Distribution-Free Distribution Regression

Barnabás Póczos[1] **Alessandro Rinaldo**[2] **Aarti Singh**[3] **Larry Wasserman**[4]
Carnegie Mellon University
Machine Learning Department[1,3,4]
Department of Statistics[2,4]
Pittsburgh, PA, USA, 15213

Abstract

'Distribution regression' refers to the situation where a response Y depends on a covariate P where P is a probability distribution. The model is $Y = f(P) + \mu$ where f is an unknown regression function and μ is a random error. Typically, we do not observe P directly, but rather, we observe a sample from P. In this paper we develop theory and methods for distribution-free versions of distribution regression. This means that we do not make strong distributional assumptions about the error term μ and covariate P. We prove that when the effective dimension is small enough (as measured by the doubling dimension), then the excess prediction risk converges to zero with a polynomial rate.

1 Introduction

In a standard regression model, we need to predict a real-valued response Y from a vector-valued covariate (or feature) $X \in \mathbb{R}^d$. Recently, there has been interest in extensions of standard regression from finite dimensional Euclidean spaces to other domains. For example, in functional regression (Ferraty and Vieu [2006]) the covariate is a function instead of a finite dimensional vector.

In this paper, we study *distribution regression* where the covariate is a probability distribution P. This differs from functional regression in two important ways. First, P is a probability measure on $\mathbb{R}^k$ rather than a one-dimensional function. Second, and more importantly, we do not observe the covariate P directly. Rather, we observe a sample from P, which means that we have a regression model with measurement error (Carroll et al. [2006], Fan and Truong [1993]).

A practical example where this framework can be useful is as follows. Suppose that we need to classify patients in a hospital and diagnose whether they are healthy or suffer from a disease. Traditional machine learning based approaches would make a couple of medical tests, and using the results of these measurements they would form a feature vector for each person and then apply a standard classifier to predict the class labels of the feature vectors. Suppose we have m patients, and these feature vectors are denoted by $X_i \in \mathbb{R}^d$, $1 \le i \le m$. Our goal is to predict the class label $Y \in \{\text{'healthy'}, \text{'diseased'}\}$ for a person. The problem with this approach is that our heart rate, blood pressure, chemical concentrations in blood, and many other medical conditions in our body are always changing, and therefore if we repeat these measurements a couple of times, then each time we might get different measurements and different feature vectors for the same person. For the ith person, let the set of these measurements be denoted by $\mathcal{X}_i = \{X_{i,1}, \ldots, X_{i,n_i}\}$, where $X_{i,n_i} \in \mathbb{R}^d$ indicates that we repeated the medical tests n_i times. Interestingly, traditional feature vector based machine learning algorithms cannot handle well such simple problems. They might construct a new feature vector as the average of the measurements ($\widetilde{X}_i \doteq \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}$), but then they lose information. If they want to keep all the measurements in a feature vector, then they cannot just simply stack the feature vectors of each person to a larger vector, because then each of these vectors could have different sizes (dn_i). In contrast to the approaches, in our framework we simply say that each person is represented by an unknown distribution P_i, and those feature vectors are samples from these distributions $X_{i,j} \sim P_i$ for $j = 1, \ldots n_i$. Our goal is to classify these unknown P_i distributions.

The formal definition of the problem is as follows.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

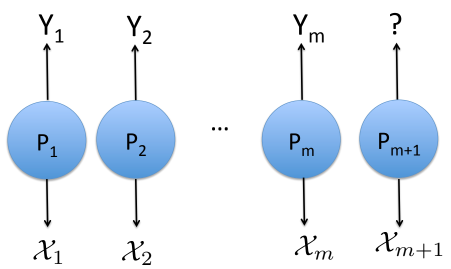


Figure 1: Illustration of the model - distributions $P_1, \ldots, P_m, P_{m+1}$ are unobserved, only the $\mathcal{X}_1, \ldots, \mathcal{X}_m, \mathcal{X}_{m+1}$ sample sets are observable.

We consider a regression problem with variables $(P_1, Y_1), \ldots, (P_m, Y_m)$ where $Y_i \in \mathbb{R}$ and each P_i is a probability distribution on a compact subset $\mathcal{K} \subset \mathbb{R}^k$. We assume that

$$Y_i = f(P_i) + \mu_i, \quad i = 1, \ldots, m,$$

for some functional f, where μ_i is a noise variable with mean 0. We do not observe P_i directly; rather we observe a sample

$$X_{i1}, \ldots, X_{in_i} \stackrel{i.i.d}{\sim} P_i. \tag{1}$$

Thus the observed data are

$$(\mathcal{X}_1, Y_1), \ldots, (\mathcal{X}_m, Y_m) \tag{2}$$

where $\mathcal{X}_i = \{X_{i1}, \ldots, X_{in_i}\}$. Our goal is to predict a new Y_{m+1} from a new batch $\mathcal{X}_{m+1}$ drawn from a new distribution P_{m+1}. This model is illustrated in Figure 1.

We model the unobservable probability distributions $P_1, \ldots, P_m$ as follows. Let $\mathbb{D}$ denote the set of all distributions on $\mathcal{K}$ that have a density with respect to the Lebesgue measure. We assume that the distributions P_i are an i.i.d. sample from a measure $\mathcal{P}$ on $\mathbb{D}$, that is[1],

$$P_1, \ldots, P_m, P_{m+1} \stackrel{i.i.d}{\sim} \mathcal{P}.$$

Note that $f : \mathbb{D} \to \mathbb{R}$. If $Q(\cdot|P)$ denotes the law of Y given P, then the joint distribution of (Y, P) is given by

$$\mathbb{P}(Y \in A, P \in B) = Q(Y \in A | P \in B)\mathcal{P}(P \in B)$$

Our main result is a theorem where we prove that when the effective dimension of $\mathcal{P}$ measured by the doubling dimension is small enough, then the estimator is consistent and the prediction risk converges to zero with a polynomial rate. Our results are *distribution free*, similar to the functional regression case Ferraty and Vieu [2006], in the sense that we do not make any strong distributional assumptions.

[1] There are some subtle technical difficulties with the definitions of measurability. Using outer expectations these issues can be resolved. In this paper, however, we ignore this question.

Outline. In Section 2 we discuss related work. We propose a specific estimator for distribution regression in Section 3. We call this *kernel-kernel estimator* since it makes use of kernels in two different ways. In Section 4 we derive an upper bound on the risk of the estimator. The proofs can be found in Section 5. In Section 6 we analyze the risk bound in terms of the doubling dimension, which is a measure of the intrinsic dimension of the space. We present numerical illustrations in Section 7. Finally, we give some concluding remarks in Section 8. The details of the proofs can be found in the Supplementary material [Póczos et al., 2013].

2 Related work

Our framework is related to functional data analysis, which is a new and steadily improving field of statistics. For comprehensive reviews and references, see Ramsay and Silverman [2005], Ferraty and Vieu [2006].

A popular approach to do machine learning, such as classification and regression, on the domain of distributions is to embed the distribution to a Hilbert space, introduce kernels between the distributions, and then use a traditional kernel machine to solve the learning problem. There are both parametric and nonparametric methods proposed in the literature.

Parametric methods, (e.g. Jebara et al. [2004], Moreno et al. [2004], Jaakkola and Haussler [1998]), usually fit a parametric family (e.g. Gaussians distributions or exponential family) to the densities, and using the fitted parameters they estimate the inner products between the distributions. The problem with parametric approaches, however, is that when the true densities do not belong to the assumed parametric families, then this method introduces some unavoidable bias during the estimation of the inner products between the densities.

A couple of nonparametric approaches exist as well. Since our covariates are represented by finite sets, reproducing kernel Hilbert space (RKHS) based set kernels can be used in these learning problems. Smola et al. [2007] proposed to embed the distributions to an RKHS using the mean map kernels. In this framework, the role of universal kernels have been studied by Christmann and Steinwart [2010]. Recently, the representer theorem has also been generalized for the space of probability distributions [Muandet et al., 2012].

Kondor and Jebara [2003] introduced Bhattacharyya's measure of affinity between finite-dimensional Gaussians in a Hilbert space. In contrast to the previous approaches, Póczos et al. [2012], Póczos et al. [2011] used nonparametric Rényi divergence estimators to solve machine learning problems on the set of distributions.

Although, there are a few algorithms designed for regression on distributions, we know very little about their theoretical properties. To the best of our knowledge, even the simplest, fundamental questions have not been studied yet. For example, we do not know how many training distributions (m) and how many samples (n_i, $i = 1, \dots, m$) we need to achieve a target prediction error. Our paper is providing an answer to this question.

3 The Kernel-Kernel Estimator

In this section we define an estimator $\widehat{f}$ for the unknown function f. Let $\widehat{P}_i$ denote an estimator of P_i based on $\mathcal{X}_i$, and let $\mathcal{X}$ be a sample from a new distribution $P = P_{m+1}$. Accordingly, we denote with $\widehat{P}$ an estimator of P based on $\mathcal{X}$. Our predictor for Y_{m+1} is then $\widehat{Y}_{m+1} = \widehat{f}(\widehat{P}_{m+1})$.

Given a bandwidth $h > 0$ and a kernel function K (whose properties will be specified later), we define

$$\begin{aligned}\widehat{f}(\widehat{P}) &= \widehat{f}(\widehat{P}; \widehat{P}_1, \dots, \widehat{P}_m) \\ &= \begin{cases} \dfrac{\sum_i Y_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right)}{\sum_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right)} & \text{if } \sum_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right) > 0 \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

To complete the definition, we need to specify $\widehat{P}_i$, $\widehat{P}$ and D. We will estimate P_i — or, more precisely, the density p_i of P_i — with a kernel density estimator

$$\widehat{p}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{b_i^k} B\left(\frac{\|x - X_{ij}\|_2}{b_i}\right) \tag{3}$$

where B is an appropriate kernel function (see, e.g. Tsybakov [2010]) with bandwidth $b_i > 0$. Here $\|x\|_2$ denotes the Euclidean norm of $x \in \mathbb{R}^k$. Accordingly, $\widehat{P}_i$ is defined by

$$\widehat{P}_i(A) = \int_A \widehat{p}_i(u) du,$$

for all Borel measurable subsets of $\mathbb{R}^k$. For any two probabilities in P and Q in $\mathbb{D}$, we take $D(P, Q)$ to be the L_1 distance of their densities: $D(P, Q) = \|p - q\| = \int |p(x) - q(x)| dx$. Hence,

$$\widehat{f}(\widehat{P}) = \widehat{f}(\widehat{P}; \widehat{P}_1, \dots, \widehat{P}_m) = \frac{\sum_{i=1}^m Y_i K\left(\frac{||\widehat{p} - \widehat{p}_i||}{h}\right)}{\sum_{i=1}^m K\left(\frac{||\widehat{p} - \widehat{p}_i||}{h}\right)} \tag{4}$$

which we call the 'kernel-kernel estimator' since it makes use of two kernels, B and K.

For simplicity, n will denote the size of the sample $\mathcal{X}$, and b will be the bandwidth in the estimator of $\widehat{p}$.

In what follows we will make the following assumptions on f, K, $\mathcal{P}$, μ_i, and Y_i.

Assumptions

- *(A1) Hölder continuous functional.* The unknown functional f belongs to the class $\mathcal{M} = \mathcal{M}(L, \beta, D)$ of Hölder continuous functionals on $\mathbb{D}$:

$$\mathcal{M} = \left\{ f : \ |f(P_i) - f(P_j)| \le L\, D(P_i, P_j)^\beta \right\},$$

  for some $L > 0$ and $0 < \beta \le 1$, where D is the above specified L_1 metric on $\mathbb{D}$. In the $\beta = 1$ special case this means that f is Lipschitz continuous.

- *(A2) Asymmetric boxed and Lipschitz kernel.* The kernel K satisfies the following properties: $K : [0, \infty] \to \mathbb{R}$ is non-negative and Lipschitz continuous with Lipschitz constant L_K. In addition, there exist constants $0 < \underline{K} < 1$ and $0 < r < R < \infty$ such that, for all $x > 0$, it holds that

$$\underline{K} I_{\{x \in \mathcal{B}(0,r)\}} \le K(x) \le I_{\{x \in \mathcal{B}(0,R)\}}.$$

- *(A3) Hölder class of distributions.* The distribution $\mathcal{P}$ is supported on the set of distributions with densities that are 1-smooth Hölder functions, as defined in Tsybakov [2010], Rigollet and Vert [2009] for example.

- *(A4) Bounded regression.* We will assume that $\sup_{P \in \mathcal{P}} |f(P)| < f_{\max}$ for some $f_{\max} > 0$. Also, μ_i has mean 0 and $\mathbb{P}(|Y_i| \le B_Y) = 1$ for some $B_Y < \infty$.

- *(A5) Lower bound on $\min_{1 \le i \le m+1} n_i$.* Let $n = \min_{1 \le i \le m+1} n_i$. We assume that $n^{\frac{k}{2+k}} \ge 3 \ln m$.

- *(A6) Requirements on regression kernel bandwidth h.* Assume that $C_* n^{-\frac{1}{2+k}} \le rh/4$ where C_* is defined in (9), and $h \le H$ where $H > 0$ is a constant.

- *(A7) Requirement on density kernel bandwidths $\{b_i\}_{i=1}^m$.* Assume the bandwidths $b_i = b := n^{-\frac{1}{k+2}}$.

4 Upper Bound on Risk

We are concerned with upper bounding the risk

$$R(m, n) = \mathbb{E}\Big[|\widehat{f}(\widehat{P}; \widehat{P}_1, \dots, \widehat{P}_m) - f(P)| \Big],$$

where the expectation is with respect to the joint distribution of the sample $(\mathcal{X}_1, Y_1), \dots, (\mathcal{X}_m, Y_m)$, the new covariate $P = P_{m+1}$ and the new observation $\mathcal{X}_{m+1}$. Note that the absolute prediction risk is $\mathbb{E}|\widehat{Y} - Y| \le R(m,n) + c$, where $c = \mathbb{E}(|\mu|)$ is a constant. So bounding the prediction risk is equivalent to bounding $R(m,n)$, which we call the excess prediction risk. In what follows, $C, c_1, c_2, \dots$ represent constants whose value can be different in different expressions.

Let $\mathcal{B}(P,h) = \{\widetilde{P} \in \mathbb{D} : D(\widetilde{P}, P) \le h\}$ denote the L_1 ball of distributions around P with radius h. We will see that the risk depends on the size of the class of probabilities $\mathbb{D}$. In particular, the risk depends on the *small ball probability*

$$\Phi_P(h) := \mathcal{P}(\mathcal{B}(P,h)),$$

where P is a fixed distribution and $\Phi_P(h)$ is a function of P.

Our first result, Theorem 1, provides a general upper bound on the risk. In our second result (Section 6) we show that when the effective dimension measured by the doubling dimension is small, then the risk converges to zero. We also derive an upper bound on the rate of convergence.

Theorem 1 *Suppose that the assumptions (A1)-(A7) stated above hold. Then*

$$\begin{aligned} R(m,n) &\le \frac{1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] C_1 n^{-\frac{1}{2+k}} + C_2 h^{\beta} \\ &+ C_3\sqrt{\frac{1}{m}}\sqrt{\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]} + \frac{C_4}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] \\ &+ (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}, \end{aligned}$$

where the constants C_i's are specified in the proof.

5 Proof of Theorem 1

In this Section we prove our main result, Theorem 1. The main idea of the proof is to use the triangle inequality to write

$$\begin{aligned} R(m,n) &= \mathbb{E}|\widehat{f}(\widehat{P}; \widehat{P}_1, \dots, \widehat{P}_m) - f(P)| \\ &\le \mathbb{E}|\widehat{f}(\widehat{P}; \widehat{P}_1, \dots, \widehat{P}_m) - \widehat{f}(P; P_1, \dots, P_m)| \quad (5) \\ &+ \mathbb{E}|\widehat{f}(P; P_1, \dots, P_m) - f(P)|. \quad (6) \end{aligned}$$

In Sections 5.2 and 5.3 we will derive upper bounds for (5) and (6), respectively. Section 5.1 contains a series of technical results needed in our proofs.

Throughout, we let $\widehat{K}_i = K\left(\frac{D(\widehat{P}_i,\widehat{P})}{h}\right)$, $K_i = K\left(\frac{D(P_i,P)}{h}\right)$ and $\epsilon_i = K_i - \widehat{K}_i$, for $i = 1, \dots, m$. Note that, for ease of readability, we have omitted the dependence on h.

5.1 Technical Results

5.1.1 L_1 Risk of Density Estimators

In this section we bound $\mathbb{E}[D(P,\widehat{P})|P] = \mathbb{E}[\int |p - \widehat{p}||P]$, the L_1 risk of the density estimator $\widehat{p}$ of p, uniformly over all P in $\mathbb{D}$. To this end, suppose that $n_i \ge n$ for all $i = 1, 2, \dots, m+1$, and let $b_i = b = n^{-\frac{1}{k+2}}$. In this case, the following lemma provides upper bound on the L_1 risk of the density estimator. Its proof can be found in the supplementary material.

Lemma 2

$$\mathbb{E}[D(\widehat{P}_i, P_i)|P_i] \le \bar{C} n^{-\frac{1}{2+k}}, \quad (7)$$
$$\mathbb{E}[D(\widehat{P}_i, P_i)] \le \bar{C} n^{-\frac{1}{2+k}},$$

where

$$\bar{C} = c_0(c_1 + c_2), \quad (8)$$

with c_0, c_1 and c_1 constants specified in the proof.

Next, we show that the terms $D(\widehat{P}_i, P_i)$ are uniformly bounded by a term of order $O(h)$, with high probability.

Lemma 3 *With probably no smaller than $1 - (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$, $D(\widehat{P}_i, P_i) < \frac{rh}{4}$ for all $i = 1, \dots, m+1$.*

Notice that by Assumption (A5), $1 - (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}} \to 1$.

Proof. From McDiarmid's inequality, for any $\epsilon > 0$ we have that

$$\mathbb{P}(||\widehat{p}_i - p_i||_1 - \mathbb{E}||\widehat{p}_i - p_i||_1 > \epsilon) \le e^{-n\epsilon^2/2}$$

(see, for example, section 2.4 of Devroye and Lugosi [2001]). Thus,

$$\mathbb{P}(||\widehat{p}_i - p_i||_1 > \mathbb{E}||\widehat{p}_i - p_i||_1 + n^{-\frac{1}{2+k}}) \le e^{-\frac{1}{2}n^{\frac{k}{2+k}}},$$

since $nn^{-\frac{2}{2+k}} = n^{\frac{k}{2+k}}$. This implies that

$$\begin{aligned} &\mathbb{P}(\max_{1 \le i \le m+1} ||\widehat{p}_i - p_i||_1 > \mathbb{E}||\widehat{p}_i - p_i||_1 + n^{-\frac{1}{2+k}}) \\ &\qquad \le (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}} \to 0, \end{aligned}$$

by assumption (A5). Therefore,

$$\begin{aligned} &1 - (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}} \\ &\le \mathbb{P}(\max_{1 \le i \le m+1} ||\widehat{p}_i - p_i||_1 \le \mathbb{E}||\widehat{p}_i - p_i||_1 + n^{-\frac{1}{2+k}}) \\ &\le \mathbb{P}(\max_{1 \le i \le m+1} ||\widehat{p}_i - p_i||_1 \le (1 + c_0(c_1 + c_2))n^{-\frac{1}{2+k}}). \end{aligned}$$

This implies that with

$$C_* = (1 + c_0(c_1 + c_2)) \quad (9)$$

and using assumption (A6), we have that

$$D(\widehat{P}_i, P_i) \leq C_* n^{-\frac{1}{k+2}} \leq \frac{rh}{4} \quad \text{for all } i \qquad (10)$$

on an event $\Omega_{m,n}$, where $\mathbb{P}(\Omega_{m,n}^c) \leq (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$. Here $\Omega_{m,n}^c$ denotes the complement of $\Omega_{m,n}$. □

5.1.2 Other Lemmata

Throughout this section we will make use of the constant $\bar{C}$, defined in (8). In what follows, we will need a few lemmas that we list below. Their proofs can be found in the supplementary material.

The following lemma provides an upper bound on $\mathbb{P}(\sum_{i=1}^m K_i = 0)$ with the help of small ball probabilities.

Lemma 4

$$\mathbb{P}\Big(\sum_{i=1}^m K_i = 0\Big) \leq \mathbb{P}\Big(\sum_{i=1}^m K_i < \underline{K}\Big) = \frac{1}{em}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right].$$

We will also need the following lemma.

Lemma 5

$$\mathbb{E}\left[\frac{1}{\sum_i K_i} I_{\{\sum_i K_i \geq \underline{K}\}}\right] \leq \frac{1+1/\underline{K}}{m\underline{K}}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right].$$

The following lemma provides an upper bound on $|\epsilon_i|$.

Lemma 6 *Assume that the kernel function K is Lipschitz continuous with Lipschitz constant L_K. We have that*

$$|\epsilon_i| \leq \frac{L_K}{h}(D(P, \widehat{P}) + D(P_i, \widehat{P}_i)).$$

By definition, $|\epsilon_i| = |K_i - \widehat{K}_i| = |K(\frac{D(P,P_i)}{h}) - K(\frac{D(\widehat{P},\widehat{P}_i)}{h})|$, which is a deterministic function of random variables P, P_i, $\widehat{P}$, and $\widehat{P}_i$. We will denote this deterministic relationship as $\epsilon_i = \epsilon_i(P, \widehat{P}, P_i, \widehat{P}_i)$. The following lemma shows that for any $\kappa > 0$,

$$\mathbb{P}\Big(\sum_i |\epsilon_i(P, \widehat{P}, P_i, \widehat{P}_i)| < \kappa | \{P_i\}_{i=1}^m, P\Big)$$

can be lower bounded by a non-trivial quantity that does not depend on P and $\{P_i\}_{i=1}^m$.

Lemma 7 *For any $\kappa > 0$ we have that*

$$\mathbb{P}(\sum_i |\epsilon_i(P, \widehat{P}, P_i, \widehat{P}_i)| < \kappa | \{P_i\}_{i=1}^m, P) \geq \eta,$$

where $\eta = \eta(\kappa, n, m) = 1 - \frac{2L_K m\bar{C}}{h\kappa} n^{-\frac{1}{2+k}}$.

The following lemma provides an upper bound on the expected value of $\sum_{i=1}^m |\epsilon_i|$.

Lemma 8

$$\mathbb{E}\left[\sum_{i=1}^m |\epsilon_i| \Big| P, \{P_i\}_{i=1}^m\right] \leq \frac{2L_K\bar{C}m}{h} n^{-\frac{1}{2+k}}.$$

The next lemma shows that $\mathbb{P}\Big(\sum_{i=1}^m \widehat{K}_i < \underline{K}\Big)$ can be upper bounded by a small quantity as well. We assume that $n_i = n$ and $b_i = b$ for all i. Define

$$\zeta = \zeta(n, m) = \frac{1}{em}\mathbb{E}\left(\frac{1}{\Phi_P\left(\frac{rh}{2}\right)}\right) + (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}.$$

Lemma 9

$$\mathbb{P}\Big(\sum_{i=1}^m \widehat{K}_i = 0\Big) \leq \mathbb{P}\Big(\sum_{i=1}^m \widehat{K}_i < \underline{K}\Big) \leq \zeta.$$

5.2 Upper bound on Equation 5

Let $\Delta\widehat{f} = |\widehat{f}(\widehat{P}; \widehat{P}_1, \ldots, \widehat{P}_m) - \widehat{f}(P; P_1, \ldots, P_m)|$. Our goal is to provide an upper bound on $\mathbb{E}[\Delta\widehat{f}]$.

Introduce the following events: $E_0 = \{\sum_i K_i = 0\}$, $E_1 = \{0 < \sum_i K_i < \underline{K}\}$, $E_2 = \{\underline{K} \leq \sum_i K_i\}$. Similarly, $\widehat{E}_0 = \{\sum_i \widehat{K}_i = 0\}$, $\widehat{E}_1 = \{0 < \sum_i \widehat{K}_i < \underline{K}\}$, $\widehat{E}_2 = \{\underline{K} \leq \sum_i \widehat{K}_i\}$. Obviously, $\mathbb{E}[\Delta\widehat{f}] = \sum_{k=0}^2 \sum_{l=0}^2 \mathbb{E}[\Delta\widehat{f} I_{E_k} I_{\widehat{E}_l}]$.

Based on the sign of $\sum_i K_i$ and $\sum \widehat{K}_i$, there are four different cases. (i) If $\sum_i K_i > 0$ and $\sum_i \widehat{K}_i > 0$, then $\Delta\widehat{f} = |\frac{\sum_i Y_i \widehat{K}_i}{\sum_i \widehat{K}_i} - \frac{\sum_i Y_i K_i}{\sum_i K_i}|$. (ii) If $\sum_i K_i > 0$ and $\sum_i \widehat{K}_i = 0$, then $\Delta\widehat{f} = |\frac{\sum_i Y_i K_i}{\sum_i K_i}|$. (iii) If $\sum_i K_i = 0$ and $\sum_i \widehat{K}_i > 0$, then $\Delta\widehat{f} = |\frac{\sum_i Y_i \widehat{K}_i}{\sum_i \widehat{K}_i}|$, and finally (iv) if $\sum_i K_i = 0$ and $\sum_i \widehat{K}_i = 0$, then $\Delta\widehat{f} = 0$. From this it immediately follows that $\mathbb{E}[\Delta\widehat{f} I_{E_0} I_{\widehat{E}_0}] = 0$.

When $\sum_i K_i > 0$, $\left|\sum_i \frac{Y_i K_i}{\sum_i K_i}\right| \leq B_Y$. Therefore,

$$\begin{aligned}
\mathbb{E}&\left[\left|\sum_i \frac{Y_i K_i}{\sum_i K_i}\right| I_{\widehat{E}_0}(I_{E_1} + I_{E_2})\right] \\
&\leq B_Y \mathbb{E}\left[I_{\{\sum_i K_i > 0 \wedge \sum_i \widehat{K}_i = 0\}}\right] \\
&= B_Y \mathbb{P}(\sum_i K_i > 0, \sum_i \widehat{K}_i = 0) \\
&\leq B_Y \mathbb{P}(\sum_{i=1}^m \widehat{K}_i = 0) \leq B_Y \zeta(n, m).
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\left|\sum_i \frac{Y_i \widehat{K}_i}{\sum_i \widehat{K}_i}\right| I_{E_0}(I_{\widehat{E}_1} + I_{\widehat{E}_2})\right] \leq \frac{B_Y}{em} \int \frac{d\mathcal{P}(P)}{\Phi_P(rh)}.$$

It is also easy to see that

$$\begin{aligned}
&\mathbb{E}\left[\Delta\widehat{f} I_{E_1}(I_{\widehat{E}_1}+I_{\widehat{E}_2})\right]\\
&\leq \mathbb{E}\left[\left(\left|\sum_i \frac{Y_iK_i}{\sum_i K_i}\right|+\left|\sum_i \frac{Y_i\widehat{K}_i}{\sum_i \widehat{K}_i}\right|\right) I_{E_1}(I_{\widehat{E}_1}+I_{\widehat{E}_2})\right]\\
&\leq \mathbb{E}\left[2B_Y I_{E_1}(I_{\widehat{E}_1}+I_{\widehat{E}_2})\right] \leq 2B_Y\mathbb{E}\left[I_{E_1}\right]\\
&= 2B_Y\mathbb{P}(0<\sum_{i=1}^m K_i<\underline{K}) \leq \frac{2B_Y}{em}\int \frac{d\mathcal{P}(P)}{\Phi_P(rh)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\Delta\widehat{f} I_{\widehat{E}_1}(I_{E_1}+I_{E_2})\right] &\leq 2B_Y\mathbb{P}(0<\sum_{i=1}^m \widehat{K}_i<\underline{K})\\
&\leq 2B_Y\zeta(n,m).
\end{aligned}$$

All that left is to upper bound $\mathbb{E}\left[\Delta\widehat{f} I_{E_2}I_{\widehat{E}_2}\right]$. The next lemma provides an upper bound for this.

Lemma 10

$$\mathbb{E}\left[\Delta\widehat{f} I_{E_2}I_{\widehat{E}_2}\right] \leq C_1\frac{1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]n^{-\frac{1}{2+k}}.$$

The proof can be found in the supplementary material.

Finally, putting the pieces together we obtain the following theorem.

Theorem 11

$$\begin{aligned}
&\mathbb{E}|\widehat{f}(\widehat{P};\widehat{P}_1,\ldots,\widehat{P}_m)-\widehat{f}(P;P_1,\ldots,P_m)|\\
&\quad\leq C_1\frac{1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]n^{-\frac{1}{2+k}}+C_2\frac{1}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]\\
&\qquad+(m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}.
\end{aligned}$$

The proof can be found in the supplementary material.

5.3 Upper bound on Equation 6

In this section we show that under the above specified conditions $\mathbb{E}|\widehat{f}(P;P_1,\ldots,P_m)-f(P)|$ can be upper bounded by

$$C_1(h^\beta)+C_2\left(\sqrt{\mathbb{E}\left[\frac{1}{m\Phi_P(rh/2)}\right]}\right)+\frac{C_3}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right],$$

where the expectation is with respect to the random probability measure P in $\mathcal{P}$.

We have to bound $\mathbb{E}|\widehat{f}(P;P_1,\ldots,P_m)-f(P)|$. Note that $Y_i=f(P_i)+\mu_i$, and

$$\begin{aligned}
&\mathbb{E}|\widehat{f}(P;P_1,\ldots,P_m)-f(P)|\\
&=\mathbb{E}\left|\frac{\sum_i Y_iK_i}{\sum_i K_i}I_{\{\sum_i K_i>0\}}-f(P)\right|\\
&=\mathbb{E}\left|\frac{\sum_i (f(P_i)+\mu_i)K_i}{\sum_i K_i}I_{\{\sum_i K_i>0\}}-f(P)\right|\\
&\leq\mathbb{E}\left[\left|\frac{\sum_i (f(P_i)-f(P))K_i}{\sum_i K_i}+\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i>0\}}\right]\\
&\quad+\mathbb{E}\left[|f(P)|\,I_{\{\sum_i K_i=0\}}\right]\\
&\leq\mathbb{E}\left[\frac{\sum_i |f(P_i)-f(P)|K_i}{\sum_i K_i}I_{\{\sum_i K_i>0\}}\right]\\
&\quad+\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i>0\}}\right]+f_{\max}\mathbb{P}(\sum_i K_i=0).
\end{aligned}$$

We will bound each of the three terms next. For the first term, since f is Hölder-β we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_i |f(P_i)-f(P)|K_i}{\sum_i K_i}I_{\{\sum_i K_i>0\}}\right]\\
&\quad\leq\mathbb{E}\left[\frac{\sum_i LD(P_i,P)^\beta K_i}{\sum_i K_i}I_{\{\sum_i K_i>0\}}\right]\leq L\,(hR)^\beta,
\end{aligned}$$

where in the last step we used the fact that

$$D(P_i,P)^\beta K_i = D(P_i,P)^\beta K\left(\frac{D(P_i,P)}{h}\right)\leq (hR)^\beta K_i,$$

since $supp(K)\subseteq B(0,R)$.

We now bound the second term.

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i>0\}}\right]\\
&=\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i\geq\underline{K}\}}+\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\underline{K}>\sum_i K_i>0\}}\right]\\
&\leq\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i\geq\underline{K}\}}\right]+B_Y\mathbb{P}(\underline{K}>\sum_i K_i)\\
&\leq\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i\geq\underline{K}\}}\right]+\frac{B_Y}{em}\int\frac{d\mathcal{P}(P)}{\Phi_P(rh)}.
\end{aligned}$$

(A4) implies that $\mathbb{P}(|\mu_i|\leq B_Y)=1$, i.e. B_Y is a bound on the noise. The last step follows from Lemma 4. For the first term in the above expression, we use the following lemma. Its proof can be found in the supplementary material.

Lemma 12

$$\mathbb{E}\left[\left|\frac{\sum_i \mu_iK_i}{\sum_i K_i}\right|I_{\{\sum_i K_i\geq\underline{K}\}}\right]\leq B_Y\sqrt{\frac{1+1/\underline{K}}{m\underline{K}}\int\frac{d\mathcal{P}(P)}{\Phi_P(rh)}}.$$

Finally, we bound the third term using Lemma 4:

$$f_{\max}\mathbb{P}(\sum_i K_i=0)\leq\frac{f_{\max}}{em}\int\frac{d\mathcal{P}(P)}{\Phi_P(rh)}.$$

Putting everything together, we have

$$\mathbb{E}|\widehat{f}(P;P_1,\ldots,P_m)-f(P)|$$

$$\leq L(hR)^\beta + B_Y\sqrt{\frac{1+1/\underline{K}}{m\underline{K}}\int \frac{d\mathcal{P}(P)}{\Phi_P(Rh)}} + \frac{B_Y}{em}\int \frac{d\mathcal{P}(P)}{\Phi_P(rh)} + \frac{f_{\max}}{em}\int \frac{d\mathcal{P}(P)}{\Phi_P(rh)}$$
$$\leq C_1 h^\beta + C_2\sqrt{\frac{1}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]} + \frac{C_3}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right].$$

Note that $\Phi_P(rh/2) \leq \Phi_P(rh) \leq \Phi_P(Rh)$.

6 Doubling Dimension

The upper bound on the risk in Theorem 1 depends on the quantity $\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]$. In future work, we will show that, without further assumptions, this quantity can be quite large which leads to very slow rates of convergence. This is because the covering number of the class $\mathcal{H}_k(1)$ is huge. For this paper, we concentrate on the more optimistic case where the support of $\mathcal{P}$ has small effective dimension.

One way to measure effective dimension is to use the doubling dimension. Following Kpotufe [2011], we say that $\mathcal{P}$ is a doubling measure with effective dimension d if, for every $r > 0$ and $0 < \epsilon < 1$,

$$\frac{\mathcal{P}(\mathcal{B}(s,r))}{\mathcal{P}(\mathcal{B}(s,\epsilon r))} \leq \left(\frac{c}{\epsilon}\right)^d, \quad \forall s. \tag{11}$$

If d denotes the doubling dimension of measure $\mathcal{P}$, then the $\sqrt{\mathbb{E}[1/(m\Phi_P(rh/2))]}$ term in Theorem 1 can be upper bounded as follows:

$$\sqrt{\mathbb{E}\left[\frac{1}{m\Phi_P(rh/2)}\right]} = \sqrt{\mathbb{E}\left[\frac{1}{m}\frac{\Phi_P(1)}{\Phi_P(rh/2)}\frac{1}{\Phi_P(1)}\right]}$$
$$\leq \sqrt{\frac{1}{m}C(rh/2)^{-d}\mathbb{E}\left[\frac{1}{\Phi_P(1)}\right]} \leq \frac{C}{\sqrt{mh^d}}.$$

Note also that when $mh^d \geq 1$, then $\frac{1}{mh^d} \leq \frac{1}{\sqrt{mh^d}}$. In this case, as a corollary of Theorem 1 and Assumptions (A5)-(A6), we now have that

$$R(m,n) \leq \frac{C_1}{h^{d+1}n^{1/(k+2)}} + C_2 h^\beta + C_3\sqrt{\frac{1}{mh^d}}, \tag{12}$$

for appropriate constants C_1, C_2 and C_3.

To derive the rates for the risk, we consider two separate cases, depending on whether the third term in the right hand side of (12) dominates the first term or not.

Thus first assume that

$$\sqrt{\frac{1}{mh^d}} = \Omega\left(\frac{C_1}{h^{d+1}n^{1/(k+2)}}\right), \tag{13}$$

so that the risk becomes, asymptotically, $O\left(h^\beta + \sqrt{\frac{1}{mh^d}}\right)$. The optimal choice for h is then $\Theta\left(m^{-1/(2\beta+d)}\right)$, yielding a rate for the risk

$$R(m,n) = O\left(m^{-\beta/(2\beta+d)}\right).$$

Notice that this choice of h ensures that our assumption (A6) is met, since in this case (13) implies that

$$n = \Omega\left(m^{\frac{\beta+d+1}{2\beta+d}(k+2)}\right),$$

from which we obtain that

$$h = \Theta\left(m^{-\frac{1}{2\beta+d}}\right) = \Omega\left(n^{-\frac{1}{(k+2)(\beta+d+1)}}\right) = \Omega\left(n^{-\frac{1}{k+2}}\right).$$

This rate is reasonable because if the number of samples per distribution n is large compared to the number m of distributions, then the learning rate is limited by the number of distributions m and is in fact precisely the same as the rate of learning a standard β-Hölder smooth regression function in d dimensions. That is, the effect of not knowing the distributions $P_1,\ldots,P_m$ exactly and only having a finite sample from the distributions is negligible.

For the second case, suppose that

$$\sqrt{\frac{1}{mh^d}} = O\left(\frac{1}{h^{d+1}n^{1/(k+2)}}\right). \tag{14}$$

Then, $R(m,n) = O\left(\frac{1}{h^{d+1}n^{1/(k+2)}} + h^\beta\right)$, which implies that the optimal choice for h is $h = \Theta\left(n^{-\frac{1}{(k+2)(\beta+d+1)}}\right)$, giving the rate

$$R(m,n) = O\left(n^{-\frac{\beta}{(k+2)(\beta+d+1)}}\right).$$

Just like before, this choice of h does not violate assumption (A6) since

$$h = \Theta\left(n^{-\frac{1}{(k+2)(\beta+d+1)}}\right) = \Omega\left(n^{-\frac{1}{k+2}}\right).$$

Notice that, (14) also implies that

$$m = \Omega\left(n^{\frac{2\beta+d}{(k+2)(\beta+d+1)}}\right).$$

In this case, the rate is limited by the number of samples per distribution n, as expected. Notice that the rate gets worse as the dimensionality of each distribution k grows and as the smoothness β of the regression function deteriorates.

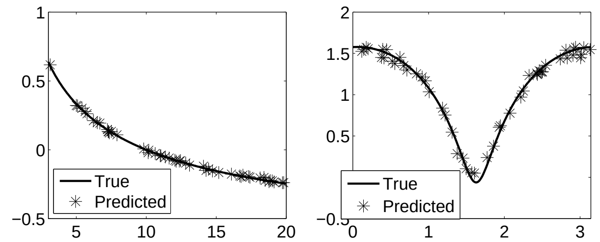
Remark. If there is no additive noise, i.e. $\mu_i = 0$, similar calculations yield that $R(m,n) = O\left(m^{-\frac{1}{\beta+d}}\right)$ when $n = \Omega\left(m^{\frac{\beta+d+1}{(\beta+d)(k+2)}}\right)$, and $R(m,n) = O\left(n^{-\frac{\beta}{(k+2)(\beta+d+1)}}\right)$ otherwise. While the rates seem reasonable, establishing optimality of the rates by demonstrating matching lower bounds is an open question that we plan to investigate in future work.

7 Numerical Illustrations

The following experiments serve as a proof of concept to demonstrate the applicability of the distribution regression estimator in Section 3. In these experiments, we used triangle kernels ($k(x) = 1 - |x|$ if $-1 \le x \le 1$, and 0 otherwise). We set all the $n, n_1, \ldots, n_m$ set sizes and $b, b_1, \ldots, b_m$ bandwidths to the same values, which will be specified below. In the first experiment, we generated 325 sample sets from $Beta(a, 3)$ distributions where a was varied between $[3, 20]$ randomly. We constructed $m = 250$ sample sets for training, 25 for validation, and 50 for testing. Each sample set contained $n = 500$ $Beta(a, 3)$ distributed i.i.d. points. Our task in this experiment was to learn the skewness of $Beta(a, b)$ distributions, $f = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$. We considered the noiseless case, i.e. μ was set to zero. Our estimator of course is not aware of that the sample sets are coming from beta distributions, and it does not know the skewness function values in the test sets either; its values are available only in the training and validation sets.

To find appropriate bandwidths b and h, we sampled 100 i.i.d. randomly and uniformly distributed values in [0,1], evaluated the MSE performance of the distribution regression estimator on the validation test using these bandwidths parameters, and then chose the bandwidth parameters that lead to the best values on the validation test. To estimate the L_2 distances between $\widehat{p}_i$ and p, we calculated their estimated values in 4096 points on a uniformly distributed grid between the min and max values in the sample sets, and then estimated the integral $\int (p(x) - \widehat{p}_i(x))^2 d(x)$ with the rectangle method for numerical integration. Figure 2(a) displays the predicted values for the 50 test sample sets, and we also show the true values of the skewness functions. As we can see the true and the estimated values are very close to each other.

In the next experiment, our task was to learn the entropy of Gaussian distributions. We chose a 2×2 covariance matrix $\Sigma = AA^T$, where $A \in \mathbb{R}^{2\times 2}$, and A_{ij} was randomly selected from the uniform distribution $U[0, 1]$. Just as in the previous experiments we constructed 325 sample sets from $\{\mathcal{N}(0, R(\alpha_i)\Sigma^{1/2})\}_{i=1}^{325}$. Where $R(\alpha_i)$ is a 2d rotation matrix with rotation angle $\alpha_i = i\pi/325$. From each $\mathcal{N}(0, R(\alpha_i)\Sigma^{1/2})$ distribution we sampled 500 2-dimensional i.i.d. points. Similarly to the previous experiment, 250 points was used for training, 25 for selecting appropriate bandwidth parameters, and 50 for training. Our goal was to learn the entropy of the first marginal distribution: $f = \frac{1}{2}\ln(2\pi e \sigma^2)$, where $\sigma^2 = M_{1,1}$ and $M = R(\alpha_i)\Sigma R^T(\alpha_i) \in \mathbb{R}^{2\times 2}$. μ was zero in this experiment as well. Figure 2(b) displays the learned entropies of the 50 test sample sets. The true and the estimated values are close to each other in this experiment as well.



(a) Skewness of Beta (b) Entropy of Gaussian

Figure 2: (a) Learned skewness of $Beta(a, 3)$ distribution. Axis x: parameter a in $[3, 20]$. Axis y: skewness of $Beta(a, 3)$. (b) Learned entropy of a 1d marginal distribution of a rotated 2d Gaussian distribution. Axes x: rotation angle in $[0, \pi]$. Axis y: entropy.

8 Discussion and Conclusion

We have presented an estimator for distribution regression which is distribution-free in the sense that the estimator makes no strong distributional assumptions on the error variables. We derived upper bounds on the risk of the estimator and, in particular, we analyzed the case with a finite doubling dimension.

We note that our rates are faster than the logarithmic rates that are sometimes obtained in measurement error nonparametric regression models as in Fan and Truong [1993]. The reason is that the logarithmic rates occur when the measurement error is Gaussian. Our measurement error corresponds to $||\widehat{p}_i - p_i||$ which is not Gaussian for finite n_i and which decreases when n_i increases. In the standard measurement error model, the error is $O(1)$ and is not decreasing.

In future work, we will prove lower bounds which show that, without further assumptions (such as assumptions about the doubling dimension), the rates can be very slow. We will also verify if the rates in the doubling dimension setting are tight or not. Also, we plan to investigate other estimators such as k-nn estimators and RKHS estimators.

Acknowledgements

This research is supported in part by NSF under grants IIS-1247658, IIS-1250350, and DMS 1149677.

References

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement error in nonlinear mod-*

els: a modern perspective, volume 105. Chapman & Hall/CRC, 2006.

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, pages 406–414, 2010.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, pages 1900–1925, 1993.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Verlag, 2006.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493. MIT Press, 1998.

T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. *JMLR*, 5:819–844, 2004.

R. Kondor and T. Jebara. A kernel between sets of vectors. In *ICML*, 2003.

S. Kpotufe. k-nn regression adapts to local intrinsic dimension. *arXiv preprint arXiv:1110.4300*, 2011.

P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *NIPS*, 2004.

K. Muandet, B. Schölkopf, K. Fukumizu, and F. Dinuzzo. Learning from distributions via support measure machines. *arXiv.org*, stat.ML, February 2012.

B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *UAI*, 2011.

B. Póczos, L. Xiong, D. Sutherland, and J. Schneider. Nonparametric kernel estimators for image classification. In *Computer Vision and Pattern Recognition*, 2012.

B. Póczos, A. Rinaldo, A. Singh, and L. Wasserman. Distribution-free distribution regression, 2013. `http://arxiv.org/abs/1302.0082`.

J.O. Ramsay and B.W Silverman. *Functional data analysis*. Springer, New York, 2nd edition, 2005.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4): 1154–1178, 2009.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*, 2007.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2010.

Supplementary material

Proof of Lemma 2

Proof. Recall that we assume that $\mathcal{P}$ is supported on the set $\mathcal{H}_k(1)$ of distributions, which are 1-smooth k-dimensional densities as defined in Rigollet and Vert [2009].

Let $\mathbb{E}\left[D_2(\widehat{P}_i, P)|P_i\right] = \mathbb{E}\left[\sqrt{\int(\widehat{p}_i - p_i)^2}\right]$ denote the integrated mean squared risk for the density estimator $\widehat{p}_i$ of a fixed density p_i. It then follows from Lemma 4.1 of Rigollet and Vert [2009] that (with an appropriate kernel function B),

$$\mathbb{E}[D_2^2(\widehat{P}_i, P_i)|P_i] \leq c_1^2 b_i^2 + \left(\frac{c_2^2}{n_i b_i^k}\right)$$

for some constants $c_1, c_2 > 0$.

From Jensen's inequality, we have that $E[X] \leq (\mathbb{E}[X^2])^{1/2}$ for any X random variable. We also know that $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ for any $a, b > 0$, therefore

$$\begin{aligned}\mathbb{E}[D_2(\widehat{P}_i, P_i)|P_i] &\leq \left(c_1^2 b_i^2 + \left(\frac{c_2^2}{n_i b_i^k}\right)\right)^{1/2}\\ &\leq c_1 b_i + \frac{c_2}{n_i^{1/2} b_i^{k/2}}.\end{aligned}$$

Since the distributions in $\mathbb{D}$ are supported on a compact set and the kernel B has also compact support, we have, for an appropriate constant $c_0 > 0$,

$$\int |p_i - \widehat{p}_i| \leq c_0 \sqrt{\int (p_i - \widehat{p}_i)^2}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[D(\widehat{P}_i, P_i)|P_i] &\leq c_0 \mathbb{E}[D_2(\widehat{P}_i, P_i)|P_i]\\ &\leq c_0(c_1 b_i + \frac{c_2}{n_i^{1/2} b_i^{k/2}})\\ &\leq c_0(c_1 + c_2) n^{-\frac{1}{2+k}},\end{aligned}$$

where the last step follows from our assumptions that $n_i^{-1/2} b_i^{-k/2} \leq n^{-\frac{1}{2}} n^{\frac{k}{2(k+2)}} = n^{-\frac{1}{k+2}}$, and thus

$$c_1 b_i + \frac{c_2}{n_i^{1/2} b_i^{k/2}} \leq (c_1 + c_2) n^{-\frac{1}{2+k}}.$$

□

Proof of Lemma 4

Proof. The proof follows the argument of **?**.

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=1}^m K_i < \underline{\mathrm{K}}\Big) &= \mathbb{P}\left(\sum_{i=1}^m K\left(\frac{D(P_i, P)}{h}\right) < \underline{\mathrm{K}}\right)\\ &\leq \mathbb{P}\left(\sum_{i=1}^m I_{\{D(P_i,P)<rh\}} = 0\right),\end{aligned}$$

since according to our assumptions on kernel K if for some i it holds that $D(P_i, P)/h \leq r$, then $K_i \geq \underline{\mathrm{K}}$. Therefore,

$$\mathbb{P}\Big(\sum_{i=1}^m K_i < \underline{\mathrm{K}}\Big) \leq \mathbb{P}\left(\sum_{i=1}^m I_{\{D(P_i,P)<rh\}} = 0\right)$$
$$= \mathbb{E}[\mathbb{P}(\sum_{i=1}^m I_{\{D(P_i,P)<rh\}} = 0\,|P)]$$
$$= \int \mathbb{P}\left(\sum_{i=1}^m I_{\{D(P_i,P)<rh\}} = 0 \,\middle|\, P\right) d\mathcal{P}(P)$$
$$= \int [1 - \mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)]^m d\mathcal{P}(P) \tag{15}$$
$$\leq \int \exp[-m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)] d\mathcal{P}(P) \tag{16}$$
$$= \int \exp[-m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)] \times \frac{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} d\mathcal{P}(P)$$
$$\leq \max_{u>0} u \exp(-u) \int \frac{d\mathcal{P}(P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} \tag{17}$$
$$\leq \frac{1}{e} \int \frac{d\mathcal{P}(P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} = \frac{1}{em} \mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right],$$

where we used in (15), (16), and (17) respectively that $\{P_i\}$ are iid, $(1-u)^m \leq \exp(-um)$ for all $0 \leq u \leq 1$, $m \geq 1$, and $\max(u \exp(-u)) = \frac{1}{e}$. □

Proof of Lemma 5

Proof.

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\sum_i K_i} I_{\{\sum_i K_i \geq \underline{K}\}}\right] &\leq \mathbb{E}\left[\frac{1 + 1/\underline{K}}{1 + \sum_i K_i}\right]\\ &\leq \mathbb{E}\left[\frac{1 + 1/\underline{K}}{1 + \underline{K}\sum_i I_{\{D(P_i,P)\leq hr\}}}\right]\\ &= \frac{1 + 1/\underline{K}}{\underline{K}}\, \mathbb{E}\left[\frac{1}{1/\underline{K} + \sum_i I_{\{D(P_i,P)\leq hr\}}}\right]\\ &\leq \frac{1 + 1/\underline{K}}{\underline{K}}\, \mathbb{E}\left[\frac{1}{1 + \sum_i I_{\{D(P_i,P)\leq hr\}}}\right]\\ &= \frac{1 + 1/\underline{K}}{\underline{K}}\, \mathbb{E}\Big[\mathbb{E}\left[\frac{1}{1 + \sum_i I_{\{D(P_i,P)\leq hr\}}}|P\right]\Big]\\ &\leq \frac{1 + 1/\underline{K}}{m\underline{K}} \mathbb{E}\Big[\frac{1}{\Phi_P(rh)}\Big],\end{aligned}$$

where the second-to-last line uses the fact that $\underline{K} < 1$ and the last line follows since for a binomial random variable $B(m,p)$, $\mathbb{E}[\frac{1}{1+B(m,p)}] \leq \frac{1}{(m+1)p} \leq \frac{1}{mp}$. □

Proof of Lemma 6

Proof. $D(P,Q)$ is a distance, therefore the triangle inequality holds, and we have that

$$\begin{aligned}|\epsilon_i| = |K_i - \widehat{K}_i| &= \left|K\left(\frac{D(P,P_i)}{h}\right) - K\left(\frac{D(\widehat{P},\widehat{P}_i)}{h}\right)\right| \\ &\le \frac{L_K}{h}|D(P,P_i) - D(\widehat{P},\widehat{P}_i)| \\ &\le \frac{L_K}{h}(D(P,\widehat{P}) + D(P_i,\widehat{P}_i)).\end{aligned}$$

Here we used that

$$\begin{aligned}&D(P,P_i) - D(\widehat{P},\widehat{P}_i) \\ &\quad\le [D(P,\widehat{P}) + D(\widehat{P},\widehat{P}_i) + D(\widehat{P}_i,P_i)] - D(\widehat{P},\widehat{P}_i) \\ &\quad= D(P,\widehat{P}) + D(\widehat{P}_i,P_i),\end{aligned}$$

and

$$\begin{aligned}&D(\widehat{P},\widehat{P}_i) - D(P,P_i) \\ &\quad\le [D(\widehat{P},P) + D(P,P_i) + D(P_i,\widehat{P}_i)] - D(P,P_i) \\ &\quad= D(\widehat{P},P) + D(P_i,\widehat{P}_i).\end{aligned}$$

□

Proof of Lemma 7

Proof. From Markov's inequality, for any X, Y and constant $\kappa > 0$,

$$1 \le \frac{\mathbb{E}[|X|\,|Y]}{\kappa} + \mathbb{P}(|X| < \kappa|Y).$$

Thus,

$$\begin{aligned}&\mathbb{P}\Big(\sum_i |\epsilon_i| < \kappa \;\Big|\; \{P_i\}_{i=1}^m, P\Big) \qquad (18)\\ &\quad\ge 1 - \frac{\mathbb{E}[\sum_i |\epsilon_i||\{P_i\}_{i=1}^m, P]}{\kappa} \\ &\quad= 1 - \frac{\sum_i \mathbb{E}[|\epsilon_i||P_i, P]}{\kappa} \\ &\quad\ge 1 - \frac{L_K}{h\kappa}\sum_{i=1}^m \mathbb{E}[(D(P,\widehat{P}) + D(P_i,\widehat{P}_i))|P_i, P] \qquad (19)\\ &\quad\ge 1 - \frac{L_K}{h\kappa} m 2\bar{C} n^{-\frac{1}{2+k}} = \eta(\kappa, n, m).\end{aligned}$$

Here (19) holds due to Lemma 6, and we also used (7).
□

Proof of Lemma 8

Proof. The term $\mathbb{E}\left[\sum_{i=1}^m |\epsilon_i|\middle| P, \{P_i\}_{i=1}^m\right]$ is upper bounded by

$$\begin{aligned}&\frac{L_K}{h}\sum_{i=1}^m \mathbb{E}\left[D(P,\widehat{P}) + D(P_i,\widehat{P}_i)\middle| P, \{P_i\}_{i=1}^m\right] \\ &\le \frac{L_K}{h} 2\bar{C} m n^{-\frac{1}{2+k}}.\end{aligned}$$

□

Proof of Lemma 9

Proof. Recall that $D(\widehat{P}_i, P_i) \le rh/4$ for all i on an event $\Omega_{m,n}$ and that $\mathbb{P}(\Omega_{m,n}^c) \le (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}$. So, on $\Omega_{m,n}$,

$$\begin{aligned}D(\widehat{P}_i,\widehat{P}) &\le D(\widehat{P}_i,P_i) + D(P,\widehat{P}) + D(P_i,P) \\ &\le D(P_i,P) + \frac{rh}{2}.\end{aligned}$$

Now, using the event $\Omega_{m,n}$ defined in Lemma 3,

$$\begin{aligned}&\mathbb{P}\Big(\sum_{i=1}^m \widehat{K}_i = 0\Big) \le \mathbb{P}\Big(\sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) \\ &= \mathbb{P}\Big(\Omega_{m,n}, \sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) + \mathbb{P}\Big(\Omega_{m,n}^c, \sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) \\ &\le \mathbb{P}\Big(\Omega_{m,n}, \sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) + \mathbb{P}\Big(\Omega_{m,n}^c\Big) \\ &\le \mathbb{P}\Big(\Omega_{m,n}, \sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) + (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\Big(\Omega_{m,n}, \sum_{i=1}^m \widehat{K}_i < \underline{\mathrm{K}}\Big) &= \mathbb{P}(\Omega_{m,n}, \sum_{i=1}^m I_{D(\widehat{P}_i,\widehat{P})<rh} = 0) \\ &\le \mathbb{P}(\sum_{i=1}^m I_{D(P_i,P)<rh/2} = 0) \\ &\le \frac{1}{em}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right].\end{aligned}$$

The result follows. □

Proof of Lemma 10

Proof.

$$\begin{aligned}&\mathbb{E}\left[\Delta\widehat{f} I_{E_2} I_{\widehat{E}_2}\right] = \mathbb{E}\left[\left|\frac{\sum_i Y_i\widehat{K}_i}{\sum_j \widehat{K}_j} - \frac{\sum_i Y_i K_i}{\sum_j K_j}\right| I_{E_2} I_{\widehat{E}_2}\right] \\ &= \mathbb{E}\left[\left|\sum_i Y_i\left(\frac{K_i}{\sum_j K_j} - \frac{\widehat{K}_i}{\sum_j \widehat{K}_j}\right)\right| I_{E_2} I_{\widehat{E}_2}\right] \\ &\le B_Y\mathbb{E}\left[\sum_i\left|\left(\frac{K_i}{\sum_j K_j} - \frac{\widehat{K}_i}{\sum_j \widehat{K}_j}\right)\right| I_{E_2} I_{\widehat{E}_2}\right] \\ &= B_Y\mathbb{E}\left[\sum_i\left(\frac{|K_i(\sum_j \widehat{K}_j) - \widehat{K}_i(\sum_j K_j)|}{(\sum_j \widehat{K}_j)(\sum_j K_j)}\right) I_{E_2} I_{\widehat{E}_2}\right] \\ &= B_Y\mathbb{E}\Bigg[\sum_i\left(\frac{|(\widehat{K}_i - \epsilon_i)(\sum_j \widehat{K}_j) - \widehat{K}_i(\sum_j (\widehat{K}_j - \epsilon_j))|}{(\sum_j \widehat{K}_j)(\sum_j K_j)}\right)\end{aligned}$$

$$
\begin{aligned}
&\times I_{E_2} I_{\widehat{E}_2}\Big] \\
&= B_Y \mathbb{E}\Bigg[\sum_i \left(\frac{|-\epsilon_i(\sum_j \widehat{K}_j) + \widehat{K}_i(\sum_j \epsilon_j)|}{(\sum_j \widehat{K}_j)(\sum_j K_j)}\right) I_{E_2} I_{\widehat{E}_2}\Bigg] \\
&\leq B_Y \mathbb{E}\Bigg[\Bigg(\frac{(\sum_i \widehat{K}_i)(\sum_j |\epsilon_j|)}{(\sum_j \widehat{K}_j)(\sum_j K_j)} \\
&\quad + \frac{(\sum_i |\epsilon_i|)(\sum_j \widehat{K}_j)}{(\sum_j \widehat{K}_j)(\sum_j K_j)}\Bigg) I_{E_2} I_{\widehat{E}_2}\Bigg] \\
&= B_Y \mathbb{E}\Bigg[\left(\frac{\sum_j |\epsilon_j|}{\sum_j K_j} + \frac{(\sum_i |\epsilon_i|)}{\sum_j K_j}\right) I_{E_2} I_{\widehat{E}_2}\Bigg] \\
&\leq 2 B_Y \mathbb{E}\Bigg[\frac{\sum_j |\epsilon_j|}{\sum_j K_j} I_{E_2}\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_j |\epsilon_j| | P, \{P_i\}_{i=1}^m\Bigg] \frac{1}{\sum_j K_j} I_{E_2}\Bigg] \\
&\leq 2 B_Y \frac{L_K}{h} 2\widehat{c} m n^{-\frac{1}{2+k}} \mathbb{E}\Bigg[\frac{1}{\sum_j K_j} I_{E_2}\Bigg] \\
&\leq 2 B_Y \frac{L_K}{h} 2\widehat{c} m n^{-\frac{1}{2+k}} \frac{1+1/\underline{K}}{m\underline{K}} \mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right] \\
&= C_1 \frac{1}{h} \mathbb{E}\Bigg[\frac{1}{\Phi_P(rh)}\Bigg] n^{-\frac{1}{2+k}}.
\end{aligned}
$$

where we used Lemma 8 and Lemma 5. □

Proof of Theorem 11

Proof.

$$
\begin{aligned}
&\mathbb{E}|\widehat{f}(\widehat{P}; \widehat{P}_1, \ldots, \widehat{P}_m) - \widehat{f}(P; P_1, \ldots, P_m)| \\
&\quad \leq \mathbb{E}\left[\Delta \widehat{f} I_{E_2} I_{\widehat{E}_2}\right] + 3 B_Y \zeta + 3 \frac{B_Y}{em} \mathbb{E}\Bigg[\frac{1}{\Phi_P(rh)}\Bigg] \\
&\quad \leq C_1 \frac{1}{h} \mathbb{E}\Bigg[\frac{1}{\Phi_P(Rh)}\Bigg] n^{-\frac{1}{2+k}} + C_2 \frac{1}{m} \mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right] \\
&\qquad + C_3 \frac{1}{m} \mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + (m+1) e^{-\frac{1}{2} n^{\frac{k}{2+k}}}.
\end{aligned}
$$

Note also that $\Phi_P(rh/2) \leq \Phi_P(rh) \leq \Phi_P(Rh)$. □

Proof of Lemma 12

Proof. Notice that if $\sum_i K_i \geq \underline{K} > 0$,

$$
\operatorname{var}\left(\frac{\sum_i \mu_i K_i}{\sum_i K_i} | P, P_1, \ldots, P_m\right) \leq B_Y^2 \frac{\sum_i K_i^2}{(\sum_i K_i)^2}.
$$

Using this and Hölder's inequality, we get:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{\sum_i \mu_i K_i}{\sum_i K_i}\right| I_{\{\sum_i K_i \geq \underline{K}\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\frac{\sum_i \mu_i K_i}{\sum_i K_i}\right| I_{\sum_i K_i \geq \underline{K}} | P, P_1, \ldots, P_m\right]\right] \\
&\leq \mathbb{E}\left[\sqrt{\operatorname{var}\left(\frac{\sum_i \mu_i K_i}{\sum_i K_i} | P, P_1, \ldots, P_m\right)} I_{\{\sum_i K_i \geq \underline{K}\}}\right] \\
&\leq \mathbb{E}\left[B_Y \frac{\sqrt{\sum_i K_i^2}}{\sum_i K_i} I_{\{\sum_i K_i \geq \underline{K}\}}\right] \\
&\leq \mathbb{E}\left[B_Y \frac{\sqrt{\sum_i K_i}}{\sum_i K_i} I_{\{\sum_i K_i \geq \underline{K}\}}\right] \\
&\leq B_Y \sqrt{\mathbb{E}\left[\frac{\sum_i K_i}{(\sum_i K_i)^2} I_{\{\sum_i K_i \geq \underline{K}\}}\right]} \\
&\leq B_Y \sqrt{\mathbb{E}\left[\frac{I_{\{\sum_i K_i \geq \underline{K}\}}}{\sum_i K_i}\right]} \\
&\leq B_Y \sqrt{\frac{1+1/\underline{K}}{m\underline{K}} \int \frac{d\mathcal{P}(P)}{\Phi_P(rh)}}.
\end{aligned}
$$

The second inequality holds since $K(x) < 1$ and the last step stems from Lemma 5. □