Carnegie Mellon University

Thesis Proposal

Estimating Probability Distributions and their Properties

Author:
Shashank Singh

Supervisor:
Dr. Barnabás Póczos

August 8, 2018

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

Contents

Abstract

This thesis studies several theoretical problems in nonparametric statistics and machine learning, mostly in the areas of estimating or generating samples from a probability distribution, estimating a real-valued functional of a probability distribution, or testing a hypothesis about a probability distribution, using IID samples from that distribution. For distribution estimation, we consider a large, novel class of losses, under which high-dimensional nonparametric distribution estimation is more tractable than under the usual $\mathcal{L}^2$ loss. These losses have with connections with recent methods such as generative adversarial modelling, helping to explain why these methods appear to perform well at problems that are intractable from traditional perspectives of nonparametric statistics. 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. Finally, we propose to study some applications of these density functional estimators to classical hypothesis testing problems such as two-sample (homogeneity) or (conditional) independence testing. A consistent theme is that, although traditional nonparametric density estimation is intractable in high-dimensions, several equally (or more) useful tasks are relatively more tractable, even with similar or weaker assumptions on the distribution.

1 Introduction

This thesis studies several different problems in nonparametric statistics. As such, we begin with a brief formal description of the problems considered.

1.1 Formal Problem Types Considered

In this section, we briefly describe, at a high-level, the formal structure that defines the problems we consider in this thesis.

Suppose we observe n IID samples $X_1, ..., X_n \overset{IID}{\sim} P$ from an unknown probability distribution P lying in a nonparametric class $\mathcal{P}$ of distributions. This thesis addresses special cases of several basic problems in statistics:

1. **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 (potentially randomized) function $\widehat{P} : \mathcal{X}^n \to \mathcal{P}$ that has small worst-case risk under ℓ:

$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_1,...,X_n \overset{IID}{\sim} P} \left[\ell \left(P, \widehat{P}(X_1, ..., X_n) \right) \right].$$

2. **Implicit Distribution Estimation (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 $P_{\widehat{f}(X_1,...,X_n,Z)|X_1,...,X_n} \in \mathcal{P}$ is the conditional distribution of $\widehat{f}(X_1, ..., X_n, Z)$ given $X_1, ..., X_n$, then

$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X \sim P} \left[\ell \left(P, P_{\widehat{f}(X_1,...,X_n,Z)|X_1,...,X_n} \right) \right].$$

3. **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 (potentially randomized) function $\widehat{F} : \mathcal{X}^n \to \mathbb{R}$ that has small worst-case $\mathcal{L}^2$ risk:

$$\sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_1,...,X_n \overset{IID}{\sim} P} \left[\left(F(P) - \widehat{F}(X_1, ..., X_n) \right)^2 \right].$$

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 an $\alpha \in (0, 1)$, we would like to compute a test statistic $\widehat{P} : \mathcal{X}^n \to \{\mathcal{P}_0, \mathcal{P}_1\}$ that has high power

$$\inf_{P \in \mathcal{P}} \Pr_{X_1,...,X_n} \left[\widehat{P}(X_1, ..., X_n) = \mathcal{P}_1 \right]$$

whenever $P \in \mathcal{P}_1$, subject to

$$\sup_{P \in \mathcal{P}_0} \Pr_{X_1,...,X_n \overset{IID}{\sim} P} \left[\widehat{P}(X_1, ..., X_n) = \mathcal{P}_1 \right] \leq \alpha.$$

In each of the above problems, two 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 null hypothesis $\mathcal{P}_0$.

1.2 Organization of this Proposal

We begin, in Section 2, by establishing some common notation and context that will be used throughout this proposal.

Section 3 motivates and discusses our past and proposed work on distribution estimation under alternative losses, beginning with a discussion of classical density estimation and its shortcomings in Section 3.1, continuing with two threads along which we have pursued this topic in Section 3.2, and finishing with proposed work that unifies these two threads in Section 3.3. It also discusses relevant connections between implicit and explicit distribution estimation.

Section 4 discusses our past and proposed work on distribution functional estimation. Since this work involved four relatively distinct projects, after a brief summary of the state-of-the-art in distribution functional estimation, we discuss each project in its own section (Sections 4.4, 4.5, 4.6, and 4.7). Since our results on distribution functional estimation are quite many, in Section 4.8 we give a condensed summary of the results in a tabular format. We then end this section with a discussion of proposed work, which unifies the four different projects, including proposed study of applications to statistical hypothesis testing.

Section 5 gives a timeline outlining when I expect to complete each piece of proposed work.

2 General Setting & Notation of This Proposal

All problems considered in this thesis begin by observing n IID observations $X_1, ..., X_n \overset{IID}{\sim} P$ from a probability distribution P on a sample space $\mathcal{X}$. P is unknown, but is assumed to lie in a family $\mathcal{P}$ of probability distributions. The sample space $\mathcal{X}$ and class $\mathcal{P}$ of distributions will vary from problem to problem; examples range from the entire class of all Borel probability distributions on an arbitrary metric space $\mathcal{X}$ to smoothness classes (e.g., balls in Sobolev, Hölder, Besov, or reproducing kernel Hilbert spaces) over the d-dimensional Euclidean unit cube $\mathcal{X} = [0, 1]^d$. In some cases, we will endeavor to unify several of these settings, which are typically analyzed using different approaches, under a single analysis framework. $\mathcal{P}$ is typically assumed to be known, although we also sometimes consider the harder ("adaptive") case in which $\mathcal{P}$ has a known form but also has some unknown parameters (such as a smoothness index or intrinsic dimension).

Our work generally focuses on point estimation in the minimax statistical framework[1] because this furnishes a general and provable notion of optimality of estimators, although we occasionally also consider construction of confidence intervals.

Finally, it is worth noting that all results in this thesis will be derived with explicit forms for leading "constant factors"; however, for brevity, in this proposal, we omit the values of these constant factors.

3 Alternative Losses for Distribution Estimation

In this section, we motivate and study a novel theoretical framework for estimating a probability distribution (with or without a density). The main novelty is in considering a larger class of losses, besides the $\mathcal{L}^2$ (or $\mathcal{L}^p$) loss typically considered in classical nonparametric density estimation. As a result, this framework subsumes that of classical nonparametric density estimation, but also allows a unified analysis of several much more general problems.

Importantly, these losses allow us to meaningfully estimate distributions that are not absolutely continuous (with respect to a base measure), or even sample spaces where no natural base measure exists.

Generative adversarial networks (GANs) and variational autoencoders (VAEs), which have become popular tools for implicit generative modeling (the problem of learning a transformation from a known latent distribution to an unknown sampling distribution given samples from the latter) implicitly use losses similar to those we consider. Hence, we show, our results have implications for these methods.

Finally, since our framework allows for distributions lacking densities, it naturally encompasses the problem of estimating a distribution supported on a manifold. Hence, we conclude this section by proposing future work that generalizes and unifies the problems of manifold learning and of learning a density with respect to the volume form on a manifold.

3.1 General Setup & Background

Density estimation, along with regression, is one of the most well-studied problems in nonparametric statistics. As such, we cannot review the literature here, and discuss only key classical results and the recent results most relevant to our work. More thorough discussion can be found in Tsybakov [2008] and Wasserman [2006].

Fix a class $\mathcal{P}$ of probability distributions on a sample space $\mathcal{X}$. Suppose that we observe n IID samples $X_1, ..., X_n \overset{IID}{\sim} P$ from some unknown distribution $P \in \mathcal{P}$.

Given a loss function $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$, we are interested in constructing an estimator $\widehat{P} : \mathcal{X}^n \to \mathcal{P}$ of P that minimizes the risk

$$R(P, \widehat{P}) := \mathop{\mathbb{E}}_{X_1,...,X_n \overset{IID}{\sim} P} \left[\ell \left(P, \widehat{P}(X_1, ..., X_n) \right) \right].$$

[1] That is, we seek estimators that minimize worst-case (over $P \in \mathcal{P}$) expected (over $X_1, ..., X_n$) error.

The minimax quantity

$$M(\mathcal{P}, \ell) := \inf_{\widehat{P}: \mathcal{X}^n \to \mathcal{P}} \sup_{P \in \mathcal{P}} R(P, \widehat{P}),$$

of interest depends on the class $\mathcal{P}$ of distributions (which implicitly encodes dependence on the sample space $\mathcal{X}$) and the loss ℓ. In the sequel, except where this causes ambiguity, we abbreviate $\widehat{P} = \widehat{P}(X_1, ..., X_n)$. Typically, the loss ℓ is strong enough that, for any $P, Q \in \mathcal{P}$, $\ell(P, Q) = 0$ implies $P = Q$. Since, in the nonparametric setting, $\mathcal{P}$ is infinite-dimensional, this makes density estimation challenging both computationally and statistically, as obtaining a consistent estimate $\widehat{P}$ requires both the representation of $\widehat{P}$ in memory and the number of parameters being estimated to grow unboundedly with the sample size n. As a result, the computational complexities of most methods are super-linear with n (although $O(n \log n)$ is often possible), and statistical convergence rates are typically strictly slower than the "parametric" rate $n^{-1/2}$.

3.1.1 Classical density estimation

The vast majority of work in nonparametric statistics has focused on the case where every $Q \in \mathcal{P}$ is absolutely continuous (i.e., $Q \ll \mu$) and hence has a density p with respect to a given base measure μ on the sample space $\mathcal{X}$ (e.g., the Lebesgue measure when $\mathcal{X} \subseteq \mathbb{R}^d$). Moreover, the loss ℓ is almost always taken to be the $\mathcal{L}^p$ distance

$$\ell_p(P, Q) = \left(\int_{\mathcal{X}} (p(x) - q(x))^p \, d\mu(x) \right)^{1/p}, \tag{1}$$

and mostly with $p = 2$ [Wasserman, 2006, p. 57]. [2]

This significantly simplifies analysis because one can study distribution estimation pointwise on $\mathcal{X}$, as well as rely on the structure (e.g., the existence of an orthonormal basis) of the function space $\mathcal{L}^2$. For simplicity, the sample space $\mathcal{X}$ is usually taken to be $\mathbb{R}^d$ or the unit cube $[0, 1]^d$, and the class $\mathcal{P}$ is typically taken to be a ball in a smooth function space, such as a Hölder, Sobolev, or Besov space. For Hölder or Sobolev classes with smoothness index s, minimax rates are typically

$$M\left(\mathcal{C}^s, \|\cdot - \cdot\|_{\mathcal{L}^2}\right) \asymp M\left(\mathcal{H}^s, \|\cdot - \cdot\|_{\mathcal{L}^2}\right) = n^{-\frac{s}{2s+d}} \gg n^{-1/2}.$$

3.2 More general losses

The assumption of absolute continuity can be quite limiting, as it excludes structured distributions such as those supported on manifolds or other low-dimensional subspaces. (Indeed,

[2] The Kullback-Leibler (KL) divergence has also been used as a loss. However, this is most natural when $\mathcal{P}$ is a (potentially non-parametric) exponential family [Wainwright et al., 2008, Sriperumbudur et al., 2017]; otherwise, since KL divergence is quite sensitive to tails of the distribution, deriving uniform convergence rates often involves assuming that p is lower bounded away from 0 (see Assumption (LB) in Section 4.8.1), in which case KL divergence becomes asymptotically equivalent to $\mathcal{L}^2$ loss anyway (as one can easily check via the fact that $-\log(1 + x) \leq x^2 - x$ for all $x \geq -0.5$).

more generally, the assumption of smoothness directly competes with concentration of the distribution, even though both are typically desirable properties.) Moreover, the widespread use of $\mathcal{L}^2$ loss is motivated primarily by simplicity of analysis, rather than any intrinsic quality of $\mathcal{L}_2$ loss as a performance measure. We consider two main alternative classes of losses: Wasserstein (optimal transport) distances and integral probability metrics (IPMs; a.k.a., adversarial losses).

Wasserstein distances:[3] Fix a metric sample space $(\mathcal{X}, \rho)$. Given two probability distributions P and Q on $\mathcal{X}$, the (r-)Wasserstein distance $W_r(P, Q)$ between P and Q is defined by

$$W_r(P, Q) = \inf_{\mu \in \Pi(P,Q)} \left(\mathop{\mathbb{E}}_{(X,Y)\sim\mu} [\rho^r(X, Y)] \right)^{1/r},$$

where $\Pi(P, Q)$ is the set of possible couplings between P and Q (i.e., the set of probability distributions over $\mathcal{X} \times \mathcal{X}$ having P and Q as marginals). $W_1(P, Q)$ can be interpreted as the average distance (under ρ) that mass must be transported to transform the distribution P into the distribution Q, according to the most efficient possible transportation scheme. $W_r(P, Q)$ generalizes to exponential weightings; the case $r = 2$ is especially fruitful because several important problems, such as K-means, PCA, and their generalizations, can be easily expressed as distribution estimation under W_2 loss, for an appropriate class $\mathcal{P}$ of distributions. As the central quantities in the field of optimal transport theory, the metrics W_r have been extensively studied in a number of contexts; see Villani [2008] for comprehensive review of the mathematical theory, although there does not yet exist a review of the numerous recent applications in machine learning and statistics.

More relevant to our work, there has been a substantial line of work, beginning with that of Dudley [1967, 1969] and continuing with Dobrić and Yukich [1995], Boissard et al. [2014], Fournier and Guillin [2015], Weed and Bach [2017], and Lei [2018], among others, studying the mean convergence of the empirical distribution

$$P_n := \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i\}}$$

to the true distribution P, in Wasserstein distance (i.e., the rate at which $\mathbb{E}[W_r^r(P, Q)] \to 0$. When the sample space $\mathcal{X} = \mathbb{R}^d$, the key problem parameters determining convergence rates are the exponent r, the dimension d, and the concentration of the distribution P, in terms of the number of its finite moments; specifically, Fournier and Guillin [2015] showed

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} [W_r^r(P, P_n)] \in O\left(n^{-1/2} + n^{-r/d} + n^{\frac{p-q}{q}}\right). \tag{2}$$

Weed and Bach [2017] considered the case of an arbitrary totally bounded metric space $(\mathcal{X}, \rho)$, in terms of the covering numbers $N(\mathcal{X}, \rho; \epsilon)$ of the space. The general upper bound

[3] The Wasserstein metric has been variously attributed to Monge, Kantorovich, Rubinstein, Gini, Mallows, and others; see Chapter 3 of [Villani, 2008] for detailed history.

in terms of covering numbers is too complex to state here, but the main consequence of interest is that, under several different notions of dimension, if $\mathcal{X}$ is a d-dimensional set, then the convergence rate is of order

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} \left[W_r^r\left(P, P_n\right)\right] \in O\left(n^{-1/2} + n^{-r/d}\right).$$

Importantly, this means that when $\mathcal{X}$ is a low-dimensional set embedded in a high-dimensional space $\mathbb{R}^D$, the convergence rate in Wasserstein distance depends only on the intrinsic dimension d, rather than the ambient dimension D.

In our paper Singh and Póczos [2018], we extended the results of Weed and Bach [2017] in two main ways. First, we considered the case of an unbounded metric space $\mathcal{X}$, making generalized finite-moment assumptions (see Assumption (ℓ-MM) in Section 4.8.1); we then proved the same upper bound (2) shown for the case $\mathcal{X} = \mathbb{R}^d$ by Fournier and Guillin [2015]. Second, whereas prior work only studied the convergence of the empirical distribution P_n to P, it remained unclear whether another estimator $\widehat{P}$ might converge more quickly; we proved a minimax lower bound, in terms of the packing number of $(\mathcal{X}, \rho)$, that implies, in many cases, that *no estimator* can converge at a faster rate than the empirical distribution.

The key feature of this analysis is that, under Wasserstein loss, the minimax rate of distribution estimation depends not on the ambient dimension D of the data, but rather on the intrinsic dimension d of the distribution P. Often, such as when the data lie along a low-dimensional manifold, $d \ll D$, and so a much faster rate of convergence can be achieved. *This kind of observation is not possible in the classical density estimation framework.* Moreover, since the estimator is simply the empirical distribution, this rate is achieved completely adaptively; no hyperparameter tuning or knowledge of d is required, resulting in a computationally efficient and realistic estimator.

Integral probability metrics (IPMs): Suppose $\mathcal{F}$ is a class of bounded[4], measurable functions on $\mathcal{X}$. The $\mathcal{F}$-IPM $\rho_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is defined for all $P, Q \in \mathcal{P}$ by

$$\rho_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] \right|. \tag{3}$$

$\rho_{\mathcal{F}}$ has also been called an *adversarial loss*, because f can be interpreted as the linear feature that an adversary (such as the discriminator in a generative adversarial network) would use to distinguish the distributions P and Q.

By choosing the class $\mathcal{F}$ appropriately, one can use the form (3) to encode a huge class of (pseudo)metrics on probability distributions, including $\mathcal{L}_p$, Sobolev, MMD, 1-Wasserstein (a.k.a. Kantorovich-Rubenstein), total variation, Kolmogorov-Smirnov, and Dudley metrics. In fact, IPMs are in fact rather classical objects in empirical process theory and statistical learning theory, and, in the case that $\mathcal{P}$ is the family of all probability distributions on $\mathcal{X}$, there exist rich theories of convergence rates under IPMs (for example those based on

[4]The boundedness assumption can be weakened for some classes $\mathcal{P}$, but is needed in general to ensure we do not subtract $\infty - \infty$ in Equation (3).

covering numbers [Dudley, 1967] or the Vapnik-Chervonenkis dimension [Vapnik, 2013] of $\mathcal{F}$).

However, when P is a more interesting class of distributions, such as a smoothness class, the results obtained from these classical methods become loose [Liang, 2017]. Recently, Liang [2017] studied the case where both $\mathcal{F}$ and $\mathcal{P}$ are s- and t-Sobolev balls, respectively, showing that a (well-tuned) orthogonal series estimate $\widehat{P}$ of P converges at the rate

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[\rho_{\mathcal{F}}\left(P,\widehat{P}\right)\right] \lesssim n^{-\frac{s+t}{2(s+t)+d}},$$

often much faster than the rate of

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[\rho_{\mathcal{F}}\left(P,P_n\right)\right] \lesssim n^{-\frac{s}{d}},$$

given for the empirical distribution by classical theory.

In our paper Singh et al. [2018b], we studied the minimax rate for quite general classes $\mathcal{P}$ and $\mathcal{F}$ defined in terms of standard sequence space representations. We showed that the upper bound of Liang [2017] for the Sobolev case is loose and that the (strictly faster, for $t > 0$) minimax rate (achieved using the same orthogonal series estimate, albeit with a different tuning) is

$$M\left(\mathcal{H}^t, \rho_{\mathcal{H}^t}\right) \asymp n^{-\frac{s+t}{2t+d}} + n^{-1/2}.$$

We also showed that the optimal tuning for this problem is the same as under the $\mathcal{L}^2$ loss, allowing us to construct a minimax estimator that adapts to unknown t, based on methods for $\mathcal{L}^2$ loss. Finally, we established rates for a number of other classes $\mathcal{F}$ and $\mathcal{P}$. For example, we showed (for the first time, it appears) that balls in reproducing kernel Hilbert spaces with translation-invariant kernels in $\mathcal{L}^2$ are $n^{-1/2}$-uniform Glivenko-Cantelli classes; i.e., even when P is the class of all probability distributions on $\mathcal{X}$, if $\mathcal{F}$ is a ball in such an RKHS, then convergence of the empirical distribution in the IPM ρ_F is of the parametric order $\asymp n^{-1/2}$.

3.3 Future Work

Our works described above intersect when $\mathcal{F}$ is the class of 1-Lipschitz functions on $\mathcal{X}$ (so that $\rho_{\mathcal{F}} = W_1$) and P is the class of all distributions on the unit cube $\mathcal{X} = [0,1]^d$. Starting from this case, the results for Wasserstein distance dictate performance when we change the sample space $\mathcal{X}$ and the results for IPMs dictate the results when we add smoothness constraints to P. Is it possible to combine these results? Of particular interest, is there a framework of smooth distribution estimation that does not require the distribution to be absolutely continuous with respect to Lebesgue measure (but perhaps with respect to another unknown measure, such as the volume measure on an unknown d-dimensional manifold embedded in $[0,1]^D$)?

One might conjecture that, if we could formulate such a model, we could obtain a minimax convergence rate of $\asymp n^{-\frac{s+t}{2t+d}}$, much faster than both the $\asymp n^{-\frac{s}{d}}$ rate given by the results for Wasserstein distances and the $\asymp n^{-\frac{s+t}{2t+D}}$ rate given by the results for IPMs.

To study this, we propose to study distribution estimation (under Wasserstein loss) in a *smooth latent variable model*; specifically, we assume X is generated according to $X = f(Z)$, where Z is a random variable with a known, nice (e.g., Gaussian or uniform) distribution and $f : \mathcal{Z} \subseteq \mathbb{R}^d \to \mathcal{X} \subseteq \mathbb{R}^D$ is an *unknown* smooth function.

We can then ask two questions:

1. In **implicit distribution estimation** or **sampling**, we ask whether we can produce a function $\widehat{f} : \mathcal{Z} \to \mathcal{X}$ such that the distribution of $\widehat{f}(Z)$ is similar to that of $f(Z)$.

2. In **explicit distribution estimation**, we ask whether we can compute a distribution $\widehat{P} \in \mathcal{P}$ that is close to the distribution $P_{f(Z)}$ of $f(Z)$.

As we discuss briefly in the next subsection, under mild conditions, implicit and explicit distribution estimation are statistically equivalent (in that a solution to either yields a solution to the other with the same convergence rate). Both problem are closely related to, but distinct from, several well-studied problems in nonparametric statistics.

First, although the goal in this task is a form of function approximation, this problem is in some ways harder, and in other ways easier, than the problem of nonparametric regression. On one hand, the loss function W_r^r is relatively weak, and there may be many globally optimal $\widehat{f}$; on the other hand, since we never observe the latent variables $Z_1, ..., Z_n$ that generated the data $X_1 = f(Z_1), ..., X_n = f(Z_n)$, the problem is unsupervised, and it is unclear, for example, how to perform cross-validation. Given this similarity, it may also be interesting to explore a hypothesis class, recently proposed by Schmidt-Hieber [2017] for nonparametric regression, in which f is a composition of many smooth functions; in this case, Schmidt-Hieber [2017] showed that sparsely-connected deep ReLU networks are nearly minimax optimal, whereas all linear wavelet regressors are sub-optimal by a factor polynomial in n.

Second, this problem is also closely related to manifold learning, in which one assumes high-dimensional data are drawn (noisily) from an embedded low-dimensional manifold, which we seek to estimate (e.g., by estimating a local chart, such as ϕ). There are two main differences from prior work in this area. First, the support of X need not be a well-behaved manifold, because we assume only that ϕ is smooth, not that it is a diffeomorphism (i.e., we do not require ϕ to be locally invertible, let alone have a smooth inverse). Second, our goal is to estimate the distribution P_X, rather than its support; in particular, in contrast to manifold learning, we are not strongly concerned with areas of low probability mass. This is implicit in our choice of Wasserstein loss, rather than the Hausdorff distance typically used as the loss in manifold learning.

It is worth noting that, when the manifold itself is known *a priori* (e.g., for structured data, such as the space of symmetric matrices), there has been work on estimating a density with respect to the manifold's volume measure. In this case, one can generalize the Fourier transform to functions on the manifold; using this, one can then generalize conventional (Hilbert-Sobolev) smoothness assumptions, $\mathcal{L}^2$ loss, and kernel density estimation to the the manifold Asta [2014]. However, in our case, the manifold is unknown, making this elegant but highly-structured approach infeasible.

3.3.1 From Implicit to Explicit Distribution Estimation

In our paper Singh et al. [2018b] on distribution estimation under IPMs, we gave conditions under which upper bounds for implicit generative modeling imply upper bounds (of the same rate) for explicit generative modeling. The conditions are as follows:

1. The loss satisfies the triangle inequality.
2. We can draw arbitrarily many IID samples of Z.
3. There exists an explicit estimator $\widehat{P}$ that is uniformly consistent over the set of possible values taken by the implicit estimator ($\{P_{\widehat{f}(x_1,...,x_n;Z)} : x_1, ..., x_n \in \mathcal{X}\}$).

Are these assumptions satisfied? Clearly, the Wasserstein distance satisfies the triangle inequality, and by construction, it is easy to draw latent random variables; the only question is about condition 3. In the case of Wasserstein metric, one can actually show that condition 3. holds; the empirical distribution itself is a uniformly consistent distribution estimator. This is a bit strange; the empirical distribution based on the original samples $X_1, ..., X_n$ is sub-optimal, as it does not benefit from smoothness, but, by supplementing with data from an appropriate implicit estimator, we can bias the estimator towards our smoothness prior.

Computationally, this is quite unsatisfying, because the explicit distribution estimate, far from compressing the data, has actually significantly expanded the representation of the data! We leave with an open question: Does there exist a computationally efficient explicit distribution estimator $\widehat{P}$ under the latent variable model of smoothness?

4 Distribution Functional Estimation

Distribution functional estimation involves estimating the value of a (known) functional $F : \mathcal{P} \to \mathbb{R}$ of the distribution at P. Note that, in the particular case that F is linear and bounded, under mild assumptions, there exists a function $f : \mathcal{X} \to \mathbb{R}$ such that

$$F(P) = \mathop{\mathbb{E}}_{X \sim P} [f(X)],$$

and hence the empirical mean $\widehat{F}(P) := \frac{1}{n} \sum_{i=1}^{n} f(X_i)$ is usually a good estimator. While one can find open questions even in this relatively simple domain (e.g., how to perform robust, computationally efficient estimation under sparsity constraints [Du et al., 2017a]), here we are interested in the more challenging setting where F is non-linear. The nonlinear functional F of interest can be quite general, and a selection of functionals of interest is given in Table 1; typically what is required is that F is smooth over $\mathcal{P}$ (e.g., in the sense of having well-behaved Fréchet derivatives).

A simple (univariate) example of $F : \mathcal{P} \to \mathbb{R}$ is the (differential) Shannon entropy

$$F(P) = - \mathop{\mathbb{E}}_{X \sim P} \left[\log \left(\frac{dP}{d\mu} \right) \right],$$

where μ is some (known) base measure (and $\mathcal{P}$ is such that $Q \ll \mu$ for every $Q \in \mathcal{P}$), but our framework will also apply to multivariate functionals, such as the KL divergence $F : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ given by

$$F(P, Q) = - \mathop{\mathbb{E}}_{X \sim P} \left[\log \left(\frac{dP}{d\mu} \right) \right].$$

The first is a semiparametric problem, in which the true probability distribution is assumed only to lie in a large nonparametric class (e.g., a smoothness class), but the estimand is a univariate quantity. As such, convergence rates are typically faster than those for density estimation, but may often be slower than the parametric rate of $\asymp n^{-1}$ (in mean squared error).

The classes of distribution functionals that can be considered, as well as the assumptions that can be made on the distribution P, are quite diverse; as such we do not list them all here, but a diverse sample is given in Table 1.

4.1 Applications of Density Functional Estimation

Estimates of dissimilarity functionals can be directly used for nonparametric goodness-of-fit, independence, and two-sample testing [Anderson et al., 1994, Dumbgen, 1998, Ingster and Suslina, 2012, Goria et al., 2005, Pardo, 2005, Chwialkowski et al., 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 et al., 2005]. Estimates of dependence functionals can be directly used for structure learning [Chow and Liu, 1968, Liu et al., 2012] and feature selection [Peng et al., 2005] and optimal error estimation [Moon et al., 2015] in supervised learning.

In machine learning, Sobolev-weighted distances can also be used in transfer learning [Du et al., 2017b] and transduction learning [Quadrianto et al., 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 et al., 2011, Póczos et al., 2012], galaxy mass estimation [Ntampaka et al., 2015], ecological inference [Flaxman et al., 2015, 2016], aerosol prediction in climate science [Szabó et al., 2015], and causal inference [Lopez-Paz et al., 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].

4.2 Related Work

Perhaps the most central results in the theory of functional estimation are those of Birgé and Massart [1995], Laurent et al. [1996] for the case of twice Fréchet-differentiable functionals

F; for distributions P having a density p in the Hölder class $\mathcal{C}^s$, they established a minimax rate of order $\asymp n^{-\frac{8s}{4s+d}} + n^{-1}$ in mean squared error. This means that the parametric rate $\asymp n^{-1}$ is achieves when $s \geq d/4$, and the slower rate of $\asymp n^{-\frac{8s}{4s+d}}$ holds otherwise. For quadratic functionals (i.e., those that can be written in the form

$$F(P) = \sum_{z \in \mathcal{Z}} a_z \widetilde{P}_z^2 \tag{4}$$

where $\mathcal{Z}$ is some countable index set, $\widetilde{P}_z := \mathbb{E}_{X \sim P}[\phi_z(X)]$ for some family $\{\phi_z\}_{z \in \mathcal{Z}}$ of bounded functions, and $\{a_z\}_{z \in \mathcal{Z}}$ is a family of real-valued weights) such as $\mathcal{L}^2$, Sobolev, or RKHS inner products, norms, and distances, the optimal rate can usually be achieved using an appropriately tuned plug-in or basis thresholding estimator [Fan, 1991, Cai et al., 2005, Singh et al., 2016]. For more general functionals, minimax convergence rates are almost always achieved by correcting plug-in estimates via the von Mises expansion of the functional F [Krishnamurthy et al., 2014, Kandasamy et al., 2015]. Informally, the idea is to expand $F(p)$ as

$$F(p) = F(\widehat{p}) + \langle \nabla F(\widehat{p}), p - \widehat{p} \rangle_{\mathcal{L}^2} + \left\langle p - \widehat{p}, (\nabla^2 F(\widehat{p}))p - \widehat{p} \right\rangle_{\mathcal{L}^2} + O\left(\|p - q\|_{\mathcal{L}^2}^3 \right), \tag{5}$$

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), 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 form Equation (4), and so, as noted above, a simple plug-in estimate achieves the minimax rate. Finally, one can show that the $\|p-q\|_{\mathcal{L}^2}^3$ term is often negligible. Thus, summing the three estimated terms gives a minimax rate-optimal estimator.

In the adaptive case, where the smoothness index s is not known beforehand, the same rate of convergence can be achieved using Lepski's method [Mukherjee et al., 2015, 2016] or, in the case of quadratic functionals, using wavelet thresholding [Cai et al., 2006]. We do not know of a method based on wavelet-thresholding for more general functionals, which motivates one of the research topics proposed later in this document.

4.3 Recent Work on Density Functional Estimation

Density functional estimation is quite an active area of research in the statistics, machine learning, and signal processing communities, and we therefore, in this section, briefly review recent advances.

Confidence Intervals: While the vast majority of work in density functional estimation has focused on studying minimax rates for point estimation, there has also been some work on obtaining confidence intervals for such estimates. One approach is based on proving asymptotic normality [Sricharan et al., 2012, Moon and Hero, 2014, Krishnamurthy et al., 2015, Singh et al., 2016] of the estimator. This is useful for obtaining an asymptotically valid confidence interval on the density functional. The other approach is to prove finite-sample concentration bounds for the estimator [Liu et al., 2012, Singh and Póczos, 2014a,b].

While useful for obtaining confidence intervals, concentration inequalities can also be used for analyzing the downstream performance of procedures that use density functional estimates as subroutines. This has consequences for applications such as structure learning and statistical testing; for example, Liu et al. [2012] showed that a concentration inequality for mutual information estimation can be used to prove minimax optimal upper bounds on using the Chow-Liu procedure [Chow and Liu, 1968] to learn a forest-shaped graphical model.

Nonsmooth Shannon Functionals: In the case of Shannon information-theoretic functionals (such as Shannon entropy, mutual information, and KL divergence), the Fréchet differentiability of F requires the assumption that the probability densities in question are lower bounded away from 0 (see Assumption (LB). Until recently, it was unclear whether this assumption was necessary or simply a proof artefact. Jiao et al. [2017] showed that, without the lower boundedness assumption, the slows to $\asymp n^{-\frac{2s}{s+d}} + n^{-1}$.

Direct Estimation of Density-Ratio Functionals: In the case of f-divergences (i.e., divergences of the form

$$D_f(P,Q) = \mathop{\mathbb{E}}_{X\sim Q}\left[f\left(\frac{dP}{dQ}(X)\right)\right], \tag{6}$$

where $f : [0,\infty) \to \mathbb{R}$ is convex with $f(1) = 0$), there has also been some work on weakening the assumptions from smoothness conditions on the individual distributions P and Q to assumptions only on the relative density $\frac{dP}{dQ}$ [Noshad et al., 2017, Kpotufe, 2017].

Computational Advances: Relatively recently, there has been a focus on developing computationally efficient functional estimators, such as the linear-time estimators of Noshad and Hero [2018], Noshad and Hero III [2018] based on hashing.

My Work: In the next several subsections, I discuss my own contributions to density functional estimation over the past few years, published in the papers Singh and Póczos [2014a,b], Singh et al. [2016], Singh and Póczos [2016, 2017] and the preprint Singh et al. [2018a].

4.4 Plugging in a Boundary-Corrected Kernel Density

Given a known density functional F and assuming the sample space to be a subset of $\mathbb{R}^d$, we first considered a simple estimator, namely the plug-in estimate $\widehat{F} = F(\widehat{p})$, where $\widehat{p}$ is a pointwise estimate of the probability density p. While these estimators are quite simple, their convergence rates under standard nonparametric assumptions were previously unknown. In two papers in 2014 (Singh and Póczos [2014a] in ICML focusing on the special case of Rényi divergences Singh and Póczos [2014b] in NIPS considering general functionals), we established the first finite-sample convergence rate guarantees for estimators of this type. Under relatively mild conditions, these papers also proved finite-sample exponential concentration inequalities for these estimators (around their expectation), which which continue to be, to the best of our knowledge, relatively unique results for density functional estimators. In this section, we briefly summarize the main results of Singh and Póczos [2014a] and Singh and Póczos [2014b]. It is worth noting that the work of Krishnamurthy et al. [2014] and Kandasamy et al. [2015] has since provided improved convergence rates (at increased com-

putational cost) for bias-corrected variants of our estimators, based on von Mises expansion of the functional F.

4.4.1 Boundary Bias

These naïve estimates are subject "boundary bias" (i.e., bias due to discontinuity of the density at the boundary of its support). Hence, to bound the finite-sample bias of such a simple estimates, the density is required to satisfy some additional assumptions near the boundary of its support. One possibility is the "periodic boundary condition" considered in Kandasamy et al. [2015], in which the density is assumed to be the restriction of a 1-periodic function (in every dimension) to $[0,1]^d$; this is equivalent to replacing the unit cube $[0,1]^d$ with the d-dimensional torus, and boundary bias can then be corrected by replacing the kernel K with its 1-periodic summation $K_{periodic} = \sum_{z \in \mathbb{Z}} K(\cdot + z)$; Alternatively, one can consider the "vanishing boundary derivative condition", in which all derivatives of the density are assumed to approach 0 at the boundary of $[0,1]^d$. In this case, boundary bias can be corrected by replacing the kernel K with the summation of its "mirrored" versions across each subset of boundary; the formal definition of this "mirrored" kernel in high dimensions is rather technical and can be found in Singh and Póczos [2014a].

Although one can construct examples satisfying either of these assumptions, both assumptions are rather artificial, and work has been done on relaxing these assumptions; see, e.g., the thesis [Moon, 2016] of Kevin Moon for both kernel and nearest neighbor methods that avoid such strong boundary assumptions.

4.4.2 Main Results

We obtained two main results:

1. Suppose F is twice Fréchet differentiable over $\mathcal{P}$, and suppose p is s-Hölder continuous with s-order boundary conditions. For an appropriately chosen (s-order, bounded support) kernel K. Then, there exists a constant $C_B > 0$ such that, for any bandwidth $h \leq 1$, the bias
$$\mathbb{B}(\widehat{F}) := \mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} \left[\widehat{F}\right] - F(p)$$
is at most $\mathbb{B}(\widehat{F}) \leq C_B \left(h^s + \frac{1}{nh^d}\right)$. Note that this is minimized (up to constant factors) by setting $h = n^{-\frac{1}{s+d}}$, in which case (for a slightly different constant C_B), $\mathbb{B}(\widehat{F}) \leq C_B \left(n^{-\frac{s}{s+d}}\right)$.

2. Suppose F is once Fréchet differentiable over $\mathcal{P}$. Then, there exists a constant $C_V > 0$ such that, regardless of the bandwidth h, the estimator satisfies the concentration inequality
$$\mathop{\mathbb{P}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} \left[\left|\widehat{F} - \mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} \left[\widehat{F}\right]\right| > \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2 n}{C_V^2}\right). \tag{7}$$

It follows that $\mathbb{V}_{X_1,\ldots,X_n \overset{IID}{\sim} P}\left[\widehat{F}\right] \leq C_V^2 n^{-1}$.

Combining these two main results, via the usual decomposition of mean squared error ($\mathcal{L}^2$ risk) into the sum of squared bias and variance, gives a bound, for some $C > 0$, of

$$\mathrm{MSE}\left[\widehat{F}\right] \leq C\left(n^{-\frac{2s}{s+d}} + n^{-1}\right).$$

It is worth noting that many functionals of interest, such as Shannon entropy, are not Fréchet differentiable at arbitrary densities. Hence, to apply our results to these functionals, we may need additional restrictions on the class $\mathcal{P}$ of permissible densities. In the case of Shannon entropy (and most other information theoretic functionals), it is sufficient to assume that the true density is both upper bounded and lower bounded away from 0. That is, $\kappa^* := \sup_{x\in\mathcal{X}} p(x) < \infty$ and $\kappa_* := \inf_{x\in\mathcal{X}} p(x) > 0$; the constant factors in the above upper bounds will then depend on κ_* and κ^*.

This work was largely motivated by the 2-dimensional entropy and mutual information estimates analyzed in Liu et al. [2012]. While this thesis focuses primarily on the minimax convergence rates of estimators in mean squared error ($\mathcal{L}^2$ risk), concentration inequalities of the form (7) are useful for analyzing (via union bounds) applications that utilize many simultaneous estimates of a density functional; for example, Liu et al. [2012] showed that, using the boundary-corrected plug-in estimator above, the greedy Chow-Liu procedure [Chow and Liu, 1968], which requires estimates of mutual information between all pairs of available variables, can be used to provide minimax optimal estimates of forest-shaped graphical models as well as of the underlying probability densities.

4.5 Bias-Corrected k-Nearest Neighbor Estimators

Next, we investigated a classical and quite popular, but relatively poorly understood, approach to estimating information theoretic quantities, based on k-nearest neighbor statistics. This approach dates back to Kozachenko and Leonenko [1987], who studied a 1-nearest neighbor estimator for differential Shannon entropy. Generalizations have since been given by Goria et al. [2005] to use $k > 1$ nearest neighbors, by Wang et al. [2009] to estimate KL divergence, by Leonenko et al. [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 et al. [2011] for a survey of these estimators and discussion of their asymptotic consistency.

As we describe below, the construction of these estimators requires a rather precise analysis specific to the density functional F of interest, and these methods therefore apply only to a select group of density functionals (namely, those listed above). Firstly, the functional of interest must have the form

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

for some $f : [0, \infty) \to \mathbb{R} \cup \{-\infty, \infty\}$, and, furthermore, as described below, to perform the bias-correction, we must be able to analytically compute a particular expectation in terms of f. However, these estimators are relatively easy to compute, and, for the functionals for which these estimators are known, these estimators often provide the best empirical performance among known estimators [Pérez-Cruz, 2009, Szabó, 2014, Berrett et al., 2016, Gao et al., 2017].

Excepting the analysis of Tsybakov and van der 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 [2016], as well as the thesis Berrett et al. [2016] of Thomas Berrett in Richard Samworth's group at Cambridge, and work [Gao et al., 2017] 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 et al. [2005]). Among these works, our paper Singh and Póczos [2016] is unique in that it provides convergence rates not only for Shannon entropy estimation, but also for KL divergence and for more general (e.g., Rényi and Tsallis) entropies and divergences. This section briefly describes the main results of that paper. First, however, we provide some intuition for the estimators considered, which we call *bias-corrected fixed-k nearest neighbor*, or BCFkNN, estimators.

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

Let $c_d := \frac{\left(2\Gamma\left(1+\frac{1}{p}\right)\right)^d}{\Gamma\left(1+\frac{n}{p}\right)}$ denote the volume of the unit ℓ^p ball in $\mathbb{R}^d$, let μ denote the Lebesgue measure, and, for any $x \in \mathcal{X}$, $r > 0$, let $B(x, r) := \{y \in \mathcal{X} : \|x - y\|_p < r\}$ denote the radius-r ℓ^p-ball centered at x. Finally, for any $k \in [n]$ and $x \in \mathcal{X}$, let $\epsilon_k(x)$ denote the distance between x and its k^{th}-nearest neighbor among the data points $X_1, ..., X_n$.

The k-NN density estimator

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

is well-studied nonparametric density estimator [Loftsgaarden et al., 1965], motivated by noting that, for small $\epsilon > 0$,

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

and that, $P(B(x, \epsilon_k(x))) \approx k/n$. 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 et al., 1965], Theorem 3.1). Thus, a natural approach for estimating $F(P)$ is the plug-in estimator

$$\widehat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f\left(\widehat{p}_k(X_i)\right). \tag{8}$$

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 et al., 2011] show a convergence rate of $O\left(n^{-\frac{2s}{s+D}} + n^{-1}\right)$ for s-Hölder continuous densities (after sample splitting and boundary correction) by setting $k \asymp n^{\frac{s}{s+d}}$.

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, suggesting that this may be sub-optimal for estimating $F(P)$. Indeed, our previous work based on kernel density estimation [Singh and Póczos, 2014b] suggested that, for plug-in functional estimation (as compared to density estimation), *under-smoothing* may be preferable, since the empirical mean effectively performs additional smoothing.

4.5.2 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

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[\mathcal{B}\left(f\left(\frac{k/n}{\mu(B(x,\epsilon_k(x))}\right)\right)\right] = \mathop{\mathbb{E}}_{X_1,\ldots,X_n}\left[f\left(\frac{P(B(x,\epsilon_k(x)))}{\mu(B(x,\epsilon_k(x))}\right)\right]. \tag{9}$$

For continuous p, the quantity

$$p_{\epsilon_k(x)}(x) := \frac{P(B(x,\epsilon_k(x)))}{\mu(B(x,\epsilon_k(x))} \tag{10}$$

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

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

that uses k/n in place of $P(B(x,\epsilon_k(x)))$. 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,

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

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

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

giving the estimator of [Kozachenko and Leonenko, 1987]. Other examples of functionals for which the bias correction is known are given in Table 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 1, as well as any estimators of this form that may be derived in the future.

4.5.3 Main Results

As with the previous results for kernel density plug-in estimators, we begin by separately bounding the bias and the variance of $\widehat{F}_{\mathcal{B}}(P)$:

1. Suppose that, for $X \sim P$, the random variables $f'(p(X))$ and $(p(X))^{-s/D}$ lie in $\mathcal{L}^2_P(\mathcal{X})$; i.e.,
$$\mathop{\mathbb{E}}_{X \sim P}\left[\left(f'(p(X))\right)^2\right] < \infty \quad \text{and} \quad \mathop{\mathbb{E}}_{X \sim P}\left[(p(X))^{-2s/D}\right] < \infty.$$
Then, for some $C_B > 0$, we have the bias bound
$$\left| \mathop{\mathbb{E}}_{X_1,\ldots,X_n \overset{IID}{\sim} P} \widehat{F}_{\mathcal{B}}(P) - F(P) \right| \in O\left(\left(\frac{k}{n}\right)^{s/D}\right).$$

2. Suppose that, for $X \sim P$, the random variable $\mathcal{B}(f(p(X)))$ lies in $\mathcal{L}^2_P(\mathcal{X})$; i.e.,
$$\mathop{\mathbb{E}}_{X \sim P}\left[\left(\mathcal{B}(f(p(X)))\right)^2\right] < \infty.$$
Suppose that the quantity $\int_0^\infty e^{-y} y^k f(y) < \infty$ is finite. Then, we have the variance bound
$$\mathop{\mathbb{V}}_{X_1,\ldots,X_n \overset{IID}{\sim} P}\left[\widehat{F}_{\mathcal{B}}(P)\right] \in O\left(n^{-1}\right).$$

Combining these two bounds and setting k to be constant (with respect to n) gives a mean squared error bound, for some constant $C > 0$, of
$$\mathop{\mathrm{MSE}}_{X_1,\ldots,X_n}\left[\widehat{F}_{\mathcal{B}}(P)\right] \leq C\left(n^{-\frac{2s}{d}} + n^{-1}\right)$$

4.6 Estimation of Sobolev Quantities and other Quadratic Fourier Functionals

The functionals discussed so far are all integral functionals, in that they depend on integrals (over the sample space) of functions of the pointwise values of the probability density p from which the data are drawn; roughly, they have the form
$$F(P) = \int_{\mathcal{X}} f(p(x))\, dx,$$

for some function $f : [0, \infty) \to \mathbb{R}$ (e.g., when F is the Shannon entropy, $f(x) = -x \log x$). This excludes functionals that cannot be computed in terms of pointwise evaluations of the density.

Here, we consider estimation of some functionals that depend on the *derivatives* of the p, or, equivalently, on the Fourier representation of p. Initially, we considered Sobolev(-Hilbert) (squared) norms, inner products, and (squared) distances, although, for simplicity, we will discuss only norms here. As an example, for an integer s, the s-order Sobolev norm $\|p\|_{\mathcal{H}^s}$ of p can be understood as the $\mathcal{L}^2$ norm of the s^{th} weak derivative of p:

$$\|p\|_{\mathcal{H}^s} = \|p^{(s)}\|_{\mathcal{L}^2};$$

$\|p\|_{\mathcal{H}^s}$ is therefore used of a measure of the smoothness of p. Standard smoothness assumptions in nonparametric statistics can be thought of as bounds on particular Sobolev norms, and these quantities thus determine the convergence rates of many nonparametric estimators (e.g., density or regression estimates) [Tsybakov, 2008]. They also appear in closed forms for the asymptotic variance of such estimators [Bickel and Ritov, 1988], as well as of robust rank-based estimators such as the Wilcoxon statistic [Hodges Jr and Lehmann, 1963, Schweder, 1975]; their estimates are therefore useful for computing confidence intervals around such estimators.

Importantly, Sobolev norms have a relatively simple representation in Fourier space:

$$\|p\|_{\mathcal{H}^t} = \sum_{z \in \mathcal{Z}} |z|^{2s} |\widetilde{p}_z|^2. \tag{11}$$

Since each $\widetilde{p}_z$ is a linear functional of p, it is straightforward to estimate by the sample mean $\widehat{p}_z := \frac{1}{n} \sum_{i=1}^n \phi_z(X_i)$. Plugging these pointwise estimates in for $\widetilde{p}_z$ in Equation (11) gives a natural estimate for $\|p\|_{\mathcal{H}^t}$.

Suppose $p \in \mathcal{H}^s$ for some $s > t$. Then, we showed in Singh et al. [2016] that the minimax convergence rate is of order $n^{-\frac{8(s-t)}{4s+d}}$.

We considered a broader class of weighted $\mathcal{L}^2$ inner products having the form

$$\langle P, Q \rangle_{a_z} = \sum_{z \in \mathcal{Z}} a_z^2 \widetilde{P}_z \overline{\widetilde{Q}_z},$$

where $\mathcal{Z}$ is some countable index set, and, for some $\mathcal{L}^2$-orthonormal family $\{\phi_z\}_{z \in \mathcal{Z}}$ of functions,

$$\widetilde{P}_z := \mathop{\mathbb{E}}_{X \sim P} [\phi_z(X)] \quad \text{and} \quad \widetilde{Q}_z := \mathop{\mathbb{E}}_{Y \sim Q} [\phi_z(X)] .$$

This class of weighted $\mathcal{L}^2$ inner products includes, of course, finite-dimensional, $\mathcal{L}^2$, and Sobolev inner products, but also, for example, the induced inner product of any reproducing kernel Hilbert space with a translation-invariant kernel in $\mathcal{L}^2$ (i.e., a kernel $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that, for some $\kappa \in \mathcal{L}^2(\mathcal{C})$, for all $x, y, \in \mathcal{H}$, $K(x, y) = \kappa(x - y)$). Namely, this includes the commonly used kernels, such as the Gaussian, Laplacian, Sobolev, and sinc kernels.

For this larger class of quadratic functionals, we recently showed in Singh et al. [2018a] that the above estimator achieves the minimax rate.

4.7 Nonparanormal Information Estimation

So far, we have 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}. Kandasamy et al. [2015] suggested that even their (minimax optimal) estimators fail to reliably converge when d is much larger than 4-6. Put simply, except in very low dimensions, these spaces are too large to perform even point estimation of nonlinear density functionals.

At the other extreme, there has been a very detailed study of the estimation of information-theoretic quantities when the data are assumed to be truly Gaussian [Ahmed and Gokhale, 1989, Misra et al., 2005, Srivastava and Gupta, 2008, Cai et al., 2015]. The most sophisticated analysis, due to Cai et al. [2015], derived the exact probability law of the log-determinant $\log|\widehat{\Sigma}|$ of the empirical covariance matrix $\widehat{\Sigma}$. From this, they derived a deterministic bias correction, giving an information estimator for which they proved an MSE upper bound of $-2\log\left(1-\frac{D}{n}\right)$ ($\approx 2D/n$ when D/n is small) and a high-dimensional central limit theorem for the case $D \to \infty$ as $n \to \infty$ (but $D < n$). However, these results rely delicately on the assumption that the data are jointly Gaussian, and the performance of these estimators can degrade very 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.

To summarize, despite substantial theoretical work on estimating information-theoretic quantities, the practical settings in which we can estimate them are quick narrow: the data dimension must either be quite low, or the data must follow an exact parametric distribution. We considered filling the gap between these two extreme settings by studying information estimation in a semiparametric compromise between the two settings, in a model known as the "nonparanormal" (a.k.a. "Gaussian copula") model. 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 far more robust than Gaussian models.

4.7.1 Multivariate Mutual Information and the Nonparanormal Model

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 1. (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

$$\begin{aligned} I(X) &:= \mathop{\mathbb{E}}_{X \sim p} \left[\log \left(\frac{p(X)}{\prod_{j=1}^{D} p_j(X_j)} \right) \right] \\ &= \sum_{j=1}^{D} H(X_j) - H(X), \end{aligned} \tag{12}$$

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" [Studenỳ 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 [Studenỳ 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. (12), but, unlike the above quantities, this relationship depends on the marginal distributions of X, and hence involves some additional considerations (namely, some fairly mild smoothness assumptions on the marginals).

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

Definition 2. (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 [5] 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 $\sigma_j = \Sigma_{j,j} = 1$). [6] Σ is called the *latent covariance* of X and f is called the *marginal transformation* of X.

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

[6] 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.

In our paper Singh and Póczos [2017], under the assumption that the data X follows a nonparanormal distribution, we proposed three estimators for $I(X)$. The first one estimator is based on normalizing the empirical marginals to be approximately Gaussian, then directly computing the covariance of the normalized data. The latter two estimators are based on rank statistics (multivariate generalizations of Spearman's ρ and Kendall's τ), which one can analytically show have bijective relationships with the covariance matrix of a multivariate Gaussian. Since rank statistics are invariant to marginal transformations of the data, applying the bijections to the rank statistics immediately gives an estimate of the latent covariance matrix Σ, which can then be used to estimate $I(X)$.

For the estimator based on Spearman's ρ, we proved a convergence rate of order $O(d^2/n)$ (assuming a lower bound on the minimum eigenvalue of Σ), a dramatic improvement over the exponential dependence of the sample complexity on d in the nonparametric case. In a number of simulations, we further explored the large-sample properties of these estimators, as well as their robustness to various forms of model misspecification.

4.8 Condensed Summary of Results on Density Functional Estimation

In this section, we provide a condensed tabular reference for all our results on density functional estimation, as well as some results due to others.

4.8.1 Assumptions

Below section, we list, for reference, all of the assumptions made in various portions of this thesis. Table 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 1: 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)

4.9 Future Work

To complete this portion of the thesis, we propose two main lines of work advancing our understanding of distribution functional estimation.

The first involves generalizing our previous works to other classes of probability distributions, in particular the Besov scale, which includes the Hölder and Sobolev classes as special cases, but also includes spaces of inhomogeneous smoothness, such as the class of bounded total variation. As noted below, this generalization can also be applied to our previously described work on distribution estimation under alternative losses.

The second involves studying a major application of estimating density functionals (especially dissimilarity functionals such as $\mathcal{L}^p$ or Sobolev distances, or information divergences), namely that of statistical hypothesis testing. To the best of our knowledge, no theoretical results are known concerning the performance of hypothesis tests based on these estimators. Thus, we wish to identify general classes of alternative hypotheses under which we can bound the Type 2 error of these tests.

4.9.1 Extending Results to Besov Spaces

Our first proposal is to extend the previous work, conducted primarily under Hölder $C^{\ell,\alpha}$ or Hilbert-Sobolev spaces $\mathcal{H}^s$, to the much larger scale of Besov spaces $B^s_{p,q}$. Besov spaces include more general Sobolev spaces, as well as the space of functions of bounded variation. Several equivalent formulations of Besov spaces can be given; here, we give the most relevant one, in terms of rates of decay of wavelet series.

Fix a wavelet basis with mother wavelet ψ and father wavelet ϕ, and fix constants $s > 0$, $p, q \geq 1$. For any function

$$f = \sum_{z \in \mathbb{Z}} \alpha_{j_0,k} \phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k}$$

(where $\{\alpha_{j,z}\}_{j \in \mathbb{N}, z \in \mathbb{Z}}$ and $\{\beta_{j,z}\}_{j \in \mathbb{N}, z \in \mathbb{Z}}$ are the coefficients of f in the wavelet basis), the (s,p,q)-Besov norm $\|f\|_{\mathcal{B}^s_{p,q}}$ of f is given by

$$\|f\|_{\mathcal{B}^s_{p,q}} := \|\alpha_0\|_{\ell^p} + \left(\sum_{j \geq 0} \left(2^{j(s+1/2-1/p)} \|\beta_j\|_{\ell^p} \right)^q \right)^{1/q}. \tag{13}$$

The radius-L (s,p,q)-Besov ball $\mathcal{B}^s_{p,q}(L)$ is then given by

$$\mathcal{B}^s_{p,q}(L) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| \|f\|_{\mathcal{B}^s_{p,q}} \leq L \right\}.$$

Here, as in the Sobolev case $\mathcal{H}^s$, s is an index of the smoothness, and as in the $\mathcal{L}^p$ case, p and q, induce different exponential weightings of the coefficients of the function. Indeed, the Hölder and Sobolev classes are special cases of Besov classes specifically, $\mathcal{B}^s_{2,2} = \mathcal{H}^s$ and $\mathcal{B}^s_{\infty,\infty} = \mathcal{C}^s$. Moreover, all of the problems we studied previously in this thesis, in which we assumed $\mathcal{P} \subseteq \mathcal{H}^s$ or $\mathcal{P} \subseteq \mathcal{C}^s$, can be extended naturally to the assumption $\mathcal{P} \subseteq \mathcal{B}^s_{p,q}$.

Indeed, there has been prior work on all of these problems, pioneered in the early 1990's by David Donoho; his paper Donoho et al. [1996] on density estimation is especially relevant. In the case of distribution estimation, it is natural to consider the case where the loss is the $\mathcal{F}$-IPM and $\mathcal{F}$ is a ball in a Besov space.

In density functional estimation, to the best our knowledge, Besov spaces have only been explored in the case of quadratic functionals [Cai et al., 2005, 2006]. Naturally, we want to investigate general smooth functional estimation over densities in a Besov ball. For these spaces, we conjecture that von Mises estimators based on plugging in optimal density estimates will continue to be able to achieve the (presently unknown) minimax rate, although the degree of the von Mises approximation required may be higher. It may also be interesting to investigate nonsmooth entropy estimation (i.e., without the assumption that the density is lower bounded away from 0), as in Jiao et al. [2017], in the Besov space setting.

4.9.2 Applications to Statistical Hypothesis Testing

Much of the work on distribution functional estimation has immediate application for nonparametric statistical hypothesis testing (a.k.a., signal detection), especially for two-sample (homogeneity) testing and, as a special case, independence testing.

For example, suppose that we observe n IID samples $X_1, ..., X_n \overset{IID}{\sim} P$ and $Y_1, ..., Y_n \overset{IID}{\sim} Q$ from each of two distributions $P, Q \in \mathcal{P}$, and we are interested in determining whether $P = Q$ (two-sample testing). If $\rho : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is any functional satisfying $\rho(P, Q) = 0$ whenever $P = Q$, then, under the null hypothesis $H_0 : P = Q$, it suffices to test whether $\rho(P, Q) = 0$. This can be done using any estimate $\widehat{\rho}$ of $\rho(P, Q)$, together with confidence intervals; moreover, confidence can be easily estimated using a permutation test (i.e., producing a sample from the null distribution $\frac{P+Q}{2}$ by permuting the samples $X_1, ..., X_n, Y_1, ..., Y_n$)).

It is well-known that no statistical test can be uniformly optimal against even a moderately large class of alternatives [Ingster and Suslina, 2012]. Therefore, given the generality of the above testing method (across over both distributional assumptions and dissimilarity functionals ρ), it is natural to wonder what sorts of alternatives such tests are effective against, and how this depends on the choice of dissimilarity functional ρ.

Ingster and Suslina [2012] thoroughly studies minimax rates for nonparametric statistical testing in a wide variety of settings. Due to its simplicity, they focus on the nonparametric Gaussian sequence model, and hence they consider some test statistics that are similar to the Sobolev distance estimators we considered in Section 4.6. However, they do not specifically study tests based on general dissimilarity functionals, statistical independence tests, or conditional tests.

There have been a few studies of the power of statistical tests based on particular dissimilarity functionals, mostly based on either MMD or classification accuracy (i.e., the accuracy of a classifier trained to distinguish samples from the two distributions. Reddi et al. [2015] provide an analysis of two-sample tests based on MMD metric, showing that its performance against Gaussian mean-shift alternatives is comparable to that of Student's t-test, which is specifically tailored to and optimal for this testing problem. Lopez-Paz and Oquab [2016] and Ramdas et al. [2016] studied the power of two-sample tests based on classifiers. In the

analysis of Lopez-Paz and Oquab [2016], the null and alternative hypotheses were expressed in terms of the accuracy of the classifier; thus the results were very general but did not elucidate the relationship between the data distribution and the testing power, at least without further analysis of a particular classifier and hypothesized distributions in question. Ramdas et al. [2016] specifically considered the case of distinguishing two Gaussians with different means and identical covariances; here, they showed that a test based on a simple classifier (Fisher's linear discriminant analysis (LDA)) is minimax rate-optimal. However, it is not clear what implications this has for nonparametric tests, especially since Fisher's LDA can distinguish only classes with different means.

Ramdas et al. [2015] studied the relationship between estimation of MMD and hypothesis testing using MMD; they showed that, although MMD can be estimated at the rate $n^{-1/2}$ independent of dimension, in many cases, statistical testing nevertheless suffers in high-dimensions because the MMD itself between the two distributions becomes small. This highlights the fact that estimating a dissimilarity metric and using it to perform statistical tests are quite different problems, requiring significantly different analysis. This difference can also have important practical consequences. For example, the experments of Pérez-Cruz [2009] suggest that, when using BCFk mutual information estimators for dependence testing, letting k scale as $\sqrt{n}$ was optimal, even though fixed k or $k \in O(\log n)$ is optimal for estimation. Intuitively, if the bias of the estimator at P and at the hypothesized null distribution are similar, then these cancel, and variance comes to dominate the error of the test, so that over-smoothing becomes preferable.

We propose to begin by lower bounding the power of two-sample tests based on plugging estimates of dissimilarity functionals into the above permutation methodology, considering basic alternatives as in Ingster and Suslina [2012], Ramdas et al. [2015], Reddi et al. [2015], as well as other novel alternatives that might be interesting for particular applications.

5 Proposed Timeline

1. **August, 2018:** Distribution estimation under IPM losses with Besov discriminator and generator classes (this work is already underway; we have already derived lower bounds that we believe to be tight, and have made progress on obtaining matching upper bounds)

2. **September-December, 2018:** distribution estimation under Wasserstein loss in the latent variable model (see Section 3.3; I have finished formulating this framework, and anticipate that I will be able to obtain upper bounds in the near future; lower bounds should be obtainable using standard techniques)

3. **Spring, 2019:** Smooth Distribution Functional Estimation over Besov Spaces

4. **Spring, 2019:** Guarantees for hypothesis testing using distribution functional estimators

5. **Summer, 2019:** Thesis writing & defense preparation

From September through December of 2018, I will be taking a leave-of-absence to complete an internship, and have therefore allocated relatively little work for this time period.

References

Nabil Ali Ahmed and DV Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. on Information Theory*, 35(3):688–692, 1989.

Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

Dena Marie Asta. Kernel density estimation on symmetric spaces. *arXiv preprint arXiv:1411.4040*, 2014.

Yannick Baraud. Confidence balls in Gaussian regression. *The Annals of statistics*, pages 528–551, 2004.

Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *arXiv preprint arXiv:1606.00304*, 2016.

Peter J Bickel and Ya'acov Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

Emmanuel Boissard, Thibaut Le Gouic, et al. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539–563, 2014.

T Tony Cai, Mark G Low, et al. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005.

T Tony Cai, Mark G Low, et al. Optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 34(5):2298–2325, 2006.

T Tony Cai, Tengyuan Liang, and Harrison H Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *J. of Multivariate Analysis*, 137:161–172, 2015.

C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

V Dobrić and Joseph E Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

Simon S Du, Sivaraman Balakrishnan, and Aarti Singh. Computationally efficient robust estimation of sparse functionals. *arXiv preprint arXiv:1702.07709*, 2017a.

Simon S Du, Jayanth Koushik, Aarti Singh, and Barnabás Póczos. Hypothesis transfer learning via transformation functions. *stat*, 1050:27, 2017b.

Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

RM Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Lutz Dumbgen. New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of statistics*, pages 288–314, 1998.

Jianqing Fan. On the estimation of quadratic functionals. *The Annals of Statistics*, pages 1273–1294, 1991.

Seth Flaxman, Dougal Sutherland, Yu-Xiang Wang, and Yee Whye Teh. Understanding the 2016 us presidential election using ecological inference and distribution regression with census microdata. *arXiv preprint arXiv:1611.03787*, 2016.

Seth R Flaxman, Yu-Xiang Wang, and Alexander J Smola. 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*, pages 289–298. ACM, 2015.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *JASA*, 76(376): 817–823, 1981.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1267–1271. IEEE, 2017.

Wendell R Garner. *Uncertainty and structure as psychological concepts.* Wiley, 1962.

Christopher R Genovese and Larry Wasserman. Confidence sets for nonparametric wavelet regression. *The Annals of statistics*, pages 698–729, 2005.

Mohammed Nawaz Goria, Nikolai N Leonenko, Victor V Mergel, and Pier Luigi Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3):277–297, 2005.

Joseph L Hodges Jr and Erich L Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, pages 598–611, 1963.

Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

Jiantao Jiao, Weihao Gao, and Yanjun Han. The nearest neighbor information estimator is adaptively near minimax rate-optimal. *arXiv preprint arXiv:1711.08824*, 2017.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *NIPS*, pages 397–405, 2015.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Samory Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328, 2017.

Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927, 2014.

Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry A Wasserman. On estimating ${}_L2^2$ divergence. In *AISTATS*, 2015.

Béatrice Laurent et al. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

Jing Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *arXiv preprint arXiv:1804.10556*, 2018.

N. Leonenko and L. Pronzato. Correction of 'a class of Rényi information estimators for mulitidimensional densities' Ann. Statist., 36(2008) 2153-2182, 2010.

N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbours. *Tatra Mt. Mathematical Publications*, 39, 2008.

Ker-Chau Li. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, pages 1001–1008, 1989.

Tengyuan Liang. How well can generative adversarial networks (gan) learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Han Liu, Larry Wasserman, and John D Lafferty. Exponential concentration for mutual information estimation with application to forests. In *Advances in Neural Information Processing Systems*, pages 2537–2545, 2012.

Don O Loftsgaarden, Charles P Quesenberry, et al. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

Neeraj Misra, Harshinder Singh, and Eugene Demchuk. Estimation of the entropy of a multivariate normal distribution. *J. Multivariate Analysis*, 92(2):324–342, 2005.

Kevin Moon and Alfred Hero. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014.

Kevin R Moon. *Nonparametric Estimation of Distributional Functionals and Applications.* PhD thesis, University of Michigan, Ann Arbor, 2016.

Kevin R Moon, Alfred O Hero, and B Véronique Delouille. Meta learning of bounds on the Bayes classifier error. In *Signal Processing and Signal Processing Education Workshop (SP/SPE), 2015 IEEE*, pages 13–18. IEEE, 2015.

Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and James Robins. Lepski's method and adaptive estimation of nonlinear integral functionals of density. *arXiv preprint arXiv:1508.00249*, 2015.

Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and James Robins. On adaptive estimation of nonparametric functionals. *arXiv preprint arXiv:1608.01364*, 2016.

Morteza Noshad and Alfred Hero. Scalable hash-based estimation of divergence measures. In *International Conference on Artificial Intelligence and Statistics*, pages 1877–1885, 2018.

Morteza Noshad and Alfred O Hero III. Scalable mutual information estimation using dependence graphs. *arXiv preprint arXiv:1801.09125*, 2018.

Morteza Noshad, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero. Direct estimation of information divergence using nearest neighbor ratios. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 903–907. IEEE, 2017.

Michelle Ntampaka, Hy Trac, Dougal J Sutherland, Nicholas Battaglia, Barnabás Póczos, and Jeff Schneider. A machine learning approach for dynamical mass measurements of galaxy clusters. *The Astrophysical Journal*, 803(2):50, 2015.

Leandro Pardo. *Statistical inference based on divergence measures*. CRC press, 2005.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

Fernando Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *Advances in neural information processing systems*, pages 1257–1264, 2009.

B. Poczos and J. Schneider. Nonparametric estimation of conditional information and divergences. In *International Conference on AI and Statistics (AISTATS)*, volume 20 of *JMLR Workshop and Conference Proceedings*, 2012.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617, 2011.

Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 599–608, Arlington, Virginia, United States, 2011. AUAI Press. ISBN 978-0-9749039-7-2. URL `http://dl.acm.org/citation.cfm?id=3020548.3020618`.

Barnabas Poczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation and its applications to machine learning. Technical report, Carnegie Mellon University, 2011.

Barnabás Póczos, Liang Xiong, Dougal J Sutherland, and Jeff Schneider. Nonparametric kernel estimators for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2989–2996. IEEE, 2012.

Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Novi Quadrianto, James Petterson, and Alex J Smola. Distribution matching for transduction. In *Advances in Neural Information Processing Systems*, pages 1500–1508, 2009.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*, 2016.

Sashank Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780, 2015.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Tore Schweder. Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics*, pages 113–126, 1975.

Shashank Singh and Barnabás Póczos. Generalized exponential concentration inequality for renyi divergence estimation. In *Proceedings of The 31st International Conference on Machine Learning*, pages 333–341, 2014a.

Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pages 3032–3040, 2014b.

Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in Neural Information Processing Systems*, pages 1217–1225, 2016.

Shashank Singh and Barnabás Póczos. Nonparanormal information estimation. In *International Conference on Machine Learning*, pages 3210–3219, 2017.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Shashank Singh, Simon S Du, and Barnabás Póczos. Efficient nonparametric smoothness estimation. In *Advances in Neural Information Processing Systems*, pages 1010–1018, 2016.

Shashank Singh, Bharath K Sriperumbudur, and Barnabás Póczos. Minimax estimation of quadratic fourier functionals. *arXiv preprint arXiv:1803.11451*, 2018a.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *arXiv preprint arXiv:1805.08836*, 2018b.

Kumar Sricharan, Raviv Raich, and Alfred O Hero. k-nearest neighbor estimation of entropies with confidence. In *IEEE International Symposium on Information Theory*, pages 1205–1209. IEEE, 2011.

Kumar Sricharan, Raviv Raich, and Alfred O Hero III. Estimation of nonlinear functionals of densities with confidence. *Information Theory, IEEE Transactions on*, 58(7):4135–4159, 2012.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

Santosh Srivastava and Maya R Gupta. Bayesian estimation of the entropy of the multivariate Gaussian. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1103–1107. IEEE, 2008.

Milan Studenỳ and Jirina Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer, 1998.

Dougal J Sutherland. *Scalable, Flexible and Active Learning on Distributions*. PhD thesis, Carnegie Mellon University, 2016.

Zoltán Szabó. Information theoretical estimators toolbox. *The Journal of Machine Learning Research*, 15(1):283–287, 2014.

Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957, 2015.

A. B. Tsybakov and E. C. van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian J. Statistics*, 23:75–83, 1996.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. of research and development*, 4(1):66–82, 1960.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

Eric Wolsztynski, Eric Thierry, and Luc Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Processing*, 85(5):937–949, 2005.