thus,

$$\mathrm{MSE}[\widehat{S}_Z] \lesssim e^{4(s-t)\zeta_n^2} + \frac{\zeta_n^D e^{s\zeta_n^2}}{n^2} + \frac{\zeta_n^{89D/20} e^{(4s-3t)\zeta_n^2}}{n} + \frac{1}{n}.$$

One can check that $e^{4(s-t)\zeta_n^2} + \frac{\zeta_n^D e^{s\zeta_n^2}}{n^2}$ is minimized when $\zeta_n \asymp \sqrt{\frac{\log n}{2t}}$, and that, for this choice of ζ_n, the $\zeta_n^{89D/20} e^{(4s-3t)\zeta_n^2}$ term is of lower order, giving an MSE convergence rate of

$$\mathrm{MSE}[\widehat{S}_Z] \lesssim n^{\frac{2(s-t)}{t}} = n^{2(s/t-1)}.$$

Lower Bound: Again, we use the bound

$$A_{\zeta_n} = \sum_{z \in Z_{\zeta_n}} e^{2s\|z\|_2^2} \lesssim \zeta_n^D e^{2s\zeta_n^2},$$

as well as the trivial lower bound $B_{\zeta_n} = \sum_{z \in Z_{\zeta_n}} e^{2s\|z\|_2^2} \geq e^{2s\zeta_n^2}$. Solving $B_{\zeta_n}^2 = \zeta_n^D n^2$ gives $\zeta_n \asymp \sqrt{\frac{\log n}{2t}}$ up to $\log\log n$ factors. Thus, ignoring $\log n$ factors, Theorem 28 gives a minimax lower bound of

$$\inf_{\widehat{S}} \sup_{\|p\|_b, \|q\|_b \leq 1} \mathbb{E}\left[\left(\widehat{S} - \langle p, q\rangle_b\right)^2\right] \gtrsim n^{\frac{2(s-t)}{t}},$$

for some $C > 0$, matching the upper bound rate. Note that the rate is parametric when $t \geq 2s$, and slower otherwise.

5.7.3 Exponential RKHS

For some $t \geq s \geq 0$, $a_z = e^{-s\|z\|_1}$ and $b_z = e^{-t\|z\|_1}$.
Upper Bound: By Proposition 24, $\mathcal{B}[\widehat{S}_Z] \lesssim e^{2(s-t)\zeta_n}$. Since, for fixed D,

$$\sum_{z \in Z} e^{r\|z\|_1} \asymp e^{r\zeta_n + D} \asymp e^{r\zeta_n},$$

by Proposition 25, we have

$$\mathbb{V}[\widehat{S}_Z] \asymp \frac{e^{4s\zeta_n}}{n^2} + \frac{e^{(4s-3t)\zeta_n}}{n} + \frac{1}{n},$$

giving a mean squared error bound of

$$\mathrm{MSE}[\widehat{S}_Z] \asymp e^{4(s-t)\zeta_n} + \frac{e^{4s\zeta_n}}{n^2} + \frac{e^{(4s-3t)\zeta_n}}{n} + \frac{1}{n}.$$

One can check that $e^{4(s-t)\zeta_n} + \frac{e^{4s\zeta_n}}{n^2}$ is minimized when $\zeta_n \asymp \frac{\log n}{2t}$, and that, for this choice of ζ_n, the $\frac{e^{(4s-3t)\zeta_n}}{n}$ term is of lower order, giving an MSE convergence rate of

$$\mathrm{MSE}[\widehat{S}_Z] \lesssim n^{\frac{2(s-t)}{t}} = n^{2(s/t-1)}.$$

Lower Bound: Note that $A_{\zeta_n} \asymp e^{2s\zeta_n}$ and $B_{\zeta_n} = e^{2t\zeta_n}$. Solving $B_{\zeta_n}^2 = \zeta_n^D n^2$ gives, up to $\log\log n$ factors, $\zeta_n \asymp \frac{\log n}{2t}$. Thus, Theorem 28 gives a minimax lower bound of

$$\inf_{\widehat{S}} \sup_{\|p\|_b, \|q\|_b \leq 1} \mathbb{E}\left[\left(\widehat{S} - \langle p, q\rangle_b\right)^2\right] \gtrsim n^{\frac{2(s-t)}{t}},$$

for some $C > 0$, matching the upper bound rate. Note that the rate is parametric when $t \geq 2s$, and slower otherwise.

5.7.4 Logarithmic decay

For some $t \geq s \geq 0$, $a_z = (\log \|z\|)^{-s}$ and $b_z = (\log \|z\|)^{-t}$. Note that, since our lower bound requires $B_{\zeta_n} \in \Omega(\zeta_n^{2D})$, we will only study the upper bound for this case.

Upper Bound: By Proposition 24, $\mathcal{B}[\widehat{S}_Z] \lesssim (\log \zeta_n)^{2(s-t)}$. By the upper bound

$$\sum_{z \in Z_{\zeta_n}} (\log \|z\|)^{\theta} \leq C_\theta \zeta_n^D (\log \zeta_n)^{\theta},$$

for any $\theta > 0$ and some $C_\theta > 0$, Proposition 25 gives

$$\mathbb{V}[\widehat{S}_Z] = \frac{\zeta_n^D (\log \zeta_n)^{4s}}{n^2} + \frac{\zeta_n^{89D/20} (\log \zeta_n)^{4s-3t}}{n} + \frac{1}{n},$$

giving a mean squared error bound of

$$\mathrm{MSE}[\widehat{S}_Z] \lesssim (\log \zeta_n)^{4(s-t)} + \frac{\zeta_n^D (\log \zeta_n)^{4s}}{n^2} + \frac{\zeta_n^{89D/20} (\log \zeta_n)^{4s-3t}}{n} + \frac{1}{n}.$$

One can check that $(\log \zeta_n)^{4(s-t)} + \frac{\zeta_n^{89D/20} (\log \zeta_n)^{4s-3t}}{n}$ is minimized when $\zeta_n^{89D/20} (\log \zeta_n)^{4t+D} \asymp n$, and one can check that, for this choice of ζ_n, the $\frac{\zeta_n^D (\log \zeta_n)^{4s}}{n^2}$ term is of lower order.

Thus, up to $\log n$ factors, $\zeta_n \asymp n^{2/D}$, and so, up to $\log \log n$ factors,

$$(\log \zeta_n)^{4(s-t)} \asymp (\log n)^{4(s-t)}.$$

5.7.5 Sinc RKHS

For any $s \in (0, \infty)^D$, the sinc_s kernel, defined by

$$K^s_{\mathrm{sinc}}(x, y) = \prod_{j=1}^{d} \frac{s_j}{\pi} \operatorname{sinc}\left(\frac{x_j - y_j}{s_j}\right),$$

where

$$\operatorname{sinc}(x) = \begin{cases} \frac{\sin(x)}{x} & \text{if } x \neq y \\ 1 & \text{else} \end{cases},$$

generates the RKHS $\mathcal{H}^s_{\mathrm{sinc}} = \left\{f \in \mathcal{L}^2 : \|f\|_{K^s_{\mathrm{sinc}}} < \infty\right\}$, of band-limited functions, where the norm is generated by the inner product $\langle f, g \rangle_{K^s_{\mathrm{sinc}}} = \langle f, g \rangle_a$, where $a_z = 1_{\{|z| \leq s\}}$ (with the convention that $\frac{0}{0} = 0$). If we assume that $p \in \mathcal{H}^t_{\mathrm{sinc}}$, where $t \leq s$, then fixing $Z := \{z \in \mathbb{Z}^D : |z| \leq s\}$, by Proposition 24, $\mathcal{B}[\widehat{S}_Z] = 0$, and, by Proposition 25, one can easily check that $\mathbb{V}[\widehat{S}_Z] \lesssim n^{-1}$. Thus, without any assumptions on P, we can always estimate $\|P\|_{K^s_{\mathrm{sinc}}}$ at the parametric rate.

5.8 Discussion

In this paper, we focused on the case of inner product weights and density coefficients in the Fourier basis, which play well-understood roles in widely used spaces such as Sobolev spaces and reproducing kernel Hilbert spaces with translation-invariant kernels.

For nearly all choices of weights $\{a_z\}_{z \in \mathcal{Z}}$ and $\{b_z\}_{z \in \mathcal{Z}}$, ignoring the parametric $1/n$ term that appears in both the upper and lower bounds, the upper bound boils

	$b_z = \log^{-t}\|z\|$	$b_z = \|z\|^{-t}$	$b_z = e^{-t\|z\|_1}$	$b_z = e^{-t\|z\|_2^2}$
$a_z = \log^{-s}\|z\|$	$\lesssim (\log n)^{4(s-t)}$	$\max\left\{n^{-1}, n^{\frac{-8t}{4t+D}}\right\}$	n^{-1}	n^{-1}
$a_z = \|z\|^{-s}$	∞	$\max\left\{n^{-1}, n^{\frac{8(s-t)}{4t+D}}\right\}$	n^{-1}	n^{-1}
$a_z = e^{-s\|z\|_1}$	∞	∞	$\max\left\{n^{-1}, n^{2\frac{s-t}{t}}\right\}$	n^{-1}
$a_z = e^{-s\|z\|_2^2}$	∞	∞	∞	$\max\left\{n^{-1}, n^{2\frac{s-t}{t}}\right\}$

TABLE 5.1: Minimax convergence rates for different combinations of a_z and b_z. Results are given up to $\log n$ factors, except the case when both a_z and b_z are logarithmic, which is given up to $\log\log n$ factors. Note that, in this last case, only the upper bound is known. A value of ∞ indicates that the estimand itself may be ∞ and consistent estimation is impossible.

down to

$$\min_{\zeta_n \in \mathbb{N}} \frac{b_{\zeta_n}^4}{a_{\zeta_n}^4} + \frac{\sum_{z \in Z_{\zeta_n}} a_z^{-4}}{n^2},$$

or, equivalently,

$$\frac{b_{\zeta_n}^4}{a_{\zeta_n}^4} \quad \text{where} \quad \frac{b_{\zeta_n}^4}{a_{\zeta_n}^4} = \frac{\sum_{z \in Z_{\zeta_n}} a_z^{-4}}{n^2}$$

and the lower bound boils down to

$$\left(\frac{\sum_{z \in Z_{\zeta_n}} a_z^{-2}}{\sum_{z \in Z_{\zeta_n}} b_z^{-2}}\right)^2, \quad \text{where} \quad \left(\sum_{z \in Z_{\zeta_n}} b_z^{-2}\right)^2 = \zeta_n^D n^2.$$

These rates match if

$$\frac{a_{\zeta_n}^{-4}}{b_{\zeta_n}^{-4}} \asymp \left(\frac{\sum_{z \in Z_{\zeta_n}} a_z^{-2}}{\sum_{z \in Z_{\zeta_n}} b_z^{-2}}\right)^2 \quad \text{and} \quad \frac{b_{\zeta_n}^4 \left(\sum_{z \in Z_{\zeta_n}} b_z^{-2}\right)^2}{a_{\zeta_n}^4 \sum_{z \in Z_{\zeta_n}} a_z^{-4}} \asymp \zeta_n^D \tag{5.20}$$

Furthermore, if the equations in (5.20) hold modulo logarithmic factors, then the upper and lower bounds match modulo logarithmic factors. This holds almost automatically if b_z decays exponentially or faster, since, then, ζ_n grows logarithmically with n. Noting that the lower bound requires $B_{\zeta_n} \in \Omega\left(\zeta_n^{2D}\right)$, this also holds automatically if $b_z = |z|^t$ with $t \geq D/2$.

Table 5.1 collects the derived minimax rates for various standard choices of a and b. For entries below the diagonal, $b_z/a_z \to \infty$ as $\|z\| \to \infty$, and so $\mathcal{H}_b \not\subseteq \mathcal{H}_a$. As a result, consistent estimation is not possible in the worst case. The diagonal entries of Table 5.1, for which a and b have the same form, are derived in Section 5.7 directly from our upper and lower bounds on $M(a, b)$. These cases exhibit a phase transition, with convergence rates depending on the parameters s and t. When t is sufficiently larger than s, the variance is dominated by the low-order terms of the estimand (5.2), giving a convergence rate of $\asymp n^{-1}$. Otherwise, the variance is dominated by the tail terms of 5.2, in which case minimax rates depend smoothly on s and t. This manifests in the $\max\{n^{-1}, n^{R(s,t)}\}$ form of the minimax rates, where R is non-decreasing in s and non-increasing in t.

Notably, the data dimension D plays a direct role in the minimax rate only in

the Sobolev case when $t < 2s + D/4$. Otherwise, the role of D is captured entirely within the assumption that $p, q \in \mathcal{H}_b$. This is consistent with known rates for estimating other functionals of densities under strong smoothness assumptions such as the RKHS assumption (Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Ramdas, Reddi, Póczos, Singh, and Wasserman, 2015).

Finally, we note some consequences for more general (non-Hilbert) Sobolev spaces $W^{s,p}$, defined for $s \geq 0$, $p \geq 1$ as the set of functions in $\mathcal{L}^p$ having weak s^{th} derivatives in $\mathcal{L}^p$. The most prominent example is that of the Hölder spaces $W^{s,\infty}$ of essentially bounded functions having essentially bounded s^{th} weak derivatives; Hölder spaces are used widely in nonparametric statistics (Bickel and Ritov, 1988; Tsybakov, 2008). Recall that, for $p \leq q$ and any $s \geq 0$, these spaces satisfy the embedding $\mathcal{W}^{s,q} \subseteq \mathcal{W}^{s,p}$ (Villani, 1985), and that $W^{s,2} = \mathcal{H}^s$. Then, for $P, Q \in \mathcal{W}^{t,p}$ our upper bound in Theorem 26 implies an identical upper bound when $p \geq 2$, and our lower bound in Theorem 28 implies an identical lower bound when $p \leq 2$.

Further work is needed to verify tightness of these bounds for $p \neq 2$. Moreover, while this paper focused on the Fourier basis, it is also interesting to consider other bases, which may be informative in other spaces. For example, wavelet bases are more natural representations in a wide range of Besov spaces (Donoho and Johnstone, 1995). It is also of interest to consider non-quadratic functionals as well as non-quadratic function classes. In these cases simple quadratic estimators such as those considered here may not achieve the minimax rate, but it may be possible to correct this with simple procedures such as thresholding, as done, for example, by Cai and Low (2005) in the case of $\mathcal{L}_p$ balls with $p < 2$. Finally, the estimators considered here require some knowledge of the function class in which the true density lies. It is currently unclear whether and how the various strategies for designing adaptive estimators, such as block-thresholding (Cai, 1999) or Lepski's method (Lepski and Spokoiny, 1997), which have been applied to estimate quadratic functionals over $\mathcal{L}_p$ balls and Besov spaces (Efromovich and Low, 1996; Cai and Low, 2006), may confer adaptivity when estimating functionals over general quadratically weighted spaces.

5.9 Proofs

In this section, we present the proofs of main results.

5.9.1 Proof of Proposition 24

We first bound the bias $\left|\mathbb{E}\left[\widehat{S}_Z\right] - \langle P, Q\rangle_a\right|$, where randomness is over the data $X_1, ..., X_n \overset{i.i.d.}{\sim} P$, and $Y_1, ..., Y_n \overset{i.i.d.}{\sim} Q$. Since

$$\widehat{S}_Z = \sum_{z \in Z} \frac{\widehat{\phi}_P(z)\overline{\widehat{\phi}_Q(z)}}{a_z^2}$$

is bilinear in $\widehat{P}$ and $\widehat{Q}$, which are independent, and

$$\widehat{\phi}_P(z) = \frac{1}{n}\sum_{i=1}^{n} \psi_z(X_i) \quad \text{and} \quad \widehat{\phi}_Q(z) = \frac{1}{n}\sum_{i=1}^{n} \psi_z(Y_i)$$

are unbiased estimators of $\phi_P(z) = \mathbb{E}_{X\sim P}[\psi_z(X)]$ and $\phi_Q(z) = \mathbb{E}_{Y\sim Q}[\psi_z(Y)]$, respectively, we have that

$$\mathbb{E}\left[\widehat{S}_Z\right] = \sum_{z\in Z} \frac{\mathbb{E}\left[\widehat{\phi}_P(z)\right]\overline{\mathbb{E}\left[\widehat{\phi}_Q(z)\right]}}{a_z^2} = \sum_{z\in Z} \frac{\phi_P(z)\overline{\phi_Q(z)}}{a_z^2}.$$

Hence, the bias is

$$\left|\mathbb{E}\left[\widehat{S}_Z\right] - \langle P, Q\rangle_a\right| = \left|\sum_{z\in\mathcal{Z}\setminus Z} \frac{\phi_P(z)\overline{\phi_Q(z)}}{a_z^2}\right|.$$

If $p, q \in \mathcal{H}_b \subset \mathcal{H}_a \subseteq \mathcal{L}^2$ defined by

$$\mathcal{H}_b := \left\{f \in \mathcal{L}^2 : \|f\|_b := \sum_{z\in\mathcal{Z}} \frac{\widetilde{f}_z^2}{b_z^2} < \infty\right\},$$

then, applying Cauchy-Schwarz followed by Hölder's inequality, we have

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{S}_Z\right] - \langle P, Q\rangle_a\right| = \left|\sum_{z\in\mathcal{Z}\setminus Z} \frac{\phi_P(z)\overline{\phi_Q(z)}}{a_z^2}\right| &\le \sqrt{\sum_{z\in\mathcal{Z}\setminus Z} \frac{|\phi_P(z)|^2}{a_z^2} \sum_{z\in\mathcal{Z}\setminus Z} \frac{|\phi_Q(z)|^2}{a_z^2}} \\ &= \sqrt{\sum_{z\in\mathcal{Z}\setminus Z} \frac{b_z^2}{a_z^2}\frac{|\phi_P(z)|^2}{b_z^2} \sum_{z\in\mathcal{Z}\setminus Z} \frac{b_z^2}{a_z^2}\frac{|\phi_Q(z)|^2}{b_z^2}} \\ &\le \|P\|_b\|Q\|_b \sup_{z\in\mathcal{Z}\setminus Z} \frac{b_z^2}{a_z^2}.\end{aligned}$$

Note that this recovers the bias bound of Singh, Du, and Póczos (2016) in the Sobolev case: If $a_z = z^{-s}$ and $b_z = z^{-t}$ with $t \ge s$, then

$$\left|\mathbb{E}\left[\widehat{S}_Z\right] - \langle P, Q\rangle_a\right| \le \|P\|_b\|Q\|_b|Z|^{2(s-t)},$$

where $|Z|$ denotes the cardinality of the index set Z.

5.9.2 Proof of Proposition 25

In this section, we bound the variance of $\mathbb{V}[S_Z]$, where, again, randomness is over the data $X_1, ..., X_n, Y_1, ..., Y_n$. The setup and first several steps of our proof are quite general, applying to arbitrary bases. However, without additional assumptions, our approach eventually hits a roadblock. Thus, to help motivate our assumptions and proof approach, we begin by explaining this general setup in Section 5.9.2, and then proceed with steps specific to the Fourier basis in Section 5.9.2.

General Proof Setup

Our bound is based on the Efron-Stein inequality (Efron and Stein, 1981). For this, suppose that we draw extra independent samples $X_1' \sim p$ and $Y_1' \sim q$, and let $\widehat{S}_Z'$ and $\widehat{S}_Z''$ denote the estimator given in Equation (5.12) when we replace X_1 with X_1'

and when we replace Y_1 with Y_1', respectively, that is

$$\widehat{S}_Z' := \sum_{z \in Z} \frac{\widehat{\phi}_P(z)' \overline{\widehat{\phi}_Q(z)}}{a_j^2} \quad \text{and} \quad \widehat{S}_Z'' := \sum_{z \in Z} \frac{\widehat{\phi}_P(z) \overline{\widehat{\phi}_Q(z)'}}{a_j^2},$$

where

$$\widehat{\phi}_P(z)' := \frac{1}{n}\left(\psi_z(X_1') + \sum_{i=2}^{n} \psi_z(X_i)\right), \text{ and } \quad \widehat{\phi}_Q(z)' := \frac{1}{n}\left(\psi_z(Y_1') + \sum_{i=2}^{n} \psi_z(Y_i')\right).$$

Then, since $X_1, ..., X_n$ and $Y_1, ..., Y_n$ are each i.i.d., the Efron-Stein inequality (Efron and Stein, 1981) gives

$$\mathbb{V}\left[\widehat{S}_Z\right] \leq \frac{n}{2}\left(\mathbb{E}\left[\left|\widehat{S}_Z - \widehat{S}_Z'\right|^2\right] + \mathbb{E}\left[\left|\widehat{S}_Z - \widehat{S}_Z''\right|^2\right]\right). \tag{5.21}$$

We now study just the first term as the analysis of the second is essentially identical. Expanding the definitions of $\widehat{S}_Z$ and $\widehat{S}_Z'$, and leveraging the fact that all terms in $\widehat{\phi}_P(z) - \widehat{\phi}_P(z)'$ not containing X_1 or X_1' cancel,

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{S}_Z - \widehat{S}_Z'\right|^2\right] &= \mathbb{E}\left[\left|\sum_{z \in Z} \frac{\left(\widehat{\phi}_P(z) - \widehat{\phi}_P(z)'\right)\overline{\widehat{\phi}_Q(z)}}{a_z^2}\right|^2\right] \\
&= \mathbb{E}\left[\sum_{z \in Z}\sum_{w \in Z}\left(\frac{\left(\widehat{\phi}_P(z) - \widehat{\phi}_P(z)'\right)\overline{\widehat{\phi}_Q(z)}}{a_z^2}\right)\left(\frac{\overline{\left(\widehat{P}_w - \widehat{P}_w'\right)}\widehat{Q}_w}{a_w^2}\right)\right] \\
&= \frac{1}{n^2}\sum_{z \in Z}\sum_{w \in Z}\mathbb{E}\left[\widehat{Q}_w\overline{\widehat{\phi}_Q(z)}\frac{(\psi_z(X_1) - \psi_z(X_1'))\,\overline{(\phi_w(X_1) - \phi_w(X_1'))}}{a_z^2 a_w^2}\right] \\
&= \frac{1}{n^2}\sum_{z \in Z}\sum_{w \in Z}\mathbb{E}\left[\overline{\widehat{\phi}_Q(z)}\widehat{Q}_w\right]\frac{\mathbb{E}\left[(\psi_z(X_1) - \psi_z(X_1'))\,\overline{(\phi_w(X_1) - \phi_w(X_1'))}\right]}{a_z^2 a_w^2} \\
&= \frac{2}{n^2}\sum_{z \in Z}\sum_{w \in Z}\mathbb{E}\left[\overline{\widehat{\phi}_Q(z)}\widehat{Q}_w\right]\frac{\mathbb{E}\left[\psi_z(X)\overline{\phi_w(X)}\right] - \mathbb{E}\left[\psi_z(X)\right]\mathbb{E}\left[\overline{\phi_w(X)}\right]}{a_z^2 a_w^2} \\
&= \frac{2}{n^2}\sum_{z \in Z}\sum_{w \in Z}\mathbb{E}\left[\overline{\widehat{\phi}_Q(z)}\widehat{Q}_w\right]\frac{\mathbb{E}\left[\psi_z(X)\overline{\phi_w(X)}\right] - \phi_P(z)\overline{\widetilde{P}_w}}{a_z^2 a_w^2}.
\end{aligned} \tag{5.22}$$

Expanding the $\mathbb{E}\left[\overline{\widehat{\phi}_Q(z)}\widehat{Q}_w\right]$ term, we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\widehat{\phi}_Q(z)}\widehat{Q}_w\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\overline{\psi_z(Y_i)}\phi_w(Y_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\overline{\psi_z(Y_i)}\phi_w(Y_i)\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j \neq i}\mathbb{E}\left[\overline{\psi_z(Y_i)}\right]\mathbb{E}\left[\phi_w(Y_j)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\overline{\psi_z(Y)}\phi_w(Y)\right] + \frac{n-1}{n}\overline{\phi_Q(z)}\widetilde{Q}_w,
\end{aligned}$$

which combined with Equation (5.22) yields

$$\mathbb{E}\left[\left|\widehat{S}_Z - \widehat{S}'_Z\right|^2\right] = \frac{2}{n^3}\sum_{z\in Z}\sum_{w\in Z}\left(\mathbb{E}\left[\overline{\psi_z(Y)}\phi_w(Y)\right] + (n-1)\overline{\phi_Q(z)}\widetilde{Q}_w\right) \times \frac{\mathbb{E}\left[\psi_z(X)\overline{\phi_w(X)}\right] - \phi_P(z)\overline{\widetilde{P}_w}}{a_z^2 a_w^2}. \tag{5.23}$$

To proceed beyond this point, it is necessary to better understand the covariance-like term $\mathbb{E}\left[\psi_z(X)\overline{\phi_w(X)}\right]$ appearing in the above equation. If $\{\psi_z\}_{z\in\mathcal{Z}}$ is an arbitrary orthonormal basis, it is difficult to argue more than that, via Cauchy-Schwarz,

$$\left|\mathbb{E}\left[\psi_z(X)\overline{\phi_w(X)}\right]\right| \leq \sqrt{|\mathbb{E}\left[|\psi_z(X)|^2\right]\mathbb{E}\left[|\phi_w(X)|^2\right]|}.$$

However, considering, for example, the very well-behaved case when P is the uniform density on $\mathcal{X}$, we would have (since $\{\psi_z\}_{z\in\mathcal{Z}}$ is orthonormal) $\mathbb{E}\left[|\psi_z(X)|^2\right] = \mathbb{E}\left[|\phi_w(X)|^2\right] = \frac{1}{\mu(\mathcal{X})}$, which does not decay as $\|z\|, \|w\| \to \infty$. If we were to follow this approach, the Efron-Stein inequality would eventually give a variance bound on $\widehat{S}_Z$ that includes a term of the form

$$\frac{(n-1)}{n^2\mu(\mathcal{X})}\sum_{z,w\in Z}\frac{\overline{\phi_Q(z)}\widetilde{Q}_w}{a_z^2 a_w^2} = \frac{(n-1)}{n^2\mu(\mathcal{X})}\left(\sum_{z\in Z}\frac{\phi_Q(z)}{a_z^2}\right)^2 \leq \frac{1}{n\mu(\mathcal{X})}\|q\|_b^2\sum_{z\in Z}\frac{b_z^2}{a_z^4}.$$

While relatively general, this bound is loose, at least in the Fourier case. Hence, we proceed along tighter analysis that is specific to the Fourier basis.

Variance Bounds in the Fourier Basis

In the case that $\{\psi_z\}_{z\in\mathcal{Z}}$ is the Fourier basis, the identities $\overline{\psi_z} = \phi_{-z}$ and $\psi_z\phi_w = \psi_{z+w}$ imply that $\mathbb{E}[\psi_z(X)\overline{\phi_w(X)}] = \widetilde{P}_{z-w}$ and $\mathbb{E}[\phi_w(Y)\overline{\psi_z(Y)}] = \widetilde{Q}_{w-z}$, thus, the expression (5.23) simplifies to

$$\mathbb{E}\left[\left|\widehat{S}_Z - \widehat{S}'_Z\right|^2\right] = \frac{2}{n^3}\sum_{z\in Z}\sum_{w\in Z}\left(\widetilde{Q}_{w-z} + (n-1)\widetilde{Q}_{-z}\widetilde{Q}_w\right)\frac{\widetilde{P}_{z-w} - \phi_P(z)\widetilde{P}_{-w}}{a_z^2 a_w^2}$$
$$= \frac{2}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\widetilde{Q}_{w-z}\widetilde{P}_{z-w} - \widetilde{Q}_{w-z}\phi_P(z)\widetilde{P}_{-w} + (n-1)\widetilde{Q}_{-z}\widetilde{Q}_w\widetilde{P}_{z-w} - (n-1)\widetilde{Q}_{-z}\widetilde{Q}_w\phi_P(z)\widetilde{P}_{-w}}{a_z^2 a_w^2}.$$

This contains four terms to bound, but they are dominated by the following three main terms:

$$\frac{2}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\widetilde{Q}_{w-z}\widetilde{P}_{z-w}}{a_z^2 a_w^2}, \tag{5.24}$$

$$\frac{2(n-1)}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\overline{\phi_Q(z)}\widetilde{Q}_w\widetilde{P}_{z-w}}{a_z^2 a_w^2}, \tag{5.25}$$

and

$$\frac{2(n-1)}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\overline{\phi_Q(z)}\widetilde{Q}_w\phi_P(z)\overline{\widetilde{P}_w}}{a_z^2 a_w^2}. \tag{5.26}$$

Bounding (5.24)**:** Applying the change of variables $k = z - w$ gives

$$\begin{aligned}\left|\frac{2}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\widetilde{P}_{z-w}\widetilde{Q}_{w-z}}{a_z^2 a_w^2}\right| &= \frac{2}{n^3}\left|\sum_{k\in z-Z}\widetilde{P}_k\widetilde{Q}_{-k}\sum_{z\in Z}\frac{1}{a_z^2 a_{z-k}^2}\right| \overset{(*)}{\leq} \frac{2}{n^3}\sum_{k\in z-Z}\left|\widetilde{P}_k\widetilde{Q}_{-k}\right|\sum_{z\in Z}\frac{1}{a_z^4} \\ &\leq \frac{2}{n^3}\sum_{k\in\mathcal{Z}}\left|\widetilde{P}_k\widetilde{Q}_{-k}\right|\sum_{z\in Z}\frac{1}{a_z^4} \leq \frac{2\|P\|_2\|Q\|_2}{n^3}\sum_{z\in Z}\frac{1}{a_z^4}, \end{aligned} \tag{5.27}$$

where in $(*)$, we use the fact that

$$f_Z(k) := \sum_{z\in Z}\frac{1}{a_z^2 a_{z-k}^2}$$

is the convolution (over $\mathbb{Z}$) of $\{a_z^{-2}\}_{z\in Z}$ with itself, which is always maximized when $k = 0$.

Bounding (5.25)**:** Applying Cauchy-Schwarz inequality twice, yields

$$\begin{aligned}\left|\sum_{z\in Z}\sum_{w\in Z}\frac{\overline{\phi_Q(z)}\widetilde{Q}_w\widetilde{P}_{z-w}}{a_z^2 a_w^2}\right| &= \left|\sum_{z\in Z}\frac{\overline{\phi_Q(z)}}{b_z}\sum_{w\in Z}\frac{b_z}{a_z^2}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right| \\ &\leq \|Q\|_b\left(\sum_{z\in Z}\left(\sum_{w\in Z}\frac{b_z}{a_z^2}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right)^2\right)^{1/2} \\ &= \|Q\|_b\left(\sum_{z\in Z}\frac{b_z^2}{a_z^4}\left(\sum_{w\in Z}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right)^2\right)^{1/2} \\ &\leq \|Q\|_b\left(\sum_{z\in Z}\frac{b_z^4}{a_z^8}\right)^{1/4}\left(\sum_{z\in Z}\left(\sum_{w\in Z}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right)^4\right)^{1/4}. \end{aligned} \tag{5.28}$$

Note that now we can view the expression

$$\left(\sum_{z\in Z}\left(\sum_{w\in Z}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right)^4\right)^{1/4} = \left\|\frac{\widetilde{Q}}{a^2} * \widetilde{P}\right\|_4$$

as the $\mathcal{L}_4$ norm of the convolution between the sequence $\widetilde{Q}/a^2$ and the sequence $\widetilde{P}$. To proceed, we apply (a discrete variant of) Young's inequality for convolutions (Beckner, 1975), which states that, for constants $\alpha, \beta, \gamma \geq 1$ satisfying $1 + 1/\gamma = 1/\alpha + 1/\beta$ and arbitrary functions $f \in \mathcal{L}^\alpha(\mathbb{R}^D), g \in \mathcal{L}^\beta(\mathbb{R}^D)$,

$$\|f * g\|_\gamma \leq \|f\|_\alpha\|g\|_\beta.$$

Applying Young's inequality for convolutions with powers[7] $\alpha = \beta = 8/5$ (so that $\alpha, \beta \geq 1$ and $1/\alpha + 1/\beta = 1 + 1/4$), gives

$$\left(\sum_{z\in Z}\left(\sum_{w\in Z}\frac{\widetilde{Q}_w\widetilde{P}_{z-w}}{a_w^2}\right)^4\right)^{1/4} \leq \left(\sum_{z\in Z}\frac{\phi_Q(z)^\alpha}{a_z^{2\alpha}}\right)^{1/\alpha}\left(\sum_{z\in Z}\phi_P(z)^\beta\right)^{1/\beta}$$
$$= \left(\sum_{z\in Z}\frac{\phi_Q(z)^\alpha}{b_z^\alpha}\frac{b_z^\alpha}{a_z^{2\alpha}}\right)^{1/\alpha}\left(\sum_{z\in Z}\frac{\phi_P(z)^\beta}{b_z^\beta}b_z^\beta\right)^{1/\beta}.$$

Since $2/\alpha = 2/\beta \geq 1$, we can now apply Hölder's inequality to each of the above summations, with powers $\left(2/\alpha, \frac{2\alpha}{2-\alpha}\right) = \left(2/\beta, \frac{2\beta}{2-\beta}\right) = \left(\frac{5}{4}, \frac{1}{8}\right)$. This gives

$$\left(\sum_{z\in Z}\frac{\phi_Q(z)^\alpha}{b_z^\alpha}\frac{b_z^\alpha}{a_z^{2\alpha}}\right)^{1/\alpha} \leq \|Q\|_b\left(\sum_{z\in Z}\left(\frac{b_z}{a_z^2}\right)^{\frac{2\alpha}{2-\alpha}}\right)^{\frac{2-\alpha}{2\alpha}} = \|Q\|_b\left(\sum_{z\in Z}\left(\frac{b_z}{a_z^2}\right)^8\right)^{1/8}$$

and

$$\left(\sum_{z\in Z}\frac{\phi_P(z)^\beta}{b_z^\beta}b_z^\beta\right)^{1/\beta} \leq \|P\|_b\left(\sum_{z\in Z}b_z^{\frac{2\beta}{2-\beta}}\right)^{\frac{2-\beta}{2\beta}} = \|P\|_b\left(\sum_{z\in Z}b_z^8\right)^{1/8}.$$

Combining these inequalities with inequality (5.28) gives

$$\left|\sum_{z\in Z}\sum_{w\in Z}\frac{\overline{\phi_Q(z)}\widetilde{Q}_w\widetilde{P}_{z-w}}{a_z^2a_w^2}\right| \leq \|Q\|_b^2\|P\|_b R_{a,b,Z},$$

where $R_{a,b,Z}$ is as in (5.16).

Bounding (5.26)**:** Applying Cauchy-Schwarz yields

$$\frac{2(n-1)}{n^3}\sum_{z\in Z}\sum_{w\in Z}\frac{\widetilde{Q}_w\overline{\phi_Q(z)}\phi_P(z)\overline{\widetilde{P}_w}}{a_z^2a_w^2} = \frac{2(n-1)}{n^3}\left(\sum_{z\in Z}\frac{\overline{\phi_Q(z)}\phi_P(z)}{a_z^2}\right)\left(\sum_{w\in Z}\frac{\widetilde{Q}_w\overline{\widetilde{P}_w}}{a_w^2}\right)$$
$$\leq \frac{2}{n^2}\left(\sum_{z\in Z}\frac{|\phi_Q(z)|^2}{a_z^2}\right)\left(\sum_{z\in Z}\frac{|\phi_P(z)|^2}{a_z^2}\right) = \frac{2\|P\|_a^2\|Q\|_a^2}{n^2}.$$

Plugging these into Efron-Stein yields the result.

5.9.3 Proof of Theorem 28

Proof: The $\Omega(n^{-1})$ term of the lower bound, reflecting parametric convergence when the tails of the estimand (5.2) are light relative to the first few terms, follows from classic information bounds (Bickel and Ritov, 1988). We focus on deriving the $\Omega(A_\zeta^2/B_\zeta^2)$ term, reflecting slower convergence when the estimand is dominated by its tail. To do this, we consider the uniform density ψ_0 and a family of $2^{|Z_\zeta|}$ small perturbations of the form

$$g_{\zeta,\tau} = \psi_0 + c_\zeta\sum_{z\in Z_\zeta}\tau_z\psi_z, \tag{5.29}$$

where $\zeta \in \mathbb{N}$, $\tau \in \{-1, 1\}^{Z_\zeta}$, and $c_\zeta = B_\zeta^{-1/2}$.

[7] This seemingly arbitrary choice of α and β arises from analytically minimizing the final bound.

We now separately consider the "smooth" case, in which $B_\zeta \in \Omega\left(\zeta^{2D}\right)$, and the "unsmooth" case, in which $B_\zeta \in o\left(\zeta^{2D}\right)$.

The smooth case ($B_\zeta \in \Omega\left(\zeta^{2D}\right)$): By Le Cam's Lemma (see, e.g., Section 2.3 of Tsybakov (2008)), it suffices to prove four main claims about the family of $g_{\zeta,\tau}$ functions defined in Equation (5.29):

1. Each $\|g_{\zeta,\tau}\|_b \leq 1$.
2. Each
$$\inf_{\tau\in\{-1,1\}^{Z_\zeta}} \|g_{\zeta,\tau}\|_a - \|\psi_0\|_a \geq \frac{A_\zeta}{B_\zeta}.$$
3. Each $g_{\zeta,\tau}$ is a density function (i.e., $\int_{\mathcal{X}} g_{\zeta,\tau} = 1$ and $g_{\zeta,\tau} \geq 0$).
4. ζ and c_ζ are chosen (depending on n) such that
$$D_{TV}\left(\psi_0^n, \frac{1}{2^{|Z_\zeta|}} \sum_{\tau\in\{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n\right) \leq \frac{1}{2}.$$

For simplicity, for now, suppose $\mathcal{Z} = \mathbb{N}^D$ and $Z_\zeta = [\zeta]^D$. For any $\tau \in \{-1,1\}^{Z_\zeta}$, let
$$g_{\zeta,\tau} = \psi_0 + c_\zeta \sum_{z\in Z_\zeta} \tau_z \psi_z.$$
By setting $c_\zeta = B_\zeta^{-1/2} = \left(\sum_{z\in Z_\zeta} b_z^{-2}\right)^{-1/2}$, we automatically ensure the first two claims:
$$\|g_{\zeta,\tau}\|_b^2 = c_\zeta^2 \sum_{z\in Z_\zeta} b_z^{-2} = 1,$$
and
$$\inf_{\tau\in\{-1,1\}^{Z_\zeta}} \|g_{\zeta,\tau}\|_a^2 - \|\psi_0\|_a^2 = c_\zeta^2 \sum_{z\in Z_\zeta} a_z^{-2} = \frac{A_\zeta}{B_\zeta}.$$
To verify that each $g_{\zeta,\tau}$ is a density, we first note that, since, for $z \neq 0$, $\int_{\mathcal{X}} \psi_z = 0$, and so
$$\int_{\mathcal{X}} g_{\zeta,\tau} = \int_{\mathcal{X}} \psi_0 = 1.$$
Also, since ψ_0 is constant and strictly positive and the supremum is taken over all $\tau \in \{-1,1\}^{Z_\zeta}$, the condition that all $g_{\zeta,\tau} \geq 0$ is equivalent to
$$\sup_{\tau\in\{-1,1\}^{Z_\zeta}} B_\zeta^{-1/2} \left\|\sum_{z\in Z_\zeta} \tau_z \psi_z\right\|_\infty = \sup_{\tau\in\{-1,1\}^{Z_\zeta}} \|g_{\zeta,\tau} - \psi_0\|_\infty \leq \|\psi_0\|_\infty.$$
For the Fourier basis, each $\|\psi_z\|_\infty = 1$,[8] and so
$$\sup_{\tau\in\{-1,1\}^{Z_\zeta}} \left\|\sum_{z\in Z_\zeta} \tau_z \psi_z\right\|_\infty \asymp \sup_{\tau\in\{-1,1\}^{Z_\zeta}} \sum_{z\in Z_\zeta} \|\psi_z\|_\infty \asymp |Z_\zeta| = \zeta^D.$$
Thus, we precisely need $B_\zeta \in \Omega(\zeta^{2D})$, and it is sufficient, for example, that $b_z \in O(\|z\|^{-D/2})$.

[8] This is the only step in the proof that uses any properties specific to the Fourier basis.

Finally, we show that

$$D_{\text{TV}}\left(\psi_0^n, \frac{1}{2^{|Z_\zeta|}} \sum_{\tau \in \{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n\right) \leq \frac{1}{2} \tag{5.30}$$

(where $h^n : \mathcal{X}^n \to [0,\infty)$ denotes the joint likelihood of n IID samples). For any particular $\tau \in \{-1,1\}^{Z_\zeta}$ and $x_1, ..., x_n \in \mathcal{X}$, the joint likelihood is

$$\begin{aligned} g_{\zeta,\tau}^n(x_1,...,x_n) &= \prod_{i=1}^n \left(1 + c_\zeta \sum_{z \in Z_\zeta} \tau_z \psi_z(x_i)\right) \\ &= 1 + \sum_{\ell=1}^n \sum_{\substack{i_1,...,i_\ell \in [n] \\ \text{distinct}}} \sum_{j_1,...,j_\ell \in Z_\zeta} \prod_{k=1}^\ell c_\zeta \tau_{j_k} \phi_{j_k}(x_{i_k}) \\ &= 1 + \sum_{\ell=1}^n c_\zeta^\ell \sum_{\substack{i_1,...,i_\ell \in [n] \\ \text{distinct}}} \sum_{z_1,...,z_\ell \in Z_\zeta} \prod_{k=1}^\ell \tau_{z_k} \psi_{z_k}(x_{i_k}). \end{aligned}$$

Thus, the likelihood of the uniform mixture over $\tau \in \{-1,1\}^{Z_\zeta}$ is

$$\begin{aligned} &\frac{1}{2^{|Z_\zeta|}} \sum_{\tau \in \{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n(x_1,...,x_n) \\ &= 1 + \frac{1}{2^{|Z_\zeta|}} \sum_{\tau \in \{-1,1\}^{Z_\zeta}} \sum_{\ell=1}^n c_\zeta^\ell \sum_{\substack{i_1,...,i_\ell \in [n] \\ \text{distinct}}} \sum_{z_1,...,z_\ell \in Z_\zeta} \prod_{k=1}^\ell \tau_{z_k} \psi_{z_k}(x_{i_k}) \\ &= 1 + \sum_{\ell=1}^{\lfloor n/2 \rfloor} c_\zeta 2\ell \sum_{\substack{i_1,...,i_{2\ell} \in [n] \\ \text{distinct}}} \sum_{z_1,...,z_\ell \in Z_\zeta} \prod_{k=1}^\ell \psi_{z_k}(x_{i_{2k-1}}) \psi_{z_k}(x_{i_{2k}}), \end{aligned}$$

where $\lfloor a \rfloor$ denotes the largest integer at most $a \in [0,\infty)$. This equality holds because, within the sum over $\tau \in \{-1,1\}^{Z_\zeta}$, any term in which any τ_z appears an odd number of times will cancel. The remaining terms each appear $2^{|Z_\zeta|}$ times. Thus, the total variation distance is

$$\begin{aligned} D_{\text{TV}}\left(\psi_0^n, \frac{1}{2^{|Z_\zeta|}} \sum_{\tau \in \{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n\right) &= \frac{1}{2} \left\| \psi_0^n - \frac{1}{2^{|Z_\zeta|}} \sum_{\tau \in \{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n \right\|_1 \\ &= \frac{1}{2} \int_{\mathcal{X}^n} \left| \sum_{\ell=1}^{\lfloor n/2 \rfloor} c_\zeta^{2\ell} \sum_{\substack{i_1,...,i_{2\ell} \in [n] \\ \text{distinct}}} \sum_{z_1,...,z_\ell \in Z_\zeta} \prod_{k=1}^\ell \psi_{z_k}(x_{i_{2k-1}}) \psi_{z_k}(x_{i_{2k}}) \right| d(x_1,...,x_n) \\ &\leq \frac{1}{2} \sum_{\ell=1}^{\lfloor n/2 \rfloor} c_\zeta^{2\ell} \int_{\mathcal{X}^n} \left| \sum_{\substack{i_1,...,i_{2\ell} \in [n] \\ \text{distinct}}} \sum_{z_1,...,z_\ell \in Z_\zeta} \prod_{k=1}^\ell \psi_{z_k}(x_{i_{2k-1}}) \psi_{z_k}(x_{2k}) \right| d(x_1,...,x_n), \end{aligned} \tag{5.31}$$

where we used the triangle inequality. By Jensen's inequality (since $\mathcal{X} = [0,1]$),

$$\int_{\mathcal{X}^n} \left| \sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ \text{distinct}}} \sum_{z_1,\dots,z_\ell \in Z_\zeta} \prod_{k=1}^{\ell} \psi_{z_k}(x_{i_{2k-1}})\psi_{z_k}(x_{i_{2k}}) \right| d(x_1,\dots,x_n)$$

$$\leq \sqrt{\int_{\mathcal{X}^n} \left(\sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ \text{distinct}}} \sum_{z_1,\dots,z_\ell \in Z_\zeta} \prod_{k=1}^{\ell} \psi_{z_k}(x_{i_{2k-1}})\psi_{z_k}(x_{i_{2k}}) \right)^2 d(x_1,\dots,x_n)}. \tag{5.32}$$

Since $\{\psi_z\}_{z\in\mathcal{Z}}$ is an orthogonal system in $\mathcal{L}^2(\mathcal{X})$, we can pull the summations outside the square, so

$$\begin{aligned}
&\int_{\mathcal{X}^n} \left(\sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ \text{distinct}}} \sum_{z_1,\dots,z_\ell \in Z_\zeta} \prod_{k=1}^{\ell} \psi_{z_k}(x_{i_{2k-1}})\psi_{z_k}(x_{i_{2k}}) \right)^2 d(x_1,\dots,x_n) \\
&= \sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ \text{distinct}}} \sum_{z_1,\dots,z_\ell \in Z_\zeta} \int_{\mathcal{X}^n} \left(\prod_{k=1}^{\ell} \psi_{z_k}(x_{i_{2k-1}})\psi_{z_k}(x_{i_{2k}}) \right)^2 d(x_1,\dots,x_n) \\
&= \sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ \text{distinct}}} \sum_{z_1,\dots,z_\ell \in Z_\zeta} 1 = \binom{n}{2\ell} \zeta^{D\ell} \leq \frac{n^{2\ell}\zeta^{D\ell}}{(\ell!)^2},
\end{aligned}$$

since

$$\binom{n}{2\ell} = \frac{n!}{(2\ell)!(n-2k)!} \leq \frac{n^{2\ell}}{(2\ell)!} \leq \frac{n^{2\ell}}{(\ell!)^2}.$$

Combining this with inequalities (5.31) and (5.32) gives

$$D_{\text{TV}}\left(\psi_0^n, \frac{1}{2^{|Z_\zeta|}} \sum_{\tau\in\{-1,1\}^{Z_\zeta}} g_{\zeta,\tau}^n \right) \leq \frac{1}{2} \sum_{\ell=1}^{\lfloor n/2 \rfloor} \frac{\left(nc_\zeta^2\zeta^{D/2}\right)^\ell}{\ell!} \leq \exp\left(nc_\zeta^2\zeta^{D/2}\right) - 1, \tag{5.33}$$

where we used the fact that the exponential function is greater than any of its Taylor approximations on $[0,\infty)$. The last expression in inequality (5.33) vanishes if $nc_\zeta^2\zeta^{D/2} \to 0$. Recalling now that we set $c_\zeta = B_\zeta^{-1/2}$, for some constant $C > 0$, the desired bound (5.30) holds by choosing ζ satisfying

$$\frac{\zeta^D}{B_\zeta^2} = \zeta^D c_\zeta^4 \leq Cn^{-2}.$$

The unsmooth case ($B_\zeta \in o\left(\zeta^{2D}\right)$): Finally, we consider the 'highly unsmooth' case, when $B_\zeta \in o(\zeta^{2D})$ In this case, we must modify the above proof to ensure that the $g_{\zeta,\tau}$ functions are all non-negative. In the Fourier case, we again wish to ensure

$$c_\zeta\zeta^D = c_\zeta|Z_\zeta| \asymp c_\zeta \sup_{\tau\in\{-1,1\}^{Z_\zeta}} \left\| \sum_{z\in Z_\zeta} \tau_z\psi_z \right\|_\infty \leq 1,$$

but this is no longer guaranteed by setting $c_\zeta = B_\zeta^{-1/2}$; instead, we use the smaller value $c_\zeta = \zeta^{-D}$. Clearly, we still have $\|g_{\zeta,\tau}\|_b^2 \leq 1$. Now, however, we have a smaller estimation error

$$\inf_{\tau \in \{-1,1\}^{Z_\zeta}} \|g_{\zeta,\tau}\|_a^2 - \|\psi_0\|_a^2 = c_\zeta^2 \sum_{z \in Z_\zeta} a_z^{-2} = \frac{A_\zeta}{\zeta^{2D}}. \tag{5.34}$$

Also, the information bound (5.33) now vanishes when $n\zeta^{-3D/2} = nc_\zeta^2 \zeta^{D/2} \to 0$, so that, for some constant $C > 0$, the desired bound (5.30) holds by choosing ζ satisfying

$$\zeta \leq C n^{2/(3D)}.$$

Plugging this into equation (5.34) gives

$$\inf_{\tau \in \{-1,1\}^{Z_\zeta}} \|g_{\zeta,\tau}\|_a^2 - \|\psi_0\|_a^2 \asymp \frac{A_{n^{2/(3D)}}}{n^{4/3}}.$$

Finally, by Le Cam's lemma, this implies the minimax rate

$$\inf_{\widehat{S}} \sup_{p,q \in \mathcal{H}_b} \mathbb{E}\left[\left(\widehat{S} - \langle p, q \rangle\right)^2\right] \geq \left(\frac{A_{n^{2/(3D)}}}{n^{4/3}}\right)^2.$$

■

Chapter 6

Wasserstein Convergence of the Empirical Measure

6.1 Introduction

The Wasserstein metric is an important measure of distance between probability distributions, with applications in machine learning, statistics, probability theory, and data analysis. This chapter provides upper and lower bounds on statistical minimax rates for the problem of estimating a probability distribution under Wasserstein loss, using only metric properties, such as covering and packing numbers, of the sample space, and weak moment assumptions on the probability distributions.

This chapter is a departure from the earlier work in this thesis in several ways. The most obvious difference is that we are interested in the conventional nonparametric problem of estimating an entire probability distribution, rather than simply a real-valued functional thereof. Secondly, whereas most of the earlier work has assumed the generating distribution is supported on a compact subset of $\mathbb{R}^d$, in this chapter, we dramatically generalize the sample space to an essentially arbitrary metric space. Finally, we drop any assumption of absolute continuity of the data distribution; instead, we will assume only that the distribution has some number of finite moments.

To motivate this final consideration, note that, in general metric spaces, even assuming the existence of a base measure may be quite restrictive. Specifically, we are most interested in situations where the distribution is supported on an (unknown) subspace of much lower "intrinsic" dimension than the known ambient space. Defining reasonable probability densities in such situations can become quite difficult and requires assuming additional well-behaved structure (e.g., a manifold). We take a simpler and more general approach, in which we discuss only probability measures (potentially without densities) and measure complexity of the sample space Ω in terms of *covering* and *packing numbers*, which are easy to define if Ω is any totally bounded metric space. When Ω is not totally bounded, we assume that it can be partitioned into a countable union of totally bounded sets (e.g., spherical shells in $\mathbb{R}^d$); by analogy to the notion of σ-finiteness in measure theory, we might call this very mild property σ-boundedness.

6.2 Background

The Wasserstein metric is an important measure of distance between probability distributions, based on the cost of transforming either distribution into the other through mass transport, under a base metric on the sample space. Originating

in the optimal transport literature,[1] the Wasserstein metric has, owing to its intuitive and general nature, been utilized in such diverse areas as probability theory and statistics, economics, image processing, text mining, robust optimization, and physics (Villani, 2008; Fournier and Guillin, 2015; Esfahani and Kuhn, 2018; Gao and Kleywegt, 2016).

In the analysis of image data, the Wasserstein metric has been used for various tasks such as texture classification and face recognition (Sandler and Lindenbaum, 2011), reflectance interpolation, color transfer, and geometry processing (Solomon, De Goes, Peyré, Cuturi, Butscher, Nguyen, Du, and Guibas, 2015), image retrieval (Rubner, Tomasi, and Guibas, 2000), and image segmentation (Ni, Bresson, Chan, and Esedoglu, 2009), and, in the analysis of text data, for tasks such as document classification (Kusner, Sun, Kolkin, and Weinberger, 2015) and machine translation (Zhang, Liu, Luan, Sun, Izuha, and Hao, 2016).

In contrast to a number of other popular notions of dissimilarity between probability distributions, such as $\mathcal{L}_p$ distances or Kullback-Leibler and other f-divergences (Morimoto, 1963; Csiszár, 1964; Ali and Silvey, 1966), which require distributions to be absolutely continuous with respect to each other or to a base measure, Wasserstein distance is well-defined between *any* pair of probability distributions over a sample space equipped with a metric.[2] As a particularly important consequence, Wasserstein distances between discrete (e.g., empirical) distributions and continuous distributions are well-defined, finite, and informative (e.g., can decay to 0 as the distributions become more similar).

Partly for this reason, many central limit theorems and related approximation results (Rüschendorf, 1985; Johnson and Samworth, 2005; Chatterjee, 2008; Rio, 2009; Rio, 2011; Chen, Goldstein, and Shao, 2010; Reitzner and Schulte, 2013) are expressed using Wasserstein distances. Within machine learning and statistics, this same property motivates a class of so-called *minimum Wasserstein distance estimates* (Barrio, Giné, and Matrán, 1999; Barrio, Giné, and Matrán, 2003; Bassetti, Bodini, and Regazzini, 2006; Bernton, Jacob, Gerber, and Robert, 2017) of distributions, ranging from exponential distributions (Baíllo, Cárcamo, and Getman, 2016) to more exotic models such as restricted Boltzmann machines (RBMs) (Montavon, Müller, and Cuturi, 2016) and generative adversarial networks (GANs) (Arjovsky, Chintala, and Bottou, 2017). This class of estimators also includes k-means and k-medians, where the hypothesis class is taken to be discrete distributions supported on at most k points (Pollard, 1982); more flexible algorithms such as hierarchical k-means (Ho, Nguyen, Yurochkin, Bui, Huynh, and Phung, 2017) and k-flats (Tseng, 2000) can also be expressed in this way, using a more elaborate hypothesis classes. PCA can also be expressed and generalized to manifolds using Wasserstein distance minimization (Boissard, Le Gouic, and Loubes, 2015). These estimators are conceptually equivalent to empirical risk minimization, leveraging the fact that Wasserstein distances between the empirical distribution and distributions in the relevant hypothesis class are well-behaved. Moreover, these estimates often perform well in practice because they are free of both tuning parameters and strong distributional assumptions.

For many of the above applications, it is important to understand how quickly the empirical distribution converges to the true distribution in Wasserstein distance,

[1] The Wasserstein metric has been variously attributed to Monge, Kantorovich, Rubinstein, Gini, Mallows, and others; see Chapter 3 of Villani (2008) for detailed history.

[2] Hence, we use "distribution estimation" in this section, rather than the more common "density estimation".

and whether there exist distribution estimators that converge more quickly. For example, Canas and Rosasco (2012) use bounds on Wasserstein convergence to prove learning bounds for k-means, while Arora, Ge, Liang, Ma, and Zhang (2017) used the slow rate of convergence in Wasserstein distance in certain cases to argue that GANs based on Wasserstein distances fail to generalize with fewer than exponentially many samples in the dimension.

To this end, the **main contribution** of this section is to identify, in a wide variety of settings, the minimax convergence rate for the problem of estimating a distribution using Wasserstein distance as a loss function. Our setting is very general, relying only on metric properties of the support of the distribution and the number of finite moments the distribution has; some diverse examples to which our results apply are given in Section 6.5. Specifically, we assume only that the distribution is has some number of finite moments in a given metric. We then prove bounds on the minimax convergence rates of distribution estimation, utilizing covering numbers of the sample space for upper bounds and packing numbers for lower bounds. It may at first be surprising that positive results can be obtained under such mild assumptions; this highlights that the Wasserstein metric is quite a weak metric (see our Lemma 42 and the subsequent remark for discussion of this). Moreover, our results imply that, without further assumptions on the population distribution, the empirical distribution is typically minimax rate-optimal. Note that, while there has been previous work on upper bounds (discussed in Section 6.4), this section is the first to study minimax lower bounds for this problem.

Organization: The remainder of this section is organized as follows. Section 6.3 provides notation required to formally state both the problem of interest and our results, while Section 6.4 reviews previous work studying convergence of distributions in Wasserstein distance. Sections 6.4.1 and 6.4.2 respectively contain our main upper and lower bound results. Since the proofs of the upper bounds, are fairly long, Sections 6.7 and 6.8 provide high-level sketches of the proofs, followed by detailed proofs in Section 6.9. The lower bound is proven in Section 6.10. Finally, in Section 6.5, we apply our upper and lower bounds to identify minimax convergence rates in a number of concrete examples. Section 6.6 concludes with a summary of our contributions and suggested avenues for future work.

6.3 Notation and Problem Setting

For any positive integer $n \in \mathbb{N}$, $[n] = \{1, 2, ..., n\}$ denotes the set of the first n positive integers. For sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ of non-negative reals, $a_n \lesssim b_n$ and, equivalently $b_n \gtrsim a_n$, indicate the existence of a constant $C > 0$ such that $\limsup_{n\to\infty} \frac{a_n}{b_n} \leq C$. $a_n \asymp b_n$ indicates $a_n \lesssim b_n \lesssim a_n$.

6.3.1 Problem Setting

For the remainder of this section, fix a metric space (Ω, ρ), over which Σ denotes the Borel σ-algebra, and let $\mathcal{P}$ denote the family of all Borel probability distributions on Ω. The main object of study in this section is the Wasserstein distance on $\mathcal{P}$, defined as follows:

Definition 32 (r-Wasserstein Distance)**.** Given two Borel probability distributions P and Q over Ω and $r \in [1, \infty)$, the *r-Wasserstein distance* $W_r(P, Q) \in [0, \infty]$ between P

and Q is defined by

$$W_r(P,Q) := \inf_{\mu \in \Pi(P,Q)} \left(\mathop{\mathbb{E}}_{(X,Y)\sim\mu} \left[\rho^r (X,Y) \right] \right)^{1/r},$$

where $\Pi(P,Q)$ denotes all couplings between $X \sim P$ and $Y \sim Q$; that is,

$$\Pi(P,Q) := \left\{ \mu : \Sigma^2 \to [0,1] \middle| \text{ for all } A \in \Sigma, \mu(A \times \Omega) = P(A) \text{ and } \mu(\Omega \times A) = Q(A) \right\},$$

is the set of joint probability measures over $\Omega \times \Omega$ with marginals P and Q.

Intuitively, $W_r(P,Q)$ quantifies the r-weighted total cost of transforming mass distributed according to P to be distributed according to Q, where the cost of moving a unit mass from $x \in \Omega$ to $y \in \Omega$ is $\rho(x,y)$. $W_r(P,Q)$ is sometimes defined in terms of equivalent (e.g., dual) formulations; these formulations will not be needed here, although they will become central in Chapter 7. W_r it is symmetric in its arguments and satisfies the triangle inequality (Clement and Desch, 2008), and, for all $P \in \mathcal{P}$, $W_r(P,P) = 0$. Thus, W_r is always a pseudometric. Moreover, it is a proper metric (i.e., $W_r(P,Q) = 0 \Rightarrow P = Q$) if and only if ρ is as well.

This section studies the following problem:

Formal Problem Statement: Suppose (Ω, ρ) is a known metric space. Suppose P is an unknown Borel probability distribution on Ω, from which we observe n IID samples $X_1, ..., X_n \overset{IID}{\sim} P$. We are interested in studying the minimax rates at which P can be estimated from $X_1, ..., X_n$, in terms of the (r^{th} power of the) r-Wasserstein loss. Specifically, we are interested in deriving finite-sample upper and lower bounds, in terms of only properties of the space (Ω, ρ), on the quantity

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_1,...,X_n \overset{IID}{\sim} P} \left[W_r^r \left(P, \widehat{P}(X_1, ..., X_n) \right) \right], \tag{6.1}$$

where the infimum is taken over all estimators $\widehat{P}$ (i.e., (potentially randomized) functions $\widehat{P} : \Omega^n \to \mathcal{P}$ of the data). In the sequel, we suppress the dependence of $\widehat{P} = \widehat{P}(X_1, ..., X_n)$ in the notation.

In particular, our upper bounds on (6.1) will utilize, as the distribution estimator, the simple empirical distribution

$$\widehat{P} := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes a Dirac delta mass at x. Our results will imply, therefore, that the empirical distribution is rate-optimal, under the types of assumptions (described in the next section) that we consider.

6.3.2 Definitions for Stating our Results

Here, we give notation and definitions needed to state our main results in Sections 6.4.1 and 6.4.2.

Let 2^Ω denote the power set of Ω. Let $\mathbb{S} \subseteq 2^{2^\Omega}$ denote the family of all Borel partitions of Ω:

$$\mathbb{S} := \left\{ \mathcal{S} \subseteq \Sigma \quad : \quad \Omega \subseteq \bigcup_{S \in \mathcal{S}} S \quad \text{and} \quad \forall S, T \in \mathcal{S}, S \cap T = \emptyset \right\}.$$

We now define some metric notions that will later be useful for bounding Wasserstein distances:

Definition 33 (Diameter and Separation of a Set, Resolution of a Partition)**.** For any set $S \subseteq \Omega$, the *diameter* $\mathrm{Diam}(S)$ *of* S is defined by $\mathrm{Diam}(S) := \sup_{x,y \in S} \rho(x, y)$, and the *separation* $\mathrm{Sep}(S)$ *of* S is defined by $\mathrm{Sep}(S) := \inf_{x \neq y \in S} \rho(x, y)$. If $\mathcal{S} \in \mathbb{S}$ is a partition of Ω, then the *resolution* $\mathrm{Res}(\mathcal{S})$ *of* $\mathcal{S}$ defined by $\mathrm{Res}(\mathcal{S}) := \sup_{S \in \mathcal{S}} \mathrm{Diam}(S)$ is the largest diameter of any set in $\mathcal{S}$.

We now define the covering and packing number of a metric space, which are classic and widely used measures of the size or complexity of a metric space (Dudley, 1967; Haussler, 1995; Zhou, 2002; Zhang, 2002). Our main convergence results will be stated in terms of these quantities, as well as the packing radius, which acts, approximately, as the inverse of the packing number.

Definition 34 (Covering Number, Packing Number, and Packing Radius of a Metric Space)**.** The *covering number* $N : (0, \infty) \to \mathbb{N}$ *of* (Ω, ρ) is defined for all $\epsilon > 0$ by

$$N(\epsilon) := \min \{ |\mathcal{S}| : \mathcal{S} \in \mathbb{S} \quad \text{and} \quad \mathrm{Res}(\mathcal{S}) \leq \epsilon \}.$$

The *packing number* $M : (0, \infty) \to \mathbb{N}$ *of* (Ω, ρ) is defined for all $\epsilon > 0$ by

$$M(\epsilon) := \max \{ |S| : S \subseteq \Omega \quad \text{and} \quad \mathrm{Sep}(S) \geq \epsilon \}.$$

Finally, the *packing radius* $R : \mathbb{N} \to [0, \infty]$ is defined for all $n \in \mathbb{N}$ by

$$R(n) := \sup \{ \mathrm{Sep}(S) : S \subseteq \Omega \quad \text{and} \quad |S| \geq n \}.$$

Sometimes, we use the covering or packing number of a metric space, say (Θ, τ), other than (Ω, ρ); in such cases, we write $N(\Theta; \tau; \epsilon)$ or $M(\Theta; \tau; \epsilon)$ rather than $N(\epsilon)$ or $M(\epsilon)$, respectively. For specific $\epsilon > 0$, we will also refer to $N(\Theta; \tau; \epsilon)$ as the *ϵ-covering number of* (Θ, τ).

Remark 35. The covering and packing numbers of a metric space are closely related. In particular, for any $\epsilon > 0$, we always have

$$M(\epsilon) \leq N(\epsilon) \leq M(\epsilon/2). \tag{6.2}$$

The packing number and packing radius also have a close approximate inverse relationship. In particular, for any $\epsilon > 0$ and $n \in \mathbb{N}$, we always have

$$R(M(\epsilon)) \geq \epsilon \quad \text{and} \quad M(R(n)) \geq n. \tag{6.3}$$

However, it is possible that $R(M(\epsilon)) > \epsilon$ or $M(R(n)) > n$.

Finally, when we consider unbounded metric spaces, we will require some sort of concentration conditions on the probability distributions of interest, to obtain useful results. Specifically, we an appropriately generalized version of the moment of the distribution:

Remark 36. We defined the covering number slightly differently from usual (using partitions rather than covers). However, the given definition is equivalent to the usual definition, since (a) any partition is itself a cover (i.e., a set $\mathcal{C} \subseteq 2^\Omega$ such that $\Omega \subseteq \bigcup_{C\in\mathcal{C}} C$), and (b), for any countable cover $\mathcal{C} := \{C_1, C_2, ...\} \subseteq 2^\Omega$, there exists a partition $\mathcal{S} \in \mathbb{S}$ with $|\mathcal{S}| \leq |\mathcal{C}|$ and each $S_i \subseteq C_i$, defined recursively by $S_i := C_i \setminus \bigcup_{j=1}^{i-1} S_i$. $\mathcal{S}$ is often called the *disjointification* of $\mathcal{C}$.

Definition 37 (Metric Moments of a Probability Distribution)**.** For any $\ell \in [0, \infty]$, probability measure $P \in \mathcal{P}$, and $x \in \Omega$, the ℓ^{th} *metric moment* $m_{\ell,x}(P)$ *of* P *around* x is defined by

$$m_{\ell,x}(P) := \left(\mathop{\mathbb{E}}_{Y\sim P} \left[(\rho(x, Y))^\ell \right] \right)^{1/\ell} \in [0, \infty],$$

using the appropriate limit if $\ell = \infty$. The chosen reference point x only affects constant factors since,

$$\text{for all } x, x' \in \Omega, \quad \left| m^\ell_{\ell,x}(P) - m^\ell_{\ell,x'}(P) \right| \leq \left(\rho(x, x')\right)^\ell.$$

Note that, if Ω has linear structure with respect to which ρ is translation-invariant (e.g., if (Ω, ρ) is a Fréchet space), we can state our results more simply in terms of $m_\ell(P) := \inf_{x\in\Omega} m_{\ell,x}(P)$. As an example, if $\Omega = \mathbb{R}$ and $\rho(x, y) = |x - y|$, then $m_2(P)$ is precisely the standard deviation of P.

6.4 Related Work

A long line of work (Dudley, 1969; Ajtai, Komlós, and Tusnády, 1984; Canas and Rosasco, 2012; Dereich, Scheutzow, and Schottstedt, 2013; Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2017; Lei, 2018) has studied the rate of convergence of the empirical distribution to the population distribution in Wasserstein distance. In terms of upper bounds, the most general and tight upper bounds are the recent works of Weed and Bach (2017) and Lei (2018). As we describe below, while these two papers overlap significantly, neither supersedes the other, and our upper bound combines the key strengths of those in Weed and Bach (2017) and Lei (2018).

The results of Weed and Bach (2017) are expressed in terms of a particular notion of dimension, which they call the *Wasserstein dimension* s, since they derive convergence rates of order $n^{-r/s}$ (matching the $n^{-r/D}$ rate achieved on the unit cube $[0, 1]^D$). The definition of s is complex (e.g., it depends on the sample size n), but Weed and Bach (2017) show that, in many cases, s converges to certain common definitions of the intrinsic dimension of the support of the distribution. Our work overcomes three main limitations of Weed and Bach (2017):

1. The upper bounds of Weed and Bach (2017) apply only to totally bounded metric spaces. In contrast, our upper bounds permit unbounded metric spaces under the assumption that the distribution P has some finite moment $m_\ell(P) < \infty$. The results of Weed and Bach (2017) correspond to the special case $\ell = \infty$.
2. Their main upper bound (their Proposition 10) only holds when $s > 2r$, with constant factors diverging to infinity as $s \downarrow 2r$. Hence, their rates are loose when r is large or when the data have low intrinsic dimension. In contrast, our upper bound is tight even when $s \leq 2r$.
3. As we discuss in our Example 8, the upper bound of Weed and Bach (2017) becomes loose as the Wasserstein dimension s approaches ∞, limiting its utility

in infinite-dimensional function spaces. In contrast, we show that our upper and lower bounds match for several standard function spaces.

Intuitively, we find that the finite-sample bounds of Weed and Bach (2017) are tight when the intrinsic dimension of the data lies in an interval $[a, b]$ with $2r < a < b < \infty$, but they can be loose outside this range. In contrast, we find our results give tight rates for a larger class of problems.

On the other hand, Lei (2018) focuses on the case where Ω is a (potentially unbounded and infinite-dimensional) Banach space, under moment assumptions on the distributions. Thus, while the results of Lei (2018) cover interesting cases such as infinite-dimensional Gaussian processes, they do not demonstrate that convergence rates improve when the intrinsic dimension of the support of P is smaller than that of Ω (unless this support lies within a *linear* subspace of Ω). As a simple example, if the distribution is in fact supported on a finite set of k linearly independent points, the bound of Lei (2018) implies only a convergence rate, whereas we give a bound of order $O(\sqrt{k/n})$. Although we do not delve into this here, our results (unlike those of Lei (2018)) should also benefit from the multi-scale behavior discussed in Section 5 of Weed and Bach (2017); namely, much faster convergence rates are often observed for small n than for large n. This may help explain why algorithms such as functional k-means (García, García-Ródenas, and Gómez, 2015) work in practice, even though the results of Lei (2018) imply only a slow convergence rate of $O\left((\log n)^{-p}\right)$, for some constant $p > 0$, in this case.

Under similarly general conditions, Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet (2010a) and Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet (2012) have studied the related problem of estimating the Wasserstein distance between two unknown distributions given samples from those two distributions. Since one can estimate Wasserstein distances by plugging in empirical distributions, our upper bounds imply upper bounds for Wasserstein distance estimation. These bounds are tighter, in several cases, than those of Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet (2010a) and Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet (2012); for example, when $\mathcal{X} = [0, 1]^D$ is the Euclidean unit cube, we give a rate of $n^{-1/D}$, whereas they give a rate of $n^{-\frac{1}{D+1}}$. Minimax rates for this problem are currently unknown, and it is presently unclear to us under what conditions recent results on estimation of $\mathcal{L}_1$ distances between discrete distributions (Jiao, Han, and Weissman, 2017) might imply an improved rate as fast as $(n \log n)^{-1/D}$ for estimation of Wasserstein distance.

To the best of our knowledge, minimax lower bounds for distribution estimation under Wasserstein loss remain unstudied, except in the very specific case when $\Omega = [0, 1]^D$ is the Euclidean unit cube and $r = 1$ (Liang, 2017). As noted above, most previous works have focused on studying convergence rate of the empirical distribution to the true distribution in Wasserstein distance. For this rate, several lower bounds have been established, matching known upper bounds in many cases. However, many distribution estimators besides the empirical distribution can be considered. For example, it is tempting (especially given the infinite dimensionality of the distribution to be estimated) to try to reduce variance by techniques such as smoothing or importance sampling (Bucklew, 2013). Our lower bound results, given in Section 6.4.2, imply that the empirical distribution is already minimax optimal, up to constant factors, in many cases.

6.4.1 Upper Bounds

We begin with our main upper bound result:

Theorem 38 (Upper Bound). *Let $x_0 \in \Omega$ and suppose $m_{\ell,x_0}(P) \in [1,\infty)$. Let $J \in \mathbb{N}$ and $\epsilon > 0$. For each $k \in \mathbb{N}$, define $B_{2^k}(x_0) := \{y \in \Omega : 2^k \leq \rho(x_0,x) < 2^{k+1}\}$. Then, for $\ell \in (r,\infty)\backslash\{2r\}$,*

$$\mathbb{E}\left[W_r^r(P,P_n)\right] \leq C_{\ell,r} m_{\ell,x_0}^{\ell}(P) \left(n^{\frac{r-\ell}{\ell}} + 2^{-2Jr} + \sum_{k\in\mathbb{N}} \sum_{j=0}^{J} 2^{kr-2jr} \min\left\{ 2^{-k\ell}, \sqrt{\frac{N\left(B_{2^k}(x_0), 2^{k-2j}\right)}{n}} \right\} \right), \tag{6.4}$$

where $C_{\ell,r}$ is a constant depending only on ℓ and r. Moreover, when $\ell = 2r$, the bound (6.4) holds with $n^{\frac{r-\ell}{\ell}}$ replaced by $\frac{\log n}{\sqrt{n}}$.

The upper bound (6.4) can be thought of as having two main terms: a "tail" term of order $n^{\frac{r-\ell}{\ell}}$ and a "dimensionality" term, which depends on how the covering numbers $N(B_w(x_0),\eta)$ of balls centered around x_0 scale with w and η, as well as on two free parameters, J and ϵ, which can be chosen (depending on the covering number N) to minimize the overall bound. Each of these terms dominates in different settings, and, as discussed below, each matches, up to constant factors, a minimax lower bound on the error of estimating P.

The proof of Theorem 38 involves two main steps, which we sketch here:

Step 1: First, consider the totally bounded case, in which $\Delta := \mathrm{Diam}(\Omega)$ and $N(\Omega,\epsilon)$ are finite for any $\epsilon > 0$. In this setting, one can prove a bound (for any $J \in \mathbb{N}$) of order

$$\Delta^r 2^{-Jr} + \frac{\Delta^r}{\sqrt{n}} \sum_{j=1}^{J} 2^{-2jr} \sqrt{N(\Omega, \Delta 2^{-2j})}; \tag{6.5}$$

this is essentially the "multi-resolution bound" of Weed and Bach (2017), wherein the parameter J, controls the number of resolutions considered can be chosen freely to minimize the bound (typically, $J \to \infty$ as $n \to \infty$, at a rate depending on how $N(\Omega,\epsilon)$ scales with ϵ).

Step 2: We now reduce the case of unbounded Ω to the totally bounded case by partitioning Ω into a sequence of "thick spherical shells" $B_{2^k}(x_0)$, of inner radius 2^k and outer radius 2^{k+1}, centered around x_0, and bounding $W_r^r(P,\widehat{P})$ by a decomposition over these shells. For small k, the covering numbers $B_{2^k}(x_0)$ are not too big, and hence we can apply the bound (6.5), leading to the "dimensionality" term in (6.4). For large k, Markov's inequality and the bounded moment assumption together imply that the probabilities $P(B_{2^k}(x_0)$ and $\widehat{P}(B_{2^k}(x_0)$ decay rapidly; this small amount of mass, which may need to be moved a relatively large distance, leads to the $C_1 n^{\frac{r-\ell}{\ell}}$ "tail" term in (6.4). This general strategy of partitioning Ω into a nested sequence of bounded subsets is similar to that used by Fournier and Guillin (2015) and Lei (2018). However, both of these works relied on the assumption that (Ω,ρ) has a linear (Banach space) structure, which enabled them to use a bound of the form $N(wB,w\epsilon) \leq N(B,\epsilon)$, where $B \subseteq \Omega$ is totally bounded and $wB = \{wx : x \in B\}$ for scalar $w > 0$. This leads to a simpler upper bound, in which the terms depending on j and k can be factored, but, as we discuss in Section 6.5, requiring Ω to have linear structure can be limiting.

6.4.2 Lower Bounds

We now turn to providing lower bounds on minimax risk of density estimation in Wasserstein distance; that is, the quantity

$$\inf_{\widehat{P}:\Omega^n\to\mathcal{P}} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{X_1^n \overset{IID}{\sim} P} \left[W_r^r\left(P, \widehat{P}\right)\right], \tag{6.6}$$

where the infimum is over all estimators $\widehat{P}$ of P (i.e., all (potentially randomized) functions $\widehat{P} : \Omega^n \to \mathcal{P}$)).

We provide two results: one in terms of packing numbers, for totally bounded metric spaces, and one in terms of the tails of the distribution. Since distributions with totally bounded support necessarily satisfy moment bounds of arbitrary order, in the general unbounded setting with moment constraints, one can apply the maximum of the two bounds.

Theorem 39 (Minimax Lower Bound in Terms of Packing Radius)**.** *Let (Ω, ρ) be a metric space, on which $\mathcal{P}$ is the set of Borel probability measures. Then,*

$$M(r, \mathcal{P}) \geq c_r \sup_{k\in[32n]} R^r(\Omega, k)\sqrt{\frac{k-1}{n}},$$

where $c_r = \frac{3\log 2}{2^{r+12}}$ depends only on r.

Theorem 40 (Minimax Lower Bound for Heavy-Tailed Distributions)**.** *Suppose $r, \ell, \mu > 0$ are constants, and fix $x_0 \in \Omega$. Let $\mathcal{P}_{\ell,x_0}(\mu)$ denote the family of distributions P on Ω with ℓ^{th} moment $\mu_{\ell,x_0}(P) \leq \mu$ around x_0 at most μ. Let $n \geq \frac{3\mu}{2}$ and assume there exists $x_1 \in \Omega$ such that $\rho(x_0, x_1) = n^{1/\ell}$. Then,*

$$M(r, \mathcal{P}_{\ell,x_0}(\mu)) \geq c_\mu n^{\frac{r-\ell}{\ell}},$$

where $c_\mu := \frac{\min\{\mu, 2/3\}}{24}$ is constant in n.

Recalling that the packing radius R is closely related to the covering number N (via Equations (6.2) and (6.3)), one can see that these two bounds correspond to the two "nonparametric" terms of the upper bound (6.4). Specifically, it is easy to see that the rate in Theorem 40 matches the "tail" term in (6.4), while it is somewhat less obvious that the simple-looking rate in Theorem 39 matches, in many cases of interest, the apparently more complex "dimension" term of (6.4). However, as we show in the next section, despite their simplicity, these bounds are indeed tight in many diverse cases of interest.

6.5 Example Applications

Our theorems in the previous sections are quite abstract and have many tuning parameters. Thus, we conclude by exploring applications of our results to cases of interest. In each of the following examples, P is an unknown Borel probability measure over the specified Ω, from which we observe n IID samples. For upper bounds, $\widehat{P}$ denotes the empirical distribution (6.3.1) of these samples.

Example 4 (Finite Space). Consider the case where Ω is a finite set, over which ρ is the discrete metric given, for some $\delta > 0$, by $\rho(x, y) = \delta 1_{\{x=y\}}$, for all $x, y \in \Omega$. Then, for

any $\epsilon \in (0, \delta)$, the covering number is $N(\epsilon) = |\Omega|$. Thus, setting $K = 1$ and sending $\epsilon_1 \to 0$ in Theorem 38 gives

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq \delta^r \sqrt{\frac{|\Omega| - 1}{n}}.$$

On the other hand, $R(|\Omega|) = \delta$, and so, setting $k = |\Omega|$ in Theorem 39 yields

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_1, \ldots, X_n \overset{IID}{\sim} P} \left[W_r^r(P, \widehat{P})\right] \gtrsim \delta^r \sqrt{\frac{|\Omega| - 1}{n}}.$$

Example 5 ($\mathbb{R}^D$, Euclidean Metric). Consider the case where $\Omega = \mathbb{R}^D$ is the unit cube and ρ is the Euclidean metric. Assuming $\ell > r$, using the fact that $N(B_k, \rho, \epsilon) \leq \left(\frac{3w_k}{\epsilon}\right)^D$ (Pollard, 1990) and plugging $\epsilon_j = 2^{-2j}$ and $w_k = 2^k$ into Theorem 38 gives (after a straightforward but very tedious calculation) a constant $C_{D,r,\ell}$ depending only on D, r, and ℓ such that

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq C_{D,\ell,r} m_\ell^\ell(P) \left(n^{\frac{\ell - r}{\ell}} + 2^{-2Jr} + \sum_{j=1}^{J} 2^{(D-2r)j} \right). \tag{6.7}$$

Of these three terms, the first depends only on the number ℓ of finite moments P is assumed to have and the order r of the Wasserstein distance, whereas the second and third terms depend on choosing the parameter J. The optimal choice of J scales with the sample size n at a rate depending on the quantity $D - 2r$. Specifically, if $D = 2r$, then setting $J \asymp \frac{1}{4r} \log_2 n$ gives a rate of $\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim n^{\frac{\ell - r}{\ell}} + n^{-1/2} \log n$. If $D \neq 2r$, then (6.7) reduces to

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq C_{D,\ell,r} m_\ell^\ell(P) \left(n^{\frac{\ell - r}{\ell}} + 2^{-2Jr} + \frac{2^{(D-2r)J} - 1}{2^{D-2r} - 1} \right).$$

Then, if $D > 2r$, sending $J \to \infty$ gives $\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim n^{\frac{\ell - r}{\ell}} + n^{-1/2}$. Finally, if $D < 2r$, then setting $J \asymp \frac{1}{2D} \log n$ gives $\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim n^{\frac{\ell - r}{\ell}} + n^{-\frac{r}{D}}$. To summarize

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim n^{\frac{\ell - r}{\ell}} + \begin{cases} n^{-1/2} & \text{if } 2r > D \\ n^{-1/2} \log n & \text{if } 2r = D \\ n^{-r/D} & \text{if } 2r < D \end{cases}$$

(reproducing Theorem 1 of (Fournier and Guillin, 2015)). On the other hand, it is easy to check that the packing radius R satisfies $R(n) \geq n^{-1/D}$ and $R(2) \geq \sqrt{D}$. Thus, Theorem 39 with $k = n$ and $k = 2$ yields

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[W_r^r(\widehat{P}, P)\right] \gtrsim \max\left\{(n+1)^{-r/D}, D^{r/2} n^{-1/2}\right\}.$$

Together, these bounds give the following minimax rates for density estimation in Wasserstein loss:

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[W_r^r(\widehat{P}, P)\right] \asymp \begin{cases} n^{-1/2} & \text{if } \ell > 2r > D \\ n^{-r/D} & \text{if } 2r < D, \ell > \frac{Dr}{D-r} \end{cases}$$

When $2r = D$ and $\ell > 2r$, our upper and lower bounds are separated by a factor of $\log n$. The main result of (Ajtai, Komlós, and Tusnády, 1984) implies that, for the case $D = 2$ and $r = 1$, the empirical distribution converges as $n^{-1/2} \log n$, suggesting that the $\log n$ factor in our upper bound may be tight. Further generalization of Theorem 39 is needed to give lower bounds when both $D, \ell \leq 2r$ or when $D > 2r$ and $\ell \leq \frac{Dr}{D-r}$.

The next example demonstrates how the rate of convergence in Wasserstein metric depends on properties of the metric space (Ω, ρ) at both large and small scales. Specifically, if we discretize Ω, then the phase transition at $2r = D$ disappears.

Example 6 (Unbounded Grid). Suppose $\Omega = \mathbb{Z}^D$ is a D-dimensional grid of integers and ρ is ℓ_∞-metric (given by $\rho(x, y) = \max_{j \in [D]} |x_j - y_j|$). Since $\mathcal{Z}^D \subseteq \mathbb{R}^D$ and the ℓ_∞ and Euclidean metrics are topologically equivalent, the upper bounds from Example 5 clearly apply, up to a factor of $\sqrt{D}$. However, we also have the fact that, whenever $\epsilon < 1$, $N(B_k, \rho, \epsilon) = w_k^D$. Therefore, setting $J = 0$, $\epsilon_0 = 0$, and $w_k = 2^k$ in Theorem 38 gives, for a constant $C_{D,\ell,r}$ depending only on D, ℓ, and r,

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq C_{D,\ell,r} m_\ell^\ell(P) \left(n^{\frac{\ell - r}{\ell}} + \sum_{k \in \mathbb{N}} \sqrt{\frac{2^{(D-\ell)k}}{n}} \right).$$

When $\ell > D$, this reduces to $\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim n^{\frac{\ell - r}{\ell}} + n^{-1/2}$, giving a tighter rate than in Example 5 when $2r \leq D < \ell$. To the best of our knowledge, no prior results in the literature imply this fact.

Example 7 (Latent Variable Models, Manifolds). This example demonstrates that the convergence rate of the empirical distribution in Wasserstein distance improves in the presence of additional structure in the data. Importantly, no *knowledge* of this structure is needed to obtain this accelerated convergence, since it is inherent to the empirical distribution itself.

Suppose that there exist a metric space $(\mathcal{Z}, \rho_\mathcal{Z})$, a L-Lipschitz mapping $\phi : \mathcal{Y} \to \Omega$, and a probability distribution Q on $\mathcal{Z}$ such that P is the pushforward on Q under ϕ; i.e., for any $A \subseteq \Omega$, $P(A) = Q(f^{-1}(A))$, where $\phi^{-1}(A)$ denotes the pre-image of A under ϕ. This setting is inherent, for example, in many latent variable models. When $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^D$ with $d < D$, this generalizes the assumption, popular in high-dimensional nonparametric statistics, that the data lie on a low-dimensional manifold.

In this setting, one can easily bound moments of P and covering numbers in Ω in terms of those of Q and in $\mathcal{Z}$, respectively. Specifically,

(a) for any $z \in \mathcal{Z}$, $\ell > 0$, $m_{\ell,\phi(z)}(P) \leq L m_{\ell,z}(Q)$, and
(b) for any $E \subseteq \Omega$, $\epsilon > 0$, $N(E, \rho, \epsilon) \leq N(\phi^{-1}(E), \rho_\mathcal{Z}, \epsilon/L)$.

This allows us to bound convergence rates over Ω in terms of moment bounds on Q and covering number bounds on $(\mathcal{Z}, \rho_\mathcal{Z})$. For example, if $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\rho_\mathcal{Z}$ is the Euclidean metric, then, for any bounded $E \subseteq \mathcal{Z}$, we necessarily have $N(E, \rho_\mathcal{Z}, \epsilon) \in O\left(\epsilon^{-d}\right)$ as $\epsilon \to 0$. If $\Omega \subseteq \mathbb{R}^D$ with $d < D$, then, via analysis similar to that in the Euclidean case above, Theorem 38 gives a convergence rate of $n^{-1/2} n^{\frac{r-\ell}{\ell}} + n^{-r/d}$, potentially much faster than the $n^{-1/2} n^{\frac{r-\ell}{\ell}} + n^{-r/D}$ minimax lower bound that can be derived without assuming this low-dimensional structure.

Finally, we consider distributions over an infinite dimensional space of smooth functions.

Example 8 (Hölder Ball, $\mathcal{L}_\infty$ Metric). Suppose that, for some $\alpha \in (0, 1]$,

$$\Omega := \left\{ f[0,1]^D \to [-1,1] \quad \middle| \quad \forall x, y \in [0,1]^D, \quad |f(x) - f(y)| \leq \|x - y\|_2^\alpha \right\}$$

is the class of unit α-Hölder functions on the unit cube and ρ is the $\mathcal{L}_\infty$-metric given by

$$\rho(f, g) = \sup_{x \in [0,1]^D} |f(x) - g(x)|, \qquad \text{for all } f, g \in \Omega.$$

The covering and packing numbers of (Ω, ρ) are well-known to be of order $\exp\left(\epsilon^{-D/\alpha}\right)$ (DeVore and Lorentz, 1993); specifically, there exist positive constants $0 < c_1 < c_2$ such that, for all $\epsilon \in (0, 1)$,

$$c_1 \exp\left(\epsilon^{-D/\alpha}\right) \leq N(\epsilon) \leq M(\epsilon) \leq c_2 \exp\left(\epsilon^{-D/\alpha}\right).$$

Since $\mathrm{Diam}(\Omega) = 2$, applying Theorem 38 with $K = 1$ and

$$\epsilon_1 = (2 \log n - (\alpha r/D) \log \log n)^{-\frac{\alpha r}{D}} \qquad \text{gives} \qquad \mathbb{E}\left[W_r^r(P, \widehat{P})\right] \lesssim (\log n)^{\frac{-\alpha r}{D}}.$$

Conversely, Inequality (6.3) implies $R(n) \geq (\log(n/c_1))^{\frac{-\alpha}{D}}$, and so setting $k = n$ in Theorem 39 gives

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_1, \ldots, X_n \overset{IID}{\sim} P} \left[W_r^r(P, \widehat{P})\right] \gtrsim \left(\frac{1}{\log(n/c_1)}\right)^{\frac{\alpha r}{D}},$$

showing that distribution estimation over $(\mathcal{P}, W_r^r)$ has the extremely slow minimax rate $(\log n)^{\frac{-\alpha r}{D}}$. Although we considered only $\alpha \in (0, 1]$ (due to the notational complexity of defining higher-order Hölder spaces), analogous rates hold for all $\alpha > 0$. Also, since our rates depend only on covering and packing numbers of Ω, identical rates can be derived for related Sobolev and Besov classes. Note that the Wasserstein dimension used in the prior work (Weed and Bach, 2017) is of order $\frac{D}{\alpha} \log n$, and so their upper bound (their Proposition 10) gives a rate of $n^{-\frac{\alpha r}{D \log n}} = \exp\left(-\frac{\alpha r}{D}\right)$, which fails to converge as $n \to \infty$.

One might wonder why we are interested in studying Wasserstein convergence of distributions over spaces of smooth functions, as in Example 8. Our motivation comes from the historical use of smooth function spaces have been widely used for modeling images and other complex naturalistic signals (Mallat, 1999; Peyré, 2011; Sadhanala, Wang, and Tibshirani, 2016). Empirical breakthroughs have recently been made in generative modeling, particularly of images, based on the principle of minimizing Wasserstein distance between the empirical distribution and a large class of models encoded by a deep neural network (Montavon, Müller, and Cuturi, 2016; Arjovsky, Chintala, and Bottou, 2017; Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville, 2017).

However, little is known about theoretical properties of these methods; while there has been some work studying the optimization landscape of such models (Nagarajan and Kolter, 2017; Liang and Stokes, 2018), we know of far less work exploring their *statistical* properties. Given the extremely slow minimax convergence rate we derived above, it must be the case that the class of distributions encoded by such models is far smaller or sparser than $\mathcal{P}$. An important avenue for further work is thus to explicitly identify stronger assumptions that can be made on distributions

over interesting classes of signals, such as images, to bridge the gap between empirical performance and our theoretical understanding.

Example 9 (Expectations of Lipschitz Functions & Monte Carlo Integration). A fundamental statistical problem is to estimate an expectation $\mathbb{E}_{X\sim P}[f(X)]$ of some function f with respect to a distribution P. A classical duality result of Kantorovich (Kantorovich, 1942) implies that

$$W_1(P,Q) = \sup_{f\in\mathcal{C}^1(\Omega)} \left| \mathop{\mathbb{E}}_{X\sim P}[f(X)] - \mathop{\mathbb{E}}_{Y\sim Q}[f(Y)] \right|, \quad \text{where} \quad \mathcal{C}^1(\Omega) := \left\{ f : \Omega \to \mathbb{R} : \sup_{x\neq y\in\Omega} \frac{|f(x)-f(y)|}{\rho(x,y)} \right.$$

denotes the class of 1-Lipschitz functions on (Ω,ρ). Our upper bound (Theorem 38) thus implies bounds, uniformly over 1-Lipschitz functions f, on the expected error of estimating an expectation $\mathbb{E}_{X\sim P}[f(X)]$ by the empirical estimate $\frac{1}{n}\sum_{i=1}^n f(X_i)$ based on $X_1^n \overset{IID}{\sim} P$. Moreover, our lower bounds (Theorems 39 and 40) imply that this empirical estimate is minimax rate-optimal over P satisfying only bounded moment assumptions.

As Weed and Bach (2017) noted, this has consequences for Monte Carlo integration, a common approach to numerical integration in which an integral $\int_\Omega f\,d\lambda$ of a function f with respect to a measure λ is estimated based on n IID samples from a probability distribution P proportional to λ; Monte Carlo integration is useful even when f and λ are known analytically, since numerically computing this integral can be challenging, especially in high dimensions or when the supports of f and λ are unbounded. In this context, the sample size n required to obtain a desired accuracy directly determines the computational demand of the integration scheme.

Our upper bounds allow one to generalize the upper bound of Weed and Bach (2017) for Monte Carlo integration (their Proposition 21) to the important case of integrals over unbounded domains Ω, and, moreover, our lower bounds imply that, at least without further knowledge of $f\in\mathcal{C}^1(\Omega)$ and $P\in\mathcal{P}_{\ell,x_0}(\mu)$, the empirical estimate above is rate-optimal among Monte Carlo estimates (i.e., among functions of X_1^n). Although improved estimates can be constructed for specific f, Ω, and λ, these worst-case results are useful when either f or λ is too complex to model analytically, as often happens, for example, in Bayesian inference problems (Geweke, 1989).

6.6 Conclusion

In this section, we derived upper and lower bounds for distribution estimation under Wasserstein loss. Our upper bounds generalize prior results and are tighter in certain cases, while our lower bounds are, to the best of our knowledge, the first minimax lower bounds for this problem. We also provided several concrete examples in which our bounds imply novel convergence rates.

6.6.1 Future Work

We studied minimax rates over the very large entire class $\mathcal{P}$ of all distributions with some number of finite moments. It would be useful to understand how minimax rates improve when additional assumptions, such as smoothness, are made (see, e.g., Liang (2017) for somewhat improved upper bounds under smoothness assumptions when (Ω,ρ) is the Euclidean unit cube). Given the slow convergence rates we found over $\mathcal{P}$ in many cases, studying minimax rates under stronger assumptions may help

to explain the relatively favorable empirical performance of popular distribution estimators based on empirical risk minimization in Wasserstein loss. Moreover, while rates over all of $\mathcal{P}$ are of interest only for very weak metrics such as the Wasserstein distance (as stronger metrics may be infinite or undefined), studying minimax rates under additional assumptions will allow for a better understanding of the Wasserstein metric in relation to other commonly used metrics.

6.7 Preliminary Lemmas and Proof Sketch of Theorem 38

In this section, we outline the proof of Theorem 38, our upper bound for the case of totally bounded metric spaces. The proof of the more general Theorem 38 for unbounded metric spaces, which is given in the next section, builds on this.

We begin by providing a few basic lemmas; these lemmas are not fundamentally novel, but they will be used in the subsequent proofs of our main upper and lower bounds, and also help provide intuition for the behavior of the Wasserstein metric and its connections to other metrics between probability distributions. The proofs of these lemmas are given later, in Section 6.9. Our first lemma relates Wasserstein distance to the notion of resolution of a partition.

Lemma 41. *Suppose $\mathcal{S} \in \mathbb{S}$ is a countable Borel partition of Ω. Let P and Q be Borel probability measures such that, for every $S \in \mathcal{S}$, $P(S) = Q(S)$. Then, for any $r \geq 1$, $W_r(P, Q) \leq \mathrm{Res}(\mathcal{S})$.*

Our next lemma gives simple lower and upper bounds on the Wasserstein distance between distributions supported on a countable subset $\mathcal{X} \subseteq \Omega$, in terms of $\mathrm{Diam}(\mathcal{X})$ and $\mathrm{Sep}(\mathcal{X})$. Since our main results will utilize coverings and packings to approximate Ω by finite sets, this lemma will provide a first step towards approximating (in Wasserstein distance) distributions on Ω by distributions on these finite sets. Indeed, the lower bound in Inequality (6.8) will suffice to prove our lower bounds, although a tighter upper bound, based on the upper bound in (6.8), will be necessary to obtain tight upper bounds.

Lemma 42. *Suppose (Ω, ρ) is a metric space, and suppose P and Q are Borel probability distributions on Ω with countable support; i.e., there exists a countable set $\mathcal{X} \subseteq \Omega$ with $P(\mathcal{X}) = Q(\mathcal{X}) = 1$. Then, for any $r \geq 1$,*

$$(\mathrm{Sep}(\mathcal{X}))^r \sum_{x \in \mathcal{X}} |P(\{x\}) - Q(\{x\})| \leq W_r^r(P, Q) \leq (\mathrm{Diam}(\mathcal{X}))^r \sum_{x \in \mathcal{X}} |P(\{x\}) - Q(\{x\})| . \tag{6.8}$$

Remark 43. Recall that the term $\sum_{x \in \mathcal{X}} |P(\{x\}) - Q(\{x\})|$ in Inequality (6.8) is the $\mathcal{L}_1$ distance

$$\|p - q\|_1 := \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

between the densities p and q of P and Q with respect to the counting measure on $\mathcal{X}$, and that this same quantity is twice the total variation distance

$$TV(P, Q) := \sup_{A \subseteq \Omega} |P(A) - Q(A)| .$$

Hence, Lemma 42 can be equivalently written as

$$\mathrm{Sep}(\Omega) \left(\|p - q\|_1\right)^{1/r} \leq W_r(P, Q) \leq \mathrm{Diam}(\Omega) \left(\|p - q\|_1\right)^{1/r}$$

and as

$$\mathrm{Sep}(\Omega)\left(2TV(P,Q)\right)^{1/r} \leq W_r(P,Q) \leq \mathrm{Diam}(\Omega)\left(2TV(P,Q)\right)^{1/r},$$

bounding the r-Wasserstein distance in terms of the $\mathcal{L}_1$ and total variation distance. As noted in Example 4, equality holds in (6.8) precisely when ρ is the unit discrete metric given by $\rho(x,y) = 1_{\{x \neq y\}}$ for all $x, y \in \Omega$.

On metric spaces that are discrete (i.e., when $\mathrm{Sep}(\Omega) > 0$), the Wasserstein metric is (topologically) at least as strong as the total variation metric (and the $\mathcal{L}_1$ metric, when it is well-defined), in that convergence in Wasserstein metric implies convergence in total variation (and $\mathcal{L}_1$, respectively). On the other hand, on bounded metric spaces, the converse is true. In either of these cases, *rates* of convergence may differ between metrics, although, in metric spaces that are both discrete *and* bounded (e.g., any finite space), we have $W_r \asymp TV^{1/r}$.

To obtain tight bounds as discussed below, we will require not only a partition of the sample space Ω, but a nested sequence of partitions, defined as follows.

Definition 44 (Refinement of a Partition, Nested Partitions)**.** Suppose $\mathcal{S}, \mathcal{T} \in \mathbb{S}$ are partitions of Ω. $\mathcal{T}$ is said to be a *refinement of* $\mathcal{S}$ if, for every $T \in \mathcal{T}$, there exists $S \in \mathcal{S}$ with $T \subseteq S$. A sequence $\{\mathcal{S}_k\}_{k\in\mathbb{N}}$ of partitions is called *nested* if, for each $k \in \mathbb{N}$, $\mathcal{S}_k$ is a refinement of $\mathcal{S}_{k+1}$,

While Lemma 42 gave a simple upper bound on the Wasserstein distance, the factor of $\mathrm{Diam}(\Omega)$ turns out to be too large to obtain tight rates for a number of cases of interest (such as the D-dimensional unit cube $\Omega = [0,1]^D$, discussed in Example 5). The following lemma gives a tighter upper bound, based on a hierarchy of nested partitions of Ω; this allows us to obtain tighter bounds (than $\mathrm{Diam}(\Omega)$) on the distance that mass must be transported between P and Q. Note that, when $K = 1$, Lemma 45 reduces to a trivial combination of Lemmas 41 and 42; indeed, these lemmas are the starting point for proving Lemma 45 by induction on K.

Note that the idea of such a "multi-resolution" upper bound has been utilized extensively before, and numerous versions have been proven before (see, e.g., Fact 6 of Do Ba, Nguyen, Nguyen, and Rubinfeld (2011), Lemma 6 of Fournier and Guillin (2015), or Proposition 1 of Weed and Bach (2017)). Most of these versions have been specific to Euclidean space; to the best of our knowledge, only Proposition 1 of Weed and Bach (2017) applies to general metric spaces. However, that result also requires that (Ω, ρ) is totally bounded (more precisely, that $m_x^\infty(P) < \infty$, for some $x \in \Omega$).

Lemma 45. *Let K be a positive integer. Suppose $\{\mathcal{S}_k\}_{k\in\mathbb{N}}$ is a nested sequence of countable Borel δ-partitions of (Ω, ρ). Then, for any $r \geq 1$ and Borel probability measures P and Q on Ω,*

$$W_r^r(P,Q) \leq (\mathrm{Res}(\mathcal{S}_0))^r + \sum_{k=1}^{\infty} (\mathrm{Res}(\mathcal{S}_k))^r \left(\sum_{S \in \mathcal{S}_{k+1}} |P(S) - Q(S)| \right). \tag{6.9}$$

Lemma 45 requires a sequence of partitions of Ω that is not only multi-resolution but also nested. While the ϵ-covering number implies the existence of small partitions with small resolution, these partitions need not be nested as ϵ becomes small. For this reason, we now give a technical lemma that, given any sequence of partitions, constructs a *nested* sequence of partitions of the same cardinality, with only a small increase in resolution.

Lemma 46. *Suppose $\mathcal{S}$ and $\mathcal{T}$ are partitions of (Ω, ρ), and suppose $\mathcal{S}$ is countable. Then, there exists a partition $\mathcal{S}'$ of (Ω, ρ) such that:*

a) $|\mathcal{S}'| \leq |\mathcal{S}|$.

b) $\mathrm{Res}(\mathcal{S}') \leq \mathrm{Res}(\mathcal{S}) + 2\,\mathrm{Res}(\mathcal{T})$.

c) $\mathcal{T}$ *is a refinement of* $\mathcal{S}'$.

Lemmas 45 and 46 are the main tools needed to bound the expected Wasserstein distance $\mathbb{E}[W_r^r(P, \widehat{P})]$ of the empirical distribution from the true distribution into a sum of its expected errors on each element of a nested partition of Ω. Then, we will need to control the total expected error across these partition elements, which we will show behaves similarly to the $\mathcal{L}_1$ error of the standard maximum likelihood (mean) estimator a multinomial distribution from its true mean. Thus, the following result of Han, Jiao, and Weissman (2015) will be useful.

Lemma 47 (Theorem 1 of (Han, Jiao, and Weissman, 2015))**.** *Suppose* $(X_1, ..., X_K) \sim \mathrm{Multinomial}(n, p_1, ..., p_K)$. *Let*

$$Z := \|X - np\|_1 = \sum_{k=1}^{K} |X_k - np_k|\,.$$

Then, $\mathbb{E}\left[Z/n\right] \leq \sqrt{(K-1)/n}$.

Finally, we are ready to prove Theorem 38.

Theorem 38. Let (Ω, ρ) be a metric space on which P is a Borel probability measure. Let $\widehat{P}$ denote the empirical distribution of n IID samples $X_1, ..., X_n \overset{IID}{\sim} P$, give by

$$\widehat{P}(S) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in S\}}, \quad \forall S \in \Sigma.$$

Then, for any sequence $\{\epsilon_k\}_{k\in[K]} \in (0, \infty)^K$ with $\epsilon_0 = \mathrm{Diam}(\Omega)$,

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq \epsilon_K^r + \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\left(\sum_{j=k-1}^{K} 2^{j-k}\epsilon_j\right)^r \sqrt{N(\epsilon_k) - 1}.$$

Proof: By recursively applying Lemma 46, there exists a sequence $\{\mathcal{S}_k\}_{k\in[K]}$ of partitions of (Ω, ρ) satisfying the following conditions:

1. for each $k \in [K]$, $|\mathcal{S}_k| = N(\epsilon_k)$.
2. for each $k \in [K]$, $\mathrm{Res}(\mathcal{S}_k) \leq \sum_{j=k}^{K} 2^{j-k}\epsilon_j$.
3. $\{S_k\}_{k\in[K]}$ is nested.

Note that, for any $k \in [K]$, the vector $n\widehat{P}(S)$ (indexed by $S \in \mathcal{S}_k$) follows an n-multinomial distribution over $|\mathcal{S}_k|$ categories, with means given by $P(S)$; i.e.,

$$(n\widehat{P}(S_1), ..., n\widehat{P}(S_k)) \sim \mathrm{Multinomial}(n, P(S_1), ..., P(S_k)).$$

Thus, by Lemma 47, for each $k \in [K]$,

$$\mathbb{E}\left[\sum_{S\in\mathcal{S}_k}\left|P(S)-\widehat{P}(S)\right|\right] \le \sqrt{\frac{|\mathcal{S}_k|-1}{n}} = \sqrt{\frac{N(\epsilon_k)-1}{n}}.$$

Thus, by Lemma 45,

$$\begin{aligned}
\mathbb{E}\left[W_r^r(P,Q)\right] &\le \mathbb{E}\left[\epsilon_K^r + \sum_{k=1}^{K}\left(\sum_{j=k}^{K} 2^{j-k}\epsilon_j\right)^r \left(\sum_{S\in\mathcal{S}_k}|P(S)-Q(S)|\right)\right] \\
&\le \epsilon_K^r + \sum_{k=1}^{K}\left(\sum_{j=k}^{K} 2^{j-k}\epsilon_j\right)^r \mathbb{E}\left[\sum_{S\in\mathcal{S}_k}|P(S)-Q(S)|\right] \\
&\le \epsilon_K^r + \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\left(\sum_{j=k}^{K} 2^{j-k}\epsilon_j\right)^r \sqrt{N(\epsilon_k)-1}
\end{aligned}$$

■

6.8 Proof Sketch of Theorem 38

In this section, we prove our more general upper bound, Theorem 38, which applies to potentially unbounded metric spaces (Ω,ρ), assuming that P is sufficiently concentrated (i.e., has at least $\ell > 0$ finite moments).

The basic idea is to partition the potentially unbounded metric space (Ω,ρ) into countably many totally bounded subsets $B_1, B_2, ...$, and to decompose the Wasserstein error into its error on each B_i, weighted by the probability $P(B_i)$. Specifically, fixing an arbitrary base point x_0, $B_1, B_2, ...$ will be spherical shells, such that $x_0 \in B_1$, and both the distance between B_i and x_0, as well as the size (covering number) of B_i, increase with i. For large i, the assumption that P has ℓ bounded moments implies (by Markov's inequality) that $P(B_i)$ is small, whereas, for small i, we adapt our previous result Theorem 38 in terms of the covering number.

To carry out this approach, we will need two new lemmas. The first decomposes Wasserstein distance into the sum of its distances on each B_i, and can be considered an adaptation of Lemma 2.2 of Lei (2018) (for Banach spaces) to general metric spaces.

Lemma 48. *Fix a reference point $x_0 \in \Omega$ and a non-decreasing real-valued sequence $\{w_k\}_{k\in\mathbb{N}}$ with $w_0 = 0$ and $\lim_{k\to\infty} w_k = \infty$. For each $k \in \mathbb{N}$, define*

$$B_k := \{x \in \Omega : w_k \le \rho(x_0, x) < w_{k+1}\}.$$

Then, there exists a constant C_r depending only on r such that, for any Borel probability measures P and Q on Ω,

$$W_r^r(P,Q) \le C_r \sum_{k=0}^{\infty} w_k^r \min\{P(B_k), Q(B_k)\} W_r^r(P_{B_k}, Q_{B_k}) + |P(B_k) - Q(B_k)|.$$

where, for any sets $A, B \subseteq \Omega$,

$$P_A(B) = \frac{P(A\cap B)}{P(B)}$$

(under the convention that $\frac{0}{0} = 0$) denotes the conditional probability of B given A, under P.

The second lemma is more nuanced variant of Lemma 47 (albeit, leading to slightly looser constants). When i is large the covering number of B_i can become quite large, but the total probability $P(B_i)$ is quite small. Whereas Lemma 47 depends only on the size of the partition, the following result will allow us to control the total error using both of these factors.

Lemma 49 (Theorem 1 of Berend and Kontorovich (2013)). *Suppose $X \sim \text{Binomial}(n, p)$. Then, we have the bound*

$$\mathbb{E}\left[|X - np|\right] \leq n \min\left\{2P(A), \sqrt{P(A)/n}\right\}. \tag{6.10}$$

on the mean absolute deviation of X.

Finally, we are ready to prove our main upper bound result for unbounded metric spaces.

Theorem 38 (General Upper Bound for Unbounded Metric Spaces). Let $x_0 \in \Omega$ and suppose $m_{\ell,x_0}(P) \in [1, \infty)$. Let J be a positive integer. Fix two non-decreasing real-valued sequences $\{w_k\}_{k\in\mathbb{N}}$ and $\{\epsilon_j\}_{j\in\mathbb{N}}$, of which $\{w_k\}_{k\in\mathbb{N}}$ is non-decreasing with $w_0 = 0$ and $\lim_{k\to\infty} w_k = \infty$ and $\{\epsilon_j\}_{j\in[J]}$ is non-increasing. For each $k \in \mathbb{N}$, define

$$B_k(x_0) := \{y \in \Omega : w_k \leq \rho(x_0, x) < w_{k+1}\}.$$

Then,

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq m_{\ell,x_0}^{\ell} \sum_{k\in\mathbb{N}} w_k^{-\ell} (\epsilon_J)^r + 2^r w_k^{r-\ell/2} \min\left\{2w_k^{-\ell/2}, \sqrt{\frac{1}{n}}\right\}$$
$$+ \sum_{j=1}^{J} \left(\sum_{t=j}^{J} 2^{J-t}\epsilon_t\right)^r \min\left\{2w_k^{-\ell}, \sqrt{\frac{w_k^{-\ell}}{n} N(B_k, \rho, \epsilon_j)}\right\}.$$

Proof: As in the proof of Theorem 38, by recursively applying Lemma 46, for each $k \in \mathbb{N}$, we can construct a nested sequence $\{\mathcal{S}_{k,j}\}_{j\in[J_k]}$ of partitions of B_k such that, for each $j \in [J_k]$,

$$|\mathcal{S}_{k,j}| = N(B_k, \rho, \epsilon_{k,j}) \quad \text{and} \quad \text{Res}(\mathcal{S}_{k,j}) \leq \sum_{t=0}^{j} 2^t \epsilon_{k,j}. \tag{6.11}$$

Since each P_{B_k} and $\widehat{P}_{B_k}$ are supported only on B_k, plugging the bound Lemma 45 into the bound in Lemma 48 gives

$$W_r^r(P, \widehat{P})$$
$$\leq \sum_{k\in\mathbb{N}} \min\left\{P(B_k), \widehat{P}(B_k)\right\} \left((\text{Res}(\mathcal{S}_{k,0}))^r + \sum_{j=1}^{J_k} (\text{Res}(\mathcal{S}_{k,j}))^r \sum_{S\in\mathcal{S}_{k,j+1}} \left|P_{B_k}(S) - \widehat{P}_{B_k}(S)\right|\right)$$
$$+ 2^r w_k^r \left|P(B_k) - \widehat{P}(B_k)\right|$$
$$\leq \sum_{k\in\mathbb{N}} 2^r w_k^r \left|P(B_k) - \widehat{P}(B_k)\right| + P(B_k) (\text{Res}(\mathcal{S}_{k,0}))^r + \sum_{j=1}^{J} (\text{Res}(\mathcal{S}_{k,j}))^r \sum_{S\in\mathcal{S}_{k,j+1}} \left|P(S) - \widehat{P}(S)\right|.$$

Since each $\widehat{P}(S) \sim \text{Binomial}(n, P(S))$, for each $k \in \mathbb{N}$ and $j \in [J_k]$, Lemma 49 followed by Cauchy-Schwarz gives

$$\mathbb{E}\left[\sum_{S\in\mathcal{S}_{k,j}} \left|P(S) - \widehat{P}(S)\right|\right] \leq \sum_{S\in\mathcal{S}_{k,j+1}} \min\left\{2P(S), \sqrt{P(S)/n}\right\}$$
$$\leq \min\left\{2P(B_k), \sqrt{\frac{P(B_k)}{n}|\mathcal{S}_{k,j}|}\right\}.$$

Therefore, taking expectations (over $X_1, ..., X_n$), applying Inequality 6.11, and applying Lemma 49 once more gives

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq \sum_{k\in\mathbb{N}} P(B_k)\left(\epsilon_{k,0}\right)^r + 2^r w_k^r \min\left\{2P(B_k), \sqrt{P(B_k)/n}\right\}$$
$$+ \sum_{j=1}^{J_k}\left(\sum_{t=0}^{j} 2^t\epsilon_{k,j}\right)^r \min\left\{2P(B_k), \sqrt{\frac{P(B_k)}{n}N(B_k, \rho, \epsilon_{k,j+1})}\right\}.$$

Now note that, by Markov's inequality,

$$P(B_k) \leq \mathbb{P}_{X\sim P}\left[\rho(x_0, X) \geq w_k\right] = \mathbb{P}_{X\sim P}\left[\rho^\ell(x_0, X) \geq w_k^\ell\right] \leq \frac{m_{\ell,x_0}^\ell(P)}{w_k^\ell}. \tag{6.12}$$

Therefore, assuming that each $m_{\ell,x_0}^\ell \geq 1$, so that $m_{\ell,x_0}^\ell \geq m_{\ell,x_0}^{\ell/2}$,

$$\mathbb{E}\left[W_r^r(P, \widehat{P})\right] \leq m_{\ell,x_0}^\ell \sum_{k\in\mathbb{N}} w_k^{-\ell}\left(\epsilon_{k,0}\right)^r + 2^r w_k^r \min\left\{2w_k^{-\ell}, \sqrt{w_k^{-\ell}/n}\right\}$$
$$+ \sum_{j=1}^{J_k}\left(\sum_{t=0}^{j} 2^t\epsilon_{k,j}\right)^r \min\left\{2w_k^{-\ell}, \sqrt{\frac{w_k^{-\ell}}{n}N(B_k, \rho, \epsilon_{k,j+1})}\right\},$$

proving the theorem. ∎

6.9 Proofs of Lemmas

Lemma 41. Suppose $\mathcal{S} \in \mathbb{S}$ is a countable Borel partition of Ω. Let P and Q be Borel probability measures such that, for every $S \in \mathcal{S}$, $P(S) = Q(S)$. Then, for any $r \geq 1$, $W_r(P, Q) \leq \text{Res}(\mathcal{S})$.

Proof: This fact is intuitively obvious; clearly, there exists a transportation map μ from P to Q that moves mass only within each $S \in \mathcal{S}$ and therefore without moving any mass further than δ. For completeness, we give a formal construction.

Let $\mu : \Sigma^2 \to [0, 1]$ denote the coupling that is conditionally independent given any set $S \in \mathcal{S}$ with $P(S) = Q(S) > 0$ (that is, for any $A, B \in \Sigma$, $\mu(A \times B \cap S \times S)P(S) = P(A \cap S)Q(B \cap S)$).[3] It is easy to verify that $\mu \in \mathcal{C}(P, Q)$. Since $\mathcal{S}$ is a

[3]The existence of such a measure can be verified by the Hahn-Kolmogorov theorem, similarly to that of the usual product measure (see, e.g., Section IV.4 of Doob (2012)).

countable partition and μ is only supported on $\bigcup_{S\in\mathcal{S}} S\times S$,

$$\begin{aligned}
W_r(P,Q) &\leq \left(\int_{\Omega\times\Omega} \rho^r(x,y)\,d\mu(x,y)\right)^{1/r} \\
&= \left(\sum_{S\in\mathcal{S}}\int_{S\times S} \rho^r(x,y)\,d\mu(x,y)\right)^{1/r} \\
&\leq \left(\sum_{S\in\mathcal{S}}\int_{S\times S} \delta^r\,d\mu(x,y)\right)^{1/r} \\
&= \delta\left(\sum_{S\in\mathcal{S}}\mu(S\times S)\right)^{1/r} = \delta\left(\sum_{S\in\mathcal{S}}\frac{P(S)Q(S)}{P(S)}\right)^{1/r} = \delta\left(\sum_{S\in\mathcal{S}}Q(S)\right)^{1/r} = \delta.
\end{aligned}$$

■

Lemma 42. Suppose (Ω,ρ) is a metric space, and suppose P and Q are Borel probability distributions on Ω with countable support; i.e., there exists a countable set $\mathcal{X}\subseteq\Omega$ with $P(\mathcal{X})=Q(\mathcal{X})=1$. Then, for any $r\geq 1$,

$$(\mathrm{Sep}(\mathcal{X}))^r\sum_{x\in\mathcal{X}}|P(\{x\})-Q(\{x\})| \leq W_r^r(P,Q) \leq (\mathrm{Diam}(\mathcal{X}))^r\sum_{x\in\mathcal{X}}|P(\{x\})-Q(\{x\})|\,.$$

Proof: The term $\sum_{x\in\mathcal{X}}|P(\{x\})-Q(\{x\})| = TV(P,Q)$ is precisely the (unweighted) amount of mass that must be transported to transform between P and Q. Hence, the result is intuitively fairly obvious; all mass moved has a cost of at least $\mathrm{Sep}(\Omega)$ and at most $\mathrm{Diam}(\Omega)$. However, for completeness, we give a more formal proof below.

To prove the lower bound, suppose $\mu\in\Pi(P,Q)$ is any coupling between P and Q. For $x\in\mathcal{X}$,

$$\mu(\{x\}\times\{x\}) + \mu(\{x\}\times(\Omega\backslash\{x\})) = \mu(\{x\}\times\Omega) = P(\{x\})$$

and, similarly,

$$\mu(\{x\}\times\{x\}) + \mu((\Omega\backslash\{x\})\times\{x\}) = \mu(\Omega\times\{x\}) = Q(\{x\}).$$

Since $P(\{x\}), Q(\{x\})\in[0,1]$, it follows that

$$\mu(\{x\}\times(\Omega\backslash\{x\})) + \mu(\mu((\Omega\backslash\{x\})\times\{x\})) \geq |P(\{x\}-Q(\{x\})|\,.$$

Therefore, since $\rho(x,y)=0$ whenever $x=y$ and $\rho(x,y)\geq\mathrm{Sep}(\Omega)$ whenever $x\neq y$,

$$\begin{aligned}
\int_{\Omega\times\Omega}\rho^r(x,y)\,d\mu(x,y) &= \int_{\mathcal{X}\times\mathcal{X}}\rho^r(x,y)\,d\mu(x,y) \\
&= \sum_{x\in\mathcal{X}}\int_{\{x\}\times(\Omega\backslash\{x\})}\rho^r(x,y)\,d\mu(x,y) + \int_{(\Omega\backslash\{x\})\times\{x\}}\rho^r(x,y)\,d\mu(x,y) \\
&\geq (\mathrm{Sep}(\Omega))^r\sum_{x\in\mathcal{X}}\mu(\{x\}\times(\Omega\backslash\{x\})) + \mu((\Omega\backslash\{x\})\times\{x\}) \\
&\geq (\mathrm{Sep}(\Omega))^r\sum_{x\in\mathcal{X}}|P(\{x\})-Q(\{x\})|\,.
\end{aligned}$$

Taking the infimum over μ on both sides gives

$$(\mathrm{Sep}(\Omega))^r \sum_{x\in\mathcal{X}} |P(\{x\}) - Q(\{x\})| \leq W_r^r(P,Q).$$

To prove the upper bound, since ρ is upper bounded by $\mathrm{Diam}(\Omega)$, it suffices to construct a coupling μ that only moves mass into or out of each given point, but not both; that is, for each $x \in \mathcal{X}$,

$$\min\{\mu(\{x\} \times (\Omega\backslash\{x\})), \mu((\Omega\backslash\{x\}) \times \{x\})\} = 0.$$

One way of doing this is as follows. Fix an ordering $x_1, x_2, ...$ of the elements of $\mathcal{X}$.

For each $i \in \mathbb{N}$, define

$$X_i := \sum_{\ell=1}^{i} (P(x_\ell) - Q(x_\ell))_+ \quad \text{and} \quad Y_i := \sum_{\ell=1}^{i} (Q(x_\ell) - P(x_\ell))_+,$$

and further define

$$j_i := \min\{j \in \mathbb{N} : X_i \leq Y_j\} \quad \text{and} \quad k_i := \min\{k \in \mathbb{N} : X_j \geq Y_i\}.$$

Then, for each $i \in \mathbb{N}$, move X_i mass from $\{x_1, ..., x_i\}$ to $\{y_1, ..., y_{j_i}\}$ and move Y_i mass from $\{y_1, ..., y_i\}$ to $\{x_1, ..., x_{k_i}\}$. As $i \to \infty$, by construction of X_i and Y_i, the total mass moved in this way is

$$\mu((\mathcal{X} \times \mathcal{X})\backslash\{(x,x) : x \in \mathcal{X}\}) = \lim_{i\to\infty} X_i + Y_i = \sum_{x\in\mathcal{X}} |P(x) - Q(x)|.$$

■

Lemma 45. Let K be a positive integer. Suppose $\{\mathcal{S}_k\}_{k\in[K]}$ is a sequence of nested countable Borel partitions of (Ω, ρ), with $\mathcal{S}_0 = \Omega$. Then, for any $r \geq 1$ and any Borel probability distributions P and Q on Ω,

$$W_r^r(P,Q) \leq (\mathrm{Res}(\mathcal{S}_K))^r + \sum_{k=1}^{K} (\mathrm{Res}(\mathcal{S}_{k-1}))^r \left(\sum_{S\in\mathcal{S}_k} |P(S) - Q(S)| \right).$$

Proof: Our proof follows the same ideas as and slightly generalizes of the proof of Proposition 1 in Weed and Bach (2017). Intuitively, to prove Lemma 45 it suffices to find a transportation map such that For each $k \in [K]$, recursively define

$$P_k := P - \sum_{j=0}^{k-1} \mu_k \quad \text{and} \quad Q_k := Q - \sum_{j=0}^{k-1} \nu_k,$$

where, for each $k \in [K]$, μ_k and ν_k are Borel measures on Ω defined for any $E \in \Sigma$ by

$$\mu_k(E) := \sum_{S\in\mathcal{S}_k : P_k(S) > 0} (P_k(S) - Q_k(S))_+ \frac{P_k(E\cap S)}{P_k(S)}$$

and

$$\nu_k(E) := \sum_{S\in\mathcal{S}_k : Q_k(S) > 0} (Q_k(S) - P_k(S))_+ \frac{Q_k(E\cap S)}{Q_k(S)}.$$

By construction of μ_k and ν_k, each μ_k and ν_k is a non-negative measure and $\sum_{k=1}^{K} \mu_k \leq P$ and $\sum_{k=1}^{K} \nu_k \leq Q$. Furthermore, for each $k \in [K-1]$, for each $S \in \mathcal{S}_k$, $\mu_{k+1}(S) = \nu_{k+1}(S)$, and

$$\mu_k(\Omega) = \nu_k(\Omega) \leq \sum_{S \in \mathcal{S}_k} |P(S) - Q(S)| .$$

Consequently, although μ and ν are not probability measures, we can slightly generalize the definition of Wasserstein distance by writing

$$W_r^r(\mu_k, \nu_k) := \mu(\Omega) \inf_{\tau \in \Pi\left(\frac{\mu_k}{\mu_k(\Omega)}, \frac{\nu_k}{\nu_k(\Omega)}\right)} \mathop{\mathbb{E}}_{(X,Y)\sim\mu} [\rho^r(X, Y)]$$

(or $W_r^r(\mu_k, \nu_k) = 0$ if $\mu_k = \nu_k = 0$). In particular, this is convenient because we one can easily show that, by construction of the sequences $\{P_k\}_{k\in[K]}$ and $\{Q_k\}_{k\in[K]}$,

$$W_r^r(P, Q) \leq W_r^r(P_K, Q_K) + \sum_{k=1}^{K} W_r^r(\mu_k, \nu_k) . \tag{6.13}$$

For each $k \in [K]$, Lemma 42 implies that

$$\begin{aligned} W_r^r(\mu_k, \nu_k) &\leq \sum_{S \in \mathcal{S}_{k-1}} (\mathrm{Diam}(S))^r \sum_{T \in \mathcal{S}_k : T \subseteq S} |P(T) - Q(T)| \\ &\leq (\mathrm{Res}(\mathcal{S}_{k-1}))^r \sum_{S \in \mathcal{S}_{k-1}} \sum_{T \in \mathcal{S}_k : T \subseteq S} |P(T) - Q(T)| \\ &= (\mathrm{Res}(\mathcal{S}_{k-1}))^r \sum_{T \in \mathcal{S}_k} |P(T) - Q(T)| . \end{aligned}$$

Furthermore, for each $S \in \mathcal{S}_K$, $P_K = Q_K$, Lemma 41 gives that

$$W_r^r(P_K, Q_K) \leq (\mathrm{Res}(\mathcal{S}_K))^r$$

Plugging these last two inequalities into Inequality (6.13) gives the desired result:

$$W_r^r(P, Q) \leq (\mathrm{Res}(\mathcal{S}_K))^r + \sum_{k=1}^{K} (\mathrm{Res}(\mathcal{S}_{k-1}))^r \sum_{S \in \mathcal{S}_k} |P(S) - Q(S)| .$$

■

Lemma 46. Suppose $\mathcal{S}$ and $\mathcal{T}$ are partitions of (Ω, ρ), and suppose $\mathcal{S}$ is countable. Then, there exists a partition $\mathcal{S}'$ of (Ω, ρ) such that:

a) $|\mathcal{S}'| \leq |\mathcal{S}|$.

b) $\mathrm{Res}(\mathcal{S}') \leq \mathrm{Res}(\mathcal{S}) + 2\,\mathrm{Res}(\mathcal{T})$.

c) $\mathcal{T}$ is a refinement of $\mathcal{S}'$.

Proof: Enumerate the elements of $\mathcal{S}$ as $S_1, S_2, \ldots$. Define $S'_0 := \emptyset$, and then, for each $i \in \{1, 2, \ldots\}$, recursively define

$$S'_i := \left(\bigcup_{T \in \mathcal{T} : T \cap S_i \neq \emptyset} T \right) \setminus \left(\bigcup_{j=1}^{i-1} S'_j \right) ,$$

and set $\mathcal{S}' = \{S'_1, S'_2, ...\}$. Clearly, $|\mathcal{S}'| \leq |\mathcal{S}|$ (equality need not hold, as we may have some $S'_i = \emptyset$). By the triangle inequality, each

$$\mathrm{Diam}(S'_i) \leq \mathrm{Diam}\left(\bigcup_{T\in\mathcal{T}:T\cap S_i\neq\emptyset} T\right) \leq \delta_{\mathcal{S}} + 2\delta_T.$$

Finally, since $\mathcal{T}$ is a partition and we can write

$$S'_i = \left(\bigcup_{T\in\mathcal{T}:T\cap S_i\neq\emptyset} T\right) \setminus \left(\bigcup_{j=1}^{i-1} \bigcup_{T\in\mathcal{T}:T\cap S'_j\neq\emptyset} T\right),$$

$\mathcal{T}$ is a refinement of $\mathcal{S}'$. ■

6.10 Proof of Lower Bound

In this section, we provide a proof of our main lower bound, Theorem 39 in the main text. The proof consists of two main steps. First, we show that the minimax error of estimation in Wasserstein distance can be lower bounded by a product of two terms, one depending on the packing radius R and the other depending on the minimax risk of estimating a particular discrete (i.e., multinomial) distribution under $\mathcal{L}_1$ loss. The second step is then to apply a minimax lower bound on the risk of estimating a multinomial distribution under $\mathcal{L}_1$ loss. These two steps respectively rely on two lemmas, Lemma 50 and Lemma 51 given below.

The first lemma implies that, when a distribution P is supported on a finite subset $\mathcal{D}$ of the sample space, then there exists an estimator $\widehat{P}_{\mathcal{D}}$ of $\widehat{P}$ that is supported on $\mathcal{D}$ is minimax optimal, up to a small constant factor. While this fact is relatively obvious for measure-theoretic metrics such as $\mathcal{L}_p$ distances, it is somewhat less obvious for Wasserstein distances, which also depend on metric properties of the space. This observation is key to lower bounding the minimax rate in terms of the minimax rate for estimating a discrete distribution.

Lemma 50 (Wasserstein Projections). *Let $(\mathcal{X}, \rho)$ be a metric space and let $\mathcal{D} \subseteq \mathcal{X}$ be finite. Let $\mathcal{P}$ denote the family of all Borel probability distributions on $\mathcal{X}$, and let*

$$\mathcal{P}_{\mathcal{D}} := \{P \in \mathcal{P} : P(\mathcal{D}) = 1\}$$

denote the set of distributions supported only on $\mathcal{D}$. Suppose $P \in \mathcal{P}_{\mathcal{D}}$ and $Q \in \mathcal{P}$. Then,

$$\operatorname*{argmin}_{\widetilde{Q}\in\mathcal{P}_{\mathcal{D}}} W_r(Q, \widetilde{Q}) \neq \emptyset \quad \textit{and, for any} \quad Q' \in \operatorname*{argmin}_{\widetilde{Q}\in\mathcal{P}'} W_r(Q, \widetilde{Q}),$$

we have $W_r(P, Q') \leq 2W_r(P, Q)$.

Proof: Let $\{\mathcal{S}_x\}_{x\in\mathcal{D}}$ denote the Voronoi diagram of $\mathcal{X}$ with respect to $\mathcal{D}$; that is, for each $x \in \mathcal{D}$, let

$$\mathcal{S}_x := \{y \in \mathcal{X} : x \in \operatorname*{argmin}_{z\in\mathcal{D}} \rho(x, y)\}.$$

Since $\{S_x\}_{x\in\mathcal{D}}$ is a finite cover of $\mathcal{X}$, we can disjointify it (see Remark 36) while retaining the property that, for every $x \in \mathcal{D}$ and every $y \in \mathcal{S}_x$, $\rho(x, y) = \min_{z\in\mathcal{D}} \rho(z, y)$; hence, we assume without loss of generality that $\{\mathcal{S}_x\}_{x\in\mathcal{D}}$ is a partition of $\mathcal{X}$. Then, there is a unique distribution $Q' \in \mathcal{P}_D$ such that, for each $x \in \mathcal{D}$, $Q'(\{x\}) = Q(\mathcal{S}_x)$. It

is easy to see by definition of the Voronoi diagram that $Q' \in \operatorname{argmin}_{\widetilde{Q} \in \mathcal{P}_{\mathcal{D}}} W_r(Q, \widetilde{Q})$; the unique transportation map $\mu_* \in \Pi(Q, Q')$ such that each $\mu(\mathcal{S}_x, \{x\}) = Q(\mathcal{S}_x)$ clearly minimizes

$$\mathop{\mathbb{E}}_{(X,Y)\sim\mu} [\rho^r(X, Y)]$$

over all $\mu \in \bigcup_{\widetilde{Q} \in \mathcal{P}_{\mathcal{D}}} \Pi(Q, \widetilde{Q})$. Moreover, since $P \in \mathcal{P}_D$, by the triangle inequality and the definition of Q', $W_r(P, Q') \leq W_r(P, Q) + W_r(Q, Q') \leq 2W_r(P, Q)$. ∎

The second lemma is a simple minimax lower bound for the problem of estimating the mean vector of a multinomial distribution, under $\mathcal{L}_1$ loss.

Lemma 51 (Minimax Lower Bound for Mean of Multinomial Distribution)**.** *Suppose $k \leq 32n$. Let $p \in \Delta^k$, and suppose $X_1, ..., X_n \overset{IID}{\sim} \text{Categorical}(p_1, ..., p_k)$ are distributed IID according to a categorical distribution on $[k]$, with mean vector p. Then, we have the following minimax lower bound for estimating p under $\mathcal{L}^1$-loss:*

$$\inf_{\widehat{p}} \sup_{p \in \Delta^k} \mathbb{E}\left[\|p - \widehat{p}\|_1\right] \geq \frac{3 \log 2}{4096} \sqrt{\frac{k-1}{n}},$$

where the infimum is taken over all estimators (i.e., all (potentially randomized) functions $\widehat{p} : [k]^n \to \Delta^k$ of the data).

Note that, while the above result is phrased for categorical distributions to simplify notation in the proof, the result is equivalent to a statement for multinomial distributions, since $\sum_{i=1}^n X_i \sim \text{Multinomial}(n, p_1, ..., p_k)$ and $X_1, ..., X_n$ are assumed to be IID and therefore exchangeable.

Proof: We follow a standard procedure for proving minimax lower bounds based on Fano's inequality, as outlined in Section 2.6 of Tsybakov (2008).

Let $p_0 = (1/k, ..., 1/k) \in \Delta^K$ denote the uniform vector in Δ^k. Let $\mathcal{I} := \left[\lfloor \frac{k}{2} \rfloor\right]$. For each $j \in \mathcal{I}$, define $\phi_j : [k] \to \mathbb{R}^k$ by

$$\phi_j := 1_{\{2j-1\}} - 1_{\{2j\}},$$

and, for each $\tau \in \{-1, 1\}^{\mathcal{I}}$, define

$$p_\tau := p_0 + \frac{c}{k} \sum_{j \in \mathcal{I}} \tau_j \phi_j,$$

where

$$c = \frac{1}{16} \sqrt{\frac{k-1}{n} \log 2} \leq \frac{1}{2}.$$

Note that, since $|c| \leq 1$ and, for each $j \in \mathcal{I}$, $\sum_{\ell \in [k]} \phi_j(\ell) = 0$, each $p_\tau \in \Delta^k$. Observe that, for any $\tau, \tau' \in \{-1, 1\}^{\mathcal{I}}$, we have

$$\|p_\tau - p_{\tau'}\|_1 = \frac{4c\omega(\tau, \tau')}{k}, \quad \text{where} \quad \omega(\tau, \tau') = \sum_{i \in \mathcal{I}} 1_{\{\tau_i \neq \tau'_i\}}$$

denotes the Hamming distance between τ and τ'. By the Varshamov-Gilbert bound (see, e.g., Lemma 2.9 of Tsybakov (2008)), there exists a subset $T \subseteq \{-1, 1\}^{\mathcal{I}}$ such that $\log |T| \geq \frac{\lfloor k/2 \rfloor \log 2}{8}$ and, for every $\tau, \tau' \in T$,

$$\omega(\tau, \tau') \geq \frac{|\mathcal{I}|}{8} = \frac{\lfloor k/2 \rfloor}{8}, \quad \text{and hence} \quad \|p_\tau - p_{\tau'}\|_1 \geq c \frac{\lfloor k/2 \rfloor}{2k}.$$

Also, for any $\tau \in T$,

$$\begin{aligned} D_{KL}(p_\tau^n, p_0^n) &= nD_{KL}(p_\tau, p_0) \\ &= n \sum_{j\in[k]} p_{\tau,j} \log\left(\frac{p_{\tau,j}}{p_{0,j}}\right) \\ &= n \sum_{j\in\mathcal{I}} p_{\tau,2j-1} \log\left(\frac{p_{\tau,2j-1}}{1/k}\right) + p_{\tau,2j} \log\left(\frac{p_{\tau,2j}}{1/k}\right) \\ &= \frac{n|\mathcal{I}|}{k}\left((1-c)\log(1-c) + (1+c)\log(1+c)\right) \end{aligned}$$

One can check (e.g., by Taylor expansion) that, for any $c \in (0, 1/2)$,

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

Thus, since $|\mathcal{I}| \le k/2$,

$$D_{KL}(p_\tau^n, p_0^n) \le \frac{2n|\mathcal{I}|c^2}{k} \le nc^2.$$

It follows that from the choice of c (and noting that, by the assumptions that $k \le 32n$, $c \in (0, 1/2)$) that

$$\frac{1}{|T|}\sum_{\tau\in T} D_{KL}(p_\tau^n, p_0^n) \le nc^2 \le \frac{\lfloor k/2\rfloor \log 2}{128} \le \frac{1}{16}\log|T|.$$

Therefore, by Fano's method for lower bounds (see, e.g., Theorem 2.5 of Tsybakov (2008), with $\alpha = 1/16$ and

$$s := \frac{c}{16} \le c\frac{\lfloor k/2\rfloor}{4k} \le \frac{1}{2}\|p_\tau - p_{\tau'}\|_1,$$

we have

$$\begin{aligned} \inf_{\widehat{p}} \sup_{p\in\Delta^k} \mathbb{E}\left[\|p - \widehat{p}\|_1\right] &\ge \inf_{\widehat{p}} \sup_{p\in\Delta^k} c\frac{\lfloor k/2\rfloor}{4k}\mathbb{P}\left[\|p-\widehat{p}\|_1 \ge c\frac{\lfloor k/2\rfloor}{4k}\right] \\ &\ge c\frac{\lfloor k/2\rfloor}{4k}\frac{3}{16} \\ &\ge \frac{3\log 2}{4096}\sqrt{\frac{k-1}{n}}. \end{aligned}$$

■

Theorem 39. Let (Ω, ρ) be a metric space, and let $\mathcal{P}$ denote the set of Borel probability measures on (Ω, ρ).

$$\inf_{\widehat{P}:\mathcal{X}^n\to\mathcal{P}} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P}\left[W_r^r(P, \widehat{P}(X_1,\ldots,X_n))\right] \ge c_r \sup_{k\in[32n]} R^r(k)\sqrt{\frac{k-1}{n}},$$

where

$$c_r = \frac{3\log 2}{4096\cdot 2^r}.$$

is independent of n and the infimum is taken over all estimators (i.e., all (potentially randomized) functions $\widehat{P} : \mathcal{X}^n \to \mathcal{P}$ of the data).

Proof: Let $k \leq 32n$, and let $\mathcal{D}$ be an $R(k)$-packing $\mathcal{D}$ of (Ω, ρ) with $|\mathcal{D}| = k$. Let $\mathcal{P}_{\mathcal{D}}$ denote the class of (discrete) distributions over $\mathcal{D}$.

By Lemma 42, Lemma 50, Lemma 51, and the definition of the packing radius (in that order)

$$\begin{aligned}
\inf_{\widehat{P}:\mathcal{X}^n\to\mathcal{P}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[W_r^r(\widehat{P}, P)\right] &\geq (\mathrm{Sep}(\mathcal{D}))^r \inf_{\widehat{P}:\mathcal{X}^n\to\mathcal{P}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\|\widehat{P} - P\|_1\right] \\
&\geq (\mathrm{Sep}(\mathcal{D}))^r \inf_{\widehat{P}:\mathcal{X}^n\to\mathcal{P}} \sup_{P\in\mathcal{P}_{\mathcal{D}}} \mathbb{E}\left[\|\widehat{P} - P\|_1\right] \\
&\geq \left(\frac{\mathrm{Sep}(\mathcal{D})}{2}\right)^r \inf_{\widehat{P}:\mathcal{X}^n\to\mathcal{P}_{\mathcal{D}}} \sup_{P\in\mathcal{P}_{\mathcal{D}}} \mathbb{E}\left[\|\widehat{P} - P\|_1\right] \\
&\geq \frac{3\log 2}{4096\cdot 2^r} (\mathrm{Sep}(\mathcal{D}))^r \sqrt{\frac{|\mathcal{D}|-1}{n}} \\
&\geq \frac{3\log 2}{4096\cdot 2^r} R^r(k) \sqrt{\frac{k-1}{n}}.
\end{aligned}$$

The theorem follows by taking the supremum over $k \leq 32n$ on both sides. ∎

6.11 Proofs of Minimax Lower Bound in terms of Moment Bounds

In this section, we prove a second lower bound theorem (Theorem 40), for the case of distributions with unbounded support and bounded moments.

Theorem 40. Suppose $r, \ell, \mu > 0$ are constants, and fix $x_0 \in \Omega$. Let $\mathcal{P}_{\ell,x_0}(\mu)$ denote the family of distributions P on Ω with ℓ^{th} moment $\mu_{\ell,x_0}(P) \leq \mu$ around x_0 at most μ. Let $n \geq \frac{3\mu}{2}$ and assume there exists $x_1 \in \Omega$ such that $\rho(x_0, x_1) = n^{1/\ell}$. Then,

$$M(r, \mathcal{P}_{\ell,x_0}(\mu)) \geq C_\mu n^{\frac{r-\ell}{\ell}},$$

where $C_\mu := \frac{\min\{\mu, 2/3\}}{24}$ is constant in n.

Proof: First, note a standard lemma for minimax lower bounds, which we reiterate in the case of Wasserstein distances:

Lemma 52 (Theorem 2.1 of Tsybakov (2009), Wasserstein Case)**.** *Assume there exist $P_0, P_1 \in \mathcal{P}$ such that $P_0 \ll P_1$ and $W_r^r(P_0, P_1) \geq 2s > 0$ such that $D_{KL}(P_0^n, P_1^n) \leq \frac{1}{2}$. Then,*

$$\inf_{\widehat{P}:\Omega\to\mathcal{P}} \sup_{P\in\mathcal{P}} \mathbb{P}\left[W_r^r\left(\widehat{P}, P\right) \geq s\right] \geq \frac{1}{2} P_1\left(\frac{dP_0}{dP_1}(x) \geq 1\right).$$

We now construct appropriate P_0 and P_1 to plug into the above lemma. Define

$$\epsilon := \frac{\min\{\mu, 2/3\}}{2n} \in (0, 1/3],$$

and consider distinguishing between two discrete distributions

$$P_0 := (1-\epsilon)\,\delta_{x_0} + \epsilon\delta_{x_1} \quad \text{and} \quad P_1 := (1-2\epsilon)\,\delta_{x_0} + 2\epsilon\delta_{x_1}$$

where δ_x denotes a unit point mass at x. Since, $\epsilon \in [0, 1/2]$, P_0 and P_1 are both probability distributions. Moreover, $\mu_{\ell,x_0}(P_0) = \epsilon n \leq \mu/2$, and $\mu_{\ell,x_0}(P_1) = 2\epsilon n \leq \mu$,

so that $P_0, P_1 \in \mathcal{P}_{\ell,x_0}(\mu)$. Note that, since $\epsilon \leq 1/3$, by the inequality $\log(1+x) \leq x$, we have

$$(1-\epsilon)\log\frac{1-\epsilon}{1-2\epsilon} = (1-\epsilon)\log\left(1+\frac{\epsilon}{1-2\epsilon}\right) \leq (1-\epsilon)\frac{\epsilon}{1-2\epsilon} \leq 2\epsilon.$$

Therefore,

$$\begin{aligned} D_{KL}\left(P_0^n, P_1^n\right) = nD_{KL}\left(P_0, P_1\right) &= n\left(P_0(x_0)\log\frac{P_0(x_0)}{P_1(x_0)} + P_0(x_1)\log\frac{P_0(x_1)}{P_1(x_1)}\right) \\ &= n\left((1-\epsilon)\log\frac{1-\epsilon}{1-2\epsilon} + \epsilon\log\frac{\epsilon}{2\epsilon}\right) \leq n\left(2\epsilon - \epsilon\log 2\right) \leq \frac{1}{2}, \end{aligned}$$

since $2 - \log 2 \leq 3/2$. Finally, note that

$$W_r^r\left(P_0, P_1\right) = \epsilon n^{r/\ell} = \min\left\{\frac{\mu}{2}, \frac{1}{3}\right\} n^{\frac{r-\ell}{\ell}}.$$

Plugging P_0 and P_1 into Lemma 52 with $s = \min\left\{\frac{\mu}{4}, \frac{1}{6}\right\} n^{\frac{r-\ell}{\ell}}$ thus gives

$$\inf_{\widehat{P}:\Omega\to\mathcal{P}_{\ell,x_0}(\mu)} \sup_{P\in\mathcal{P}_{\ell,x_0}(\mu)} \mathbb{P}\left[W_r^r\left(\widehat{P}, P\right) \geq s\right] \geq \frac{1}{2}P_1\left(x_0\right) = \frac{1-2\epsilon}{2} \geq 1/6.$$

Thus,

$$M(r, \mathcal{P}_{\ell,x_0}(\mu)) = \inf_{\widehat{P}:\Omega\to\mathcal{P}_{\ell,x_0}(\mu)} \sup_{P\in\mathcal{P}_{\ell,x_0}(\mu)} \mathop{\mathbb{E}}_{X_1^n \overset{IID}{\sim} P}\left[W_r^r\left(\widehat{P}, P\right)\right] \geq \frac{s}{6} = \frac{\min\{\mu, 2/3\}}{24} n^{\frac{r-\ell}{\ell}}.$$

■

Chapter 7

Distribution Estimation under Adversarial Losses

7.1 Introduction

This chapter, as with the previous chapter, studies the fundamental problem of estimating a probability distribution, with an emphasis on using a variety of different measures of losses. While the previous chapter studied r-Wasserstein distances, which directly measures the average distance that we must transport mass to transform the estimate into the true distribution, the present chapter studies the maximum, over functions f in a pre-selected "discriminator class" $\mathcal{F}$, of the difference in the expectations of f under the estimated and true distributions, a distance which we refer to as an "adversarial loss". Both Wasserstein and adversarial losses are quite flexible, and can be tailored to be used with many different types of data and notions of error – for Wasserstein distances, this flexibility is achieved is through specification of the underlying metric ρ, whereas, for adversarial losses, this is achieved through specification of the class $\mathcal{F}$. In adversarial losses, the role of distance in the sample space can be indirectly but precisely captured by properties, such as smoothness, of the functions in $\mathcal{F}$ – smoother classes $\mathcal{F}$ are more tolerant of "short-distance" or "local" errors.

As in the previous chapter, we focus on minimax rates of distribution estimation. The main difference in this chapter is the incorporation of distributional (density) smoothness – we obtain improved rates when the true data distribution is assumed to be smooth. To achieve this, rather than simply using the empirical distribution, we use series estimators, essentially regularizing the Fourier series of the estimate. Interestingly, in the Fourier basis, contributions of smoothness in the distribution class and smoothness in the discriminator class $\mathcal{F}$ interact very nicely, leading to a rather straightforward analysis.

This chapter also includes a careful mathematical formalization of the notion of "implicit density estimation" (traditionally known simply as "sampling"), recently popularized by deep generative models such as variational autoencoders (VAEs) and generative adversarial networks (GANs). We formally show that convergence rates, in both minimax and other senses, for density estimation and for implicit generative modeling, are, under very mild conditions, essentially identical. Importantly, this means that the recent empirical successes of implicit generative models such as GANs and VAEs cannot be attributed, at least from a statistical perspective, to the ease of sampling over estimation. The difference may, in our view, have more to do with the computational tractability of these models.

Note that, since this work was published (in NeurIPS 2018), the term Integral Probability Metric (IPM) had superseded "adversarial loss" as the most common name for these losses. In this chapter, we stick with "adversarial loss", as in the

original paper, but, in the remainder of the thesis, we often use the term "IPM" to mean the same thing.

7.2 Background

Generative modeling, that is, modeling the distribution from which data are drawn, is a central task in machine learning and statistics. Often, prior information is insufficient to guess the form of the data distribution. In statistics, generative modeling in these settings is usually studied from the perspective of nonparametric density estimation, in which histogram, kernel, orthogonal series, and nearest-neighbor methods are popular approaches with well-understood statistical properties (Wasserman, 2006; Tsybakov, 2009; Efromovich, 2010; Biau and Devroye, 2015b).

Recently, machine learning has made significant empirical progress in generative modeling, using such tools as generative adversarial networks (GANs) and variational autoencoders (VAEs). Computationally, these methods are quite distinct from classical density estimators; they usually rely on deep neural networks, fit by black-box optimization, rather than a mathematically prescribed smoothing operator, such as convolution with a kernel or projection onto a finite-dimensional subspace.

Ignoring the implementation of these models, from the perspective of statistical analysis, these recent methods have at least two main differences from classical density estimators. First, they are *implicit*, rather than *explicit* (or *prescriptive*) generative models (Diggle and Gratton, 1984; Mohamed and Lakshminarayanan, 2016); that is, rather than an estimate of the probability of a set or the density at a point, they return novel samples from the data distribution. Second, in many recent models, loss is measured not with $\mathcal{L}^p$ distances (as is conventional in nonparametric statistics (Wasserman, 2006; Tsybakov, 2009)), but rather with weaker losses, such as

$$d_{\mathcal{F}_D}(P, Q) = \sup_{f \in \mathcal{F}_D} \left| \mathop{\mathbb{E}}_{X \sim P}[f(X)] - \mathop{\mathbb{E}}_{X \sim Q}[f(X)] \right|, \tag{7.1}$$

where $\mathcal{F}_D$ is a *discriminator class* of bounded, Borel-measurable functions, and P and Q lie in a *generator class* $\mathcal{F}_G$ of Borel probability measures on a sample space $\mathcal{X}$. Figure 7.1 shows two examples of discriminator functions between distributions, corresponding to the Wasserstein and Gaussian MMD distances. Importantly, GANs often use losses of this form because the function class $\mathcal{F}$ in (7.1) can be approximated by a discriminator neural network.

This work attempts to help bridge the gap between traditional nonparametric statistics and these recent advances by studying these two differences from a statistical minimax perspective. Specifically, under traditional statistical smoothness assumptions, we identify (i.e., prove matching upper and lower bounds on) minimax convergence rates for density estimation under several losses of the form (7.1). We also discuss some consequences this has for particular neural network implementations of GANs based on these losses. Finally, we study connections between minimax rates for explicit and implicit generative modeling, under a plausible notion of risk for implicit generative models.

7.2.1 Adversarial Losses

The quantity (7.1) has been extensively studied, in the case that $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS) under the name *maximum mean discrepancy* (MMD; (Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Tolstikhin, Sriperumbudur,

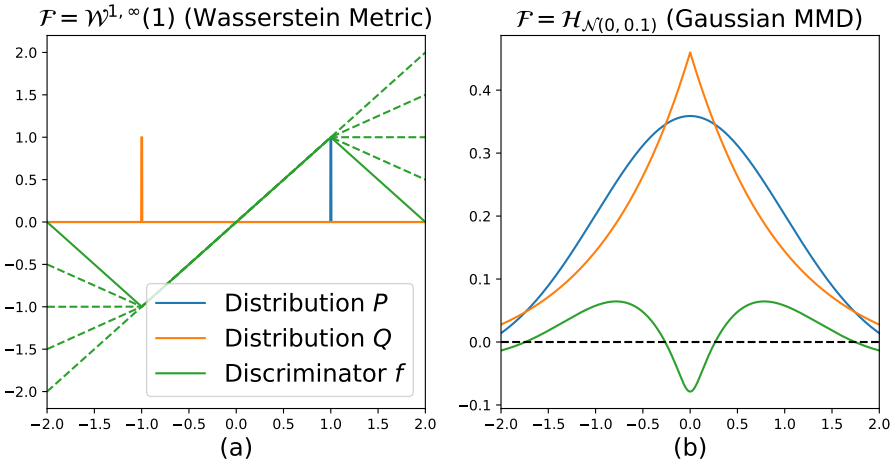


FIGURE 7.1: Examples of probability distributions P and Q and corresponding discriminator functions f^*. In (a), P and Q are single Dirac masses at $+1$ and -1, respectively, and $\mathcal{F}$ is the 1-Lipschitz class, so that $d_{\mathcal{F}}$ is the Wasserstein metric. In (b), P and Q are standard Gaussian and standard Laplace distributions, respectively, and $\mathcal{F}$ is a ball in an RKHS with a Gaussian kernel, so that $d_{\mathcal{F}}$ is the Gaussian Maximum Mean Discrepancy (MMD).

and Muandet, 2017)), and, in a wider context under the name *integral probability metric* (IPM; (Müller, 1997; Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet, 2010b; Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet, 2012; Bottou, Arjovsky, Lopez-Paz, and Oquab, 2018)). (Arora, Ge, Liang, Ma, and Zhang, 2017) also called (7.1) the *$\mathcal{F}_D$-distance*, or, when $\mathcal{F}_D$ is a family of functions that can be implemented by a neural network, the *neural network distance*. We settled on the name "adversarial loss" because, without assuming any structure on $\mathcal{F}_D$, this matches the intuition of the expression (7.1), namely that of an adversary selecting the most distinguishing linear projection $f \in \mathcal{F}_D$ between the true density P and our estimate $\widehat{P}$ (e.g., by the discriminator network in a GAN).

One can check that $d_{\mathcal{F}_D} : \mathcal{F}_G \times \mathcal{F}_G \to [0, \infty]$ is a pseudometric (i.e., it is nonnegative and satisfies the triangle inequality, and $d_{\mathcal{F}_D}(P, Q) > 0 \Rightarrow P \neq Q$, although $d_{\mathcal{F}_D}(P, Q) = 0 \not\Rightarrow P = Q$ unless $\mathcal{F}_D$ is sufficiently rich). Many popular (pseudo)metrics between probability distributions, including $\mathcal{L}^p$ (Wasserman, 2006; Tsybakov, 2009), Sobolev (Leoni, 2017; Mroueh, Li, Sercu, Raj, and Cheng, 2017), maximum mean discrepancy (MMD; (Tolstikhin, Sriperumbudur, and Muandet, 2017))/energy (Székely, Rizzo, and Bakirov, 2007; Ramdas, Trillos, and Cuturi, 2017), total variation (Villani, 2008), (1-)Wasserstein/Kantorovich-Rubinstein (Kantorovich and Rubinstein, 1958; Villani, 2008), Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1948), and Dudley (Dudley, 1972; Abbasnejad, Shi, and Hengel, 2018) metrics can be written in this form, for appropriate choices of $\mathcal{F}_D$.

The **main contribution of this chapter** is a statistical analysis of the problem of estimating a distribution P from n IID observations using the loss $d_{\mathcal{F}_D}$, in a minimax sense over $P \in \mathcal{F}_G$, for fairly general nonparametric smoothness classes $\mathcal{F}_D$ and $\mathcal{F}_G$. General upper and lower bounds are given in terms of decay rates of coefficients of functions in terms of an (arbitrary) orthonormal basis of $\mathcal{L}^2$ (including, e.g., Fourier or wavelet bases); note that this does *not* require $\mathcal{F}_D$ or $\mathcal{F}_G$ to have any inner product structure, only that $\mathcal{F}_D \subseteq \mathcal{L}^1$. We also discuss some consequences for density estimators based on neural networks (such as GANs), and consequences for the closely

related problem of implicit generative modeling (i.e., of generating novel samples from a target distribution, rather than estimating the distribution itself), in terms of which GANs and VAEs are usually cast.

Chapter Organization: Section 7.3 provides our formal problem statement and required notation. Section 7.4 discusses related work on nonparametric density estimation. Sections 7.5 and 7.6 contain our main theoretical upper and lower bound results, with proofs in Sections 7.12 and 7.13, respectively. Section 7.7 develops our general results from Sections 7.5 and 7.6 into concrete minimax convergence rates for some important special cases. Section 7.8 uses our theoretical results to upper bound the error of perfectly optimized GANs. Section 7.9 establishes some theoretical relationships between the convergence of optimal density estimators and optimal implicit generative models. The final sections provide proofs of our theoretical results, further applications, further discussion of related and future work on the theory of GANs, and small experiments on simulated data that validate our theoretical results.

7.3 Problem Statement and Notation

We now provide a formal statement of the problem studied in this chapter in a very general setting, and then define notation required for our specific results.

Formal Problem Statement: Let $P \in \mathcal{F}_G$ be an unknown probability measure on a sample space $\mathcal{X}$, from which we observe n IID samples $X_{1:n} = X_1, ..., X_n \overset{IID}{\sim} P$. We are interested in using the samples $X_{1:n}$ to estimate the measure P, with error measured using the adversarial loss $d_{\mathcal{F}_D}$. Specifically, for various choices of spaces $\mathcal{F}_D$ and $\mathcal{F}_G$, we seek to bound the minimax rate

$$M(\mathcal{F}_D, \mathcal{F}_G) := \inf_{\widehat{P}} \sup_{P \in \mathcal{F}_G} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D} \left(P, \widehat{P}(X_{1:n}) \right) \right]$$

of estimating distributions assumed to lie in a class $\mathcal{F}_G$, where the infimum is taken over all estimators $\widehat{P}$ (i.e., all (potentially randomized) functions $\widehat{P} : \mathcal{X}^n \to \mathcal{F}_G$). We will discuss both the case when $\mathcal{F}_G$ is known *a priori* and the *adaptive* case when it is not.

7.3.1 Notation

For a non-negative integer n, we use $[n] := \{1, 2, ..., n\}$ to denote the set of positive integers at most n. For sequences $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$ of non-negative reals, $a_n \lesssim b_n$ and, similarly $b_n \gtrsim a_n$, indicate the existence of a constant $C > 0$ such that $\limsup_{n \to \infty} \frac{a_n}{b_n} \leq C$. $a_n \asymp b_n$ indicates $a_n \lesssim b_n \lesssim a_n$. For functions $f : \mathbb{R}^d \to \mathbb{R}$, we write

$$\lim_{\|z\| \to \infty} f(z) := \sup_{\{z_n\}_{n \in \mathbb{N}} : \|z_n\| \to \infty} \lim_{n \to \infty} f(z_n),$$

where the supremum is taken over all diverging $\mathbb{R}^d$-valued sequences. Note that, by equivalence of finite-dimensional norms, the exact choice of the norm $\|\cdot\|$ does not matter here. We will also require summations of the form $\sum_{z \in \mathcal{Z}} f(z)$ in cases where $\mathcal{Z}$ is a (potentially infinite) countable index set and $\{f(z)\}_{z \in \mathcal{Z}}$ is summable but not necessarily absolutely summable. Therefore, to ensure that the summation is well-defined, the order of summation will need to be specified, depending on the application (as in, e.g., Section 7.7).

Fix the sample space $\mathcal{X} = [0,1]^d$ to be the d-dimensional unit cube, over which λ denotes the usual Lebesgue measure. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, let, for any Borel measure μ on $\mathcal{X}$, $p \in [1, \infty]$, and $L > 0$,

$$\|f\|_{\mathcal{L}^p_\mu} := \left(\int_{\mathcal{X}} |f|^p \, d\mu \right)^{1/p} \quad \text{and} \quad \mathcal{L}^p_\mu(L) := \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{L}^p_\mu} < L \right\}$$

(taking the appropriate limit if $p = \infty$) denote the Lebesgue norm and ball of radius L, respectively.

Fix an orthonormal basis $\mathcal{B} = \{\phi_z\}_{z \in \mathcal{Z}}$ of $\mathcal{L}^2_\lambda$ indexed by a countable family $\mathcal{Z}$. To allow probability measures P without densities (i.e., $P \not\ll \mu$), we assume each basis element $\phi_z : \mathcal{X} \to \mathbb{R}$ is a bounded function, so that $\widetilde{P}_z := \mathbb{E}_{X \sim P}[\phi_z(X)]$ is well-defined. For constants $L > 0$ and $p \geq 1$ and real-valued net $\{a_z\}_{z \in \mathcal{Z}}$, our results pertain to generalized ellipses of the form

$$\mathcal{H}_{p,a}(L) = \left\{ f \in \mathcal{L}^1(\mathcal{X}) : \left(\sum_{z \in \mathcal{Z}} a_z^p |\widetilde{f}_z|^p \right)^{1/p} \leq L \right\}.$$

(where $\widetilde{f}_z := \int_{\mathcal{X}} f \phi_z \, d\mu$ is the z^{th} coefficient of f in the basis $\mathcal{B}$). We sometimes omit dependence on L (e.g., $\mathcal{H}_{p,a} = \mathcal{H}_{p,a}(L)$) when its value does not matter (e.g., when discussing *rates* of convergence).

A particular case of interest is the scale of the Sobolev spaces defined for $s, L \geq 0$ and $p \geq 1$ by

$$\mathcal{W}^{s,p}(L) = \left\{ f \in \mathcal{L}^1(\mathcal{X}) : \left(\sum_{z \in \mathcal{Z}} |z|^{sp} |\widetilde{f}_z|^p \right)^{1/p} \leq L \right\}.$$

For example, when $\mathcal{B}$ is the standard Fourier basis and s is an integer, for a constant factor c depending only on s and the dimension d,

$$\mathcal{W}^{s,p}(cL) := \left\{ f \in \mathcal{L}^p_\lambda \middle| \left\| f^{(s)} \right\|_{\mathcal{L}^p_\lambda} < L \right\}$$

corresponds to the natural standard smoothness class of $\mathcal{L}^p_\lambda$ functions having s^{th}-order (weak) derivatives $f^{(s)}$ in $\mathcal{L}^p_\lambda(L)$ (Leoni, 2017)).

7.4 Related Work

Our results apply directly to many of the losses that have been used in GANs, including 1-Wasserstein distance (Arjovsky, Chintala, and Bottou, 2017; Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville, 2017), MMD (Li, Chang, Cheng, Yang, and Póczos, 2017), Sobolev distances (Mroueh, Li, Sercu, Raj, and Cheng, 2017), and the Dudley metric (Abbasnejad, Shi, and Hengel, 2018). As discussed in Section 7.15.2, slightly different assumptions are required to obtain results for the Jensen-Shannon divergence (used in the original GAN formulation of (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio, 2014)) and other f-divergences (Nowozin, Cseke, and Tomioka, 2016).

Given their generality, our results relate to many prior works on distribution estimation, including classical work in nonparametric statistics and empirical process theory, as well as more recent work studying Wasserstein distances and MMD.

Here, we briefly survey known results for these problems. There have also been a few other statistical analyses of the GAN framework; we discuss these works further in Section 7.11.

$\mathcal{L}^2_\lambda$ distances: Classical work on nonparametric statistics has typically focused on the problem of smooth density estimation under $\mathcal{L}^2_\lambda$ loss, corresponding the adversarial loss $d_{\mathcal{F}_D}$ with $\mathcal{F}_D = \mathcal{L}^2_\lambda(L_D)$ (the Hölder dual) of $\mathcal{L}^2$ (Wasserman, 2006; Tsybakov, 2009). In this case, when $\mathcal{F}_G = \mathcal{W}^{t,2}(L_G)$ is a Sobolev class, then the minimax rate is typically $M(\mathcal{F}_D, \mathcal{F}_G) \asymp n^{-\frac{t}{2t+d}}$, matching the rates given by our main results.

Maximum Mean Discrepancy (MMD): When $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS), the adversarial loss $d_{\mathcal{F}_D}$ has been widely studied under the name *maximum mean discrepancy (MMD)* (Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Tolstikhin, Sriperumbudur, and Muandet, 2017). When the RKHS kernel is translation-invariant, one can express $\mathcal{F}_D$ in the form $\mathcal{H}_{2,a}$, where a is determined by the spectrum of the kernel, and so our analysis holds for MMD losses with translation-invariant kernels (see Example 11). To the best of our knowledge, minimax rates for density estimation under MMD loss have not been established in general; our analysis suggests that density estimation under an MMD loss is essentially equivalent to the problem of estimating kernel mean embeddings studied in (Tolstikhin, Sriperumbudur, and Muandet, 2017), as both amount to density estimation while ignoring bias, and both typically have a parametric $n^{-1/2}$ minimax rate. Note that the related problems of estimating MMD itself, and of using it in statistical tests for homogeneity and dependence, have received extensive theoretical treatment (Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Ramdas, Reddi, Póczos, Singh, and Wasserman, 2015).

Wasserstein Distances: When $\mathcal{F}_D = \mathcal{W}^{1,\infty}(L)$ is the class of 1-Lipschitz functions, $d_{\mathcal{F}_D}$ is equivalent to the *(order-1) Wasserstein* (also called *earth-mover's* or *Kantorovich-Rubinstein*) distance. In this case, when $\mathcal{F}_G$ contains all Borel measurable distributions on $\mathcal{X}$, minimax bounds have been established under very general conditions (essentially, when the sample space $\mathcal{X}$ is an arbitrary totally bounded metric space) in terms of covering numbers of $\mathcal{X}$ (Weed and Bach, 2017; Singh and Póczos, 2018; Lei, 2018). In the particular case that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ of full dimension (i.e., having non-empty interior, comparable to the case $\mathcal{X} = [0,1]^d$ that we study here), these results imply a minimax rate of $M(\mathcal{F}_D, \mathcal{F}_G) = n^{-\min\left\{\frac{1}{2}, \frac{1}{d}\right\}}$, matching our rates. Notably, these upper bounds are derived using the empirical distribution, which *cannot* benefit from smoothness of the true distribution (see (Weed and Bach, 2017)). At the same time, it is obvious to generalize smoothing estimators to sample spaces that are not sufficiently nice subsets of $\mathbb{R}^d$.

Sobolev IPMs: The closest work to the present is (Liang, 2017), which we believe was the first work to analyze how convergence rates jointly depend on (Sobolev) smoothness restrictions on both $\mathcal{F}_D$ and $\mathcal{F}_G$. Specifically, for Sobolev spaces $\mathcal{F}_D = \mathcal{W}^{s,p}$ and $\mathcal{F}_G = \mathcal{W}^{t,q}$ with $p, q \geq 2$ (compare our Example 10), they showed

$$n^{-\frac{s+t}{2t+d}} \lesssim M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \lesssim n^{-\frac{s+t}{2(s+t)+d}}. \tag{7.2}$$

Our main results in Sections 7.5 and 7.6 improve on this in two main ways. First, our results generalize to and are tight for many spaces besides Sobolev spaces. Examples include when $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS) with translation-invariant kernel, or when $\mathcal{F}_G$ is the class of all Borel probability measures. Our bounds also allow other (e.g., wavelet) estimators, whereas the bounds of (Liang,

2017) are for the (uniformly $\mathcal{L}_\lambda^\infty$-bounded) Fourier basis. Second, the lower and upper bounds in (7.2) diverge by a factor polynomial in n. We tighten the upper bound to match the lower bound, identifying, for the first time, minimax rates for many problems of this form (e.g., $M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \asymp n^{-\frac{s+t}{2t+d}}$ in the Sobolev case above). Our analysis has several interesting implications:

1. When $s > d/2$, the convergence becomes *parametric*: $M(W^{s,2}, \mathcal{F}_G) \asymp n^{-1/2}$, for *any class of distributions* $\mathcal{F}_G$. This highlights that the loss $d_{\mathcal{F}_D}$ is quite weak for large s, and matches known minimax results for the Wasserstein case $s = 1$ (Canas and Rosasco, 2012; Singh and Póczos, 2018).
2. Our upper bounds, as in (Liang, 2017), are for smoothing estimators (namely, the orthogonal series estimator 7.3). In contrast, previous analyses of Wasserstein loss focused on convergence of the (unsmoothed) empirical distribution $\widehat{P}_E$ to the true distribution, which typically occurs at rate of $\asymp n^{-1/d} + n^{-1/2}$, where d is the intrinsic dimension of the support of P (Canas and Rosasco, 2012; Weed and Bach, 2017; Singh and Póczos, 2018). Moreover, if $\mathcal{F}_G$ includes all Borel probability measures, this rate is minimax optimal (Singh and Póczos, 2018). The loose upper bound of (Liang, 2017) left open the questions of whether (when $s < d/2$) a very small amount ($t \in \left(0, \frac{2s^2}{d-2s}\right]$) of smoothness improves the minimax rate and, more importantly, whether smoothed estimators are outperformed by $\widehat{P}_E$ in this regime. Our results imply that, for $s < d/2$, the minimax rate strictly improves with smoothness t, and that, as long as the support of P has full dimension, the smoothed estimator *always* converges faster than $\widehat{P}_E$. An important open problem is to simultaneously leverage when P is smooth *and* has support of low intrinsic dimension; many data (e.g., images) likely enjoy both these properties.
3. (Liang, 2017) suggested over-smoothing the estimate (the smoothing parameter ζ discussed in Equation (7.3) below was set to $\zeta \asymp n^{\frac{1}{2(s+t)+d}}$) compared to the case of $\mathcal{L}_\lambda^2$ loss, and hence it was not clear how to design estimators that adapt to unknown smoothness under losses $d_{W^{s,p}}$. We show that the optimal smoothing ($\zeta \asymp n^{\frac{1}{2t+d}}$) under $d_{W^{s,p}}$ loss is identical to that under $\mathcal{L}_\lambda^2$ loss, and we use this to design an adaptive estimator (see Corollary 56).
4. Our bounds imply improved performance bounds for optimized GANs, discussed in Section 7.8.

7.5 Upper Bounds for Orthogonal Series Estimators

This section gives upper bounds on the adversarial risk of the following density estimator. For any finite set $Z \subseteq \mathcal{Z}$, let $\widehat{P}_Z$ be the truncated series estimate

$$\widehat{P}_Z := \sum_{z \in Z} \widehat{P}_z \phi_z, \quad \text{where, for any } z \in \mathcal{Z}, \quad \widehat{P}_z := \frac{1}{n} \sum_{i=1}^{n} \phi_z(X_i). \tag{7.3}$$

Z is a tuning parameter that typically corresponds to a smoothing parameter; for example, when $\mathcal{B}$ is the Fourier basis and $Z = \{z \in \mathbb{Z}^d : \|z\|_\infty \leq \zeta\}$ for some $\zeta > 0$, $\widehat{P}_Z$ is equivalent to a kernel density estimator using a sinc product kernel $K_h(x) = \prod_{j=1}^d \frac{2}{h} \frac{\sin(2\pi x/h)}{2\pi x/h}$ with bandwidth $h = 1/\zeta$ (Owen, 2007).

We now present our main upper bound on the minimax rate of density estimation under adversarial losses. The upper bound is given by the orthogonal series

estimator given in Equation (7.3), but we expect kernel and other standard linear density estimators to converge at the same rate.

Theorem 53 (Upper Bound). *Suppose that $\mu(\mathcal{X}) < \infty$ and there exist constants $L_D, L_G > 0$, real-valued nets $\{a_z\}_{z\in\mathcal{Z}}$, $\{b_z\}_{z\in\mathcal{Z}}$ such that $\mathcal{F}_D = \mathcal{H}_{p,a}(\mathcal{X}, L_D)$ and $\mathcal{F}_G = \mathcal{H}_{q,b}(\mathcal{X}, L_G)$, where $p, q \geq 1$. Let $p' = \frac{p}{p-1}$ denote the Hölder conjugate of p. Then, for any $P \in \mathcal{F}_G$,*

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] \leq L_D \frac{c_{p'}}{\sqrt{n}} \left\| \left\{ \frac{\|\phi_z\|_{\mathcal{L}^\infty_P}}{a_z} \right\}_{z\in Z} \right\|_{p'} + L_D L_G \left\| \left\{ \frac{1}{a_z b_z} \right\}_{z\in\mathcal{Z}\setminus Z} \right\|_{\frac{1}{1-1/p-1/q}} \tag{7.4}$$

The two terms in the bound (7.4) demonstrate a bias-variance tradeoff, in which the first term (*variance*) increases with the truncation set Z and is typically independent of the class $\mathcal{F}_G$ of distributions, while the second term (*bias*) decreases with Z at a rate depending on the complexity of $\mathcal{F}_G$.

Corollary 54 (Sufficient Conditions for Parametric Rate). *Consider the setting of Theorem 53. If*

$$A := \sum_{z\in\mathcal{Z}} \frac{\|\phi_z\|^2_{\mathcal{L}^\infty_P}}{a_z^2} < \infty \quad \text{and} \quad \max\{a_z, b_z\} \to \infty.$$

whenever $\|z\| \to \infty$, then, the minimax rate is parametric; specifically, $M(\mathcal{F}_D, \mathcal{F}_G) \leq L_D\sqrt{A/n}$. In particular, letting $c_z := \sup_{x\in\mathcal{X}} |\phi_z(x)|$ for each $z \in \mathcal{Z}$, this occurs whenever $\sum_{z\in\mathcal{Z}} \frac{c_z^2}{a_z^2} < \infty$.

In many contexts (e.g., if $P \ll \lambda$ and $\lambda \ll P$), the simpler condition $\sum_{z\in\mathcal{Z}} \frac{c_z^2}{a_z^2} < \infty$ suffices. The first, and slightly weaker condition in terms of $\|\phi_z\|^2_{\mathcal{L}^\infty_P}$ is useful when we restrict $\mathcal{F}_G$; e.g., if $\mathcal{B}$ is the Haar wavelet basis and $\mathcal{F}_G$ contains only discrete distributions supported on at most k points, then $\|\phi_{i,j}\|^2_{\mathcal{L}^\infty_P} = 0$ for all but k values of $j \in [2^i]$, at each resolution $i \in \mathbb{N}$. The assumption $\max\left\{\lim_{\|z\|\to\infty} a_z, \lim_{\|z\|\to\infty} b_z\right\} = \infty$ is quite mild; for example, the Riemann-Lebesgue lemma and the assumption that $\mathcal{F}_D$ is bounded in $\mathcal{L}^\infty_\lambda \subseteq \mathcal{L}^1_\lambda$ together imply that this condition always holds if $\mathcal{B}$ is the Fourier basis.

7.6 Minimax Lower Bound

In this section, we lower bound the minimax risk $M(\mathcal{F}_D, \mathcal{F}_G)$ of distribution estimation under $d_{\mathcal{F}_D}$ loss over $\mathcal{F}_G$, for the case when $\mathcal{F}_D = \mathcal{H}_{p,a}$ and $\mathcal{F}_G := \mathcal{H}_{q,b}$ are generalized ellipses. As we show in some examples in Section 7.7, our lower bound rate matches our upper bound rate in Theorem 53 for many spaces $\mathcal{F}_D$ and $\mathcal{F}_G$ of interest. Our lower bound also suggests that the assumptions in Corollary 54 are typically necessary to guarantee the parametric convergence rate $n^{-1/2}$.

Theorem 55 (Minimax Lower Bound). *Fix $\mathcal{X} = [0,1]^d$, and let p_0 denote the uniform density (with respect to Lebesgue measure) on $\mathcal{X}$. Suppose $\{p_0\} \cup \{\phi_z\}_{z\in\mathcal{Z}}$ is an orthonormal basis in $\mathcal{L}^2_\mu$, and $\{a_z\}_{z\in\mathcal{Z}}$ and $\{b_z\}_{z\in\mathcal{Z}}$ are two real-valued nets. Let $L_D, L_G \geq 0$ and $p, q \geq 2$. For any $Z \subseteq \mathcal{Z}$, let*

$$A_Z := |Z|^{1/2} \sup_{z\in Z} a_z \quad \text{and} \quad B_Z := |Z|^{1/2} \sup_{z\in Z} b_z.$$

Then, for $\mathcal{F}_D = \mathcal{H}_{p,a}(L_D)$ *and* $\mathcal{F}_G := \mathcal{H}_{q,b}(L_G)$*, for any* $Z \subseteq \mathcal{Z}$ *satisfying*

$$B_Z \geq 16 L_G \sqrt{\frac{n}{\log 2}} \quad \text{and} \quad 2\frac{L_G}{B_Z} \sum_{z \in Z} \|\phi_z\|_{\mathcal{L}^\infty_\mu} \leq 1, \tag{7.5}$$

we have $M(\mathcal{F}_D, \mathcal{F}_G) \geq \dfrac{L_G L_D |Z|}{64 A_Z B_Z} = \dfrac{L_G L_D}{64 \left(\sup_{z \in Z} a_z\right) \left(\sup_{z \in Z} b_z\right)}$.

As in most minimax lower bounds, our proof relies on constructing a finite set Ω_G of "worst-case" densities in $\mathcal{F}_G$, lower bounding the distance $d_{\mathcal{F}_D}$ over Ω_G, and then letting elements of Ω_G shrink towards the uniform distribution p_0 at a rate such that the average information (here, Kullback-Leibler) divergence between each $p \in \Omega_G$ and p_0 does not grow with n. The first condition in (7.5) ensures that the information divergence between each $p \in \Omega_G$ and p_0 is sufficiently small, and typically results in tuning of Z identical (in rate) to its optimal tuning in the upper bound (Theorem 53).

The second condition in (7.5) is needed to ensure that the "worst-case" densities we construct are everywhere non-negative. Hence, this condition is not needed for lower bounds in the Gaussian sequence model, as in Theorem 2.3 of (Liang, 2017). However, failure of this condition (asymptotically) corresponds to the breakdown point of the asymptotic equivalence between the Gaussian sequence model and the density estimation model in the regime of very low smoothness (e.g., in the Sobolev setting, when $t < d/2$; see (Brown and Zhang, 1998)), and so finer analysis is needed to establish lower bounds here.

7.7 Examples

In this section, we apply our bounds from Sections 7.5 and 7.6 to compute concrete minimax convergence rates for two examples choices of $\mathcal{F}_D$ and $\mathcal{F}_G$, namely Sobolev spaces and reproducing kernel Hilbert spaces.

For the purpose of this section, suppose that $\mathcal{X} = [0, 2\pi]^d$, $\mathcal{Z} = \mathbb{Z}^d$, and, for each $z \in \mathcal{Z}$, ϕ_z is the z^{th} standard Fourier basis element given by $\phi_z(x) = e^{i\langle z, x\rangle}$ for all $x \in \mathcal{X}$. In this case, we will always choose the truncation set Z to be of the form $Z := \{z \in \mathcal{Z} : \|z\|_\infty \leq \zeta\}$, for some $\zeta > 0$, so that $|Z| \leq \zeta^d$. Moreover, for every $z \in Z$, $\|\phi_z\|_{\mathcal{L}^\infty_\mu} = 1$, and hence $C_Z \leq 1$.

Example 10 (Sobolev Spaces). Suppose that, for some $s, t \geq 0$, $a_z = \|z\|^s_\infty$ and $b_z = \|z\|^t_\infty$. Then, setting $\zeta = n^{\frac{1}{2t+d}}$ in Theorems 53 and 55 gives that there exist constants $C > c > 0$ such that

$$c n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}} \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}\right) \leq C n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}. \tag{7.6}$$

Combining the observation that the s-Hölder space $\mathcal{W}^{s,\infty} \subseteq \mathcal{W}^{s,2}$ with the lower bound (over $\mathcal{W}^{s,\infty}$) in Theorem 3.1 of (Liang, 2017), we have that (7.6) also holds when $\mathcal{W}^{s,2}$ is replaced with $\mathcal{W}^{s,p}$ for any $p \in [2, \infty]$ (e.g., in the case of the Wasserstein metric $d_{\mathcal{W}^{1,\infty}}$).

So far, we have assumed the smoothness t of the true distribution P is known, and used that to tune the parameter ζ of the estimator. However, in reality, t is not known. In the next result, we leverage the fact that the rate-optimal choice $\zeta = n^{\frac{1}{2t+d}}$ above does not rely on the loss parameters s, together with Theorem 53 to construct an *adaptively minimax estimator*, i.e., one that is minimax and fully-data dependent. There is a large literature on adaptive nonparametric density estimation under $\mathcal{L}^2_\mu$

loss; see (Efromovich, 2010) for accessible high-level discussion and (Goldenshluger and Lepski, 2014) for a technical but comprehensive review.

Corollary 56 (Adaptive Upper Bound for Sobolev Spaces). *There exists an adaptive choice $\widehat{\zeta} : \mathcal{X}^n \to \mathbb{N}$ of the hyperparameter ζ (independent of s, t), such that, for any $s, t \geq 0$, there exists a constant $C > 0$ (independent of n), such that*

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_{Z_{\widehat{\zeta}(X_{1:n})}} \right) \right] \leq M \left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2} \right) \tag{7.7}$$

The actual construction of the adaptive $\widehat{\zeta}$ is presented in Section 7.14, but, in brief, it is a standard construction based on leave-one-out cross-validation under $\mathcal{L}^2_\mu$ loss which is known (e.g., see Sections 7.2.1 and 7.5.2 of (Massart, 2007)) to be adaptively minimax under $\mathcal{L}^2_\mu$ loss. Using the fact that our upper bound Theorem 53 uses a choice of ζ is independent of the loss parameter s, we show that the $d_{\mathcal{W}^{s,\infty}}$ risk of $\widehat{P}_\zeta$ can be factored into its $\mathcal{L}^2_\mu$ risk and a component (ζ^{-s}) that is independent of t. Since $\mathcal{L}^2_\mu$ risk can be rate-minimized in independently of t, it follows that the $d_{\mathcal{W}^{s,\infty}}$ risk can be rate-minimized independently of t. Adaptive minimaxity then follows from Theorem 55.

Example 11 (Reproducing Kernel Hilbert Space/MMD Loss). Suppose $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Aronszajn, 1950; Berlinet and Thomas-Agnan, 2011). If k is translation invariant (i.e., there exists $\kappa \in \mathcal{L}^2_\mu$ such that, for all $x, y \in \mathcal{X}$, $k(x, y) = \kappa(x - y)$), then Bochner's theorem (see, e.g., Theorem 6.6 of (Wendland, 2004)) implies that, up to constant factors,

$$\mathcal{H}_k(L) := \{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq L \} = \left\{ f \in \mathcal{H}_k : \sum_{z \in \mathcal{Z}} |\widetilde{\kappa}_z|^2 |\widetilde{f}_z|^2 < L^2 \right\}.$$

Thus, in the setting of Theorem 53, we have $\mathcal{H}_k = \mathcal{H}_{2,a}$, where $a_z = |\widetilde{\kappa}_z|$ satisfies $\sum_{z \in \mathcal{Z}} a_z^{-2} = \|\kappa\|^2_{\mathcal{L}^2_\mu} < \infty$. Corollary 54 then gives $M(\mathcal{H}_k(L_D), \mathcal{F}_G) \leq L_D \|\kappa\|_{\mathcal{L}^2_\mu} n^{-1/2}$ for *any class* $\mathcal{F}_G$. It is well-known known that MMD can always be *estimated* at the parametric rate $n^{-1/2}$ (Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012); however, to the best of our knowledge, only recently has it been shown that any probability distribution can be estimated at the rate $n^{-1/2}$ under MMD loss(Sriperumbudur, 2016), emphasizing the fact that MMD is a very weak metric. This has important implications for applications such as two-sample testing (Ramdas, Reddi, Póczos, Singh, and Wasserman, 2015).

7.8 Consequences for Generative Adversarial Neural Networks (GANs)

This section discusses implications of our minimax bounds for GANs. Neural networks in this section are assumed to be fully-connected, with rectified linear unit (ReLU) activations. (Liang, 2017) used their upper bound result (7.2) to prove a similar theorem, but, since their upper bound was loose, the resulting theorem was also loose. The following results are immediate consequences of our improvement (Theorem 53) over the upper bound (7.2) of (Liang, 2017), and so we refer to that paper for the proof. Key ingredients are an oracle inequality proven in (Liang, 2017), an

upper bound such as Theorem 53, and bounds of (Yarotsky, 2017) on the size of a neural network needed to approximate functions in a Sobolev class.

In the following, $\mathcal{F}_D$ denotes the set of functions that can be encoded by the discriminator network and $\mathcal{F}_G$ denotes the set of distributions that can be encoded by the generator network. $P_n := \frac{1}{n}\sum_{i=1}^n 1_{\{X_i\}}$ denotes the empirical distribution of the observed data $X_{1:n} \overset{IID}{\sim} P$.

Theorem 57 (Improvement of Theorem 3.1 in Liang (2017))**.** *Let $s, t > 0$, and fix a desired approximation accuracy $\epsilon > 0$. Then, there exists a GAN architecture, in which*

1. *the discriminator $\mathcal{F}_D$ has at most $O(\log(1/\epsilon))$ layers and $O(\epsilon^{-d/s}\log(1/\epsilon))$ parameters,*
2. *and the generator $\mathcal{F}_G$ has at most $O(\log(1/\epsilon))$ layers and $O(\epsilon^{-d/t}\log(1/\epsilon))$ parameters,*

such that, if $\widehat{P}_*(X_{1:n}) := \underset{\widehat{P}\in\mathcal{F}_G}{\operatorname{argmin}} d_{\mathcal{F}_D}\left(P_n, \widehat{P}\right)$, is the optimized GAN estimate of P,

$$\text{then} \quad \sup_{P\in\mathcal{W}^{t,2}} \mathbb{E}_{X_{1:n}}\left[d_{\mathcal{W}^{s,2}}\left(P, \widehat{P}_*(X_{1:n})\right)\right] \leq C\left(\epsilon + n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}\right).$$

The discriminator and generator in the above theorem can be implemented as described in (Yarotsky, 2017). The assumption that the GAN is perfectly optimized may be strong; see (Nagarajan and Kolter, 2017; Liang and Stokes, 2018) for discussion of this.

Though we do not present this result here, we can similarly improve the upper bound of (Liang, 2017) (their Theorem 3.2) for very deep neural networks, further improving on the previous state-of-the-art bounds of (Anthony and Bartlett, 2009) (which did not leverage smoothness assumptions on P).

7.9 Minimax Comparison of Explicit and Implicit Generative Models

In this section, we draw formal connections between our work on density estimation (explicit generative modeling) and the problem of implicit generative modeling under an appropriate measure of risk. In the sequel, we fix a class $\mathcal{F}_G$ of probability measures on a sample space $\mathcal{X}$ and a loss function $\ell : \mathcal{F}_G \times \mathcal{F}_G \to [0, \infty]$ measuring the distance of an estimate $\widehat{P}$ from the true distribution P. ℓ need not be an adversarial loss $d_{\mathcal{F}_D}$, but our discussion does apply to all ℓ of this form.

7.9.1 A Minimax Framework for Implicit Generative Models

Thus far, we have analyzed the *minimax risk of density estimation*, namely

$$M_D(\mathcal{F}_G, \ell, n) = \inf_{\widehat{P}} \sup_{P\in\mathcal{F}_G} R_D(P, \widehat{P}), \text{ where } R_D(P, \widehat{P}) = \mathbb{E}_{X_{1:n}\overset{IID}{\sim}P}\left[\ell(P, \widehat{P}(X_{1:n}))\right] \quad (7.8)$$

denotes the *density estimation risk of $\widehat{P}$ at P* and the infimum is taken over all estimators (i.e., (potentially randomized) functions $\widehat{P} : \mathcal{X}^n \to \mathcal{F}_G$). Whereas density estimation is a classical statistical problem to which we have already contributed novel results, our motivations for studying this problem arose from a desire to better understand recent work on implicit generative modeling.

Implicit generative models, such as GANs (Arjovsky, Chintala, and Bottou, 2017; Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio, 2014) and VAEs (Kingma and Welling, 2014; Rezende, Mohamed, and Wierstra,

2014), address the problem of *sampling*, in which we seek to construct a *generator* that produces novel samples from the distribution P (Mohamed and Lakshminarayanan, 2016). In our context, a generator is a function $\widehat{X} : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ that takes in n IID samples $X_{1:n} \sim P$ and a source of randomness (a.k.a., *latent variable*) $Z \sim Q_Z$ with known distribution Q_Z (independent of $X_{1:n}$) on a space $\mathcal{Z}$, and returns a novel sample $\widehat{X}(X_{1:n}, Z) \in \mathcal{X}$.

The evaluating the performance of implicit generative models, both in theory and in practice, is difficult, with solutions continuing to be proposed Sutherland, Tung, Strathmann, De, Ramdas, Smola, and Gretton, 2017, some of which have proven controversial. Some of this controversy stems from the fact that many of the most straightforward evaluation objectives are optimized by a trivial generator that 'memorizes' the training data (e.g., $\widehat{X}(X_{1:n}, Z) = X_Z$, where Z is uniformly distributed on $[n]$). One objective that can avoid this problem is as follows. For simplicity, fix the distribution Q_Z of the latent random variable $Z \sim Q_Z$ (e.g., $Q_Z = \mathcal{N}(0, I)$). For a fixed training set $X_{1:n} \overset{IID}{\sim} P$ and latent distribution $Z \sim Q_Z$, we define the *implicit distribution of a generator* $\widehat{X}$ as the conditional distribution $P_{\widehat{X}(X_{1:n},Z)|X_{1:n}}$ over $\mathcal{X}$ of the random variable $\widehat{X}(X_{1:n}, Z)$ given the training data. Then, for any $P \in \mathcal{F}_G$, we define the *implicit risk of* $\widehat{X}$ *at* P by

$$R_I(P, \widehat{X}) := \mathop{\mathbb{E}}_{X_{1:n} \sim P} \left[\ell(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}}) \right].$$

We can then study the *minimax risk of sampling*,

$$M_I(\mathcal{F}_G, \ell, n) := \inf_{\widehat{X}} \sup_{P \in \mathcal{F}_G} R_I(P, \widehat{X}).$$

A few remarks about $M_I(\mathcal{F}, \ell, n)$: First, we implicitly assumed $\ell(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}})$ is well-defined, which is not obvious unless $P_{\widehat{X}(X_{1:n},Z)} \in \mathcal{F}_G$. We discuss this assumption further below. Second, since the risk $R_I(P, \widehat{X})$ depends on the unknown true distribution P, we cannot calculate it in practice. Third, for the same reason (because $R_P(P, \widehat{X})$ depends directly on P rather than particular data $X_{1:n}$), it detect lack-of-diversity issues such as mode collapse. As we discuss later, these latter two points are distinctions from the recent work of (Arora, Ge, Liang, Ma, and Zhang, 2017) on generalization in GANs.

7.9.2 Comparison of Explicit and Implicit Generative Models

Algorithmically, sampling is a very distinct problem from density estimation; for example, many computationally efficient Monte Carlo samplers rely on the fact that a function *proportional* to the density of interest can be computed much more quickly than the exact (normalized) density function (Chib and Greenberg, 1995). In this section, we show that, given unlimited computational resources, the problems of density estimation and sampling are equivalent in a minimax statistical sense. Since exactly minimax estimators ($\operatorname{argmin}_{\widehat{P}} \sup_{P \in \mathcal{F}_G} R_D(P, \widehat{P})$) often need not exist, the following weaker notion is useful for stating our results:

Definition 58 (Nearly Minimax Sequence)**.** A sequence $\{\widehat{P}_k\}_{k \in \mathbb{N}}$ of density estimators (resp., $\{\widehat{X}_k\}_{k \in \mathbb{N}}$ of generators) is called *nearly minimax over* $\mathcal{F}_G$ if $\lim_{k \to \infty} \sup_{P \in \mathcal{F}_G} R_{P,D}(\widehat{P}_k) = M_D(\mathcal{F}_G, \ell, n)$ (resp., $\lim_{k \to \infty} \sup_{P \in \mathcal{F}_G} R_{P,I}(\widehat{X}_k) = M_I(\mathcal{F}_G, \ell, n)$).

The following theorem identifies sufficient conditions under which, in the statistical minimax framework described above, density estimation is no harder than sampling. The idea behind the proof is as follows: If we have a good sampler $\widehat{X}$ (i.e., with $R_I(\widehat{X})$ small), then we can draw m 'fake' samples from $\widehat{X}$. We can use these 'fake' samples to construct a density estimate $\widehat{P}$ of the implicit distribution of $\widehat{X}$ such that, under the technical assumptions below, $R_D(\widehat{P}) - R_I(\widehat{X}) \to 0$ as $m \to \infty$.

Theorem 59 (Conditions under which Density Estimation is Statistically no harder than Sampling)**.** *Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Suppose*

(A1) *$\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is non-negative, and there exists $C_\triangle > 0$ such that, for all $P_1, P_2, P_3 \in \mathcal{F}_G$, $\ell(P_1, P_3) \leq C_\triangle (\ell(P_1, P_2) + \ell(P_2, P_3))$.*

(A2) *$M_D(\mathcal{F}_G, \ell, m) \to 0$ as $m \to \infty$.*

(A3) *For all $m \in \mathbb{N}$, we can draw m IID samples $Z_1, ..., Z_m \overset{IID}{\sim} Q_Z$ of the latent variable Z.*

(A4) *there exists a nearly minimax sequence of samplers $\widehat{X}_k : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ such that, for each $k \in \mathbb{N}$, almost surely over $X_{1:n}$, $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}} \in \mathcal{F}_G$.*

Then, $M_D(\mathcal{F}_G, \ell, n) \leq C_\triangle M_I(\mathcal{F}_G, \ell, n)$.

Assumption (A1) is a generalization of the triangle inequality (and reduces to the triangle inequality when $C_\triangle = 1$). This weaker assumption applies, for example, when ℓ is the Jensen-Shannon divergence (with $C_\triangle = 2$) used in the original GAN formulation of (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio, 2014), even though this does not satisfy the triangle inequality (Endres and Schindelin, 2003)). Assumption (A2) is equivalent to the existence of a uniformly ℓ-risk-consistent estimator over $\mathcal{F}_G$, a standard property of most distribution classes $\mathcal{F}_G$ over which density estimation is studied (e.g., our Theorem 53).

Assumption (A3) is a natural design criterion of implicit generative models; usually, Q_Z is a simple parametric distribution such as a standard normal.

Finally, Assumption (A4) is the most mysterious, because, currently, little is known about the minimax theory of samplers when $\mathcal{F}_G$ is a large space. On one hand, since $M_I(\mathcal{F}_G, \ell, n)$ is an infimum over $\widehat{X}$, Theorem 59 continues to hold if we restrict the class of samplers (e.g., to those satisfying Assumption (A4) or those we can compute). On the other hand, even without restricting $\widehat{X}$, this assumption may not be too restrictive, because nearly minimax samplers are necessarily close to $P \in \mathcal{F}_G$. For example, if $\mathcal{F}_G$ contains only smooth distributions but $\widehat{X}$ is the trivial empirical sampler described above, then $\ell(P, P_{\widehat{X}})$ should be large and $\widehat{X}$ is unlikely to be minimax optimal.

Finally, in practice, we often do not know estimators that are nearly minimax for finite samples, but may have estimators that are rate-optimal (e.g., as given by Theorem 53), i.e., that satisfy

$$C := \limsup_{n \to \infty} \frac{\sup_{P \in \mathcal{F}_G} R_I(P, \widehat{X})}{M_I(\mathcal{F}_G, \ell, n)} < \infty.$$

Under this weaker assumption, it is straightforward to modify our proof to conclude that

$$\limsup_{n \to \infty} \frac{M_D(\mathcal{F}_G, \ell, n)}{M_I(\mathcal{F}_G, \ell, n)} \leq C_\triangle C.$$

The converse result ($M_D(\mathcal{F}_G, \ell, n) \geq M_I(\mathcal{F}_G, \ell, n)$) is simple to prove in many cases, and is related to the well-studied problem of Monte Carlo sampling (Robert, 2004); we discuss this briefly later.

7.10 Conclusions

Given the recent popularity of implicit generative models in many applications, it is important to theoretically understand why these models appear to outperform classical methods for similar problems. This chapter provided new minimax bounds for density estimation under adversarial losses, both with and without adaptivity to smoothness, and gave several applications, including both traditional statistical settings and perfectly optimized GANs. We also gave simple conditions under which minimax bounds for density estimation imply bounds for the problem of implicit generative modeling, suggesting that sampling is typically not *statistically* easier than density estimation. Thus, for example, the strong curse of dimensionality that is known to afflict to nonparametric density estimation Wasserman (2006) should also limit the performance of implicit generative models such as GANs. Section 7.17 describes several specific avenues for further investigation, including whether the curse of dimensionality can be avoided when data lie on a low-dimensional manifold.

7.11 Further Related Work

As noted previously, our problem setting is quite general, and thus overlaps with several previous settings that have been studied. First, we note the analysis of (Liu, Bousquet, and Chaudhuri, 2017), which also studied convergence of distribution estimation under adversarial losses. Considering a somewhat broader class of non-metric losses (including, e.g., Jensen-Shannon divergence), which they call *adversarial divergences*, (Liu, Bousquet, and Chaudhuri, 2017) provided consistency results (in distribution) for a number of GAN formulations, assuming convergence of the min-max GAN optimization problem to a generator-optimal equilibrium. However, they did not study rates of convergence.

Our results can also be viewed as a refinement of several results from empirical process and learning theory, especially the wealth of literature on the case where $\mathcal{F}_D$ is a Glivenko-Cantelli (GC, a.k.a., Vapnik-Chervonenkis (VC)) class (Pollard, 1990). Corollary 54 can be interpreted as showing that spaces $\mathcal{F}_D$ that are sufficiently small in terms of orthonormal basis expansions are $n^{-1/2}$-uniformly GC/VC classes (Alon, Ben-David, Cesa-Bianchi, and Haussler, 1997; Vapnik and Chervonenkis, 2015). In particular, this gives a simple functional-analytic proof of this property for the general case when $\mathcal{F}_D$ is a ball in a translation-invariant RKHS. On the other hand, some related results, cast in terms of fat-shattering dimensions (Mendelson, 2002; Dziugaite, Roy, and Ghahramani, 2015), appear to lead to slower rates for RKHSs.

Glivenko-Cantelli classes are defined without regards to the class $\mathcal{F}_G$ of possible distributions. However, the more interesting consequences of our results are for the case that $\mathcal{F}_G$ is restricted, as in Theorem 53. In Example 10 this allowed us to characterize the interaction between smoothness constraints on the discriminator class $\mathcal{F}_D$ and the generator class $\mathcal{F}_G$, showing in particular, that, when $\mathcal{F}_D$ is large, restricting $\mathcal{F}_G$ improves convergence rates. Aside for the results of (Liang, 2017) and many results for the specific case $\mathcal{F}_D = \mathcal{L}^2_\lambda$, we do not know of any results that show this.

Several prior works have studied the closely related problem of estimating certain adversarial metrics, including $\mathcal{L}^2$ distance (Krishnamurthy, Kandasamy, Poczos, and Wasserman, 2015), MMD (Gretton, Borgwardt, Rasch, Schölkopf, and Smola,

2012), Sobolev distances (Singh, Sriperumbudur, and Póczos, 2018b), and others (Sriperumbudur, Fukumizu, Gretton, Schölkopf, and Lanckriet, 2012). In some cases, these metrics can themselves be estimated far more efficiently than the underlying distribution under that loss, and these estimators have various applications including two-sample/homogeneity and independence testing (Anderson, Hall, and Titterington, 1994; Gretton, Borgwardt, Rasch, Schölkopf, and Smola, 2012; Ramdas, Reddi, Póczos, Singh, and Wasserman, 2015), and distributional (Sutherland, 2016), transfer (Du, Koushik, Singh, and Póczos, 2017), and transductive (Quadrianto, Petterson, and Smola, 2009) learning.

There has also been some work studying the min-max optimization problem in terms of which GANs are typically cast (Nagarajan and Kolter, 2017; Liang and Stokes, 2018). However, in this work, as in (Liu, Bousquet, and Chaudhuri, 2017; Liang, 2017), we implicitly assume the optimization procedure has converged to a generator-optimal equilibrium. Another work that studies adversarial losses is (Bottou, Arjovsky, Lopez-Paz, and Oquab, 2018), which focuses on a comparison of Wasserstein distance and MMD in the context of implicit generative modeling.

7.11.1 Other statistical analyses of GANs

Our results are closely related to some previous work studying the *generalization error* of GANs under MMD (Dziugaite, Roy, and Ghahramani, 2015) or Jensen-Shannon divergence, Wasserstein, or other adversarial losses (Arora, Ge, Liang, Ma, and Zhang, 2017).

Assume, for simplicity, that ℓ satisfies a weak triangle inequality (Assumption (A1) above), and let P denote the true distribution from which the data are drawn IID. Then, we can bound the true loss $\ell(P, \widehat{P})$ of an estimator $\widehat{P}$ in terms of the approximation error $\ell(P, P_*)$ (corresponding to bias) and generalization error $\ell(P_*, \widehat{P})$ (i.e., corresponding to variance):

$$\ell(P, \widehat{P}) \leq C_{\triangle} \left(\ell(P, P_*) + \ell(P_*, \widehat{P}) \right),$$

where $P_* := \operatorname{argmin}_{Q \in \widehat{\mathcal{F}}} \ell(P, Q)$ denotes the optimal approximation of P in some restricted class $\widehat{\mathcal{F}} \subseteq \mathcal{F}_G$ of estimators in which $\widehat{P}$ lies.

Bounding the approximation error $\ell(P, P_*)$ typically requires restricting the space $\mathcal{F}_G$ in which P lies. Theorem 1 of (Dziugaite, Roy, and Ghahramani, 2015) and Theorem 3.1 of (Arora, Ge, Liang, Ma, and Zhang, 2017) focus on bounding the generalization error $\ell(P_*, \widehat{P})$, and thus avoid making such assumptions on P. However, our Theorem 53 shows that, when $\mathcal{F}_D$ is sufficiently small (e.g., an RKHS, as in (Dziugaite, Roy, and Ghahramani, 2015)), $\ell = d_{\mathcal{F}_D}$ is so weak that $\ell(P, P_*)$ can be bounded even when $\mathcal{F}_G$ includes *all* probability measures. In particular, while (Dziugaite, Roy, and Ghahramani, 2015) gave only high-probability bounds of order $n^{-1/2}$ on the *generalization error* $\ell(P_*, \widehat{P})$ in terms of the fat-shattering dimension of the RKHS, we show that, for any RKHS with a translation-invariant kernel, the *total* risk $\mathbb{E}[\ell(P, \widehat{P})]$ can be bounded at the parametric rate of $n^{-1/2}$.

(Arora, Ge, Liang, Ma, and Zhang, 2017) also showed that, if $\widehat{\mathcal{F}}$ is too large (specifically, if $\widehat{\mathcal{F}}$ contains the empirical distribution), then the generalization error $\ell(P_*, \widehat{P})$ (or, specifically, an empirical estimate thereof) need not vanish as the sample size increases, or, in the case of Wasserstein distance, if the dimension d grows faster than logarithmically with the sample size n. Our Theorem 53 showed that,

if $\widehat{F}$ contains only (e.g., orthogonal series) estimates of a fixed smoothness (e.g., orthogonal series estimates with a fixed ζ), then the generalization error decays at the rate $\asymp \zeta^{d/2} n^{-1/2}$ (the first term on the right-hand side of 7.4), so that $d \in o(\log n)$ is still necessary[1]. Our minimax lower bound 55 suggests that, without making significantly stronger assumptions, we cannot hope to avoid this curse of dimensionality, at least without sacrificing approximation error (bias).

7.12 Proof of Upper Bound

In this section, we prove our main upper bound, Theorem 53. We begin with a simple lemma showing that, under mild assumptions, we can write an adversarial loss in terms of an $\mathcal{L}^2_\lambda$ basis expansion.

Lemma 60 (Basis Expansion of Adversarial Loss)**.** *Consider a class $\mathcal{F}_D$ of discriminator functions, two probability distributions P and Q, and an orthonormal basis $\{\phi_z\}_{z\in\mathcal{Z}}$ of $\mathcal{L}^2_\lambda(\mathcal{X})$. Moreover, suppose that either of the following conditions holds:*

1. *$P, Q \ll \lambda$ have densities $p, q \in \mathcal{L}^2_\lambda$.*
2. *For every $f \in \mathcal{F}_D$, the expansion of f in the basis $\mathcal{B}$ converges uniformly (over $\mathcal{X}$) to f. That is,*

$$\lim_{Z \uparrow \mathcal{Z}} \sup_{x\in\mathcal{X}} \left| f(x) - \sum_{z\in Z} \widetilde{f}_z(x)\phi_z(x) \right| \to 0.$$

Then, we can expand the adversarial loss $d_{\mathcal{F}_D}$ over $\mathcal{P}$ as

$$d_{\mathcal{F}_D}(P, Q) = \sup_{f\in\mathcal{F}_D} \sum_{z\in\mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z \right).$$

Condition 1 above is quite straightforward, and would be taken for granted in most classical non-parametric analysis. When $\mathcal{B}$ is the Fourier basis, the assumption that $p, q \in \mathcal{L}^r_\mu$ for $r = 2$ can be weakened to any $r > 1$ using Hölder's inequality together with the facts that $f \in \mathcal{L}^{r'}$ and that Fourier series converge in $\mathcal{L}^{r'}$ (where $r' = \frac{r}{r-1}$ denote the Hölder conjugate of r).

Since we are also interested in probability distributions that lack density functions, we provide the fairly mild Condition 2 as an alternative. As an example of this condition in the Fourier case, suppose $\mathcal{F}_D$ is uniformly equi-continuous, say, with modulus of continuity $\omega : [0, \infty) \to [0, \infty)$ satisfying $\omega(\epsilon) \in o\left(\frac{1}{\log 1/\epsilon}\right)$. Then, there exists a constant $C > 0$ such that

$$\sup_{x\in\mathcal{X}} \left| f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x) \right| \le K(\log\zeta)\omega\left(\frac{2\pi}{\zeta}\right). \tag{7.9}$$

As a concrete example of this, it suffices if every f is α_f-Hölder continuous for some $\alpha_f > 0$. Finally, we note that, if P and Q are allowed to be arbitrary, then the above uniform convergence assumption is essentially also necessary.

[1] Note that the case of Jensen-Shannon divergence requires an additional uniform lower boundedness assumption.

Proof: First note that it suffices to show that, for all $f \in \mathcal{F}_D$,

$$\mathop{\mathbb{E}}_{X\sim P}[f(X)] - \mathop{\mathbb{E}}_{X\sim Q}[f(X)] = \sum_{z\in\mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z\right).$$

We show this separately for the two sets of assumptions considered:

1. **Case 1: P, Q have a densities $p, q \in \mathcal{L}^2_\mu$.** Then $\widetilde{P}_z = \langle p, \phi_z\rangle_{\mathcal{L}^2}$, and so, by the Plancherel Theorem, since $f \in \mathcal{L}^\infty_\mu(\mathcal{X}) \subseteq \mathcal{L}^2_\mu(\mathcal{X})$,

$$\mathop{\mathbb{E}}_{X\sim P}[f(X)] = \int_{\mathcal{X}} fp\, d\mu = \langle f, p\rangle_{\mathcal{L}^2_\mu} = \sum_{z\in\mathcal{Z}} \widetilde{f}_z \widetilde{P}_z < \infty.$$

   Similarly, $\mathbb{E}_{X\sim Q}[f(X)] = \sum_{z\in\mathcal{Z}} \widetilde{f}_z \widetilde{Q}_z < \infty$. Since these quantities are finite, we can split the sum of differences

$$\sum_{z\in\mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z\right) = \sum_{z\in\mathcal{Z}} \widetilde{f}_z \widetilde{P}_z - \sum_{z\in\mathcal{Z}} \widetilde{f}_z \widetilde{Q}_z = \mathop{\mathbb{E}}_{X\sim P}[f(X)] - \mathop{\mathbb{E}}_{X\sim Q}[f(X)].$$

2. **Case 2: For every $f \in \mathcal{F}_D$, the basis expansion of f in $\mathcal{B}$ converges uniformly (over $\mathcal{X}$) to f.** Then,

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{X\sim P}[f(X)] - \mathop{\mathbb{E}}_{X\sim Q}[f(X)] - \sum_{|z|\le\zeta} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z\right)\right| \\
&= \left|\int_{\mathcal{X}} f(x)\, dP - \int_{\mathcal{X}} f(x)\, dQ - \sum_{|z|\le\zeta} \widetilde{f}_z \left(\int_{\mathcal{X}} \phi_z(x)\, dP - \int_{\mathcal{X}} \phi_z(x)\, dQ\right)\right| \\
&= \left|\int_{\mathcal{X}} f(x)\, dP - \int_{\mathcal{X}} f(x)\, dQ - \int_{\mathcal{X}} \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\, dP - \int_{\mathcal{X}} \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\, dQ\right| \\
&= \left|\int_{\mathcal{X}} f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\, dP + \int_{\mathcal{X}} f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\, dQ\right| \\
&\le \int_{\mathcal{X}} \left|f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\right| dP + \int_{\mathcal{X}} \left|f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\right| dQ \\
&\le 2 \sup_{x\in\mathcal{X}} \left|f(x) - \sum_{|z|\le\zeta} \widetilde{f}_z \phi_z(x)\right| \to 0 \quad \text{as } \zeta \to \infty.
\end{aligned}$$

■

Theorem 53. Suppose that $\mu(\mathcal{X}) < \infty$ and there exist constants $L_D, L_G > 0$, real-valued nets $\{a_z\}_{z\in\mathcal{Z}}$, $\{b_z\}_{z\in\mathcal{Z}}$ such that $\mathcal{F}_D = \mathcal{H}_{p,a}(\mathcal{X}, L_D)$ and $\mathcal{F}_G = \mathcal{H}_{q,b}(\mathcal{X}, L_G)$, where $p, q \ge 1$. Let $p' = \frac{p}{p-1}$ denote the Hölder conjugate of p. Then, for any $P \in \mathcal{F}_G$,

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] \le L_D \frac{c_{p'}}{\sqrt{n}} \left\|\left\{\frac{\|\phi_z\|_{\mathcal{L}^\infty_P}}{a_z}\right\}_{z\in Z}\right\|_{p'} + L_D L_G \left\|\left\{\frac{1}{a_z b_z}\right\}_{z\in\mathcal{Z}\setminus Z}\right\|_{1/(1-1/p-1/q)}.$$

Proof: By Lemma 60,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P,\widehat{P}\right)\right] &= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in\mathcal{Z}}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|\right] \\
&= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in Z}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|+\sum_{z\in\mathcal{Z}\setminus Z}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|\right] \\
&= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in Z}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|+\sum_{z\in\mathcal{Z}\setminus Z}|\widetilde{f}_z\widetilde{P}_z|\right] \\
&\leq \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in Z}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|\right]+\sup_{f\in\mathcal{F}_D}\sum_{z\in\mathcal{Z}\setminus Z}|\widetilde{f}_z\widetilde{P}_z|.
\end{aligned}$$

Note that we have decomposed the risk into two terms, the first comprising estimation error (variance) and the second comprising approximation error (bias). Indeed, in the case that $\mathcal{F}_D = \mathcal{L}^2(\mathcal{X})$, the above becomes precisely the usual bias-variance decomposition of mean squared error.

To bound the first term, applying the Holder's inequality, the fact that $f \in \mathcal{F}_D$, and Jensen's inequality (in that order), we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in Z}|\widetilde{f}_z\left(\widetilde{P}_z-\widehat{P}_z\right)|\right] &= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\sum_{z\in Z}a_z|\widetilde{f}_z|\frac{|\widetilde{P}_z-\widehat{P}_z|}{a_z}\right] \\
&\leq \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}_D}\left(\sum_{z\in Z}a_z^p|\widetilde{f}_z|^p\right)^{\frac{1}{p}}\left(\sum_{z\in Z}\left(\frac{|\widetilde{P}_z-\widehat{P}_z|}{a_z}\right)^{p'}\right)^{\frac{1}{p'}}\right] \\
&\leq L_D\mathop{\mathbb{E}}_{X_{1:n}}\left[\left(\sum_{z\in Z}\left(\frac{|\widetilde{P}_z-\widehat{P}_z|}{a_z}\right)^{p'}\right)^{\frac{1}{p'}}\right] \\
&\leq L_D\left(\sum_{z\in Z}\frac{\mathbb{E}_{X_{1:n}}\left[\left|\widetilde{P}_z-\widehat{P}_z\right|^{p'}\right]}{a_z^{p'}}\right)^{\frac{1}{p'}} \leq \frac{L_D}{\sqrt{n}}\left(\sum_{z\in Z}\frac{\|\phi_z\|_{\mathcal{L}_P^\infty}^{p'}}{a_z^{p'}}\right)^{\frac{1}{p'}},
\end{aligned}$$

where $p' = \frac{p}{p-1}$ is the Hölder conjugate of p. In the last inequality we have used Rosenthal's inequality i.e.,

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[\left|\widetilde{P}_z-\widehat{P}_z\right|^{p'}\right] \leq c_{p'}\frac{\|\phi_z\|_{\mathcal{L}_P^\infty}^{p'}}{n^{p'/2}}.$$

For the second term, by Holder's inequality,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_D}\sum_{z\in\mathcal{Z}\setminus Z}|\widetilde{f}_z\widetilde{P}_z| &\le \sup_{f\in\mathcal{F}_D}\left(\sum_{z\in\mathcal{Z}\setminus Z}\left(a_z|\widetilde{f}_z|\right)^p\right)^{1/p}\left(\sum_{z\in\mathcal{Z}\setminus Z}\left(\frac{|\widetilde{P}_z|}{a_z}\right)^{p'}\right)^{1/p'} \\
&\le L_D\left\|\left\{\frac{b_z\widetilde{P}_z}{b_za_z}\right\}_{z\in\mathcal{Z}\setminus Z}\right\|_{p'} \\
&\le L_D\left\|\{b_z\widetilde{P}_z\}_{z\in\mathcal{Z}\setminus Z}\right\|_q\left\|\left\{\frac{1}{b_za_z}\right\}_{z\in\mathcal{Z}\setminus Z}\right\|_{\frac{p'q}{q-p'}} \quad \text{by Holder} \\
&= L_DL_G\left\|\left\{\frac{1}{a_zb_z}\right\}_{z\in\mathcal{Z}\setminus Z}\right\|_{\frac{1}{1-(1/p+1/q)}}
\end{aligned}$$

■

7.13 Proof of Lower Bound

Theorem 55 (Minimax Lower Bound). Let $\lambda(\mathcal{X}) = 1$, and let p_0 denote the uniform density (with respect to Lebesgue measure) on $\mathcal{X}$. Suppose $\{p_0\} \cup \{\phi_z\}_{z\in\mathcal{Z}}$ is an orthonormal basis in $\mathcal{L}^2_\lambda$, suppose $\{a_z\}_{z\in\mathcal{Z}}$ and $\{b_z\}_{z\in\mathcal{Z}}$ are two real-valued nets, and let $L_D, L_G \ge 0$. For any $Z \subseteq \mathcal{Z}$, define

$$A_Z := |Z|^{1/p}\sup_{z\in Z}a_z \quad \text{and} \quad B_Z := |Z|^{1/q}\sup_{z\in Z}b_z.$$

Then, for $\mathcal{H}_D = \mathcal{H}_{p,a}(L_D)$ and $\mathcal{H}_G := \mathcal{H}_{b,q}(L_G)$, for any $Z \subseteq \mathcal{Z}$ satisfying

$$B_Z \ge 16L_G\sqrt{\frac{n}{\log 2}} \tag{7.10}$$

and

$$2\frac{L_G}{B_Z}\sum_{z\in Z}\|\phi_z\|_{\mathcal{L}^\infty_\mu} \le 1, \tag{7.11}$$

we have

$$M(\mathcal{H}_D,\mathcal{H}_G) \ge \frac{L_GL_D|Z|}{64A_ZB_Z} = \frac{L_GL_D|Z|^{1-1/p-1/q}}{64\left(\sup_{z\in Z}a_z\right)\left(\sup_{z\in Z}b_z\right)}.$$

Proof: We will follow a standard procedure for proving minimax lower bounds based on the Varshamov-Gilbert bound and Fano's lemma (as outlined, e.g., Chapter 2 of Tsybakov (2009)). The proof is quite similar to a standard proof for the case of $\mathcal{L}^2_\lambda$-loss, based on constructing a finite "worst-case" subset $\Omega_G \subseteq \mathcal{F}_G$ of densities over which estimation is difficult. The main difference is that we also construct a similar finite "worst-case" subset $\Omega_D \subseteq \mathcal{F}_D$ of the discriminator class $\mathcal{F}_D$, which we use to lower bound $d_{\mathcal{F}_D} \ge d_{\Omega_D}$ over Ω_G. Specifically, we will use the following result:

Lemma 61 (Simplified Form of Theorem 2.5 of Tsybakov (2009))**.** *Fix a family $\mathcal{P}$ of distributions over a sample space $\mathcal{X}$ and fix a pseudo-metric $\rho : \mathcal{P}\times\mathcal{P} \to [0,\infty]$ over $\mathcal{P}$.*

Suppose there exists a set $T \subseteq \mathcal{P}$ such that

$$s := \inf_{p,p' \in T} \rho(p, p') > 0 \quad \textit{and} \quad \sup_{p \in T} D_{KL}(p, p_0) \leq \frac{\log |T|}{16},$$

where $D_{KL} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ denotes Kullback-Leibler divergence. Then,

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}\left[\rho(p, \widehat{p})\right] \geq \frac{s}{16},$$

where the inf *is taken over all estimators $\widehat{p}$ (i.e., (potentially randomized) functions of $\widehat{p} : \mathcal{X} \to \mathcal{P}$).*

Note that, compared to Theorem 2.5 of Tsybakov (2009), we have loosened some of the constants in order to provide a simpler finite-sample statement.

Suppose $Z \subseteq \mathcal{Z}$ satisfies condition (7.10) and (7.11). For each $\tau \in \{-1, 1\}^Z$ define

$$p_\tau := p_0 + c_G \sum_{z \in Z} \tau_z \phi_z,$$

where $c_G = \frac{L_G}{B_Z}$, and let $\Omega_G := \left\{p_\tau : \tau \in \{-1, 1\}^Z\right\}$.

Since each ϕ_z is orthogonal to p_0, each $p \in \Omega_G$ has unit mass $\int_{\mathcal{X}} p \, d\lambda = 1$, and, by assumption (7.11),

$$\|p_\tau - p_0\|_{\mathcal{L}^\infty_\lambda} = \left\| \frac{L_G}{B_Z} \sum_{z \in Z} \tau_z \phi_z \right\|_{\mathcal{L}^\infty_\lambda} \leq \frac{L_G}{B_Z} \sum_{z \in Z} \|\phi_z\|_{\mathcal{L}^\infty_\lambda} \leq 0.5,$$

which implies that each $p \in \Omega_G$ is lower bounded on $\mathcal{X}$ by 0.5. Thus, each $p \in \Omega_G$ is a probability density. Note that, if we had worked with Gaussian sequences, as in Liang (2017), we would not need to check this, and could hence omit assumption (7.11). Finally, by construction, for each $p \in \Omega_G$,

$$\|p\|_b^q = \sum_{z \in Z} b_z^q |p_z|^q = c^q \sum_{z \in Z} b_z^q \leq c^q |Z| \sup_{z \in Z} b_z^q = L_G^q$$

so that $\Omega_G \subseteq \mathcal{H}_{b,q}(L_G)$. Also, for $c_D := \frac{L_D}{A_Z}$ and for each $\tau \in \{-1, 1\}^Z$, let

$$f_\tau := \frac{L_D}{A_Z} \sum_{z \in Z} \tau_z \phi_z,$$

and define $\Omega_D := \left\{f_\tau : \tau \in \{-1, 1\}^Z\right\}$. By construction, for each $f_\tau \in \Omega_D$,

$$\|f_\tau\|_a^p = \frac{L_D^p}{A_Z^p} \sum_{z \in Z} a_z^p \leq \frac{L_D^p}{A_Z^p} |Z| \sup_{z \in Z} a_z^p = L_D^p,$$

so that $\Omega_D \subseteq \mathcal{H}_{p,a}(L_D)$. Then, for any $\tau, \tau' \in \{-1, 1\}^Z$,

$$d_{\mathcal{F}_D}(p_\tau, p_{\tau'}) \geq d_{\Omega_D}(p_\tau, p_{\tau'}) = \sup_{\tau'' \in \{-1,1\}^Z} \sum_{z \in Z} f_{\tau'', z} c_G (\tau_z - \tau'_z) = 2 c_G c_D \omega\left(\tau, \tau'\right),$$

where $\omega(\tau, \tau') := \sum_{z \in Z} 1_{\{\tau_z \neq \tau'_z\}}$ denotes the Hamming distance between τ and τ'. By the Varshamov-Gilbert bound (Lemma 2.9 of Tsybakov (2009)), we can select

$T \subseteq \{-1,1\}^Z$ such that $\log|T| \geq \frac{|Z|\log 2}{8}$ and, for each $\tau, \tau' \in T$,

$$\omega\left(\tau,\tau'\right) \geq \frac{|Z|}{8}, \quad \text{so that} \quad d_{\mathcal{F}}\left(\theta_\tau, \theta_{\tau'}\right) \geq \frac{c_G c_D |Z|}{4}.$$

Moreover, for any $\tau \in \{-1,1\}^Z$, using the facts that $-\log(1+x) \leq x^2 - x$ for all $x \geq -0.5$ and that $\int_{\mathcal{X}} p_\tau \, dx = 1 = \int_{\mathcal{X}} p_0 \, dx$,

$$\begin{aligned}
D_{KL}(p_\tau^n, p_0^n) &= nD_{KL}(p_\tau, p_0) \\
&= n\int_{\mathcal{X}} p_\tau(x) \log \frac{p_\tau(x)}{p_0(x)} \, dx \\
&= -n\int_{\mathcal{X}} p_\tau(x) \log\left(1 + \frac{p_0(x) - p_\tau(x)}{p_\tau(x)}\right) dx \\
&\leq n\int_{\mathcal{X}} p_\tau(x) \left(\left(\frac{p_0(x) - p_\tau(x)}{p_\tau(x)}\right)^2 - \frac{p_0(x) - p_\tau(x)}{p_\tau(x)}\right) dx \\
&= n\int_{\mathcal{X}} \frac{(p_0(x) - p_\tau(x))^2}{p_\tau(x)} \, dx \\
&\leq 2n\int_{\mathcal{X}} (p_0(x) - p_\tau(x))^2 \, dx \\
&= 2n\|p_0 - p_\tau\|_{\mathcal{L}^2_\lambda} = 2n\frac{L_G^2}{B_Z^2}|Z| \leq n\frac{L_G^2}{B_Z^2}\frac{16}{\log 2}\log|T| \leq \frac{\log|T|}{16},
\end{aligned}$$

where the last two inequalities follow from the Varshamov-Gilbert bound and assumption (7.10), respectively. Combining the above results, Lemma 61 gives a minimax lower bound of

$$M(\mathcal{F}_D, \mathcal{F}_G) \geq \frac{c_G c_D |Z|}{64} = \frac{L_G L_D |Z|}{64 A_Z B_Z}.$$

■

7.14 Proofs and Further Discussion of Applications in Section 7.7

Example 10 (Sobolev Spaces, Oracle and Adaptive estimators in Fourier basis). Suppose that, for some $s, t \geq 0$, $a_z = \left(1 + \|z\|_\infty^2\right)^{s/2}$ and $b_z = \left(1 + \|z\|_\infty^2\right)^{t/2}$. Then, one can check that, for $c = \frac{2^{d-2s}d}{d-2s}$,

$$\sum_{z\in Z} a_z^{-2} \leq 1 + c\left(\zeta^{d-2s} - 1\right), \quad \sup_{z\in\mathcal{Z}\setminus Z} a_z^{-1} \leq \zeta^{-s}, \quad \text{and} \quad \sup_{z\in\mathcal{Z}\setminus Z} b_z^{-1} \leq \zeta^{-t},$$

so that Theorem 53 gives

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] \leq \frac{L_D}{\sqrt{n}}\left(1 + c\zeta^{d/2-s}\right) + L_D L_G \zeta^{-(s+t)}. \tag{7.12}$$

Setting $\zeta = n^{\frac{1}{2t+d}}$ gives

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] \leq Cn^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}, \quad \text{where} \quad C := L_D\left(2\sqrt{c} + L_G\right).$$

On the other hand, as long as $t > d/2$, setting

$$\zeta = \left(256 L_G^2 \frac{n}{\log 2}\right)^{\frac{1}{2t+d}}$$

satisfies the conditions of Theorem 55, giving the minimax lower bound

$$M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \geq \frac{L_G L_D}{64 \zeta^{s+t}} = c_1 n^{-\frac{s+t}{2t+d}} \quad \text{where} \quad c_1 = \frac{L_G L_D}{64}\left(\frac{\log 2}{256 L_G^2}\right)^{\frac{t+s}{2t+d}}.$$

Classical methods can also be used to show that, for all values of s and t, $M(\mathcal{H}_{s,2}, \mathcal{H}_{t,2}) \geq c_2 n^{-1/2}$. Thus, we conclude, there exist constants $C, c > 0$ such that

$$c n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}} \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}\right) \leq C n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}. \tag{7.13}$$

Combining the observation that the s-Hölder space $\mathcal{W}^{s,\infty} \subseteq \mathcal{W}^{s,2}$ with the lower bound in Theorem 3.1 of Liang (2017), we have that (7.13) also holds when $\mathcal{H}_{s,2}$ is replaced with $\mathcal{W}^{s,\infty}$ (e.g., in the case of the Wasserstein metric $d_{\mathcal{W}^{1,\infty}}$), or indeed $\mathcal{W}^{s,q}$ for any $q \geq 2$.

Corollary 62 (Adaptive Upper Bound for Sobolev Spaces). *For any $t, \zeta \geq 0$ and $s \in (0, d/2)$,*

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P}\left[d_{\mathcal{W}^{s,2}}\left(P, \widehat{P}_{Z_\zeta}\right)\right] \leq C \zeta^{-s} \sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P}\left[d_{\mathcal{L}^2_\mu}\left(P, \widehat{P}_{Z_\zeta}\right)\right], \tag{7.14}$$

where $C := \sqrt{2}\left(1 + \frac{2^{d-2s} d}{d-2s}\right)$ does not depend on n or ζ. Hence, if $\widehat{\zeta}(X_{1:n})$ is any adaptive scheme for choosing ζ (i.e., if computing $\widehat{\zeta}$ does not require knowledge of t), then $\widehat{P}_{\widehat{\zeta}}$ is adaptively minimax under the loss $d_{\mathcal{W}^{s,2}}$; that is, for all $t > 0$, there exists $C > 0$ such that

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P}\left[d_{\mathcal{W}^{s,2}}\left(P, \widehat{P}_{Z_{\widehat{\zeta}}}\right)\right] \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}\right).$$

One common scheme for choosing $\widehat{\zeta}$ is to use a leave-one-out cross-validation scheme. Specifically, for

$$\widehat{J}(\zeta) := \|\widehat{P}_\zeta\|_2^2 - \frac{2}{n}\sum_{i=1}^n \widehat{P}_{\zeta,-i}(X_i), \quad \text{where} \quad \widehat{P}_{\zeta,-i} := \sum_{z \in Z_\zeta}\left(\frac{1}{n-1}\sum_{j \in [n]\setminus\{i\}} \phi_z(X_j)\right)\phi_z$$

is a computation of the estimate $\widehat{P}_\zeta$ omitting the i^{th} sample X_i, one can show that $\mathbb{E}_{X_{1:n} \overset{IID}{\sim} P}\left[\widehat{J}(\zeta)\right] = \mathbb{E}_{X_{1:n} \overset{IID}{\sim} P}\left[d^2_{\mathcal{L}^2_\mu}\left(P, \widehat{P}_\zeta\right)\right] - \|P\|^2_{\mathcal{L}^2_\mu}$, so that, up to an additive constant independent of ζ, $\widehat{J}(\zeta)$ is an unbiased estimate of the squared $\mathcal{L}^2_\mu$-risk using the parameter ζ. Based on this, setting

$$\widehat{\zeta} := \mathop{\operatorname{argmin}}_{\zeta \in [0, n^{-1/d}]} J(\zeta),$$

one can show that $\widehat{P}_{\widehat{\zeta}}$ is adaptively minimax over all Sobolev spaces $\mathcal{W}^{t,2}$ with $t > 0$; that is, for all $t > 0$,

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P}\left[d_{\mathcal{L}^2_\mu}\left(P, \widehat{P}_{\widehat{\zeta}}\right)\right] \asymp M\left(\mathcal{L}^2_\mu, \mathcal{W}^{t,2}\right). \tag{7.15}$$

This equivalence (7.14) implies that we can generalize the adaptive minimaxity bound (7.15) to

$$\sup_{P\in\mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P}\left[d_{\mathcal{W}^{s,2}}\left(P,\widehat{P}_{\widehat{\zeta}}\right)\right] \asymp M\left(\mathcal{W}^{s,2},\mathcal{W}^{t,2}\right). \tag{7.16}$$

for all $s\in[0,d/2]$.

Proof: A proof of the adaptive minimaxity of the cross-validation estimator in $d_{\mathcal{L}^2_\mu}$ can be found in Sections 7.2.1 and 7.5.1 of Massart (2007). Therefore, we prove only Inequality (7.14) here. To do this, we combine Theorem 53 with a lower bound on the worst-case performance of the orthogonal series estimator under $\mathcal{L}^2_\mu$ loss, which we establish by explicitly constructing a worst-case true distribution as follows.

Define $P_\zeta := 1+L_G\zeta^{-t}\phi_\zeta$ (where ϕ_ζ is any ϕ_z satisfying $\|z\|_\infty=\zeta$), one can easily check that $P_\zeta\in\mathcal{W}^{t,2}$, and that, for any z with $\|z\|<\zeta$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P_\zeta}\left[\left(\widetilde{(P_\zeta)}_z-\widehat{P}_z\right)^2\right] &= \mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P_\zeta}\left[\left(\frac{1}{n}\sum_{i=1}^n\phi_z(X_i)\right)^2\right]\\
&= \frac{1}{n}\mathop{\mathbb{E}}_{X\sim P_\zeta}\left[\phi_z^2(X)\right]\\
&= \frac{1}{n}\int_{\mathcal{X}}\phi_z^2(x)\left(1+L_G\zeta^{-t}\phi_\zeta(x)\right)dx\\
&\geq \frac{1}{n}\int_{\mathcal{X}}\phi_z^2(x)\,dx=\frac{1}{n}
\end{aligned}$$

(with equality if $\zeta\neq 2z$). Also, let

$$f := \frac{L_D}{\sqrt{2}}\sum_{\|z\|<\zeta}\frac{\left(\widetilde{P_{\zeta_z}}-\widehat{P}_z\right)}{\sqrt{|Z_\zeta|}}\phi_z+\frac{L_D}{\sqrt{2}}\phi_\zeta$$

so that

$$\|f\|_2^2 = \frac{L_D^2}{2}\sum_{\|z\|<\zeta}\frac{\left(\widetilde{P_{\zeta_z}}-\widehat{P}_z\right)^2}{|Z_\zeta|}+\frac{L_D^2}{2}\leq\frac{L_D^2}{2}\sum_{\|z\|<\zeta}|Z_\zeta|^{-1}+\frac{L_D^2}{2}\leq L_D^2,$$

and hence $f\in\mathcal{L}^2_\mu(1)$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P_\zeta}\left[d_{\mathcal{L}^2_\mu}\left(P_\zeta,\widehat{P}_{Z_\zeta}\right)\right] &\geq \mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P}\left[\sum_{\|z\|<\zeta}\widetilde{f}_z\left(\widetilde{P_{\zeta_z}}-\widehat{P}_z\right)^2+\widetilde{f}_\zeta\widetilde{P_{\zeta_z}}\right]\\
&= \frac{L_D}{\sqrt{2|Z_\zeta|}}\sum_{\|z\|<\zeta}\mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P}\left[\left(\widetilde{P_{\zeta_z}}-\widehat{P}_z\right)^2\right]+\frac{L_DL_G}{\sqrt{2}}\zeta^{-t}\\
&\geq \frac{L_D}{\sqrt{2|Z_\zeta|}}\sum_{\|z\|<\zeta}\frac{1}{\sqrt{n}}+\frac{L_DL_G}{\sqrt{2}}\zeta^{-t}=\frac{L_D}{\sqrt{2}}\left(\sqrt{\frac{\zeta^d}{n}}+L_G\zeta^{-t}\right)
\end{aligned}$$

It follows that

$$\sup_{P\in\mathcal{W}^{t,2}}\mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P}\left[d_{\mathcal{L}^2_\mu}\left(P,\widehat{P}_{Z_\zeta}\right)\right]\geq\frac{L_D}{\sqrt{2}}\left(\sqrt{\frac{\zeta^d}{n}}+\zeta^{-t}\right).$$

On the other hand, as we already saw, Theorem 53 gives

$$\sup_{P\in\mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{W}^{s,2}}\left(P,\widehat{P}\right)\right] \leq \left(1+\frac{2^{d-2s}d}{d-2s}\right) L_D \left(\sqrt{\frac{\zeta^d}{n}} + L_G\zeta^{-t}\right)\zeta^{-s}.$$

Combining these two inequalities gives

$$\sup_{P\in\mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{W}^{s,2}}\left(P,\widehat{P}\right)\right] \leq C\zeta^{-s} \sup_{P\in\mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P} \left[d_{\mathcal{L}^2_\mu}\left(P,\widehat{P}_{Z_\zeta}\right)\right].$$

■

7.14.1 Wavelet Basis

Our previous applications were given in terms of the Fourier basis. In this section, we demonstrate that our upper and lower bounds can give tight minimax results using other bases (in this case, the Haar wavelet basis).

Suppose that $\mathcal{X} = [0,1]^D$, and suppose that a function $f : \mathcal{X} \to \mathbb{R}$ has Haar wavelet basis coefficients $\widetilde{f}_{i,j}$, indexed by $z \in \mathcal{Z} := \{(i,j) \in \mathbb{N}\times\mathbb{N} : j \in [2^i]\}$, where $i \in \mathbb{N}$ is the order and $j \in [2^i]$ is the index within that order.

One can show (see, e.g., Donoho, Johnstone, Kerkyacharian, and Picard (1996)) that the Besov seminorm $\|\cdot\|_{\mathcal{B}^q_{p,q}}$ satisfies

$$\|f\|^q_{\mathcal{B}^r_{p,q}} = \sum_{i\in\mathbb{N}} 2^{iqs}\left(\sum_{j\in[2^i]} |\widetilde{f}_{i,j}|^p\right)^{q/p} = \sum_{i\in\mathbb{N}} 2^{iqs}\|\widetilde{f}_i\|^q_p,$$

where $s = r + \frac{1}{2} - \frac{1}{p}$. In particular, when $p = q = 2$, $s = r$, and one can show that $\mathcal{B}^r_{p,q} = \mathcal{W}^r_2$, and

$$\|f\|^q_{\mathcal{B}^r_{p,q}} = \sum_{(i,j)\in\mathcal{Z}} 2^{2is}|\widetilde{f}_{i,j}|^2,$$

For some $\zeta > 0$, we will choose the truncation set Z to be of the form

$$Z = \{(i,j) \in \mathcal{Z} : i \leq \zeta\}.$$

Note that, for each $i \in \mathbb{N}$, since $\phi_{i,1}, ..., \phi_{i,2^i}$ have disjoint supports

$$\sup_{x\in\mathcal{X}} \sum_{j\in[2^i]} |\phi_{i,j}(x)| = \sup_{x\in\mathcal{X}} \sup_{j\in[2^i]} |\phi_{i,j}(x)| = 2^{i/2}.$$

Thus,

$$\sum_{j\in[2^i]} \|\phi_{i,j}\|^2_{\mathcal{L}^2_P} = \sum_{j\in[2^i]} \int_{\mathcal{X}} \phi^2_{i,j}(x)dP \leq \int_{\mathcal{X}} \left(\sum_{j\in[2^i]} \phi_{i,j}(x)\right)^2 dP = 2^i.$$

Example 12 (Sobolev Space, Wavelet Basis). Suppose that, for some $s, t \geq 0$, $a_{i,j} = 2^{is}$ and $b_{i,j} = 2^{it}$. Then, one can check that, for some $c > 0$

$$\sum_{z\in\mathcal{Z}} \frac{\|\phi_z\|^2_{\mathcal{L}^2_P}}{a^2_z} = \sum_{i\leq\zeta}\sum_{j\in[2^i]} \frac{\|\phi_{i,j}\|^2_{\mathcal{L}^2_P}}{2^{2is}} = \sum_{i\leq\zeta} \frac{2^i}{2^{2is}} = \frac{2^{(\zeta+1)(1-2s)}-1}{2^{1-2s}-1} \asymp 2^{\zeta(1-2s)}.$$

Also, $\sup_{z\in\mathcal{Z}\setminus Z} a_z^{-1} \leq 2^{-s\zeta}$ and $\sup_{z\in\mathcal{Z}\setminus Z} b_z^{-1} \leq 2^{-t\zeta}$. Thus, Theorem 53 gives

$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P,\widehat{P}\right)\right] \lesssim L_D\left(\sqrt{\frac{c}{n}}2^{(d/2-s)\zeta} + L_G 2^{-(s+t)\zeta}\right).$$

By letting $\zeta = \log_2 \xi$, we can easily see that this is identical, up to constants, to the bound for the Sobolev case. In contrast to Fourier basis, a larger variety of function spaces (such as inhomogeneous Besov spaces) can be expressed in terms of wavelet basis. The classical work of Donoho, Johnstone, Kerkyacharian, and Picard (1996) showed that, under $\mathcal{L}^p_\mu$ losses, linear estimators, such as that analyzed in our Theorem 53 are sub-optimal in these spaces, but that relatively simple thresholding estimators can recover the minimax rate. We leave it to future work to understand how this phenomenon extends to more general adversarial losses.

7.15 Proofs and Applications of Explicit & Implicit Generative Modeling Results (Section 7.9)

Here, we prove Theorem 59 from the main text, provide some discussion of when the converse direction $M_I(\mathcal{P},\ell,n) \leq M_D(\mathcal{P},\ell,n)$ holds, and also provide some concrete applications.

7.15.1 Proofs of Theorem 59 and Converse

Theorem 59 (Conditions under which Density Estimation is Statistically no harder than Sampling). Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Assume the following:

(A1) $\ell : \mathcal{P}\times\mathcal{P} \to [0,\infty]$ is non-negative, and there exists $C_\triangle > 0$ such that, for all $P_1, P_2, P_3 \in \mathcal{F}_G$,
$$\ell(P_1,P_3) \leq C_\triangle\left(\ell(P_1,P_2)+\ell(P_2,P_3)\right).$$

(A2) $M_D(\mathcal{F}_G,\ell,m)\to 0$ as $m\to\infty$.

(A3) For all $m\in\mathbb{N}$, we can draw m IID samples $Z_{1:m} = Z_1,...,Z_m \overset{IID}{\sim} Q_Z$ of the latent variable Z.

(A4) there exists a nearly minimax sequence of samplers $\widehat{X}_k : \mathcal{X}^n\times\mathcal{Z}\to\mathcal{X}$ such that, for each $k\in\mathbb{N}$, almost surely over $X_{1:n}$, $P_{\widehat{X}_k(X_{1:n},Z)|X_{1:n}} \in \mathcal{F}_G$.

Then, $M_D(\mathcal{F}_G,\ell,n) \leq C_\triangle M_I(\mathcal{F}_G,\ell,n)$.

Proof: The assumption (A2) implies that there exists a sequence $\{\widehat{P}_m\}_{m\in\mathbb{N}}$ of density estimators $\widehat{P}_m : \mathcal{X}^m \to \mathcal{P}$ that is uniformly consistent in ℓ over $\mathcal{P}$; that is,

$$\lim_{m\to\infty}\sup_{P\in\mathcal{P}}\mathop{\mathbb{E}}_{Y_{1:m}\overset{IID}{\sim}P}\left[\ell\left(P,\widehat{P}_m(Y_{1:m})\right)\right]. \tag{7.17}$$

For brevity, we use the abbreviation $P_{\widehat{X}_k} = P_{\widehat{X}_k(X_{1:n},Z)|X_{1:n}}$ in the rest of this proof to denote the conditional distribution of the 'fake data' generated by $\widehat{X}_k$ given the true data. Recalling that the minimax risk is at most the risk of any particular

sampler, we have

$$M_D(\mathcal{P},\ell,n) := \inf_{\widehat{P}} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P,\widehat{P}(X_{1:n})\right)\right]$$
$$\leq \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P,\widehat{P}_m(X_{n+1:n+m})\right)\right].$$

Taking $\lim_{m\to\infty}$ gives, by Tonelli's theorem and non-negativity of ℓ,

$$\begin{aligned}
&M_D(\mathcal{P},\ell,n)\\
&\leq \lim_{m\to\infty}\sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P,\widehat{P}_m(X_{n+1:n+m})\right)\right]\\
&\leq C_\triangle \lim_{m\to\infty}\sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P,P_{\widehat{X}_k}\right)+\ell\left(P_{\widehat{X}_k},\widehat{P}_m(X_{n+1:n+m})\right)\right]\\
&\leq C_\triangle \lim_{m\to\infty}\sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P,P_{\widehat{X}_k}\right)+\ell\left(P_{\widehat{X}_k},\widehat{P}_m(X_{n+1:n+m})\right)\right]\\
&\leq C_\triangle \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{X_{1:n}\overset{IID}{\sim}P} \left[\ell\left(P,P_{\widehat{X}_k}\right)\right] \qquad (7.18)\\
&+ C_\triangle \lim_{m\to\infty}\sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P_{\widehat{X}_k},\widehat{P}_m(X_{n+1:n+m})\right)\right]. \qquad (7.19)
\end{aligned}$$

In the above, we upper bounded $M_D(\mathcal{P},\ell,n)$ by the sum of two terms, (7.18) and (7.19). Since the sequence $\{\widehat{X}_k\}_{k\in\mathbb{N}}$ is nearly minimax, if we were to take an infimum over $k\in\mathbb{N}$ on both sides, the term (7.18) would become precisely $C_\triangle M_I(\mathcal{P},\ell,n)$. Therefore, it suffices to observe that the second term (7.19) is 0. Indeed, by the assumption that $P_{\widehat{X}_k}\in\mathcal{P}$ for all $X_{1:n}\in\mathcal{X}$ and the uniform consistency assumption (7.17),

$$\begin{aligned}
&\lim_{m\to\infty}\sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n}\overset{IID}{\sim}P \\ Z_{1:m}\overset{IID}{\sim}Q_Z}} \left[\ell\left(P_{\widehat{X}_k},\widehat{P}_m(X_{n+1:n+m})\right)\right]\\
&\leq \lim_{m\to\infty}\sup_{P\in\mathcal{P},X_{1:n}\overset{IID}{\sim}P} \mathop{\mathbb{E}}_{Z_{1:m}\overset{IID}{\sim}Q_Z} \left[\ell\left(P_{\widehat{X}_k},\widehat{P}_m(X_{n+1:n+m})\right)\right]\\
&\leq \lim_{m\to\infty}\sup_{P'\in\mathcal{P}} \mathop{\mathbb{E}}_{X_{n+1:n+m}\overset{IID}{\sim}P'} \left[\ell\left(P,\widehat{P}_m(X_{n+1:n+m})\right)\right] = 0.
\end{aligned}$$

■

For completeness, we provide a very simple result on the converse of Theorem 59:

Theorem 63 (Conditions under which Sampling is Statistically no harder than Density Estimation)**.** *Suppose that, there exists as nearly minimax sequence $\{\widehat{P}_k\}_{k\in\mathbb{N}}$ such that, for any $k\in\mathbb{N}$, we can draw a random sample $\widehat{X}$ from $\widehat{P}_k(X_{1:n})$. Then,*

$$M_D(\mathcal{F}_G,\ell,n)\geq M_I(\mathcal{F}_G,\ell,n).$$

The assumption above that we can draw samples from a nearly minimax sequence of estimators if not particularly insightful, but techniques for drawing such samples have been widely studied in the vast literature of Monte Carlo sampling (Robert, 2004). As an example, if $\widehat{P}$ is a kernel density estimator with kernel K, then, recalling that K is itself a probability density, of which $\widehat{P}$ is a mixture, we can sample from $\widehat{P}$ simply by choosing a sample uniformly from $X_{1:n}$ and adding noise $\epsilon \sim K$. Alternatively, if $\widehat{P}$ is bounded and has bounded support, then one can perform rejection sampling.

Proof: Since, by definition of the implicit distribution of $\widehat{X}$,

$$P_{\widehat{X}(X_{1:n},Z)|X_{1:n}} = \widehat{P}(X_{1:n})$$

is precisely the implicit distribution of $\widehat{X}$, we trivially have

$$M_I(\mathcal{F}_G, \ell, n) \leq \sup_{P \in \mathcal{F}_G} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\ell \left(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}} \right) \right]$$

■

7.15.2 Applications

Example 13 (Density Estimation and Sampling in Sobolev families under Dual-Sobolev Loss). There exist constants $C > c > 0$ such that, for all $n \in \mathbb{N}$,

$$cn^{-\min\left\{\frac{s+t}{2s+d}, \frac{1}{2}\right\}} \leq M_I\left(\mathcal{W}^{t,2}, d_{\mathcal{W}^{s,2}}, n\right) \leq Cn^{-\min\left\{\frac{s+t}{2s+d}, \frac{1}{2}\right\}}.$$

Proof: Since adversarial losses always satisfy the triangle inequality, the first inequality follows Theorems 59 and the discussion in Example 10. For the second inequality, since we have already established that the orthogonal series estimator $\widehat{P}_Z$ is nearly minimax, by Theorem 63 it suffices to give a scheme for sampling from the distribution $\widehat{P}_Z(X_{1:n})$. Since the sample space $\mathcal{X} = [0,1]^d$ is bounded and the estimator $\widehat{P}_Z(X_{1:n})$ has a bounded density $p : \mathcal{X} \to [0, \infty)$, we can simply perform rejection sampling; that is, repeatedly sample $Z \times Y$ uniformly from $\mathcal{X} \times [0, \sup_{x \in \mathcal{X}} p(x)]$. Let Z^* denote the first Z sample satisfying $Y < p(Z)$. Then, we Z^* will necessarily have the density p. ■

Example 14 (Density Estimation and Sampling in Exponential Families under Jensen-Shannon, $\mathcal{L}^q$, Hellinger, and RKHS losses). Let $\mathcal{H}$ be an RKHS over a compact sample space $\mathcal{X} \subseteq \mathbb{R}^d$, and let

$$\mathcal{F}_G := \left\{ p_f : \mathcal{X} \to [0, \infty) \middle| p_f(x) = e^{f(x) - A(f)} \text{ for all } x \in \mathcal{X}, f \in \mathcal{H} \right\},$$

in which $A(f) := \log \int_{\mathcal{X}} e^{f(x)} \, d\mu$ denotes the log-partition function.

The Jensen-Shannon divergence $J : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined by

$$J(P, Q) := \frac{1}{2} \left(D_{KL}\left(P, \frac{P+Q}{2}\right) + D_{KL}\left(Q, \frac{P+Q}{2}\right) \right),$$

where $\frac{P+Q}{2}$ denotes the uniform mixture of P and Q, and, noting that we always have $P \ll \frac{P+Q}{2}$ and $Q \ll \frac{P+Q}{2}$,

$$D_{KL}(P, Q) := \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP$$

denotes the Kullback-Leibler divergence. Although J does not satisfy the triangle inequality, one can show that $\sqrt{J}$ is a metric on $\mathcal{P}$ (Endres and Schindelin, 2003), and hence, for all $P, Q \in \mathcal{P}$, by Cauchy-Schwarz,

$$J(P, Q) = \left(\sqrt{J(P, Q)}\right)^2 \leq \left(\sqrt{J(P, R)} + \sqrt{J(R, Q)}\right)^2 \leq 2J(P, R) + 2J(R, Q). \tag{7.20}$$

Also, under mild regularity conditions on $\mathcal{H}$, Sriperumbudur, Fukumizu, Gretton, Hyvärinen, and Kumar (2017) (in their Theorem 7) provides uniform convergence guarantees for a particular density estimator over $\mathcal{P}$. Combining this the inequality (7.20), our Theorem 59 implies

$$M_D(\mathcal{P}, J, n) \leq 2M_I(\mathcal{P}, J, n).$$

For the same class $\mathcal{P}$, the convergence results of Sriperumbudur, Fukumizu, Gretton, Hyvärinen, and Kumar (2017) (their Theorems 6 and 7) also imply similar guarantees under several other losses, including the parameter estimation loss $\|f_P - f_{\widehat{P}}\|_H$ in the RKHS metric, as well as the $\mathcal{L}^q_\mu$ and Hellinger metrics H (on the density), so that we have $M_D(\mathcal{P}, \rho, n) \leq M_I(\mathcal{P}, \rho, n)$ when ρ is any of these metrics.

Perhaps more interestingly, in the case of Jensen-Shannon divergence, under certain regularity conditions, we can altogether drop the assumption that $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}} \in \mathcal{P}$ using uniform convergence bounds shown in Section 5 of Sriperumbudur, Fukumizu, Gretton, Hyvärinen, and Kumar (2017) for the mis-specified case; the density estimator described therein converges (uniformly over P_*) to the projection P_* of $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}}$ onto $\mathcal{P}$ even when samples are drawn from $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}}$.

It is also worth pointing out that, when densities in $\mathcal{F}_G$ are additionally assumed to be lower bounded by a positive constant $\kappa > 0$ (i.e.,

$$\kappa := \inf_{p \in \mathcal{F}_G} \inf_{x \in \mathcal{X}} p(x) > 0,$$

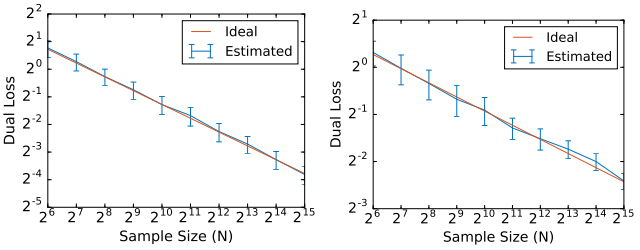
then, by the inequality $-\log(1 + x) \leq x^2 - x$ that holds for all $x \geq -0.5$, for all densities $p, q \in \mathcal{F}_G$,

$$\begin{aligned}
\int_{\mathcal{X}} p(x) \log\left(\frac{2p(x)}{p(x) + q(x)}\right) dx &= -\int_{\mathcal{X}} p(x) \log\left(1 + \frac{q(x) - p(x)}{2p(x)}\right) dx \\
&\leq \int_{\mathcal{X}} p(x) \left(\left(\frac{q(x) - p(x)}{2p(x)}\right)^2 - \left(\frac{q(x) - p(x)}{2p(x)}\right)\right) dx \\
&= \int_{\mathcal{X}} \frac{(q(x) - p(x))^2}{2p(x)} dx \leq \frac{1}{2\kappa} \|P - Q\|^2_{\mathcal{L}^2_\mu},
\end{aligned}$$

and, therefore, $J(P, Q) \leq \frac{1}{2\kappa}\|P - Q\|_{\mathcal{L}^2_\mu}$. Thus, under this additional assumption of uniform lower-boundedness, standard results for density estimation under $\mathcal{L}^2_\mu$ apply (Tsybakov, 2009).

7.16 Experimental Results

This section presents some empirical results supporting the theoretical bounds above. First, we consider an example with a finite basis, which should yield the parametric $n^{-1/2}$ rate. In particular, we construct the true distribution P to consist of 6 randomly chosen basis functions in the Fourier basis. We employ the truncated series estimator $\widehat{P}$ of (7.3) in the same basis using different number of samples n and compute the distance $d_{\mathcal{F}_D}\left(P, \widehat{P}\right)$. Under this setting, the maximization problem of (7.1) needed to evaluate this distance can be solved in closed form. The risk empirically appears to closely follow our derived minimax rate of $n^{-1/2}$, as shown in Figure 7.2(a). Next, we consider a nonparametric case, in which the number of active basis elements increases as function of n, weighted such that Inequality (7.6) predicts a rate of $n^{-1/3}$. As expected, the estimated risk, shown in Figure 7.2(b), closely resembles the rate of $n^{-1/3}$.



(A) Parametric Regime (B) Nonparametric Regime

FIGURE 7.2: Simple synthetic experiments to showcase the tightness of our bound on convergence rates under adversarial losses in the Sobolev case.

7.17 Future Work

In this chapter, we showed that minimax convergence rates for distribution estimation under certain adversarial losses can improve when the probability distributions are assumed to be smooth, using an orthogonal series estimator that smooths the observed empirical distribution. On the other hand, recent work has also shown that, at least under Wasserstein losses, minimax convergence rates improve when the distribution is assumed to have support of low intrinsic dimension, even within a high-dimensional ambient space (Singh and Póczos, 2018). In any case, further work is needed to understand whether minimax rates further improve when distributions are simultaneously smooth and supported on a set of low intrinsic dimension. It is easy to see that the empirical distribution does *not* benefit from assumed smoothness (see, e.g., Proposition 6 of Weed and Bach (2017)). Whether an orthogonal series estimate benefits from low intrinsic dimension may depend on the basis used; the Fourier basis is not likely to benefit, but a wavelet basis, which is spatially localized, may. Nearest neighbor methods have also been shown to benefit from both smoothness and low intrinsic dimensionality, under $\mathcal{L}^2_\mu$ loss, and may therefore be promising (Kpotufe and Garg, 2013).

In Chapter 8, we briefly discuss extension of the present chapter's key results to larger classes of spaces, such as inhomogeneous Besov spaces. Over these spaces, we extend the classic work of Donoho, Johnstone, Kerkyacharian, and Picard (1996), which showed that simple linear density estimators such as the orthogonal series estimator studied in this chapter cease to be minimax rate-optimal, but simple nonlinear estimators such as wavelet thresholding estimators continue to be minimax optimal.

The results of Yarotsky (2017), on uniform approximation of smooth functions (over Sobolev spaces) by neural networks, were crucial to the result Theorem 57

bounding the error of perfectly optimized GANs. If these approximation-theoretic results can be generalized to other spaces (e.g., RKHSs), then our Theorem 53 can be used to derive performance bounds for perfectly optimized GANs over these spaces.

Finally, it has been widely observed that, in practice, optimization of GANs can be quite difficult (Nagarajan and Kolter, 2017; Liang and Stokes, 2018; Arora, Ge, Liang, Ma, and Zhang, 2017). This limits the practical implications of our performance bounds on GANs, which assumed perfect optimization (i.e., convergence to a generator-optimal equilibrium). Conversely, most work studying the optimization landscape of GANs is specific to the noiseless (i.e., "infinite sample size") case, whereas our lower bounds suggest that the sample complexity of training GANs may be substantial. Hence, it is important to generalize these statistical results to the case of imperfect optimization, and, conversely, to understand the effects of statistical noise on the optimization procedure.

Chapter 8

Open Questions, Preliminary Results, and Future Work

While I hope that the previous chapters have shed light on a few problems in density and density functional estimation, far more problems in this domain remain open. In this chapter, I will discuss, at a high level, a few pieces of additional work that are closely related to, but not formally part of, this thesis, either because they are primarily collaborative in nature, or because they are still preliminary.

8.1 Distribution estimation under Besov IPM losses

In Chapter 7, we discussed the nonparametric estimation of probability distributions under a new class of losses, integral probability metrics (IPMs), which are indexed by a family $\mathcal{F}$ of discriminator functions. We focused, in particular, the case in which $\mathcal{F}$ is a weighted $\mathcal{L}^2$-type ellipse, such as a ball in a Hilbert-Sobolev space. Hilbert-Sobolev IPMs are mathematically nice to work with, and can be used to upper bound a number of more commonly used IPMs, such as Wasserstein and total variation distances. However, these one-sided bounds do not, *a priori*, imply tight bounds on minimax convergence rates under these latter losses.

Recently, Ananya Uppal, myself, and Barnabás Póczos, considered a very broad class of IPMs, indexed by balls in Besov spaces, which generalize Hilbert-Sobolev spaces and, moreover include Hölder spaces and $\mathcal{L}^p$ spaces for general $p \geq 1$. In addition to the Hilbert-Sobolev IPMs and certain types of MMD we previously studied (in the $\mathcal{L}^2$ case), the class of IPMs indexed by Besov balls also includes (constant-factor approximations of) Wasserstein, total variation, Kolmogorov-Smirnov distances between probability distributions. For data drawn from probability densities that themselves lie in Besov classes, we have derived minimax estimation rates under the entire range of Besov IPMs; for some Besov IPMs, these are the first known minimax rates, and, for most Besov IPMs, these are the first minimax rates derived under smoothness assumptions on the distribution (although the concurrent work of Weed and Berthet (2019) has explored the case of smoothness under Wasserstein distances). The upper bounds utilize a non-linear wavelet thresholding density estimator originally proposed by Donoho, Johnstone, Kerkyacharian, and Picard (1996). We show, moreover, that, for densities with relatively inhomogenous smoothness (e.g., the emission spectrum of a metal halide lamp pictured in Figure 8.1), non-linearity in the density estimator is necessary to obtain an optimal worst-case convergence rate; specifically, we prove a sharper lower bound on the minimax error of linear estimators, generalizing results of Donoho, Johnstone, Kerkyacharian, and Picard (1996) for $\mathcal{L}^p$ losses to general Besov IPMs.

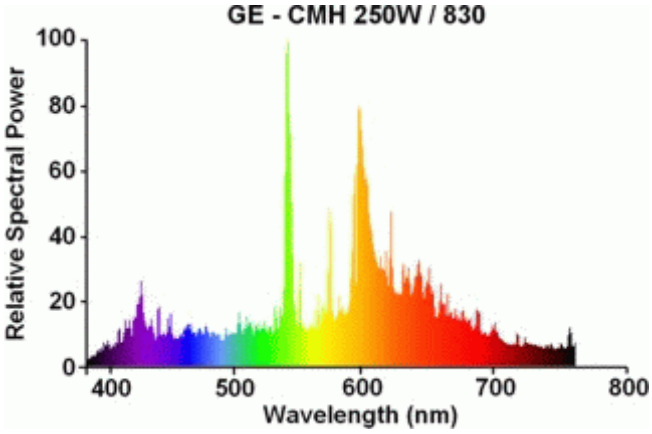


FIGURE 8.1: Emission spectrum of a metal halide lamp, shown as an example of a density function with very inhomogeneous smoothness. Any linear density estimator, depending on its tuning, will either over-fit the data in smooth regions (e.g., around 500nm or 700nm) or under-fit the data in spiky regions (e.g., around 550nm or 600nm). On the other hand, a non-linear wavelet-thresholding estimator can adaptively allocate a finer representation to the spikier portions of the density. *Image credits: Philips Lighting (https://commons.wikimedia.org/wiki/File:MHL.png), "MHL", https://creativecommons.org/licenses/by-sa/2.5/nl/deed.en*

As we did in Chapter 7, using the oracle inequality of Liang (2017), our own results on the relationship between implicit and explicit generative modeling (e.g., Theorem 59), we were able extend our upper bounds to perfectly-optimized neural network GANs (according to a recent scheme for approximating Besov functions by deep fully-connected ReLU networks (Suzuki, 2018)), after applying a new form of regularization based on wavelet-thresholding. The resulting GAN construction, tuned appropriately, is also minimax optimal over all Besov spaces. Interestingly, this one of the first theoretical results showing a clear advantage of neural networks over classical (linear) density estimators.

This work is available on arXiv (Uppal, Singh, and Póczos, 2019) and is currently under review. In the next section, we provide a more technical overview of our results, although the details are still simplified compared to the full paper.

8.1.1 Summary of Results for Besov IPMs

As noted in the previous section, the advantage of Besov spaces (over more classical Hölder or Sobolev spaces) is the ability simultaneously model smoothness at different spatial scales. While there exist many equivalent definitions of Besov spaces (see, e.g., Leoni (2017) or Burenkov (1998)), the definition under which this advantage is most intuitive, and the definition which we utilize, is in terms of wavelet bases, a simple example of which is illustrated in Figure 8.2. A Besov space $\mathcal{B}^s_{p,q}$, parametrized by a triple $(s, p, q) \in (0, \infty) \times [1, \infty] \times [1, \infty]$, is a space of smooth functions, in which smoothness is characterized by rapid decay of functions' wavelet

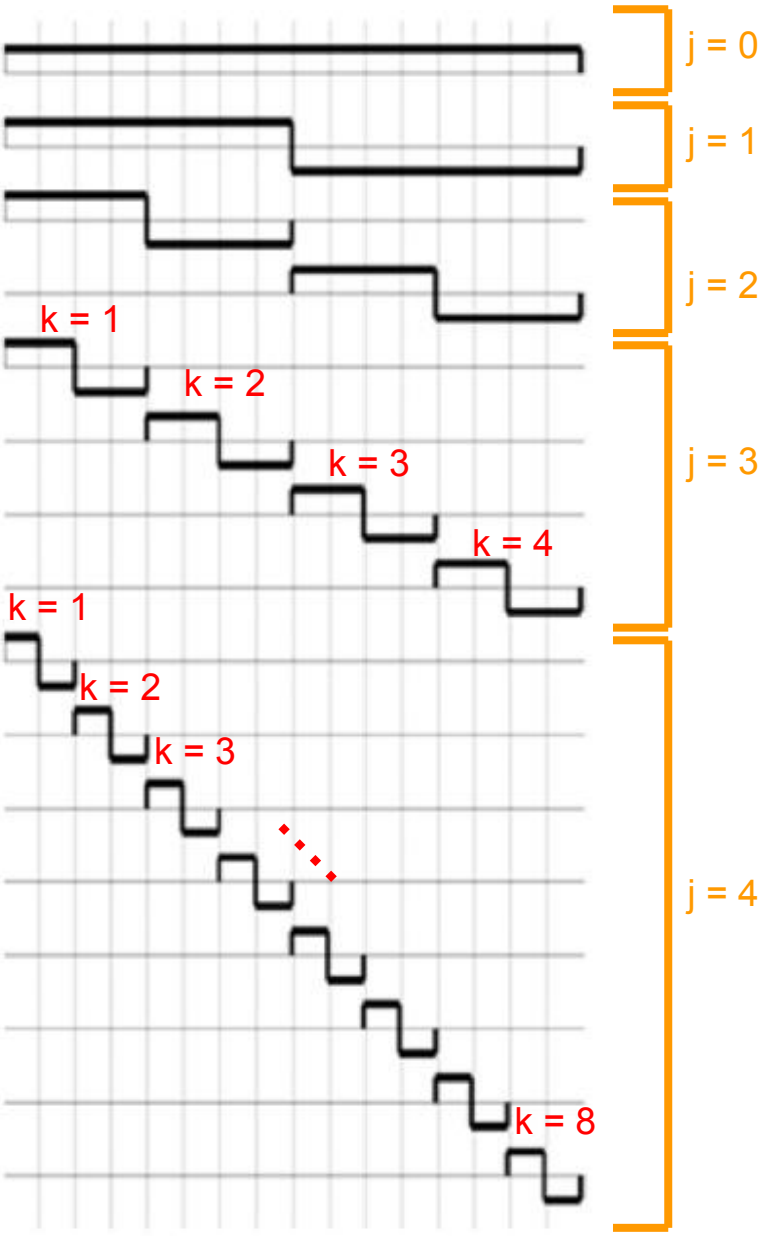


FIGURE 8.2: The first five levels of the Haar wavelet basis, the simplest of wavelet basis. The level (or scale) is indexed by $j \in \mathbb{N}$, while the shift (or offset) is indexed by $k \in [2^j]$. Note that our results actually require using slightly smoother wavelets.

basis coefficients; specifically, the norm

$$\|f\|_{B^s_{p,q}} := \left(\sum_{j \in \mathbb{N}} \left(2^{j(s+D(1/2-1/p))} \left(\sum_{k=1}^{2^j} \beta_{j,k}^p \right)^{1/p} \right)^q \right)^{1/q},$$

in which $\beta_{j,k} := \int f \phi_{j,k}$ denotes the projection of f onto the basis element $\phi_{j,k}$ of level $j \in \mathbb{N}$ and shift $k \in [2^j]$ (the indexing scheme is also illustrated in Figure 8.2). The parameter s measures smoothness, corresponding roughly to the number of well-behaved derivatives of f, much as in the Hölder or Sobolev cases. The parameter p measures the homogeneity of the smoothness constraint over the domain (in our case, the sample space). When p is large, the smoothness constraint is applied in a strong, worst-case sense over the domain; for example, the space $\mathcal{B}^s_{\infty,\infty}$ is topologically equivalent to an s-Holder space, in which the derivative $f^{(s)}$ lies in $\mathcal{L}^\infty$ (i.e., is bounded, except perhaps on a set of measure zero). When p is small, the smoothness constraint is applied in a more relaxed average sense over the domain; for example, the space $\mathcal{B}^1_{1,1}$ corresponds (at least in 1 dimension) to the space of functions of bounded variation. The parameter q affects only polylogarithmic factors in our results, and we don't examine its effect in detail here.

Having defined Besov spaces, and omitting some technical conditions on the wavelet basis, our main technical contributions are now easy to summarize:

1. We prove lower and upper bounds on minimax convergence rates of distribution estimation under IPM losses when the distribution class $\mathcal{P} = \mathcal{B}^{\sigma_g}_{p_g,q_g}$ and the discriminator class $\mathcal{F} = \mathcal{B}^{\sigma_d}_{p_d,q_d}$ are Besov spaces; these rates match up to polylogarithmic factors in the sample size n. Our upper bounds use the wavelet-thresholding estimator proposed in Donoho, Johnstone, Kerkyacharian, and Picard (1996), which we show converges at the optimal rate for a much wider range of losses than previously known. Specifically, if $M(\mathcal{F}, \mathcal{P})$ denotes minimax risk, we show that for $p'_d \geq p_g$, $\sigma_g \geq D/p_g$,

$$M\left(\mathcal{B}^{\sigma_d}_{p_d,q_d}, \mathcal{B}^{\sigma_g}_{p_g,q_g}\right) \asymp \max\left\{ n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}}, n^{-\frac{\sigma_g+\sigma_d+D\left(1-1/p_g-1/p_d\right)}{2\sigma_g+D\left(1-2/p_g\right)}} \right\},$$

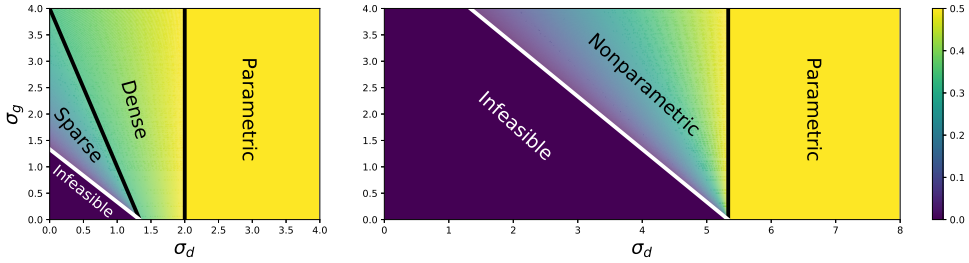
where $\asymp$ denotes equality up to factor of $\sqrt{\log n}$.

2. We show that, for $p'_d \geq p_g$ and $\sigma_g \geq D/p_g$, no estimator in a large class of distribution estimators, called "linear estimators", can converge at a rate faster than

$$M_{\text{lin}}\left(\mathcal{B}^{\sigma_d}_{p_d,q_d}, \mathcal{B}^{\sigma_g}_{p_g,q_g}\right) \asymp n^{-\frac{\sigma_g+\sigma_d-D/p_g+D/p'_d}{2\sigma_g+D\left(1-2/p_g\right)+2D/p'_d}}.$$

"Linear estimators" include the empirical distribution, kernel density estimates with uniform bandwidth, and the orthogonal series estimators we discussed in Chapter 7. Importantly, this effect becomes larger when the data dimension D is large and the distribution P has relatively sparse support (e.g., if P is supported near a low-dimensional manifold).

3. In analogy to Theorem 57 in Chapter 7, we show that the minimax convergence rate can be achieved by a GAN with (fully-connected, ReLU) generator and discriminator networks of bounded size, after some regularization. This result is perhaps more interesting that Theorem 57, in that it is one of the first theoretical results separating performance of GANs from that of classic nonparametric tools such as kernel methods; this may help explain GANs' successes with high-dimensional



(A) (B)

FIGURE 8.3: Minimax convergence rates as functions of discriminator smoothness σ_d and distribution function smoothness σ_g, for (a) general and (b) linear estimators, in the case dimension $D = 4$, and Besov parameters $p_d = 1.2$, $p_g = 2$. Color shows exponent α of minimax convergence rate $n^{-\alpha}$, ignoring polylogarithmic factors.

data such as images. While the proof of Theorem 57 relied on the scheme of Yarotsky (2017) for uniformly approximating Hölder functions by fully-connected ReLU networks, this result relies instead on a similar result of Suzuki (2018) for Besov functions.

As a visual aid for understanding our results, Figure 8.3 show phase diagrams of minimax convergence rates, as functions of discriminator smoothness σ_d and distribution smoothness σ_g, in the illustrative case $D = 4$, $p_d = 1.2$, $p_g = 2$. When $1/p_g + 1/p_d > 1$, a minimum total smoothness $\sigma_d + \sigma_g \geq D(1/p_d + 1/p_g - 1)$ is needed for consistent estimation to be possible – this fails in the "Infeasible" region of the phase diagrams. Intuitively, this occurs because $\mathcal{F}_d$ is not contained in the topological dual $\mathcal{F}_g'$ of $\mathcal{F}_g$. For linear estimators, even greater smoothness $\sigma_d + \sigma_g \geq D(1/p_d + 1/p_g)$ is needed. At the other extreme, for highly smooth discriminator functions, both linear and nonlinear estimators converge at the parametric rate $O\left(n^{-1/2}\right)$, corresponding to the "Parametric" region. In between, rates for linear estimators vary smoothly with σ_d and σ_g, while rates for nonlinear estimators exhibit another phase transition on the line $\sigma_g + 3\sigma_d = D$; to the left lies the "Sparse" case, in which estimation error is dominated by a small number of large errors at locations where the distribution exhibits high local variation; to the right lies the "Dense" case, where error is relatively uniform on the sample space.

8.2 Some further implications of convergence rates for density estimation under IPMs

8.2.1 Monte Carlo Integration

Numerical integration is a ubiquitous problem, not only in machine learning and statistics, but also in engineering problems. Unfortunately, conventional deterministic numerical integration schemes can quickly become intractable in high dimensions. A common approach, therefore, is Monte Carlo integration, which proceeds as follows.

For simplicity, suppose one wants to compute an integral of the form $I_f = \int_{\mathcal{X}} f \, d\lambda$, where λ is some finite, non-negative measure (these conditions can be relaxed in many cases). Let $P = \frac{\lambda}{\lambda(\mathcal{X})}$ denote the probability measure on $\mathcal{X}$ that is proportional to λ, and assume we are able to easily sample from P. One can then draw a

large number of samples $X_1, \cdots, X_n \overset{IID}{\sim} P$ and estimate I_f by the empirical mean $\widehat{I}_f := \frac{1}{n}\sum_{i=1}^n f(X_i)$. Our results in Chapters 6 and 7 on distribution estimation under IPMs imply convergence bounds on

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|I_f - \widehat{I}_f\right|\right]$$

for many classes $\mathcal{F}$. In the case of Monte Carlo integration, these convergence rates can directly determine the computation time needed to achieve a target accuracy. Moreover, our minimax lower bounds imply that, without additional constraints on f or λ, these rates are optimal. While improved schemes can be devised for particular cases of f and λ, the worst-case (sup over f) bounds that we provide are useful when I_f needs to be computed for many distinct f, or when either f or λ is too complex to model analytically, as often happens, for example, in Bayesian inference problems (Geweke, 1989).

8.2.2 Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) is an approach to regularizing optimization of functions depending partly on data. Suppose we want to minimize (in θ) the expectation $\mathbb{E}_X[f(X,\theta)]$ of a function f depending on a random quantity $X \sim P$ with distribution P. We do not know P, but we observe IID samples $X_1, ..., X_n \overset{IID}{\sim} P$; let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denote the resulting empirical distribution.

One often uses the empirical optimizer

$$\widehat{\theta} := \operatorname*{argmin}_\theta \sum_{i=1}^n f(X_i,\theta) = \operatorname*{argmin}_\theta \mathop{\mathbb{E}}_{X\sim P_n}[f(X,\theta)].$$

However, this this is susceptible of over-fitting to the data, especially if the function f is not smooth or otherwise well-behaved. Rather than optimizing purely with respect the empirical distribution, DRO typically optimizes the objective with respect to the worst case over a set $\mathcal{P}$ of "candidate distributions"; specifically,

$$\widehat{\theta}_{\mathcal{P}} := \operatorname*{argmin}_\theta \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{X}[f(X,\theta)]. \tag{8.1}$$

Most often, $\mathcal{P}$ is taken to be a ball $\mathcal{P} = \{Q : \rho(P_n, Q) \le \delta\}$, in some distance measure ρ, centered around P_n, with a radius δ acting as a regularization hyperparameter.

The choice of distance ρ in (8.1) is up to the user, and requires considering both computational and statistical factors, about which I won't go into detail here. Examples of distances that have been used in practice are Wasserstein distances (Esfahani and Kuhn, 2018; Gao and Kleywegt, 2016; Gao, Xie, Xie, and Xu, 2018), ϕ-divergences (Hu and Hong, 2013; Namkoong and Duchi, 2016), and, recently, MMD (Staib and Jegelka, 2019). On the other hand, the choice of the parameter δ is primarily guided by how quickly the distance $\rho(P_n, P)$ decays with n. This is where our results on density estimation under diverse losses comes in.

To see this, consider the following approach to proving regret bounds for DRO. Suppose that $P \in \mathcal{P}$ (I'll return to this assumption in a moment). Then, by construction of the DRO estimate $\widehat{\theta}_{\mathcal{P}}$,

$$\mathop{\mathbb{E}}_{X\sim P}[f(X,\widehat{\theta}_{\mathcal{P}})] \le \sup_{Q\in\mathcal{P}} \mathop{\mathbb{E}}_{X\sim Q}[f(X,\widehat{\theta}_{\mathcal{P}})] \le \sup_{Q\in\mathcal{P}} \mathop{\mathbb{E}}_{X\sim Q}[f(X,\widehat{\theta})] \tag{8.2}$$

Moreover, since, by the triangle inequality, $\sup_{Q in \mathcal{P}} \rho(P, Q) \leq 2\delta$, as long as $\mathbb{E}_{X\sim Q}[f(X, \widehat{\theta})]$ is smooth in Q (with respect to ρ), then the right-hand side of (8.2) can be bounded in terms of δ. This latter smoothness condition is fairly mild, and so the key question becomes: how small can δ be such that $P \in \mathcal{P}$ (with high probability).

Standard concentration of measure arguments typically imply that $\rho(P_n, P)$ is exponentially concentrated around $\mathbb{E}[\rho(P_n, P)]$. Thus, the punchline: δ should typically be of order $\mathbb{E}[\rho(P_n, P)]$ (plus a small high-probability term). For this reason, our results in Chapters 6 and 7 on distribution estimation, which essentially calculate the decay rate of $\mathbb{E}[\rho(P_n, P)]$, fairly immediately enable the formulation of a whole array of new DRO schemes. Moreover, besides the use of new distances, our results may enable the use of smoothing to accelerate convergence rates when P can be assumed to be somewhat smooth – to the best of our knowledge, this has not been previously considered.

We end this note with the important caveat that we haven't discussed *computation* of the DRO estimate 8.1, which appears in general to be quite difficult. However, since somewhat practical approximation schemes have been devised for the cases when ρ is 1-Wasserstein distance or MMD (see references above), we hope that our results may motivate someone to come up with an approximate scheme for computing DRO estimates based on other IPM or Wasserstein distances.

8.3 Asymptotic Distributions for BCF-k Estimators

In this section, I will briefly describe some preliminary on the asymptotic distributions of BCF-k Estimators. Currently, my results are limited to estimation of entropy and KL divergence, but I have begun trying to generalize these results to other functionals.

Lemma 64. *Suppose that we observe $2n$ IID samples $X_1, ..., X_n, Y_1, ..., Y_n \overset{IID}{\sim} P$ from a probability distribution P on $[0, \infty)$. Let $Z_n := \frac{\min\{X_1,...,X_n\}}{\min\{Y_1,...,Y_n\}}$ denote the ratio of the minima of the two samples. Assume the right-sided limit*

$$p_0 := \lim_{r\downarrow 0} \frac{P([0, r))}{r} \in (0, \infty)$$

exists and is positive and finite, and that P has sub-exponential tails. Then, Z_n converges in distribution to a half-Cauchy random variable; specifically,

$$Z_n \xrightarrow{D} |Z|, \quad \text{where} \quad Z \sim \text{Cauchy}(0, \pi).$$

Corollary 65. *Suppose that we observe $2n$ IID samples $X_1, ..., X_n, Y_1, ..., Y_n \overset{IID}{\sim} P$ from a probability distribution P on $\mathbb{R}^D$. Fix any $x \in \mathbb{R}^D$ and $k \in [n]$, and let $\epsilon_k(x)$ denote the distance from x to its k^{th}-nearest neighbor in $\{X_1, ..., X_n\}$, and let $\delta_k(x)$ denote the distance from x to its k^{th}-nearest neighbor in $\{Y_1, ..., Y_n\}$. Suppose that the distributions of $\epsilon_k(x)$ and $\delta_k(x)$ are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$. Then, assuming P has sub-exponential tails and has intrinsic dimension d around x (i.e., that $P(B_r(x)) \asymp r^d$ for small r), letting $Z_n := \epsilon_1^d(x)/\epsilon_1^d(x)$,$Z_n$ converges in distribution to a half-Cauchy random variable; specifically,*

$$Z_n \xrightarrow{D} |Z|, \quad \text{where} \quad Z \sim \text{Cauchy}(0, \pi).$$

This simple asymptotic distribution suggests a fast (both permutation-free and parameter-free) nonparametric approach to testing whether two densities are equal at a particular point $x \in \mathcal{X}$. However, it is not immediately clear how to extend this to testing equality of entire distributions, as the different Z-statistics above will be dependent, given a particular data set. One option is to split the samples into subsamples, using, say, $X_1, ..., X_k$ to compute the k-NN statistic around a test point x_1, $X_{k+1}, ..., X_{2k}$, to compute the k-NN statistic around a test point x_2, etc., and then aggregate the resulting statistics according to a product of $\lfloor n/k \rfloor$ Cauchy distributions to compute a final p-value. Alternatively, for sufficiently sparsely selected test points x_j, one may be able to formalize the intuitive idea that these statistics are asymptotically independent even when the full sample $X_1, ..., X_n$ is used to calculate the final statistic. This would in turn suggest a computationally efficient test for global equality. Further work would be needed to understand the statistical properties of such a test.

Appendix A

Other Projects

Besides the estimation of probability distributions and their functionals, I've studied a few other topics during the course of my doctoral studies at CMU. These are quite unrelated to the core topics of this thesis, and are primarily collaborations with applied researchers in other fields, although my contributions are primarily from the direction of mathematical modeling, machine learning, and statistics. Hence, although they are not meant to be part of this thesis work, I have included brief sections describing these project.

A.1 Distributed Gradient Descent and Bacterial Foraging

Communication and coordination play a major role in the ability of bacterial cells to adapt to ever changing environments and conditions. Recent work has shown that such coordination underlies several aspects of bacterial responses including their ability to develop antibiotic resistance. In this work, we developed a novel distributed gradient descent algorithm to model how bacterial cells collectively search for food in harsh environments using extremely limited communication and computational complexity. We explored the behavior of the model in simulated physical environments with obstacles, as illustrated in Figure A.1. Such an algorithm can also be used for computational tasks when agents are facing similarly restricted conditions. We formalized the communication and computation assumptions required for successful coordination and proved that the proposed algorithm converges to a stationary point even under a dynamically changing interaction network. The proposed method elaborates on upon a prior model of Shklarsh, Ariel, Schneidman, and Ben-Jacob (2011) for bacterial foraging, while making fewer assumptions on the inherent computational abilities of bacterial cells. Simulation studies and analysis of experimental data illustrate the ability of the method to predict several aspects of bacterial swarm food search.

This work was supervised by Professor Ziv Bar-Joseph in the Computational Biology Department (CBD) at CMU, with considerable input from Saket Navlakha (now and assistant professor at the Salk Institute), when he was doing a post-doc with Ziv. Later parts of the project, including comparison with biological lab experiments, were completed by Sabrina Rashid, a doctoral student in CBD under Ziv.

Early parts of this work were published and presented at RECOMB 2016 (Singh, Rashid, Long, Navlakha, Salman, Oltvai, and Bar-Joseph, 2016). More recently Sabrina Rashid has continued work based on my initial model, in two lines of work. The first, published in *Swarm and Evolutionary Computation* (Rashid, Singh, Navlakha, and Bar-Joseph, 2019), compared distributed optimization in bacterial swarms with that in human crowds; a Java simulation I implemented supporting this work can be

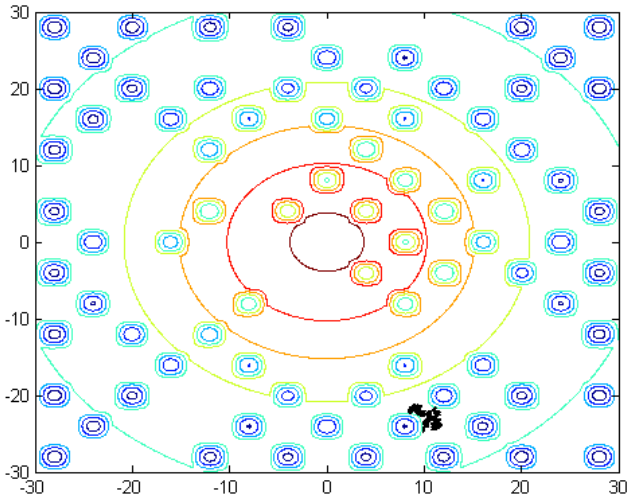


FIGURE A.1: Terrain model for bacterial food search. Obstacles are placed are regular intervals, and the food source is at the center of the region; contours display the diffusion of the food source gradient. The bacterial swarm, in the bottom right area, is depicted as a set of black points, each corresponding to an individual cell.

found here. The second, published in the *Proceedings of the National Academy of Sciences* (Rashid, Long, Singh, Kohram, Vashistha, Navlakha, Salman, Oltvai, and Bar-Joseph, 2019) compared the predictions of our model with experiments (performed by Zhicheng Long, Maryam Kohram, Harsh Vashistha, Hanna Salman, and Zoltan Oltvai at the University of Pittsburgh) using real *E. coli* swarms in custom-fabricated microfluidic devices; in the context of our model, results indicate that bacteria are able to accelerate chemotactic foraging by adjusting tumbling rates in response to the presence of environmental obstacles. While no biological mechanism is currently known to directly modulate tumbling rates, discovering such a mechanism would imply that *E. coli* possess some form of operational short-term memory.

A.2 Sequence-based Prediction of Enhancer-Promoter Interactions

In the human genome, a large number of distal enhancers and proximal promoters form enhancer-promoter interactions (EPI) to regulate target genes. Although recent high-throughput genome-wide mapping approaches have allowed us to more comprehensively identify potential EPI, it is still largely unknown whether sequence-based features alone are sufficient to predict such interactions. In this work, we developed a new computational methods (named PEP and SPEID, pronounced "speed") to predict EPI based on sequence-based features only, when the locations of putative enhancers and promoters in a particular cell type are given. PEP utilizes two modules (PEP-Motif and PEP-Word) based on different but complementary strategies for extracting feature from sequence information. SPEID, which is illustrated in Figure A.2, utilizes a deep learning model, based on convolutional and recurrent neural network layers. Results across six different cell types demonstrated that our methods are effective in predicting EPI, as compared to the state-of-the-art methods that

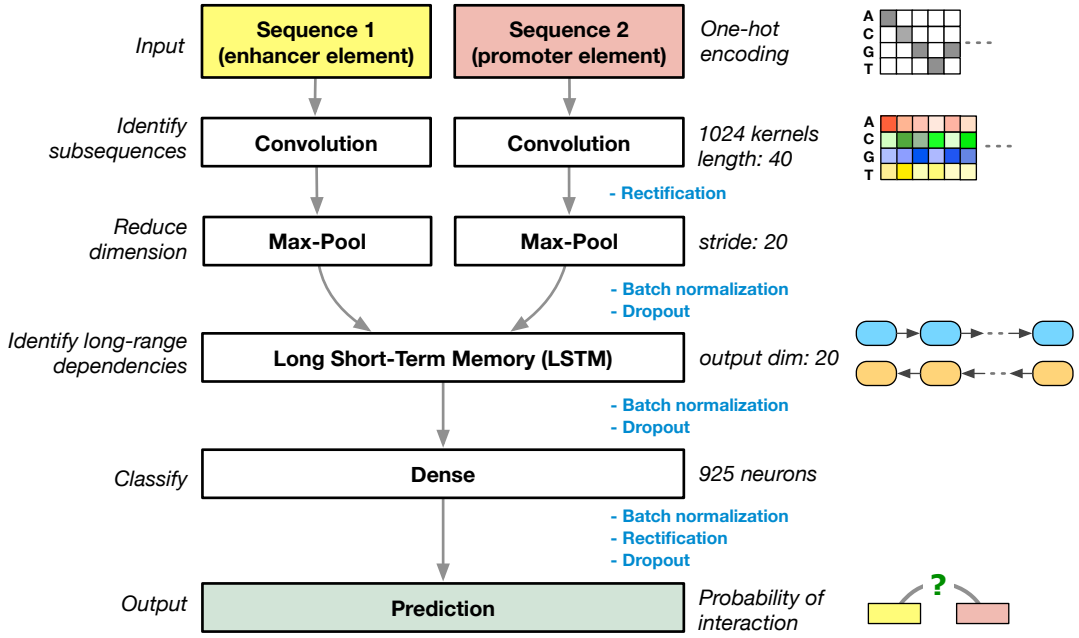


FIGURE A.2: Schematic of the deep learning model, SPEID, that we designed to predict enhancer-promoter interactions from DNA sequence. Key steps involving rectification, batch normalization, and dropout are annotated. Note that the final output step is essentially a logistic regression in SPEID which provides a probability to indicate whether the input enhancer element and promoter element interact.

use functional genomic signals. Our work demonstrated that sequence-based features alone can reliably predict enhancer-promoter interactions genome-wide, which could potentially facilitate the discovery of important sequence determinants for long-range gene regulation. The source code of PEP is available here, and source code of SPEID is available here.

This work was supervised by Professor Jian Ma in CBD. PEP was implemented and tested by Yang Yang and Ruochi Zhang, doctoral students in CBD under Jian. SPEID was implemented and tested by me, but Yang Yang performed substantial parts of the downstream analysis. PEP was published in ISMB 2017 (Yang, Zhang, Singh, and Ma, 2017), and SPEID was published in *Quantitative Biology* (Singh, Yang, Póczos, and Ma, 2019).

A.3 Reconstruction Risk of Convolutional Sparse Dictionary Learning

Sparse dictionary learning (SDL) has become a popular method for learning parsimonious representations of data, a fundamental problem in machine learning and signal processing. While most work on SDL assumes a training dataset of independent and identically distributed (IID) samples, a variant known as convolutional sparse dictionary learning (CSDL) relaxes this assumption to allow dependent, non-stationary sequential data sources (e.g., audio, language, DNA/(epi)genomic features), as illustrated in Figure A.3. Recent work has explored statistical properties of IID SDL, however, the statistical properties of CSDL remain largely unstudied. In this paper, we identified minimax rates of CSDL in terms of reconstruction risk, providing both lower and upper bounds, in a variety of settings. Our results made

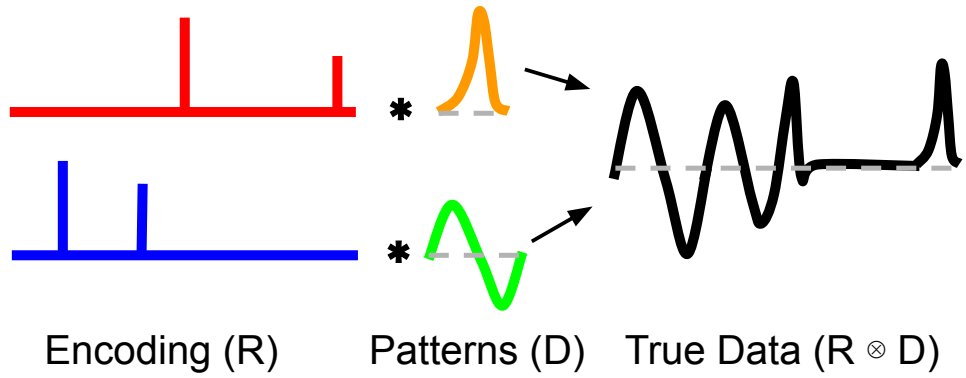


FIGURE A.3: Illustration of how, in a convolutional dictionary model, a long, rich signal ("True Data", black) can be decomposed into a sum of convolutions of long, sparse signals ("Encoding", red/blue) with short, simple signals ("Patterns", orange/green).

minimal assumptions, allowing arbitrary dictionaries and showing that CSDL is robust to dependent noise. We also compared our results to similar results for IID SDL and verify our theory with synthetic experiments.

Specific technical results explored the exact dependence of (minimax) convergence rates on the assumptions of noise structure; both empirically and theoretically, we showed that minimax rates improve significantly when noise is independent across time/space, but that convergence rates decay as noise is allowed to exhibit dependence over longer distances. These distinctions are quite important, for example, in image processing applications, where noise (e.g., in the form of discolorations, scratches, or stains) is usually quite structured and hence likely to much more strongly correlated between nearby pixels than between distant pixels.

This work was supervised by Professor Jian Ma in CBD, as well as by my advisor, Professor Póczos. These results were published in AISTATS 2018 (Singh, Póczos, and Ma, 2018). Work is ongoing to apply CSDL to discover patterns in functional genomics data, and to extend our theoretical results to hierarchical convolutional models (resembling multi-layer convolutional neural networks).

A.4 A Hidden Markov Model for Eye-Tracking Data Analysis

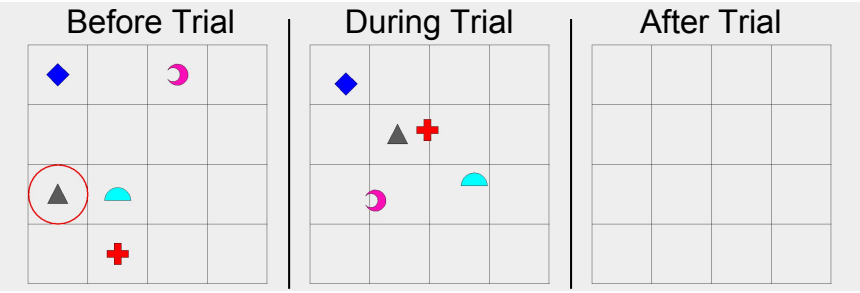


FIGURE A.4: An example trial of the standard TrackIt task (endogenous condition), on a 4×4 grid with 4 distractor objects. The target object here is the grey triangle, as indicated before the trial. Videos of example TrackIt trials can be found at https://github.com/sss1/eyetracking/tree/master/videos.

Eye-tracking provides an opportunity to generate and analyze high-density data relevant to understanding cognition. However, while events in the real world are often dynamic, eye-tracking paradigms are typically limited to assessing gaze toward static objects.In this study, we propose a generative framework, based on a hidden Markov model (HMM), for using eye-tracking data to analyze behavior in the context of multiple moving objects of interest. We apply this framework to analyze data from a recent visual object tracking task paradigm, TrackIt (shown in Figure A.4), for studying selective sustained attention in children. Within this paradigm, we present two validation experiments to show that the HMM (illustrated in Figure A.5) provides a viable approach to studying eye-tracking data with moving stimuli, and to illustrate the benefits of the HMM approach over some more naive possible approaches. The first experiment utilizes a novel 'supervised' variant of TrackIt, while the second compares directly with judgments made by human coders using data from the original TrackIt task. Our results suggest that the HMM-based method provides a robust analysis of eye-tracking data with moving stimuli, both for adults and for children as young as 3.5-6 years old.

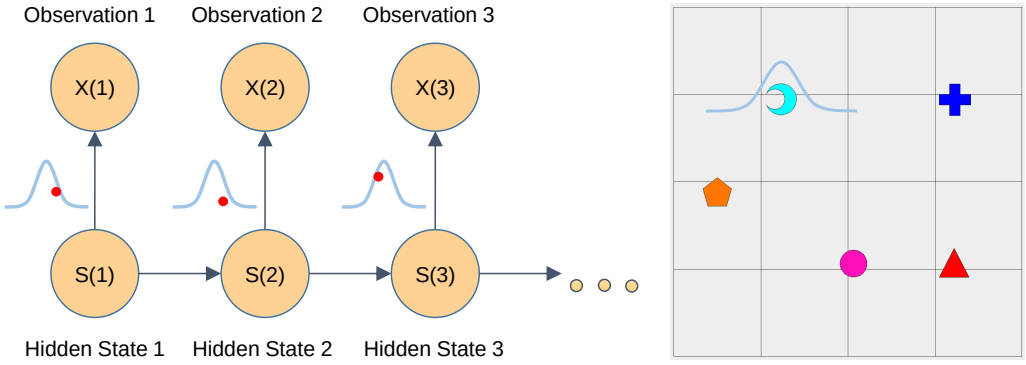


FIGURE A.5: (a) Graphical model schematic of HMM. The initial state (object) $S(1)$ is sampled uniformly at random. At each time point t, we observe a gaze data point $X(t)$, distributed according to a Gaussian centered around the state $S(t)$. At the next time point $t+1$, a new state $S(t+1)$ is sampled according to a distribution depending on $S(t)$, and the process repeats. (b) Example conditional distribution of $E(t)$ given $S(t) =$ "Blue Moon".

This work was supervised by Professors Erik Thiessen and Anna Fisher in the Psychology Department at CMU. The experiment and analyses were designed, and the paper written, in collaboration with Jaeah Kim, a doctoral student in Psychology under Erik and Anna. Anna Vande Velde and Emily Keebler contributed to the experiment design and data collection. A recent version of this work (Kim, Singh, Thiessen, and Fisher, 2019) has been submitted to *Behavior Research Methods* and is under revision, while early versions were presented at CogSci 2018 and 2019 (Kim, Singh, Vande Velde, Thiessen, and Fisher, 2018). Further work is underway, applying the HMM framework to study attention development in children, and to generalize the framework beyond the TrackIt environment to arbitrary video stimuli (including, e.g., videos collected by head-mounted eye-trackers), leveraging recent advances in dynamic object detection based on deep learning.

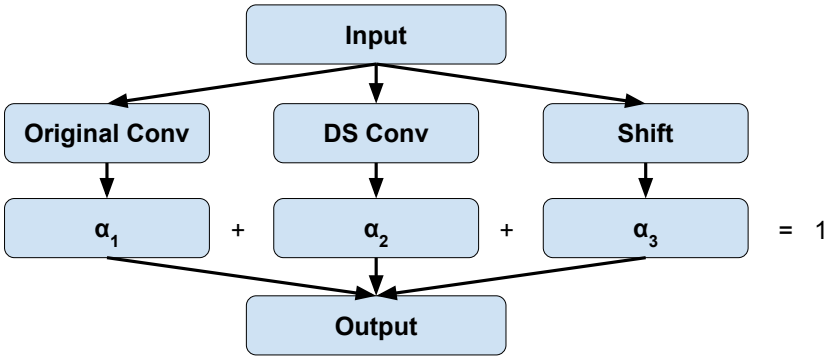


FIGURE A.6: Example of a "cell" (here, a mixture of a full convolution, a depthwise separable (DS) convolution, and a shift operation), used as the basic network building block in DARC. A cost-weighted ℓ_1 penalty is placed on α during training. After training, operations with corresponding $\alpha_j = 0$ are removed from the network.

A.5 DARC: Differentiable Architecture Compression

In many learning situations, resources at inference time are significantly more constrained than resources at training time. This paper studies a general paradigm, called Differentiable ARchitecture Compression (DARC), that combines model compression and DARTS, a neural architecture search method proposed by Liu, Simonyan, and Yang (2018), to learn models that are resource-efficient at inference time. Given a resource-intensive base architecture, DARC utilizes the training data to learn which sub-components can be replaced by cheaper alternatives. For example, convolutions, which dominate the computation and memory demands of most deep networks for computer vision tasks, can often be approximated by smaller, faster operations, such as depthwise-separable convolutions (Jaderberg, Vedaldi, and Zisserman, 2014) or shifts Wu, Wan, Yue, Jin, Zhao, Golmant, Gholaminejad, Gonzalez, and Keutzer (2018). The basic mechanism of DARC is simple: in each layer, we train not only a single operation (e.g., a convolution), but rather a mixture of operations, called a "cell", which, besides the original convolution, can contain depthwise-separable convolutions, shifts, perhaps identity operations (which can then be removed entirely), etc.); an example of such a cell is given in Figure A.6. During training, we impose an ℓ_1 penalty, weighted by the computational (e.g., time or memory) cost of that operation, encouraging the network to assign 0 mixture weights to unnecessarily expensive operations, which can then be removed from the network. The high-level technique can be applied to any neural architecture, and we report experiments on state-of-the-art convolutional neural networks for image classification. We also give theoretical Rademacher complexity bounds in simplified cases, showing how DARC avoids overfitting despite over-parameterization.

This research was performed with Ashish Khetan and Zohar Karnin, while I was an intern hosted by Zohar at Amazon (New York). The paper (Singh, Khetan, and Karnin, 2019) has been submitted to a machine learning conference.

Appendix B

A Condensed Summary of Results on Density Functional Estimation

This appendix, written primarily for my own organizational needs, provides a condensed tabular reference for all our results on density functional estimation, as well as some results due to others.

Below, we list, for reference, all of the assumptions made in the various portions of this thesis devoted to density functional estimation. Table B.1 indicates which of these assumptions we utilize, for each functional and estimator of interest.

(D) The probability distribution P has a **d**ensity $p : \mathcal{X} \to [0, \infty)$.

(s-H) p is s-**H**ölder continuous ($s > 0$). Specifically, is t is the greatest integers strictly less than s, then p is t-times (strongly) differentiable, and $f^{(t)} \in \mathcal{L}^\infty$ for .any This is equivalent to the Sobolev space condition $f \in W^{s,\infty}$.

(s-S) p lies in the s-**S**obolev-Hilbert spaces H^s ($s > 0$). This is equivalent to the Sobolev space condition $f \in W^{s,2}$.

(LB) p is **l**ower **b**ounded away from 0; i.e., $\inf_{x \in \mathcal{X}} p(x) > 0$.

(B) p is well-behaved near the **b**oundary of $\mathcal{X}$; typically, this means either a periodic or vanishing-derivative boundary condition. Usually, it is also required that the sample space $\mathcal{X}$ is known.

(Fr2) The functional $F : \mathcal{P} \to \mathbb{R}$ is **twice-Fr**échet differentiable.

(NPN) p is a nonparanormal distribution (i.e., has a Gaussian copula)

(s-SM) The 1-dimensional **m**arginals of p are s-**S**obolev (see assumption s-S above).

(d-PCN) The ϵ-**c**overing **n**umber of r-bounded subsets of the metric space $(\mathcal{X}, \rho)$ grows at most **p**olynomially, of order d, with $(r/\epsilon)^d$. Specifically, for any $x \in \mathcal{X}$, the covering number $N_{B_x(r)} : (0, \infty) \to \mathbb{N}$ of the ball $B_x(r) := \{y \in \mathcal{X} : \rho(x, y) < r\}$ of radius $r \in (0, \infty)$ centered at x is of order

$$N_{B_x(r)}(\epsilon) \in O\left(\left(\frac{r}{\epsilon}\right)^d\right),$$

where

$$N_{B_x(r)}(\epsilon) := \inf\{|S| : S \subseteq \mathcal{X} \text{ such that, } \forall z \in B_x(r), \exists y \in S \text{ with } \rho(z, y) < \epsilon\}$$

denotes the size of the smallest ϵ-cover of $B_x(r)$. Note that this assumption holds whenever, $\mathcal{X} \subseteq \mathbb{R}^d$, although it may also hold when $\mathcal{X} \subseteq \mathbb{R}^D$ (if the support of P has lower intrinsic dimension d) or for non-Euclidean metric spaces.

Our results on convergence in Wasserstein distance actually hold for more general covering numbers, but it is more difficult to express a closed form for the convergence rate, and we thus consider this simplified form here.

(ℓ-MM) P has a finite ℓ^{th} metric moment

$$m_\ell(P) := \inf_{x \in \mathcal{X}} \left(\mathop{\mathbb{E}}_{Y \sim P} \left[(\rho(x, Y))^\ell \right] \right)^{1/\ell} < \infty.$$

When $(\mathcal{X}, \rho)$ is Euclidean, m_ℓ corresponds to the usual centered ℓ^{th} moment of P.

Functional Name	Estimators	Assumptions	Convergence Rate	Notes
Differential Shannon Entropy $H(P)$	Kernel Plug-in	D, s-H, LB, B	$n^{-\frac{2s}{s+d}} + n^{-1}$	CI
	von Mises	D, s-H, LB, B	$n^{-\frac{8s}{4s+d}} + n^{-1}$	Minimax
	BCFkNN	D, s-H ($s \leq 2$), LB, B	$n^{-\frac{2s}{d}} + n^{-1}$	s-Adaptive, Intrinsic d
		D, s-H ($s \leq 2$), B	$n^{-\frac{2s}{s+d}} + n^{-1}$	Minimax, s-Adaptive
	Nonparanormal	NPN, s-SM ($s \geq 1/2$)	d^2/n	CI
Multivariate Differential Shannon Mutual Information $I(P)$	Kernel Plug-in	D, s-H, LB, B	$n^{-\frac{2s}{s+d}} + n^{-1}$	CI
	BCFkNN	D, s-H ($s \leq 2$), B	$n^{-\frac{2s}{d}} + n^{-1}$	
	Nonparanormal	NPN	d^2/n	CI
General Density Functionals $F(P)$	Kernel Plug-in	D, s-H, Fr2	$n^{-\frac{2s}{s+d}} + n^{-1}$	CI
	kNN Plug-in	D, s-H ($s \leq 2$), Fr2	$n^{-\frac{2s}{s+d}} + n^{-1}$	CLT
	Ensemble	D, s-H, Fr2	$n^{-\frac{2s}{d}} + n^{-1}$	CLT
	von Mises	D, s-H, Fr2	$n^{-\frac{8s}{4s+d}} + n^{-1}$	CLT, Minimax
Sobolev Quantities ($\langle P, Q\rangle_{\mathcal{H}^t}$, $\|P\|^2_{\mathcal{H}^t}$, $\|P-Q\|^2_{\mathcal{H}^t}$)	Series Plug-in	s-S ($s > t$), B	$n^{-\frac{8(s-t)}{4s+d}} + n^{-1}$	Minimax, CLT
	von Mises	s-S ($s > t$, $s, t \in \mathbb{N}$, $d = 1$), B	$n^{-\frac{8(s-t)}{4s+d}} + n^{-1}$	Minimax
RKHS Quantities ($\langle P, Q\rangle_{\mathcal{H}_K}$, $\|P\|^2_{\mathcal{H}_K}$, $\|P-Q\|^2_{\mathcal{H}_K}$)	Fourier Series	t-Exp Kernel, s-Exp RKHS ($s > t$)	$n^{2(t/s-1)} + n^{-1}$	Minimax
Wasserstein Dist. $W_r^r(P, Q)$	Min. Matching	ℓ-MM, d-PCN	$n^{\frac{2(\ell-r)}{\ell}} + n^{-\frac{2r}{d}} + n^{-1}$	d-Adaptive, Intrinsic d

TABLE B.1: Table of density functionals studied in this thesis. 'CI' indicates the existence of a concentration inequality around the estimator's mean. 'CLT' indicates the existence of a central limit theorem (under additional assumptions). 'Minimax' indicates that the convergence rate matches known minimax lower bounds (up to polylogarithmic factors), for all values for s and d. 's-Adaptive' (resp., 'd-Adaptive') indicates that the estimator does not require knowledge of the true smoothness s (resp., the true support dimension d) of the density. Results in green are novel contributions of this thesis. 'Intrinsic d' indicates that d denotes the *intrinsic* dimension of the support of the density (which is often much smaller than the *ambient* data dimension)

Bibliography

Abbasnejad, Ehsan, Javen Shi, and Anton van den Hengel (2018). *Deep Lipschitz networks and Dudley GANs*. URL: https://openreview.net/forum?id=rkw-jlb0W.

Adami, Christoph (2004). "Information theory in molecular biology". In: *Physics of Life Reviews* 1.1, pp. 3–22.

Aghagolzadeh, Mehdi, Hamid Soltanian-Zadeh, B Araabi, and Ali Aghagolzadeh (2007). "A hierarchical clustering based on mutual information maximization". In: *Image Processing, 2007. ICIP 2007. IEEE International Conference on*. Vol. 1. IEEE, pp. I–277.

Aghakouchak, Amir (2014). "Entropy–copula in hydrology and climatology". In: *J. Hydrometeorology* 15.6, pp. 2176–2189.

Ahmed, Nabil Ali and DV Gokhale (1989). "Entropy expressions and their estimators for multivariate distributions". In: *IEEE Trans. on Information Theory* 35.3, pp. 688–692.

Ajtai, Miklós, János Komlós, and Gábor Tusnády (1984). "On optimal matchings". In: *Combinatorica* 4.4, pp. 259–264.

Alemany, P. A. and D. H. Zanette (1994). "Fractal random walks from a variational formalism for Tsallis entropies". In: *Phys. Rev. E* 49.2, R956–R958. DOI: 10.1103/PhysRevE.49.R956.

Ali, Syed Mumtaz and Samuel D Silvey (1966). "A general class of coefficients of divergence of one distribution from another". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142.

Alon, Noga, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler (1997). "Scale-sensitive dimensions, uniform convergence, and learnability". In: *Journal of the ACM (JACM)* 44.4, pp. 615–631.

Anderson, Niall H, Peter Hall, and D Michael Titterington (1994). "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates". In: *Journal of Multivariate Analysis* 50.1, pp. 41–54.

Anderson, TW (1984). "Multivariate statistical analysis". In: *Wi1ey and Sons, New York, NY*.

Anthony, Martin and Peter L Bartlett (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International Conference on Machine Learning*, pp. 214–223.

Aronszajn, Nachman (1950). "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3, pp. 337–404.

Arora, Sanjeev, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang (2017). "Generalization and Equilibrium in Generative Adversarial Nets (GANs)". In: *International Conference on Machine Learning*, pp. 224–232.

Ba, Amadou Diadie and Gane Samb Lo (2019). "Non parametric estimation of Joint entropy and Shannon mutual information, Asymptotic limits: Application to statistic tests". In: *arXiv preprint arXiv:1906.06484*.

Baíllo, Amparo, Javier Cárcamo, and Konstantin V Getman (2016). "The estimation of Wasserstein and Zolotarev distances to the class of exponential variables". In: *arXiv preprint arXiv:1603.06806*.

Baraud, Yannick (2004). "Confidence balls in Gaussian regression". In: *The Annals of Statistics*, pp. 528–551.

Barrio, Eustasio del, Evarist Giné, and Carlos Matrán (1999). "Central limit theorems for the Wasserstein distance between the empirical and the true distributions". In: *Annals of Probability*, pp. 1009–1071.

— (2003). "Correction: Central limit theorems for the Wasserstein distance between the empirical and the true distributions". In: *The Annals of Probability* 31.2, pp. 1142–1143.

Bassetti, Federico, Antonella Bodini, and Eugenio Regazzini (2006). "On minimum Kantorovich distance estimators". In: *Statistics & Probability Letters* 76.12, pp. 1298–1302.

Beckner, William (1975). "Inequalities in Fourier analysis". In: *Annals of Mathematics*, pp. 159–182.

Beirlant, Jan, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen (1997). "Nonparametric entropy estimation: An overview". In: *International Journal of Mathematical and Statistical Sciences* 6.1, pp. 17–39.

Berend, Daniel and Aryeh Kontorovich (2013). "A sharp estimate of the binomial mean absolute deviation with applications". In: *Statistics & Probability Letters* 83.4, pp. 1254–1259.

Berkes, Pietro, Frank Wood, and Jonathan W Pillow (2009). "Characterizing neural dependencies with copula models". In: *Advances in Neural Information Processing Systems*, pp. 129–136.

Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bernton, Espen, Pierre E Jacob, Mathieu Gerber, and Christian P Robert (2017). "Inference in generative models using the Wasserstein distance". In: *arXiv preprint arXiv:1701.05146*.

Berrett, Thomas B and Richard J Samworth (2019). "Efficient two-sample functional estimation and the super-oracle phenomenon". In: *arXiv preprint arXiv:1904.09347*.

Berrett, Thomas B, Richard J Samworth, Ming Yuan, et al. (2019). "Efficient multivariate entropy estimation via k-nearest neighbour distances". In: *The Annals of Statistics* 47.1, pp. 288–318.

Biau, Gérard and Luc Devroye (2015a). "Entropy estimation". In: *Lectures on the Nearest Neighbor Method*. Springer, pp. 75–91.

— (2015b). *Lectures on the nearest neighbor method*. Springer.

Bickel, Peter J and Elizaveta Levina (2008). "Regularized estimation of large covariance matrices". In: *Annals of Stat.* Pp. 199–227.

Bickel, Peter J and Ya'acov Ritov (1988). "Estimating integrated squared density derivatives: sharp best order of convergence estimates". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 381–393.

Birgé, Lucien and Pascal Massart (1995). "Estimation of integral functionals of a density". In: *The Annals of Statistics*, pp. 11–29.

Boissard, Emmanuel and Thibaut Le Gouic (2014). "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 50. 2. Institut Henri Poincaré, pp. 539–563.

Boissard, Emmanuel, Thibaut Le Gouic, and Jean-Michel Loubes (2015). "Distribution's template estimate with Wasserstein metrics". In: *Bernoulli* 21.2, pp. 740–759.

Bos, Adriaan Van den (2007). *Parameter estimation for scientists and engineers*. John Wiley & Sons.

Bottou, Leon, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab (2018). "Geometrical insights for implicit generative modeling". In: *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State*. Springer, pp. 229–268.

Bouezmarni, T., J. Rombouts, and A. Taamouti (2009). *A nonparametric copula based test for conditional independence with applications to Granger causality*. Technical report, Universidad Carlos III, Departamento de Economia.

Brown, Lawrence D and Cun-Hui Zhang (1998). "Asymptotic nonequivalence of nonparametric experiments when the smoothness index is $1/2$". In: *The Annals of Statistics* 26.1, pp. 279–287.

Bucklew, James (2013). *Introduction to rare event simulation*. Springer Science & Business Media.

Bulinski, Alexander and Denis Dimitrov (2018). "Statistical estimation of the Shannon entropy". In: *arXiv preprint arXiv:1801.02050*.

Bulinski, Alexander and Alexey Kozhevin (2018). "Statistical Estimation of Conditional Shannon Entropy". In: *arXiv preprint arXiv:1804.08741*.

Burenkov, Victor I (1998). *Sobolev spaces on domains*. Vol. 137. Springer.

Cai, T Tony (1999). "Adaptive wavelet estimation: A block thresholding and oracle inequality approach". In: *The Annals of statistics*, pp. 898–924.

Cai, T Tony, Tengyuan Liang, and Harrison H Zhou (2015). "Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions". In: *J. of Multivariate Analysis* 137, pp. 161–172.

Cai, T Tony and Mark G Low (2005). "Nonquadratic estimators of a quadratic functional". In: *The Annals of Statistics* 33.6, pp. 2930–2956.

— (2006). "Optimal adaptive estimation of a quadratic functional". In: *The Annals of Statistics* 34.5, pp. 2298–2325.

Cai, T Tony and Ming Yuan (2012). "Adaptive covariance matrix estimation through block thresholding". In: *Annals of Stat.* 40.4, pp. 2014–2042.

Calsaverini, Rafael S and Renato Vicente (2009). "An information-theoretic approach to statistical dependence: Copula information". In: *EPL (Europhysics Letters)* 88.6, p. 68003.

Canas, Guillermo and Lorenzo Rosasco (2012). "Learning probability measures with respect to optimal transport metrics". In: *Advances in Neural Information Processing Systems*, pp. 2492–2500.

Chai, B., D. B. Walther, D. M. Beck, and L. Fei-Fei (2009). "Exploring Functional Connectivity of the Human Brain using Multivariate Information Analysis". In: *NIPS*.

Chatterjee, Sourav (2008). "A new method of normal approximation". In: *The Annals of Probability* 36.4, pp. 1584–1610.

Chaudhuri, Kamalika and Sanjoy Dasgupta (2014). "Rates of convergence for nearest neighbor classification". In: *Advances in Neural Information Processing Systems*, pp. 3437–3445.

Chaudhuri, Kamalika, Sanjoy Dasgupta, Samory Kpotufe, and Ulrike von Luxburg (2014). "Consistent procedures for cluster tree estimation and pruning". In: *IEEE Trans. on Information Theory* 60.12, pp. 7900–7912.

Chen, Louis HY, Larry Goldstein, and Qi-Man Shao (2010). *Normal approximation by Stein's method*. Springer Science & Business Media.

Chib, Siddhartha and Edward Greenberg (1995). "Understanding the Metropolis-Hastings algorithm". In: *The American Statistician* 49.4, pp. 327–335.

Chwialkowski, Kacper P, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton (2015). "Fast two-sample testing with analytic representations of probability measures". In: *Advances in Neural Information Processing Systems*, pp. 1981–1989.

Clement, Philippe and Wolfgang Desch (2008). "An elementary proof of the triangle inequality for the Wasserstein metric". In: *Proceedings of the American Mathematical Society* 136.1, pp. 333–339.

Csiszár, Imre (1964). "Eine informationstheoretische Ungleichung und ihre Anwendung auf Beweis der Ergodizitaet von Markoffschen Ketten". In: *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8, pp. 85–108.

Dereich, Steffen, Michael Scheutzow, and Reik Schottstedt (2013). "Constructive quantization: Approximation by empirical measures". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 49. 4. Institut Henri Poincaré, pp. 1183–1203.

DeVore, Ronald A and George G Lorentz (1993). *Constructive approximation*. Vol. 303. Springer Science & Business Media.

Diggle, Peter J and Richard J Gratton (1984). "Monte Carlo methods of inference for implicit statistical models". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 193–227.

Do Ba, Khanh, Huy L Nguyen, Huy N Nguyen, and Ronitt Rubinfeld (2011). "Sublinear time algorithms for earth mover's distance". In: *Theory of Computing Systems* 48.2, pp. 428–442.

Dong, Hao-Wen and Yi-Hsuan Yang (2019). "Towards a Deeper Understanding of Adversarial Losses". In: *arXiv preprint arXiv:1901.08753*.

Donoho, David L and Iain M Johnstone (1995). "Adapting to unknown smoothness via wavelet shrinkage". In: *Journal of the American Statistical Association* 90.432, pp. 1200–1224.

Donoho, David L, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard (1996). "Density estimation by wavelet thresholding". In: *The Annals of Statistics*, pp. 508–539.

Donoho, David L and Michael Nussbaum (1990). "Minimax quadratic estimation of a quadratic functional". In: *Journal of Complexity* 6.3, pp. 290–323.

Doob, Joseph L (2012). *Measure theory*. Vol. 143. Springer Science & Business Media.

Du, Simon S, Jayanth Koushik, Aarti Singh, and Barnabás Póczos (2017). "Hypothesis Transfer Learning via Transformation Functions". In: *Advances in Neural Information Processing Systems*, pp. 574–584.

Dudley, Richard M (1967). "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes". In: *Journal of Functional Analysis* 1.3, pp. 290–330.

Dudley, RM (1969). "The speed of mean Glivenko-Cantelli convergence". In: *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

— (1972). "Speeds of metric probability convergence". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 22.4, pp. 323–332.

Dumbgen, Lutz (1998). "New goodness-of-fit tests and their application to nonparametric confidence sets". In: *The Annals of statistics*, pp. 288–314.

Dziugaite, Gintare Karolina, Daniel M Roy, and Zoubin Ghahramani (2015). "Training generative neural networks via maximum mean discrepancy optimization". In: *arXiv preprint arXiv:1505.03906*.

Efromovich, Sam (2010). "Orthogonal series density estimation". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 467–476.

Efromovich, Sam and Mark Low (1996). "On optimal adaptive estimation of a quadratic functional". In: *The Annals of Statistics* 24.3, pp. 1106–1125.

Efron, Bradley and Charles Stein (1981). "The jackknife estimate of variance". In: *The Annals of Statistics*, pp. 586–596.

Elidan, Gal (2013). "Copulas in machine learning". In: *Copulae in mathematical and quantitative finance*. Springer, pp. 39–60.

Endres, Dominik Maria and Johannes E Schindelin (2003). "A new metric for probability distributions". In: *IEEE Transactions on Information theory* 49.7, pp. 1858–1860.

Esfahani, Peyman Mohajerin and Daniel Kuhn (2018). "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". In: *Mathematical Programming* 171.1-2, pp. 115–166.

Evans, Dafydd (2008). "A law of large numbers for nearest neighbour statistics". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 464. 2100. The Royal Society, pp. 3175–3192.

Fan, Jianqing (1991). "On the estimation of quadratic functionals". In: *The Annals of Statistics*, pp. 1273–1294.

Flaxman, Seth, Dougal Sutherland, Yu-Xiang Wang, and Yee Whye Teh (2016). "Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata". In: *arXiv preprint arXiv:1611.03787*.

Flaxman, Seth R, Yu-Xiang Wang, and Alexander J Smola (2015). "Who supported Obama in 2012?: Ecological inference through distribution regression". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 289–298.

Fournier, Nicolas and Arnaud Guillin (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3-4, pp. 707–738.

Friedman, Jerome H and Werner Stuetzle (1981). "Projection pursuit regression". In: *JASA* 76.376, pp. 817–823.

Fukumizu, K., A. Gretton, X. Sun, and B. Schoelkopf (2008). "Kernel measures of conditional dependence". In: *Neural Information Processing Systems (NIPS)*.

Gao, Rui and Anton J Kleywegt (2016). "Distributionally robust stochastic optimization with Wasserstein distance". In: *arXiv preprint arXiv:1604.02199*.

Gao, Rui, Liyan Xie, Yao Xie, and Huan Xu (2018). "Robust hypothesis testing using Wasserstein uncertainty sets". In: *Advances in Neural Information Processing Systems*, pp. 7902–7912.

Gao, Shuyang, Greg Steeg, and Aram Galstyan (2015a). "Efficient Estimation of Mutual Information for Strongly Dependent Variables". In: *The 18th International Conference on Artificial Intelligence and Statistics(AISTATS)*.

Gao, Shuyang, Greg Ver Steeg, and Aram Galstyan (2015b). "Estimating mutual information by local Gaussian approximation". In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 278–287.

Gao, Shuyang, Greg Ver Steeg, and Aram Galstyan (2015). "Efficient Estimation of Mutual Information for Strongly Dependent Variables." In: *AISTATS*.

Gao, Weihao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath (2017). "Estimating mutual information for discrete-continuous mixtures". In: *Advances in Neural Information Processing Systems*, pp. 5982–5993.

Gao, Weihao, Sewoong Oh, and Pramod Viswanath (2017a). "Demystifying fixed k-nearest neighbor information estimators". In: *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, pp. 1267–1271.

— (2017b). "Density functional estimators with k-nearest neighbor bandwidths". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1351–1355.

García, María Luz López, Ricardo García-Ródenas, and Antonia González Gómez (2015). "K-means algorithms for functional data". In: *Neurocomputing* 151, pp. 231–245.

Garner, Wendell R (1962). "Uncertainty and structure as psychological concepts." In:

Genovese, Christopher R and Larry Wasserman (2005). "Confidence sets for nonparametric wavelet regression". In: *The Annals of statistics*, pp. 698–729.

Gershgorin, Semyon Aranovich (1931). "Uber die abgrenzung der eigenwerte einer matrix". In: 6, pp. 749–754.

Geweke, John (1989). "Bayesian inference in econometric models using Monte Carlo integration". In: *Econometrica: Journal of the Econometric Society*, pp. 1317–1339.

Giné, Evarist and Richard Nickl (2008). "A simple adaptive estimator of the integrated square of a density". In: *Bernoulli*, pp. 47–61.

Goldenshluger, Alexander and Oleg Lepski (2014). "On adaptive minimax density estimation on R^d". In: *Probability Theory and Related Fields* 159.3-4, pp. 479–543.

Goldfeld, Ziv, Kristjan Greenewald, Yury Polyanskiy, and Jonathan Weed (2019). "Convergence of Smoothed Empirical Measures with Applications to Entropy Estimation". In: *arXiv preprint arXiv:1905.13576*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

Goria, Mohammed Nawaz, Nikolai N Leonenko, Victor V Mergel, and Pier Luigi Novi Inverardi (2005). "A new class of random vector entropy estimators and its applications in testing statistical hypotheses". In: *Journal of Nonparametric Statistics* 17.3, pp. 277–297.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test". In: *Journal of Machine Learning Research* 13.Mar, pp. 723–773.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). "Improved training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*, pp. 5769–5779.

Han, Insu, Dmitry Malioutov, and Jinwoo Shin (2015). "Large-scale log-determinant computation through stochastic Chebyshev expansions." In: *ICML*, pp. 908–917.

Han, Yanjun, Jiantao Jiao, and Tsachy Weissman (2015). "Minimax Estimation of Discrete Distributions Under ℓ_1 Loss". In: *IEEE Transactions on Information Theory* 61.11, pp. 6343–6354.

Han, Yanjun, Jiantao Jiao, Tsachy Weissman, and Yihong Wu (2017). "Optimal rates of entropy estimation over Lipschitz balls". In: *arXiv preprint arXiv:1711.02141*.

Haussler, David (1995). "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension". In: *Journal of Combinatorial Theory, Series A* 69.2, pp. 217–232.

Henrion, Didier and Jérôme Malick (2012). "Projection methods in conic optimization". In: *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer, pp. 565–600.

Hernández-Lobato, José Miguel, James R Lloyd, and Daniel Hernández-Lobato (2013). "Gaussian process conditional copulas with applications to financial time series". In: *Advances in Neural Information Processing Systems*, pp. 1736–1744.

Hero, A. O., B. Ma, O. Michel, and J. Gorman (2002a). *Alpha-Divergence for Classification, Indexing and Retrieval*. Communications and Signal Processing Laboratory Technical Report CSPL-328.

Hero, Alfred O, Bing Ma, Olivier JJ Michel, and John Gorman (2002b). "Applications of entropic spanning graphs". In: *IEEE signal processing magazine* 19.5, pp. 85–95.

Hlaváckova-Schindler, K., M. Paluŝb, M. Vejmelkab, and J. Bhattacharya (2007). "Causality detection based on information-theoretic approaches in time series analysis". In: *Physics Reports* 441, pp. 1–46.

Ho, Nhat, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung (2017). "Multilevel Clustering via Wasserstein Means". In: *arXiv preprint arXiv:1706.03883*.

Hoff, Peter D (2007). "Extending the rank likelihood for semiparametric copula estimation". In: *The Annals of Applied Statistics*, pp. 265–283.

Hu, Zhaolin and L Jeff Hong (2013). "Kullback-Leibler divergence constrained distributionally robust optimization". In: *Available at Optimization Online*.

Hulle, M. M. Van (2008). "Constrained Subspace ICA Based on Mutual Information Optimization Directly". In: *Neural Computation* 20, pp. 964–973.

Ince, Robin AA, Bruno L Giordano, Christoph Kayser, Guillaume A Rousselet, Joachim Gross, and Philippe G Schyns (2017). "A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula". In: *Human brain mapping* 38.3, pp. 1541–1573.

Ingster, Yuri and Irina A Suslina (2012). *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media.

Jaderberg, Max, Andrea Vedaldi, and Andrew Zisserman (2014). "Speeding up Convolutional Neural Networks with Low Rank Expansions". In: *Proceedings of the British Machine Vision Conference. BMVA Press*.

Jiao, Jiantao, Weihao Gao, and Yanjun Han (2018). "The nearest neighbor information estimator is adaptively near minimax rate-optimal". In: *Advances in neural information processing systems*, pp. 3156–3167.

Jiao, Jiantao, Yanjun Han, and Tsachy Weissman (2017). "Minimax Estimation of the L_1 Distance". In: *arXiv preprint arXiv:1705.00807*.

Johnson, Oliver and Richard Samworth (2005). "Central limit theorem and convergence to stable laws in Mallows distance". In: *Bernoulli* 11.5, pp. 829–845.

Kandasamy, Kirthevasan, Akshay Krishnamurthy, Barnabas Poczos, and Larry Wasserman (2015). "Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations". In: *NIPS*, pp. 397–405.

Kantorovich, Leonid Vasilevich and Gennady S Rubinstein (1958). "On a space of completely additive functions". In: *Vestnik Leningrad. Univ* 13.7, pp. 52–59.

Kantorovich, Leonid Vitalievich (1942). "On the translocation of masses". In: *Dokl. Akad. Nauk. USSR (NS)*. Vol. 37, pp. 199–201.

Kim, Jaeah, Shashank Singh, Erik Thiessen, and Anna Fisher (2019). *A Hidden Markov Model for Analyzing Eye-Tracking of Moving Objects*. DOI: 10.31234/osf.io/mqpnf. URL: psyarxiv.com/mqpnf.

Kim, Jaeah, Shashank Singh, Anna Vande Velde, Erik D. Thiessen, and Anna V. Fisher (2018). "A Hidden Markov Model for Analyzing Eye-Tracking of Moving Objects". In: *Proceedings of the 2018 Annual Conference of the Cognitive Science Society (CogSci)*.

Kingma, Diederik P and Max Welling (2014). "Auto-encoding variational Bayes". In: *ICLR*.

Klaassen, Chris AJ and Jon A Wellner (1997). "Efficient estimation in the bivariate normal copula model: normal margins are least favourable". In: *Bernoulli* 3.1, pp. 55–77.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.

Kolmogorov, Andrey (1933). "Sulla determinazione empirica di una lgge di distribuzione". In: *Inst. Ital. Attuari, Giorn.* 4, pp. 83–91.

Kontorovich, Aryeh and Roi Weiss (2015). "A Bayes consistent 1-NN classifier". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 480–488.

Kozachenko, LF and Nikolai N Leonenko (1987). "Sample estimate of the entropy of a random vector". In: *Problemy Peredachi Informatsii* 23.2, pp. 9–16.

Kpotufe, Samory and Vikas Garg (2013). "Adaptivity to local smoothness and dimension in kernel regression". In: *Advances in neural information processing systems*, pp. 3075–3083.

Kpotufe, Samory and Ulrike V Luxburg (2011). "Pruning nearest neighbor cluster trees". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 225–232.

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger (2004). "Estimating mutual information". In: *Physical review E* 69.6, p. 066138.

Krishnamurthy, Akshay, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman (2014). "Nonparametric estimation of renyi divergence and friends". In: *International Conference on Machine Learning*, pp. 919–927.

Krishnamurthy, Akshay, Kirthevasan Kandasamy, Barnabas Poczos, and Larry A Wasserman (2015). "On Estimating L_2^2 Divergence." In: *AISTATS*.

Kruskal, William H (1958). "Ordinal measures of association". In: *JASA* 53.284, pp. 814–861.

Kullback, S. and R.A. Leibler (1951). "On Information and Sufficiency". In: *Annals of Mathematical Statistics* 22, pp. 79–86.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015). "From word embeddings to document distances". In: *International Conference on Machine Learning*, pp. 957–966.

Laurent, Béatrice (1996). "Efficient estimation of integral functionals of a density". In: *The Annals of Statistics* 24.2, pp. 659–681.

Laurent, Beatrice and Pascal Massart (2000). "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics*, pp. 1302–1338.

Learned-Miller, E. G. and J. W. Fisher (2003). "ICA Using Spacings Estimates of Entropy". In: *Journal of Machine Learning Research* 4, pp. 1271–1295.

Lebesgue, Henri (1910). "Sur l'intégration des fonctions discontinues". In: *Annales scientifiques de l'École normale supérieure*. Vol. 27. Société mathématique de France, pp. 361–450.

Lei, Jing (2018). "Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces". In: *arXiv preprint arXiv:1804.10556*.

Leonenko, N. and L. Pronzato (2010). *Correction Of 'A Class Of Rényi Information Estimators For MuliTidimensional Densities' Ann. Statist., 36(2008) 2153-2182*.

Leonenko, N., L. Pronzato, and V. Savani (2008). "Estimation of Entropies and Divergences via Nearest Neighbours". In: *Tatra Mt. Mathematical Publications* 39.

Leoni, Giovanni (2017). *A first course in Sobolev spaces*. American Mathematical Soc.

Lepski, Oleg, Arkady Nemirovski, and Vladimir Spokoiny (1999). "On estimation of the L_r norm of a regression function". In: *Probability theory and related fields* 113.2, pp. 221–253.

Lepski, Oleg V and VG Spokoiny (1997). "Optimal pointwise adaptive methods in nonparametric estimation". In: *The Annals of Statistics*, pp. 2512–2546.

Lewi, J., R. Butera, and L. Paninski (2007). "Real-time adaptive information-theoretic optimization of neurophysiology experiments". In: *Advances in Neural Information Processing Systems*. Vol. 19.

Li, Chun-Liang, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos (2017). "MMD GAN: Towards deeper understanding of moment matching network". In: *Advances in Neural Information Processing Systems*, pp. 2200–2210.

Li, Ker-Chau (1989). "Honest confidence regions for nonparametric regression". In: *The Annals of Statistics*, pp. 1001–1008.

Liang, Tengyuan (2017). "How Well Can Generative Adversarial Networks (GAN) Learn Densities: A Nonparametric View". In: *arXiv preprint arXiv:1712.08244*.

Liang, Tengyuan and James Stokes (2018). "Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks". In: *arXiv preprint arXiv:1802.06132*.

Liu, Han, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman (2012). "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4, pp. 2293–2326.

Liu, Han, John Lafferty, and Larry Wasserman (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs". In: *Journal of Machine Learning Research* 10.Oct, pp. 2295–2328.

Liu, Han, Larry Wasserman, and John D Lafferty (2012). "Exponential concentration for mutual information estimation with application to forests". In: *Advances in Neural Information Processing Systems*, pp. 2537–2545.

Liu, Hanxiao, Karen Simonyan, and Yiming Yang (2018). "DARTS: Differentiable architecture search". In: *arXiv preprint arXiv:1806.09055*.

Liu, Shuang, Olivier Bousquet, and Kamalika Chaudhuri (2017). "Approximation and convergence properties of generative adversarial learning". In: *Advances in Neural Information Processing Systems*, pp. 5551–5559.

Loftsgaarden, Don O and Charles P Quesenberry (1965). "A nonparametric estimate of a multivariate density function". In: *The Annals of Mathematical Statistics* 36.3, pp. 1049–1051.

Lopez-Paz, David, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin (2015). "Towards a learning theory of cause-effect inference". In: *International Conference on Machine Learning*, pp. 1452–1461.

Luukkainen, Jouni and Eero Saksman (1998). "Every complete doubling metric space carries a doubling measure". In: *Proceedings of the American Mathematical Society* 126.2, pp. 531–534.

Ma, Jian and Zengqi Sun (2011). "Mutual information is copula entropy". In: *Tsinghua Science & Tech.* 16.1, pp. 51–54.

Mack, YP and M Rosenblatt (1979). "Multivariate k-nearest neighbor density estimates". In: *J. Multivar. Analysis*.

Malevergne, Yannick and Didier Sornette (2003). "Testing the Gaussian copula hypothesis for financial assets dependences". In: *Quantitative Finance* 3.4, pp. 231–250.

Mallat, Stéphane (1999). *A wavelet tour of signal processing*. Academic press.
Massart, Pascal (2007). *Concentration inequalities and model selection*. Vol. 6. Springer.
McDiarmid, C. (1989). "On the method of bounded differences". In: *Surveys in Combinatorics* 141. Ed. by J. Siemons, pp. 148–188.
Mendelson, Shahar (2002). "Learnability in Hilbert spaces with reproducing kernels". In: *journal of complexity* 18.1, pp. 152–170.
Mescheder, Lars, Andreas Geiger, and Sebastian Nowozin (2018). "Which training methods for GANs do actually Converge?" In: *arXiv preprint arXiv:1801.04406*.
Misra, Neeraj, Harshinder Singh, and Eugene Demchuk (2005). "Estimation of the entropy of a multivariate normal distribution". In: *J. Multivariate Analysis* 92.2, pp. 324–342.
Mitra, Ritwik and Cun-Hui Zhang (2014). "Multivariate analysis of nonparametric estimates of large correlation matrices". In: *arXiv preprint arXiv:1403.6195*.
Mohamed, Shakir and Balaji Lakshminarayanan (2016). "Learning in implicit generative models". In: *arXiv preprint arXiv:1610.03483*.
Montavon, Grégoire, Klaus-Robert Müller, and Marco Cuturi (2016). "Wasserstein training of restricted Boltzmann machines". In: *Advances in Neural Information Processing Systems*, pp. 3718–3726.
Montgomery, D. (2005). *Design and Analysis of Experiments*. John Wiley and Sons.
Moon, Kevin and Alfred Hero (2014a). "Multivariate f-divergence estimation with confidence". In: *Advances in Neural Information Processing Systems*, pp. 2420–2428.
Moon, Kevin R (2016). "Nonparametric Estimation of Distributional Functionals and Applications." In:
Moon, Kevin R and Alfred O Hero (2014b). "Ensemble estimation of multivariate f-divergence". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, pp. 356–360.
Moon, Kevin R, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero (2016). "Improving convergence of divergence functional ensemble estimators". In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1133–1137.
Moon, Kevin R, Kumar Sricharan, and Alfred O Hero III (2017). "Ensemble Estimation of Mutual Information". In: *arXiv preprint arXiv:1701.08083*.
Morimoto, Tetsuzo (1963). "Markov processes and the H-theorem". In: *Journal of the Physical Society of Japan* 18.3, pp. 328–331.
Mroueh, Youssef, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng (2017). "Sobolev GAN". In: *arXiv preprint arXiv:1711.04894*.
Mukherjee, Rajarshi, Eric Tchetgen Tchetgen, and James Robins (2015). "Lepski's Method and Adaptive Estimation of Nonlinear Integral Functionals of Density". In: *arXiv preprint arXiv:1508.00249*.
— (2016). "On Adaptive Estimation of Nonparametric Functionals". In: *arXiv preprint arXiv:1608.01364*.
Müller, Alfred (1997). "Integral probability metrics and their generating classes of functions". In: *Advances in Applied Probability* 29.2, pp. 429–443.
Nagarajan, Vaishnavh and J Zico Kolter (2017). "Gradient descent GAN optimization is locally stable". In: *Advances in Neural Information Processing Systems*, pp. 5591–5600.
Namkoong, Hongseok and John C Duchi (2016). "Stochastic gradient methods for distributionally robust optimization with f-divergences". In: *Advances in Neural Information Processing Systems*, pp. 2208–2216.
Nguyen, X., M.J. Wainwright, and M.I. Jordan. (2010). "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Trans. on Information Theory.*

Nguyen, XuanLong, Martin J Wainwright, and Michael I Jordan (2010). "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory* 56.11, pp. 5847–5861.

Ni, Kangyu, Xavier Bresson, Tony Chan, and Selim Esedoglu (2009). "Local histogram based segmentation using the Wasserstein distance". In: *International journal of computer vision* 84.1, pp. 97–111.

Noshad, Morteza and Alfred O Hero III (2018). "Scalable Mutual Information Estimation using Dependence Graphs". In: *arXiv preprint arXiv:1801.09125*.

Noshad, Morteza, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero (2017). "Direct estimation of information divergence using nearest neighbor ratios". In: *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, pp. 903–907.

Nowozin, Sebastian, Botond Cseke, and Ryota Tomioka (2016). "f-GAN: Training generative neural samplers using variational divergence minimization". In: *Advances in Neural Information Processing Systems*, pp. 271–279.

Ntampaka, Michelle, Hy Trac, Dougal J Sutherland, Nicholas Battaglia, Barnabás Póczos, and Jeff Schneider (2015). "A machine learning approach for dynamical mass measurements of galaxy clusters". In: *The Astrophysical Journal* 803.2, p. 50.

Oliva, J., B. Poczos, and J. Schneider (2013). "Distribution to Distribution Regression". In: *International Conference on Machine Learning (ICML)*.

Owen, Mark (2007). *Practical signal processing*. Cambridge university press.

Pál, Dávid, Barnabás Póczos, and Csaba Szepesvári (2010). "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs". In: *Advances in Neural Information Processing Systems*, pp. 1849–1857.

Pardo, Leandro (2005). *Statistical inference based on divergence measures*. CRC press.

Pearl, J. (1998). *Why there is no statistical test for confounding, why many think there is, and why they are almost right*. UCLA Computer Science Department Technical Report R-256.

Peng, Hanchuan, Fuhui Long, and Chris Ding (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27.8, pp. 1226–1238.

Pérez-Cruz, Fernando (2009). "Estimation of information theoretic measures for continuous random variables". In: *Advances in neural information processing systems*, pp. 1257–1264.

Peyré, Gabriel (2011). "The numerical tours of signal processing". In: *Computing in Science & Engineering* 13.4, pp. 94–97.

Póczos, B. and A. Lőrincz (2005). "Independent Subspace Analysis Using Geodesic Spanning Trees". In: *ICML*, pp. 673–680.

Póczos, B. and A. Lőrincz (2009). "Identification of Recurrent Neural Networks by Bayesian Interrogation Techniques". In: *J. Machine Learning Research* 10, pp. 515–554.

Poczos, B. and J. Schneider (2012). "Nonparametric Estimation of Conditional Information and Divergences". In: *International Conference on AI and Statistics (AISTATS)*. Vol. 20. JMLR Workshop and Conference Proceedings.

Póczos, Barnabás and Jeff Schneider (2011). "On the estimation of alpha-divergences". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 609–617.

Poczos, Barnabas, Liang Xiong, and Jeff Schneider (2011). *Nonparametric divergence estimation and its applications to machine learning*. Tech. rep. Carnegie Mellon University.

Póczos, Barnabás, Liang Xiong, and Jeff Schneider (2011). "Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. UAI'11. Barcelona, Spain: AUAI Press, pp. 599–608. ISBN: 978-0-9749039-7-2. URL: http://dl.acm.org/citation.cfm?id=3020548.3020618.

Póczos, Barnabás, Liang Xiong, Dougal J Sutherland, and Jeff Schneider (2012). "Nonparametric kernel estimators for image classification". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pp. 2989–2996.

Pollard, David (1982). "Quantization and the method of k-means". In: *IEEE Transactions on Information theory* 28.2, pp. 199–205.

— (1990). "Empirical processes: theory and applications". In: *NSF-CBMS regional conference series in probability and statistics*. JSTOR, pp. i–86.

Principe, Jose C (2010). *Information theoretic learning: Rényi's entropy and kernel perspectives*. Springer Science & Business Media.

Quadrianto, Novi, James Petterson, and Alex J Smola (2009). "Distribution matching for transduction". In: *Advances in Neural Information Processing Systems*, pp. 1500–1508.

Ramdas, Aaditya, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman (2015). "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions". In: *AAAI*, pp. 3571–3577.

Ramdas, Aaditya, Nicolás García Trillos, and Marco Cuturi (2017). "On Wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2, p. 47.

Rashid, Sabrina, Zhicheng Long, Shashank Singh, Maryam Kohram, Harsh Vashistha, Saket Navlakha, Hanna Salman, Zoltán N Oltvai, and Ziv Bar-Joseph (2019). "Adjustment in tumbling rates improves bacterial chemotaxis on obstacle-laden terrains". In: *Proceedings of the National Academy of Sciences*, p. 201816315.

Rashid, Sabrina, Shashank Singh, Saket Navlakha, and Ziv Bar-Joseph (2019). "A bacterial based distributed gradient descent model for mass scale evacuations". In: *Swarm and Evolutionary Computation* 46, pp. 97–103.

Reddi, S. J. and B. Poczos (2013). "Scale Invariant Conditional Dependence Measures". In: *International Conference on Machine Learning (ICML)*.

Reitzner, Matthias and Matthias Schulte (2013). "Central limit theorems for U-statistics of Poisson point processes". In: *The Annals of Probability* 41.6, pp. 3879–3909.

Renard, Benjamin and Michel Lang (2007). "Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology". In: *Advances in Water Resources* 30.4, pp. 897–912.

Rényi, A. (1970). *Probability Theory*. North-Holland Publishing Company, Amsterdam.

Rey, Mélanie and Volker Roth (2012). "Meta-Gaussian information bottleneck". In: *Advances in Neural Information Processing Systems*, pp. 1916–1924.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *ICML*.

Rio, Emmanuel (2009). "Upper bounds for minimal distances in the central limit theorem". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 45. 3. Institut Henri Poincaré, pp. 802–817.

— (2011). "Asymptotic constants for minimal distance in the central limit theorem". In: *Electronic Communications in Probability* 16, pp. 96–103.

Robert, Christian P (2004). *Monte Carlo methods*. Wiley Online Library.

Rubenstein, Paul K, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya Tolstikhin (2019). "Practical and Consistent Estimation of f-Divergences". In: *arXiv preprint arXiv:1905.11112*.

Rubner, Yossi, Carlo Tomasi, and Leonidas J Guibas (2000). "The earth mover's distance as a metric for image retrieval". In: *International journal of computer vision* 40.2, pp. 99–121.

Rüschendorf, Ludger (1985). "The Wasserstein distance and approximation theorems". In: *Probability Theory and Related Fields* 70.1, pp. 117–129.

Sadhanala, Veeranjaneyulu, Yu-Xiang Wang, and Ryan J Tibshirani (2016). "Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers". In: *Advances in Neural Information Processing Systems*, pp. 3513–3521.

Sandler, Roman and Michael Lindenbaum (2011). "Nonnegative matrix factorization with earth mover's distance metric for image analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, pp. 1590–1602.

Schweder, Tore (1975). "Window estimation of the asymptotic variance of rank estimators of location". In: *Scandinavian Journal of Statistics*, pp. 113–126.

Sekeh, Salimeh Yasaei and Alfred O Hero (2019). "Geometric estimation of multivariate dependency". In: *arXiv preprint arXiv:1905.08594*.

Sekeh, Salimeh Yasaei, Brandon Oselio, and Alfred O Hero (2018). "Multi-class Bayes error estimation with a global minimal spanning tree". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, pp. 676–681.

Shan, C., S. Gong, and P. W. Mcowan (2005). "Conditional Mutual Information Based Boosting for Facial Expression Recognition". In: *British Machine Vision Conference (BMVC)*.

Shklarsh, Adi, Gil Ariel, Elad Schneidman, and Eshel Ben-Jacob (2011). "Smart swarms of bacteria-inspired agents with performance adaptable interactions". In: *PLoS computational biology* 7.9, e1002177.

Singh, Shashank, Simon S Du, and Barnabás Póczos (2016). "Efficient Nonparametric Smoothness Estimation". In: *Advances in Neural Information Processing Systems*, pp. 1010–1018.

Singh, Shashank, Ashish Khetan, and Zohar Karnin (2019). "DARC: Differentiable ARchitecture Compression". In: *arXiv preprint arXiv:1905.08170*.

Singh, Shashank and Barnabás Póczos (2014a). "Exponential concentration of a density functional estimator". In: *Advances in Neural Information Processing Systems*, pp. 3032–3040.

— (2014b). "Generalized Exponential Concentration Inequality for Rényi Divergence Estimation". In: *Proceedings of The 31st International Conference on Machine Learning*, pp. 333–341.

— (2016a). "Analysis of k-Nearest Neighbor Distances with Application to Entropy Estimation". In: *arXiv preprint arXiv:1603.08578*.

— (2016b). "Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functional Estimators". In: *Advances in Neural Information Processing Systems*, pp. 1217–1225.

— (2017). "Nonparanormal Information Estimation". In: *International Conference on Machine Learning*, pp. 3210–3219.

— (2018). "Minimax Distribution Estimation in Wasserstein Distance". In: *arXiv preprint arXiv:1802.08855*.

Singh, Shashank, Barnabás Póczos, and Jian Ma (2018). "Minimax reconstruction risk of convolutional sparse dictionary learning". In: *International Conference on Artificial Intelligence and Statistics*, pp. 1327–1336.

Singh, Shashank, Sabrina Rashid, Zhicheng Long, Saket Navlakha, Hanna Salman, Zoltán N Oltvai, and Ziv Bar-Joseph (2016). "Distributed Gradient Descent in Bacterial Food Search". In: *arXiv preprint arXiv:1604.03052*.

Singh, Shashank, Bharath K Sriperumbudur, and Barnabás Póczos (2018a). "Minimax Estimation of Quadratic Fourier Functionals". In: *arXiv preprint arXiv:1803.11451*.

— (2018b). "Minimax Estimation of Quadratic Fourier Functionals". In: *arXiv preprint arXiv:1803.11451*.

Singh, Shashank, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos (2018). "Nonparametric density estimation under adversarial losses". In: *Advances in Neural Information Processing Systems*, pp. 10225–10236.

Singh, Shashank, Yang Yang, Barnabás Póczos, and Jian Ma (2019). "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks". In: *Quantitative Biology*, pp. 1–16.

Smirnov, Nickolay (1948). "Table for estimating the goodness of fit of empirical distributions". In: *The annals of mathematical statistics* 19.2, pp. 279–281.

Solomon, Justin, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas (2015). "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics (TOG)* 34.4, p. 66.

Sønderby, Casper Kaae, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár (2016). "Amortised map inference for image super-resolution". In: *arXiv preprint arXiv:1610.04490*.

Sricharan, Kumar, Raviv Raich, and Alfred O Hero (2011). "k-nearest neighbor estimation of entropies with confidence". In: *IEEE International Symposium on Information Theory*. IEEE, pp. 1205–1209.

Sricharan, Kumar, Raviv Raich, and Alfred O Hero III (2012). "Estimation of nonlinear functionals of densities with confidence". In: *Information Theory, IEEE Transactions on* 58.7, pp. 4135–4159.

Sricharan, Kumar, Dennis Wei, and Alfred O Hero (2013). "Ensemble estimators for multivariate entropy estimation". In: *IEEE Transactions on Information Theory* 59.7, pp. 4374–4388.

Sriperumbudur, Bharath (2016). "On the optimal estimation of probability measures in weak and strong topologies". In: *Bernoulli* 22.3, pp. 1839–1893.

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2017). "Density estimation in infinite dimensional exponential families". In: *The Journal of Machine Learning Research* 18.1, pp. 1830–1888.

Sriperumbudur, Bharath K, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet (2010a). "Non-parametric estimation of integral probability metrics". In: *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, pp. 1428–1432.

— (2010b). "Non-parametric estimation of integral probability metrics". In: *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, pp. 1428–1432.

— (2012). "On the empirical estimation of integral probability metrics". In: *Electronic Journal of Statistics* 6, pp. 1550–1599.

Srivastava, Santosh and Maya R Gupta (2008). "Bayesian estimation of the entropy of the multivariate Gaussian". In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1103–1107.

Staib, Matthew and Stefanie Jegelka (2019). "Distributionally Robust Optimization and Generalization in Kernel Methods". In: *arXiv preprint arXiv:1905.10943*.

Studenỳ, Milan and Jirina Vejnarová (1998). "The multiinformation function as a tool for measuring stochastic dependence". In: *Learning in graphical models*. Springer, pp. 261–297.

Su, L. and H. White (2008). "A nonparametric Hellinger metric test for conditional independence". In: *Econometric Theory* 24, pp. 829–864.

Sutherland, D., H-Y Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton (2017). "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". In: *ICLR*. URL: https://arxiv.org/abs/1611.04488.

Sutherland, Dougal J (2016). "Scalable, Flexible and Active Learning on Distributions". PhD thesis. Carnegie Mellon University.

Suzuki, Taiji (2018). "Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality". In: *arXiv preprint arXiv:1810.08033*.

Szabó, Zoltán (2014). "Information theoretical estimators toolbox". In: *The Journal of Machine Learning Research* 15.1, pp. 283–287.

Szabó, Zoltán, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur (2015). "Two-stage sampled learning theory on distributions". In: *Artificial Intelligence and Statistics*, pp. 948–957.

Szabó, Zoltán, Barnabás Póczos, and András Lőrincz (2007). "Undercomplete blind subspace deconvolution". In: *Journal of Machine Learning Research* 8.May, pp. 1063–1095.

Szabó, Zoltán, Barnabás Póczos, Gábor Szirtes, and András Lőrincz (2007). "Post nonlinear independent subspace analysis". In: *International Conference on Artificial Neural Networks*. Springer, pp. 677–686.

Székely, Gábor J, Maria L Rizzo, and Nail K Bakirov (2007). "Measuring and testing dependence by correlation of distances". In: *The annals of statistics* 35.6, pp. 2769–2794.

Tolstikhin, Ilya, Bharath K Sriperumbudur, and Krikamol Muandet (2017). "Minimax estimation of kernel mean embeddings". In: *The Journal of Machine Learning Research* 18.1, pp. 3002–3048.

Tolstikhin, Ilya O, Bharath K Sriperumbudur, and Bernhard Schölkopf (2016). "Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels". In: *Advances in Neural Information Processing Systems*, pp. 1930–1938.

Tseng, Paul (2000). "Nearest q-flat to m points". In: *Journal of Optimization Theory and Applications* 105.1, pp. 249–252.

Tsybakov, A. B. and E. C. van der Meulen (1996). "Root-n Consistent Estimators of Entropy for Densities with Unbounded Support". In: *Scandinavian J. Statistics* 23, pp. 75–83.

Tsybakov, A.B. (2008). *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated. ISBN: 0387790519, 9780387790510.

Tsybakov, Alexandre B (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.

Uppal, Ananya, Shashank Singh, and Barnaás Póczos (2019). "Nonparametric Density Estimation under Besov IPM Losses". In: *arXiv preprint arXiv:1902.03511*.

Vapnik, Vladimir N and A Ya Chervonenkis (2015). "On the uniform convergence of relative frequencies of events to their probabilities". In: *Measures of complexity*. Springer, pp. 11–30.

Varga, Richard S (2009). *Matrix Iterative Analysis*. Vol. 27. Springer Science & Business Media.

Villani, Alfonso (1985). "Another note on the inclusion $L_p(\mu) \subset L_q(\mu)$". In: *The American Mathematical Monthly* 92.7, pp. 485–C76.

Villani, Cédric (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Villmann, T. and S. Haase (2010). *Mathematical Aspects of Divergence Based Vector Quantization Using Frechet-Derivatives*. University of Applied Sciences Mittweida.

Wang, Qing, Sanjeev R Kulkarni, and Sergio Verdú (2009). "Divergence estimation for multidimensional densities via k-nearest-neighbor distances". In: *IEEE Transactions on Information Theory* 55.5, pp. 2392–2405.

Wasserman, Larry (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

Watanabe, Satosi (1960). "Information theoretical analysis of multivariate correlation". In: *IBM J. of research and development* 4.1, pp. 66–82.

Weed, Jonathan and Francis Bach (2017). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In: *arXiv preprint arXiv:1707.00087*.

Weed, Jonathan and Quentin Berthet (2019). "Estimation of smooth densities in Wasserstein distance". In: *arXiv preprint arXiv:1902.01778*.

Wendland, Holger (2004). *Scattered data approximation*. Vol. 17. Cambridge university press.

Wilson, Andrew and Zoubin Ghahramani (2010). "Copula processes". In: *Advances in Neural Information Processing Systems*, pp. 2460–2468.

Wisler, Alan, Kevin Moon, and Visar Berisha (2017). "Direct ensemble estimation of density functionals". In: *arXiv preprint arXiv:1705.06315*.

Wolsztynski, E., E. Thierry, and L. Pronzato (2005a). "Minimum-entropy estimation in semi-parametric models". In: *Signal Process.* 85.5, pp. 937–949. ISSN: 0165-1684.

Wolsztynski, Eric, Eric Thierry, and Luc Pronzato (2005b). "Minimum-entropy estimation in semi-parametric models". In: *Signal Processing* 85.5, pp. 937–949.

Wu, Bichen, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer (2018). "Shift: A zero flop, zero parameter alternative to spatial convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9127–9135.

Yang, Yang, Ruochi Zhang, Shashank Singh, and Jian Ma (2017). "Exploiting sequence-based features for predicting enhancer–promoter interactions". In: *Bioinformatics* 33.14, pp. i252–i260.

Yang, Yuhong and Andrew Barron (1999). "Information-theoretic determination of minimax rates of convergence". In: *The Annals of Statistics*, pp. 1564–1599.

Yarotsky, Dmitry (2017). "Error bounds for approximations with deep ReLU networks". In: *Neural Networks* 94, pp. 103–114.

Zellinger, Werner, Bernhard A Moser, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz (2019). "Robust unsupervised domain adaptation for neural networks via moment alignment". In: *Information Sciences*.

Zhang, K., J. Peters, D. Janzing, and B. Scholkopf (2011). "Kernel-based conditional independence test and application in causal discovery". In: *Uncertainty in Artificial Intelligence (UAI)*.

Zhang, Meng, Yang Liu, Huan-Bo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao (2016). "Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation." In: *AAAI*, pp. 2870–2876.

Zhang, Tong (2002). "Covering number bounds of certain regularized linear function classes". In: *Journal of Machine Learning Research* 2.Mar, pp. 527–550.

Zhou, Ding-Xuan (2002). "The covering number in learning theory". In: *Journal of Complexity* 18.3, pp. 739–767.