More on Kernel Density Estimation

These notes follow Larry Wasserman's notes (available at `https://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf`) closely, while supplementing them with additional details and related material.

1 Introduction

Let $X_1, \ldots, X_n$ be a sample from a distribution P with density p. The goal of nonparametric density estimation is to estimate p with as few assumptions about p as possible. We denote the estimator by $\hat{p}$. The estimator will depend on a smoothing parameter h and choosing h carefully is crucial. To emphasize the dependence on h we sometimes write $\hat{p}_h$.

Density estimation has been widely studied (e.g., Stone (1984) [6]; Tsybakov (2009) [7]), and is used for regression, classification, clustering, among others. For example, if $\hat{p}(x, y)$ is an estimate of $p(x, y)$ then we get the following estimate of the regression function $m(x) = \mathbb{E}[Y \mid X = x]$:

$$\hat{m}(x) = \int y\, \hat{p}(y|x)\, dy$$

where $\hat{p}(y|x) = \hat{p}(y, x)/\hat{p}(x)$. For classification, recall that the Bayes rule is

$$h^*(x) = I(p_1(x)\pi_1 > p_0(x)\pi_0)$$

where $\pi_1 = P(Y = 1)$, $\pi_0 = P(Y = 0)$, $p_1(x) = p(x|y = 1)$ and $p_0(x) = p(x|y = 0)$. Inserting sample estimates of π_1 and π_0, and density estimates for p_1 and p_0 yields an estimate of the Bayes rule. For clustering, we look for the high density regions, based on an estimate of the density. Many familiar classifiers can be re-expressed this way.

Perhaps the simplest density estimators are histograms, but we focus on kernel density estimators, which are smoother and converge to the true density faster, in these notes. Also, we use the L_2 loss

$$L(p, \hat{p}) = \int (\hat{p}(x) - p(x))^2\, dx.$$

2 Kernel Density Estimation

A one-dimensional smoothing kernel is any smooth function K such that $\int K(x)\, dx = 1$, $\int xK(x)\, dx = 0$ and $\sigma_K^2 \equiv \int x^2 K(x)\, dx > 0$. Some commonly used kernels are the following:

Boxcar:	$K(x) = \frac{1}{2} I(x)$	**Gaussian:**	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Epanechnikov:	$K(x) = \frac{3}{4}(1 - x^2) I(x)$	**Tricube:**	$K(x) = \frac{70}{81}(1 - \lvert x\rvert^3)^3 I(x)$

where $I(x) = 1$ if $|x| \leq 1$ and $I(x) = 0$ otherwise. Two commonly used multivariate kernels are $\prod_{j=1}^{d} K(x_j)$ and $K(\|x\|)$.

Suppose that $X \in \mathbb{R}^d$. Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right). \tag{1}$$

More generally, we define

$$\hat{p}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where H is a positive definite bandwidth matrix and $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$. For simplicity, we will take $H = h^2 I$ and we get back the previous formula.

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel satisfying

$$\int_{\mathbb{R}^d} K(u)\, du = 1.$$

Definition 2.1. Given a bandwidth $h > 0$, the kernel density estimator is defined to be

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Its expectation is

$$p_h(x) := \mathbb{E}[\hat{p}_h(x)] = \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x - u}{h}\right) p(u)\, du.$$

The kernel estimator places a smoothed out lump of mass of size $1/n$ over each data point X_i. The choice of kernel K is not crucial, but the choice of bandwidth h is important. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates.

Theorem 2.2 (Pointwise consistency of KDE)**.** *Assume that p is continuous at x and that p is bounded. If $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$, then*

$$\hat{p}_{h_n}(x) \xrightarrow{P} p(x).$$

Proof. We show that the mean squared error $E[(\hat{p}_{h_n}(x) - p(x))^2] \to 0$, which implies convergence in probability by Markov's inequality. Since

$$E\big[(\hat{p}_{h_n}(x) - p(x))^2\big] = \big(p_{h_n}(x) - p(x)\big)^2 + \mathrm{Var}\big(\hat{p}_{h_n}(x)\big),$$

it suffices to show each term tends to zero.

Bias. Substituting $v = (x - u)/h_n$ in the definition of $p_{h_n}(x)$ gives

$$p_{h_n}(x) = \int_{\mathbb{R}^d} K(v)\, p(x - h_n v)\, dv.$$

Since $\int K(v)\, dv = 1$, subtracting $p(x)$ yields

$$p_{h_n}(x) - p(x) = \int_{\mathbb{R}^d} K(v)\big[p(x - h_n v) - p(x)\big]\, dv.$$

Let $\varepsilon > 0$. By continuity of p at x, there exists $\delta > 0$ such that $|p(x - h_n v) - p(x)| < \varepsilon$ whenever $h_n\|v\| < \delta$. Since K has bounded support and $h_n \to 0$, for all sufficiently large n every v in the support of K satisfies $h_n\|v\| < \delta$. Therefore

$$|p_{h_n}(x) - p(x)| \le \int_{\mathbb{R}^d} K(v)\, |p(x - h_n v) - p(x)|\, dv < \varepsilon \int_{\mathbb{R}^d} K(v)\, dv = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, $p_{h_n}(x) - p(x) \to 0$.

Variance. Write $\hat{p}_{h_n}(x) = n^{-1} \sum_{i=1}^{n} Z_i$ where $Z_i = h_n^{-d} K((x - X_i)/h_n)$. The Z_i are i.i.d., so

$$\mathrm{Var}\big(\hat{p}_{h_n}(x)\big) = \frac{1}{n}\mathrm{Var}(Z_1) \le \frac{1}{n}E[Z_1^2] = \frac{1}{nh_n^{2d}} \int_{\mathbb{R}^d} K^2\left(\frac{x-u}{h_n}\right) p(u)\, du.$$

Substituting $v = (x - u)/h_n$ gives

$$\mathrm{Var}\big(\hat{p}_{h_n}(x)\big) \le \frac{1}{nh_n^d} \int_{\mathbb{R}^d} K^2(v)\, p(x - h_n v)\, dv \le \frac{\sup_u p(u)}{nh_n^d} \int_{\mathbb{R}^d} K^2(v)\, dv = \frac{c}{nh_n^d},$$

where $c = \sup_u p(u) \int K^2(v)\, dv < \infty$ since p is bounded and $K \in L^2$. Since $nh_n^d \to \infty$ by assumption, $\mathrm{Var}(\hat{p}_{h_n}(x)) \to 0$.

Finally, for any $\varepsilon > 0$, Markov's inequality applied to the MSE gives

$$P(|\hat{p}_{h_n}(x) - p(x)| > \varepsilon) \le \frac{(p_{h_n}(x) - p(x))^2 + \mathrm{Var}(\hat{p}_{h_n}(x))}{\varepsilon^2} \to 0. \qquad \square$$

2.1 Risk Analysis

In this section, we study the accuracy of kernel density estimation. Some additional smoothness or structural condition must hold in order for density estimation to be possible. If the density is completely unconstrained, then it can wiggle or oscillate so quickly that there is no hope of estimating its value at any given point. One of the most widely studied smoothness conditions is Hölder smoothness. This requires that a function has sufficiently many derivatives, which themselves have to be appropriately smooth.

We begin with some definitions and regularity assumptions. Assume that $X_1, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$, where $\mathcal{X}$ is compact. Let $\beta > 0$, and write

$$\beta = k + \alpha, \qquad k := \lceil \beta \rceil - 1 \in \mathbb{N}_0, \qquad \alpha \in (0, 1].$$

For a multi-index $s = (s_1, \dots, s_d) \in \mathbb{N}_0^d$, define

$$|s| = s_1 + \cdots + s_d, \qquad s! = s_1! \cdots s_d!, \qquad x^s = x_1^{s_1} \cdots x_d^{s_d}, \qquad D^s = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

The Hölder class $\Sigma(\beta, L)$ consists of all functions $g : \mathcal{X} \to \mathbb{R}$ such that all mixed partial derivatives $D^s g$ exist for $|s| \le k$, and

$$|D^s g(x) - D^s g(y)| \le L\|x - y\|^\alpha \qquad \text{for all } |s| = k \text{ and all } x, y \in \mathcal{X}.$$

Equivalently, functions in $\Sigma(\beta, L)$ have k derivatives, and all derivatives of order k are α-Hölder continuous.

When β is an integer, we have $\alpha = 1$ and $k = \beta - 1$, so this reduces to the usual definition: all derivatives up to order $\beta - 1$ exist, and the derivatives of order $\beta - 1$ are Lipschitz.

If $g \in \Sigma(\beta, L)$ and $\beta = k + \alpha$, then for each $x \in \mathcal{X}$,

$$\big|g(u) - g_{x,k}(u)\big| \le C\|u - x\|^\beta,$$

where

$$g_{x,k}(u) = \sum_{|s| \le k} \frac{(u - x)^s}{s!} D^s g(x),$$

and C depends only on (β, L, d).

In the important case $\beta = 2$, this becomes

$$\big|g(u) - \big[g(x) + (u-x)^T \nabla g(x)\big]\big| \le C\|u-x\|^2.$$

Now assume that the kernel has product form

$$K(u) = G(u_1) \cdots G(u_d),$$

where G is supported on $[-1, 1]$, satisfies

$$\int G(t)\,dt = 1, \qquad \int |G(t)|^p\,dt < \infty \quad \text{for every } p \ge 1,$$

and such that

$$\int |u|^\beta |K(u)|\,du < \infty, \qquad \int K(u)^2\,du < \infty.$$

We also assume that K has vanishing moments up to order k, namely

$$\int u^s K(u)\,du = 0 \qquad \text{for every multi-index } s \text{ with } 1 \le |s| \le k.$$

When $\beta = 2$, a standard example is the Epanechnikov kernel

$$G(t) = \frac{3}{4}(1 - t^2), \qquad |t| \le 1.$$

For higher-order kernels, the moment conditions generally require kernels that take negative values.

Let $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$. The next lemma bounds the bias $p_h(x) - p(x)$.

Lemma 2.3. *Suppose $p \in \Sigma(\beta, L)$, where $\beta > 0$ and $\beta = k + \alpha$ with $k = \lceil \beta \rceil - 1$ and $\alpha \in (0, 1]$. Assume that the kernel $K : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$\int K(u)\,du = 1, \qquad \int \|u\|^\beta |K(u)|\,du < \infty,$$

and the moment conditions

$$\int u^s K(u)\,du = 0 \qquad \text{for every multi-index } s \text{ with } 1 \le |s| \le k.$$

Then

$$|p_h(x) - p(x)| \le Ch^\beta$$

for all $x \in \mathcal{X}$, where C depends only on (β, L, d, K).

Proof. Recall that

$$p_h(x) = \int \frac{1}{h^d} K\left(\frac{x-u}{h}\right) p(u)\,du.$$

With the change of variables $v = (x-u)/h$, equivalently $u = x - hv$, we get

$$p_h(x) - p(x) = \int K(v)\big(p(x - hv) - p(x)\big)\,dv.$$

Let

$$p_{x,k}(u) = \sum_{|s|\le k} \frac{(u-x)^s}{s!} D^s p(x)$$

denote the Taylor polynomial of order $k = \lceil \beta \rceil - 1$ around x. Add and subtract $p_{x,k}(x - hv)$:

$$p_h(x) - p(x) = \int K(v)\big(p(x-hv) - p_{x,k}(x-hv)\big)\,dv + \int K(v)\big(p_{x,k}(x-hv) - p(x)\big)\,dv.$$

Since $p \in \Sigma(\beta, L)$, the Taylor remainder satisfies

$$|p(x-hv) - p_{x,k}(x-hv)| \le Ch^\beta \|v\|^\beta.$$

Therefore

$$\left| \int K(v)\big(p(x-hv) - p_{x,k}(x-hv)\big)\,dv \right| \le Ch^\beta \int \|v\|^\beta |K(v)|\,dv \lesssim h^\beta.$$

For the second term, note that $p_{x,k}(x-hv) - p(x)$ is a polynomial in v of degree at most k with no constant term. Hence, by the vanishing-moment assumptions,

$$\int K(v)\big(p_{x,k}(x-hv) - p(x)\big)\,dv = 0.$$

Combining the two bounds gives

$$|p_h(x) - p(x)| \le Ch^\beta.$$

□

Lemma 2.4. *For $p \in \Sigma(\beta, L)$,*

$$\mathrm{Var}(\hat{p}_h(x)) \le \frac{C}{nh^d}.$$

Proof. Write

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n Z_i, \qquad Z_i = \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Since the Z_i are iid,

$$\mathrm{Var}(\hat{p}_h(x)) = \frac{1}{n}\mathrm{Var}(Z_1).$$

Also,

$$\mathrm{Var}(Z_1) \le \mathbb{E}[Z_1^2] = \frac{1}{h^{2d}} \int K^2\left(\frac{x-u}{h}\right) p(u)\,du.$$

Using the change of variables $v = (x-u)/h$, equivalently $u = x - hv$, gives

$$\mathbb{E}[Z_1^2] = \frac{1}{h^d} \int K^2(v) p(x-hv)\,dv.$$

Since p is uniformly bounded on the class $\Sigma(\beta, L)$,

$$\mathbb{E}[Z_1^2] \le \frac{\sup_y p(y)}{h^d} \int K^2(v)\,dv \le \frac{C}{h^d}.$$

Hence

$$\mathrm{Var}(\hat{p}_h(x)) \le \frac{C}{nh^d}.$$

□

Since the mean squared error is equal to the variance plus the bias squared we have:

Theorem 2.5. *Suppose $p \in \Sigma(\beta, L)$, where $\beta > 0$, and assume that the kernel K satisfies the assumptions of the previous lemma as well as*

$$\int K(u)^2 \, du < \infty.$$

Then

$$\sup_{p \in \Sigma(\beta, L)} \mathbb{E} \int_{\mathcal{X}} (\hat{p}_h(x) - p(x))^2 \, dx \lesssim h^{2\beta} + \frac{1}{nh^d}.$$

In particular, if

$$h \asymp n^{-1/(2\beta+d)},$$

then

$$\sup_{p \in \Sigma(\beta, L)} \mathbb{E} \int_{\mathcal{X}} (\hat{p}_h(x) - p(x))^2 \, dx \lesssim n^{-2\beta/(2\beta+d)}.$$

Proof. For each fixed x,

$$\hat{p}_h(x) - p(x) = \big(\hat{p}_h(x) - p_h(x)\big) + \big(p_h(x) - p(x)\big),$$

where $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$. Since $\mathbb{E}[\hat{p}_h(x) - p_h(x)] = 0$, we have

$$\mathbb{E}(\hat{p}_h(x) - p(x))^2 = \operatorname{Var}(\hat{p}_h(x)) + \big(p_h(x) - p(x)\big)^2.$$

Integrating over $x \in \mathcal{X}$ gives

$$\mathbb{E} \int_{\mathcal{X}} (\hat{p}_h(x) - p(x))^2 \, dx = \int_{\mathcal{X}} \operatorname{Var}(\hat{p}_h(x)) \, dx + \int_{\mathcal{X}} (p_h(x) - p(x))^2 \, dx.$$

For the bias term, Lemma 2.1 gives

$$\sup_{x \in \mathcal{X}} |p_h(x) - p(x)| \lesssim h^{\beta}.$$

Since $\mathcal{X}$ has finite Lebesgue measure,

$$\int_{\mathcal{X}} (p_h(x) - p(x))^2 \, dx \le |\mathcal{X}| \sup_{x \in \mathcal{X}} |p_h(x) - p(x)|^2 \lesssim h^{2\beta}.$$

For the variance term, write

$$\hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} Z_i(x), \qquad Z_i(x) := \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Since the $Z_i(x)$ are iid,

$$\operatorname{Var}(\hat{p}_h(x)) = \frac{1}{n} \operatorname{Var}(Z_1(x)) \le \frac{1}{n} \mathbb{E}[Z_1(x)^2].$$

Now

$$\mathbb{E}[Z_1(x)^2] = \frac{1}{h^{2d}} \int K\left(\frac{x - u}{h}\right)^2 p(u) \, du.$$

Therefore

$$\int_{\mathcal{X}} \operatorname{Var}(\hat{p}_h(x)) \, dx \le \frac{1}{nh^{2d}} \int_{\mathcal{X}} \int K\left(\frac{x - u}{h}\right)^2 p(u) \, du \, dx.$$

By Fubini,

$$\int_{\mathcal{X}} \mathrm{Var}(\hat{p}_h(x))\,dx \le \frac{1}{nh^{2d}} \int p(u) \left[\int_{\mathcal{X}} K\left(\frac{x-u}{h}\right)^2 dx \right] du.$$

Using the change of variables $v = (x-u)/h$, so that $dx = h^d\,dv$, we obtain

$$\int_{\mathcal{X}} K\left(\frac{x-u}{h}\right)^2 dx \le h^d \int_{\mathbb{R}^d} K(v)^2\,dv.$$

Hence

$$\int_{\mathcal{X}} \mathrm{Var}(\hat{p}_h(x))\,dx \le \frac{1}{nh^d} \left(\int K(v)^2\,dv \right) \left(\int p(u)\,du \right) \lesssim \frac{1}{nh^d}.$$

Combining the bias and variance bounds yields

$$\mathbb{E} \int_{\mathcal{X}} (\hat{p}_h(x) - p(x))^2\,dx \lesssim h^{2\beta} + \frac{1}{nh^d}.$$

To optimize the rate, balance the two terms:

$$h^{2\beta} \asymp \frac{1}{nh^d},$$

which implies

$$h \asymp n^{-1/(2\beta+d)}.$$

Substituting this choice gives

$$\mathbb{E} \int_{\mathcal{X}} (\hat{p}_h(x) - p(x))^2\,dx \lesssim n^{-2\beta/(2\beta+d)}.$$

□

2.2 Minimax Bound

A fundamental way to evaluate performance in mathematical statistics and statistical decision theory is the minimax risk. This is the best performance that any method can achieve in the worst case over a specified problem class.

According to the next theorem, there does not exist an estimator that converges faster than $O(n^{-2\beta/(2\beta+d)})$. We state the result for integrated L_2 loss although similar results hold for other loss functions and other function spaces.

Theorem 2.6 (Minimax lower bound for density estimation over Hölder classes)**.** *Assume $\beta > 0$ is noninteger, and let $\mathcal{P}(\beta, L)$ denote the class of probability densities p on $[0,1]^d$ such that $p \in \Sigma(\beta, L)$. Then there exists a constant $c > 0$, depending only on β, L, d, such that*

$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}(\beta,L)} \mathbb{E}_p \|\hat{p} - p\|_2^2 \;\ge\; c\, n^{-2\beta/(2\beta+d)}.$$

Here the infimum is over all estimators $\hat{p} = \hat{p}(X_1, \ldots, X_n)$ based on iid observations $X_1, \ldots, X_n \sim p$, and

$$\|f\|_2^2 := \int_{[0,1]^d} f(x)^2\,dx.$$

Proof. We construct a finite family of densities inside $\mathcal{P}(\beta, L)$ and reduce estimation to a testing problem.

Step 1: Construction of localized perturbations.

Choose a nonzero function $\psi \in C_c^\infty((0,1)^d)$ such that

$$\int_{[0,1]^d} \psi(x)\, dx = 0.$$

Since $\psi \in C_c^\infty((0,1)^d)$, there exists $\delta \in (0, 1/2)$ such that

$$\operatorname{supp}(\psi) \subset [\delta, 1-\delta]^d.$$

Let $m \in \mathbb{N}$ and define

$$h := \frac{1}{m}.$$

Partition $[0,1]^d$ into the $M = m^d = h^{-d}$ disjoint cubes

$$Q_j = t_j + h[0,1]^d, \qquad j = 1, \dots, M,$$

where t_j denotes the lower-left corner of Q_j.

For each j, define

$$\psi_j(x) := \begin{cases} h^\beta \psi\left(\dfrac{x - t_j}{h}\right), & x \in Q_j, \\ 0, & x \notin Q_j. \end{cases}$$

Because $\operatorname{supp}(\psi) \subset [\delta, 1-\delta]^d$, we have

$$\operatorname{supp}(\psi_j) \subset t_j + h[\delta, 1-\delta]^d,$$

so each ψ_j is supported strictly inside the interior of Q_j. In particular, each ψ_j is C^∞ on all of $[0,1]^d$, and the supports of the ψ_j are pairwise disjoint.

Step 2: Definition of the hypercube family.

For $\theta = (\theta_1, \dots, \theta_M) \in \{-1, +1\}^M$, define

$$p_\theta(x) := 1 + \varepsilon \sum_{j=1}^{M} \theta_j \psi_j(x),$$

where $\varepsilon > 0$ is a constant to be chosen later.

We verify that, for sufficiently small ε, all p_θ belong to $\mathcal{P}(\beta, L)$.

(a) Normalization. Since ψ_j is supported in Q_j,

$$\int_{[0,1]^d} \psi_j(x)\, dx = h^\beta \int_{Q_j} \psi\left(\frac{x - t_j}{h}\right) dx.$$

With the change of variables $u = (x - t_j)/h$, so that $dx = h^d\, du$, this becomes

$$\int_{[0,1]^d} \psi_j(x)\, dx = h^{\beta+d} \int_{[0,1]^d} \psi(u)\, du = 0.$$

Therefore

$$\int_{[0,1]^d} p_\theta(x)\,dx = \int_{[0,1]^d} 1\,dx + \varepsilon \sum_{j=1}^{M} \theta_j \int_{[0,1]^d} \psi_j(x)\,dx = 1.$$

(b) Nonnegativity. Since the supports of the ψ_j are disjoint, at each point x at most one summand is nonzero. Hence

$$\left| \sum_{j=1}^{M} \theta_j \psi_j(x) \right| \le \max_{1 \le j \le M} |\psi_j(x)| \le h^\beta \|\psi\|_\infty.$$

Thus

$$p_\theta(x) \ge 1 - \varepsilon h^\beta \|\psi\|_\infty.$$

Since $h \le 1$, it suffices to choose

$$\varepsilon \le \frac{1}{2\|\psi\|_\infty}.$$

Then $p_\theta(x) \ge 1/2$ for all x, so each p_θ is nonnegative.

(c) Membership in the Hölder class. Write

$$k = \lfloor \beta \rfloor, \qquad \alpha = \beta - k \in (0,1).$$

We first show that each ψ_j has Hölder norm bounded uniformly in j and h.

Let s be a multi-index with $|s| \le k$. On Q_j,

$$D^s \psi_j(x) = h^{\beta - |s|} (D^s \psi)\left(\frac{x - t_j}{h} \right),$$

and outside Q_j we have $D^s \psi_j(x) = 0$. Therefore

$$\|D^s \psi_j\|_\infty \le h^{\beta - |s|} \|D^s \psi\|_\infty \le \|D^s \psi\|_\infty,$$

because $\beta - |s| \ge 0$ and $h \le 1$.

Now let $|s| = k$. We bound the α-Hölder seminorm of $D^s \psi_j$.

If $x, y \in Q_j$, then

$$|D^s \psi_j(x) - D^s \psi_j(y)| = h^{\beta - k} \left| (D^s \psi)\left(\frac{x - t_j}{h} \right) - (D^s \psi)\left(\frac{y - t_j}{h} \right) \right|.$$

Since $D^s \psi$ is α-Hölder continuous,

$$|D^s \psi_j(x) - D^s \psi_j(y)| \le h^{\beta - k} [D^s \psi]_\alpha \left\| \frac{x - y}{h} \right\|^\alpha = [D^s \psi]_\alpha \|x - y\|^\alpha.$$

If one of x, y lies in $\operatorname{supp}(\psi_j)$ and the other lies outside Q_j, then

$$\|x - y\| \ge \delta h.$$

Also,

$$\|D^s \psi_j\|_\infty \le C h^{\beta - k}$$

for a constant C depending only on ψ and β. Hence

$$\frac{|D^s \psi_j(x) - D^s \psi_j(y)|}{\|x - y\|^\alpha} \le \frac{2C h^{\beta - k}}{(\delta h)^\alpha} = \frac{2C}{\delta^\alpha},$$

because $\alpha = \beta - k$. If both x and y lie outside $\mathrm{supp}(\psi_j)$, then the numerator is zero. It follows that

$$[D^s \psi_j]_\alpha \le C_\psi$$

for some constant C_ψ independent of j and h. Therefore

$$\|\psi_j\|_{\Sigma(\beta)} \le C_\psi \qquad \text{for all } j.$$

Because the supports of the ψ_j are pairwise disjoint and separated at scale h, the same uniform bound holds for the sum:

$$\left\| \sum_{j=1}^{M} \theta_j \psi_j \right\|_{\Sigma(\beta)} \le C_\psi.$$

Hence

$$\|p_\theta\|_{\Sigma(\beta)} \le \|1\|_{\Sigma(\beta)} + \varepsilon C_\psi.$$

By choosing $\varepsilon > 0$ sufficiently small, depending only on (β, L, ψ), we ensure

$$p_\theta \in \Sigma(\beta, L) \qquad \text{for all } \theta.$$

Thus the whole family $\{p_\theta : \theta \in \{-1, +1\}^M\}$ is contained in $\mathcal{P}(\beta, L)$.

Step 3: L_2-separation of neighboring densities.

Fix $\theta \in \{-1, +1\}^M$ and $j \in \{1, \dots, M\}$. Let $\theta^{(j)}$ denote the vector obtained from θ by flipping only the j-th coordinate:

$$\theta_i^{(j)} = \begin{cases} -\theta_j, & i = j, \\ \theta_i, & i \ne j. \end{cases}$$

Then

$$p_\theta - p_{\theta^{(j)}} = 2\varepsilon \theta_j \psi_j,$$

and therefore

$$\|p_\theta - p_{\theta^{(j)}}\|_2^2 = 4\varepsilon^2 \|\psi_j\|_2^2.$$

Using again the change of variables $u = (x - t_j)/h$,

$$\|\psi_j\|_2^2 = \int_{[0,1]^d} \psi_j(x)^2 \, dx = h^{2\beta} \int_{Q_j} \psi\left(\frac{x - t_j}{h}\right)^2 dx = h^{2\beta + d} \|\psi\|_2^2.$$

Thus

$$\|p_\theta - p_{\theta^{(j)}}\|_2^2 = 4\varepsilon^2 h^{2\beta + d} \|\psi\|_2^2.$$

Step 4: Kullback–Leibler divergence between neighboring models.

Let P_θ denote the probability measure on $[0, 1]^d$ with density p_θ, and let P_θ^n denote the joint law of n iid observations from p_θ.

We claim that

$$\mathrm{KL}(P_\theta, P_{\theta^{(j)}}) \le \int \frac{(p_\theta - p_{\theta^{(j)}})^2}{p_{\theta^{(j)}}} \, dx.$$

Indeed, using $\log u \le u - 1$ for $u > 0$,

$$p_\theta \log \frac{p_\theta}{p_{\theta^{(j)}}} \le p_\theta \left(\frac{p_\theta}{p_{\theta^{(j)}}} - 1 \right) = \frac{p_\theta (p_\theta - p_{\theta^{(j)}})}{p_{\theta^{(j)}}}.$$

Write

$$p_\theta(p_\theta - p_{\theta^{(j)}}) = (p_\theta - p_{\theta^{(j)}})^2 + p_{\theta^{(j)}}(p_\theta - p_{\theta^{(j)}}).$$

After dividing by $p_{\theta^{(j)}}$ and integrating, the second term vanishes because both densities integrate to one. Hence

$$\mathrm{KL}(P_\theta, P_{\theta^{(j)}}) \le \int \frac{(p_\theta - p_{\theta^{(j)}})^2}{p_{\theta^{(j)}}}\, dx.$$

Since $p_{\theta^{(j)}} \ge 1/2$, it follows that

$$\mathrm{KL}(P_\theta, P_{\theta^{(j)}}) \le 2\|p_\theta - p_{\theta^{(j)}}\|_2^2 \le 8\varepsilon^2 h^{2\beta+d}\|\psi\|_2^2.$$

For product measures,

$$\mathrm{KL}(P_\theta^n, P_{\theta^{(j)}}^n) = n\,\mathrm{KL}(P_\theta, P_{\theta^{(j)}}),$$

so

$$\mathrm{KL}(P_\theta^n, P_{\theta^{(j)}}^n) \le 8n\varepsilon^2 h^{2\beta+d}\|\psi\|_2^2.$$

Now choose

$$m = \left\lfloor c_1^{-1} n^{1/(2\beta+d)} \right\rfloor, \qquad h = \frac{1}{m},$$

with $c_1 > 0$ sufficiently small. Then

$$h \asymp n^{-1/(2\beta+d)}, \qquad nh^{2\beta+d} \le C c_1^{2\beta+d}$$

for a universal constant C. Therefore, by choosing c_1 small enough, we may ensure that

$$\mathrm{KL}(P_\theta^n, P_{\theta^{(j)}}^n) \le \kappa$$

for some fixed constant $\kappa > 0$ as small as we wish, uniformly in θ and j.

By Pinsker's inequality,

$$\mathrm{TV}(P_\theta^n, P_{\theta^{(j)}}^n) \le \sqrt{\frac{1}{2}\mathrm{KL}(P_\theta^n, P_{\theta^{(j)}}^n)} \le \rho$$

for some constant $\rho < 1$, uniformly in θ and j.

Step 5: Reduction from estimation to sign recovery.

Let $\hat{p} = \hat{p}(X_1, \dots, X_n)$ be any estimator. For each j, define

$$\hat{\theta}_j := \mathrm{sign}\langle \hat{p} - 1, \psi_j \rangle,$$

with an arbitrary convention if the inner product is zero.

Since the supports of the ψ_j are disjoint,

$$\langle p_\theta - 1, \psi_j \rangle = \left\langle \varepsilon \sum_{i=1}^{M} \theta_i \psi_i, \psi_j \right\rangle = \varepsilon \theta_j \|\psi_j\|_2^2.$$

If $\hat{\theta}_j \ne \theta_j$, then $\langle \hat{p} - 1, \psi_j \rangle$ has the opposite sign, so

$$|\langle \hat{p} - p_\theta, \psi_j \rangle| = \left|\langle \hat{p} - 1, \psi_j \rangle - \varepsilon\theta_j \|\psi_j\|_2^2\right| \ge \varepsilon \|\psi_j\|_2^2.$$

Now use orthogonality. Since the supports of the ψ_j are disjoint,

$$\langle \psi_i, \psi_j \rangle = 0 \qquad (i \ne j).$$

Since $\{\psi_j/\|\psi_j\|_2\}_{j=1}^M$ is an orthonormal family in L^2, Bessel's inequality implies that the orthogonal projection of $\hat{p} - p_\theta$ onto $\text{span}\{\psi_1, \dots, \psi_M\}$ has squared norm

$$\sum_{j=1}^{M} \frac{\langle \hat{p} - p_\theta, \psi_j \rangle^2}{\|\psi_j\|_2^2},$$

and hence

$$\|\hat{p} - p_\theta\|_2^2 \geq \sum_{j=1}^{M} \frac{\langle \hat{p} - p_\theta, \psi_j \rangle^2}{\|\psi_j\|_2^2}.$$

On the event $\{\hat{\theta}_j \neq \theta_j\}$ we have

$$|\langle \hat{p} - p_\theta, \psi_j \rangle| \geq \varepsilon \|\psi_j\|_2^2,$$

so

$$\frac{\langle \hat{p} - p_\theta, \psi_j \rangle^2}{\|\psi_j\|_2^2} \geq \varepsilon^2 \|\psi_j\|_2^2 \, \mathbf{1}\{\hat{\theta}_j \neq \theta_j\}.$$

Summing over j yields

$$\|\hat{p} - p_\theta\|_2^2 \geq \varepsilon^2 \sum_{j=1}^{M} \|\psi_j\|_2^2 \mathbf{1}\{\hat{\theta}_j \neq \theta_j\}.$$

Since all $\|\psi_j\|_2^2 = h^{2\beta+d}\|\psi\|_2^2$, we obtain

$$\|\hat{p} - p_\theta\|_2^2 \geq \varepsilon^2 h^{2\beta+d} \|\psi\|_2^2 \sum_{j=1}^{M} \mathbf{1}\{\hat{\theta}_j \neq \theta_j\}.$$

Taking expectations under P_θ^n,

$$\mathbb{E}_\theta \|\hat{p} - p_\theta\|_2^2 \geq \varepsilon^2 h^{2\beta+d} \|\psi\|_2^2 \sum_{j=1}^{M} \mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j).$$

Step 6: Lower bound on the average sign error.

Fix j. For each θ, consider the paired vector $\theta^{(j)}$ obtained by flipping the j-th coordinate. The random variable $\hat{\theta}_j$ is a test for distinguishing P_θ^n from $P_{\theta^{(j)}}^n$. By the standard two-point testing bound,

$$\mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j) + \mathbb{P}_{\theta^{(j)}}(\hat{\theta}_j \neq \theta_j^{(j)}) \geq 1 - \text{TV}(P_\theta^n, P_{\theta^{(j)}}^n) \geq 1 - \rho.$$

Averaging over all $\theta \in \{-1, +1\}^M$, we get

$$\frac{1}{2^M} \sum_{\theta} \mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j) \geq \frac{1-\rho}{2} =: c_0 > 0.$$

Summing over $j = 1, \dots, M$,

$$\frac{1}{2^M} \sum_{\theta} \sum_{j=1}^{M} \mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j) \geq c_0 M.$$

Step 7: Final steps.

Average the lower bound from Step 5 over all θ:

$$\frac{1}{2^M}\sum_{\theta}\mathbb{E}_\theta\|\hat{p}-p_\theta\|_2^2 \geq \varepsilon^2 h^{2\beta+d}\|\psi\|_2^2 \cdot \frac{1}{2^M}\sum_{\theta}\sum_{j=1}^{M}\mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j).$$

Using the bound from Step 6,

$$\frac{1}{2^M}\sum_{\theta}\mathbb{E}_\theta\|\hat{p}-p_\theta\|_2^2 \geq c_0\varepsilon^2 h^{2\beta+d}\|\psi\|_2^2 M.$$

Since $M = h^{-d}$,

$$h^{2\beta+d}M = h^{2\beta}.$$

Therefore

$$\frac{1}{2^M}\sum_{\theta}\mathbb{E}_\theta\|\hat{p}-p_\theta\|_2^2 \geq c\,h^{2\beta}$$

for some constant $c > 0$ depending only on β, L, d.

Hence

$$\sup_{\theta\in\{-1,+1\}^M}\mathbb{E}_\theta\|\hat{p}-p_\theta\|_2^2 \geq c\,h^{2\beta}.$$

Since the family $\{p_\theta : \theta \in \{-1,+1\}^M\}$ is contained in $\mathcal{P}(\beta, L)$, it follows that

$$\inf_{\hat{p}}\sup_{p\in\mathcal{P}(\beta,L)}\mathbb{E}_p\|\hat{p}-p\|_2^2 \geq c\,h^{2\beta}.$$

Finally, since $h \asymp n^{-1/(2\beta+d)}$, we obtain

$$h^{2\beta} \asymp n^{-2\beta/(2\beta+d)}.$$

Therefore

$$\inf_{\hat{p}}\sup_{p\in\mathcal{P}(\beta,L)}\mathbb{E}_p\|\hat{p}-p\|_2^2 \geq c\,n^{-2\beta/(2\beta+d)}.$$

This completes the proof. □

The upper and lower bounds together show that the minimax rate over Hölder classes $\Sigma(\beta, L)$ is

$$n^{-2\beta/(2\beta+d)},$$

and kernel density estimators attain this rate under the stated kernel and smoothness assumptions.

2.3 Concentration Analysis of Kernel Density Estimator

Now we state a result which says how fast $\hat{p}_h(x)$ concentrates around $p(x)$. First, recall Bernstein's inequality (see Chapter 1.2.3 in [1]): Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with mean μ, $\mathrm{Var}(Y_i) \leq \sigma^2$ and $|Y_i| \leq M$ a.s.. Then, for $\epsilon \geq 0$,

$$P(|\bar{Y}-\mu| > \epsilon) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right). \tag{2}$$

Theorem 2.7. *Assume that the kernel K is bounded, so that $\|K\|_\infty < \infty$, and that*

$$\int K(u)^2\,du < \infty.$$

Then there exist constants $c, C > 0$ such that for every sufficiently small $\varepsilon > 0$,

$$\mathbb{P}\big(|\hat{p}_h(x) - p_h(x)| > \varepsilon\big) \leq 2\exp\big(-cnh^d\varepsilon^2\big).$$

Consequently, for every $\delta \in (0,1)$,

$$\sup_{p\in\Sigma(\beta,L)} \mathbb{P}\left(|\hat{p}_h(x) - p(x)| > \sqrt{\frac{C\log(2/\delta)}{nh^d}} + Ch^\beta \right) \le \delta.$$

In particular, if $h \asymp n^{-1/(2\beta+d)}$, then

$$\sup_{p\in\Sigma(\beta,L)} \mathbb{P}\Big(|\hat{p}_h(x) - p(x)| > Cn^{-\beta/(2\beta+d)}\sqrt{\log(2/\delta)} \Big) \le \delta,$$

after adjusting constants.

Proof. Write

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n Z_i, \qquad Z_i := \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Then

$$\mathbb{E}[Z_i] = p_h(x).$$

Moreover,

$$|Z_i| \le \frac{\|K\|_\infty}{h^d}.$$

Also,

$$\mathrm{Var}(Z_i) \le \mathbb{E}[Z_i^2] = \frac{1}{h^{2d}} \int K\left(\frac{x-u}{h}\right)^2 p(u)\,du.$$

By the change of variables $v = (x-u)/h$, this becomes

$$\mathbb{E}[Z_i^2] = \frac{1}{h^d}\int K(v)^2 p(x - hv)\,dv \lesssim \frac{1}{h^d},$$

so

$$\mathrm{Var}(Z_i) \lesssim h^{-d}.$$

Applying Bernstein's inequality to $n^{-1}\sum_{i=1}^n Z_i$, we obtain

$$\mathbb{P}\big(|\hat{p}_h(x) - p_h(x)| > \varepsilon\big) \le 2\exp\left(-\frac{n\varepsilon^2}{2c_1 h^{-d} + c_2 h^{-d}\varepsilon}\right)$$

for suitable constants $c_1, c_2 > 0$. Hence, for all sufficiently small ε,

$$\mathbb{P}\big(|\hat{p}_h(x) - p_h(x)| > \varepsilon\big) \le 2\exp(-cnh^d\varepsilon^2).$$

Now by the triangle inequality,

$$|\hat{p}_h(x) - p(x)| \le |\hat{p}_h(x) - p_h(x)| + |p_h(x) - p(x)|.$$

By Lemma 2.1,

$$|p_h(x) - p(x)| \lesssim h^\beta.$$

Therefore,

$$\mathbb{P}\Big(|\hat{p}_h(x) - p(x)| > \varepsilon + Ch^\beta\Big) \le \mathbb{P}\big(|\hat{p}_h(x) - p_h(x)| > \varepsilon\big) \le 2\exp(-cnh^d\varepsilon^2).$$

Choosing

$$\varepsilon = \sqrt{\frac{C\log(2/\delta)}{nh^d}}$$

with C large enough yields

$$\mathbb{P}\left(|\hat{p}_h(x) - p(x)| > \sqrt{\frac{C\log(2/\delta)}{nh^d}} + Ch^\beta\right) \leq \delta.$$

This proves the second claim.

Finally, if $h \asymp n^{-1/(2\beta+d)}$, then

$$h^\beta \asymp n^{-\beta/(2\beta+d)}, \qquad (nh^d)^{-1/2} \asymp n^{-\beta/(2\beta+d)},$$

so both terms have the same order. □

2.4 Uniform Concentration in Supremum Norm

We now strengthen the pointwise concentration result to a uniform one. Our goal is to control

$$\|\hat{p}_h - p_h\|_\infty := \sup_{x\in\mathcal{X}} |\hat{p}_h(x) - p_h(x)|.$$

Combining such a bound with the bias estimate will then give a uniform bound on $\|\hat{p}_h - p\|_\infty$.

For this subsection, assume that $\mathcal{X} \subset \mathbb{R}^d$ is compact, that the density p is supported on $\mathcal{X}$ and bounded, and that the kernel K is bounded, Lipschitz, and compactly supported. Recall that

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right), \qquad p_h(x) = \mathbb{E}[\hat{p}_h(x)].$$

Theorem 2.8 (Uniform concentration for the kernel density estimator)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is compact, that p is supported on $\mathcal{X}$ and satisfies $\|p\|_\infty < \infty$, and that K is bounded, Lipschitz, and compactly supported. Then there exist constants $c_1, c_2, c_3 > 0$, depending only on K, $\|p\|_\infty$, $\mathcal{X}$, and d, such that for every $h > 0$ and every $\epsilon > 0$ satisfying*

$$0 < \epsilon \leq c_3 h^{-d},$$

we have

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \leq c_1 \left(\frac{1}{h^{d+1}\epsilon}\right)^d \exp\left(-c_2 n h^d \epsilon^2\right). \tag{3}$$

Proof. We divide the proof into three steps.

Step 1: A pointwise Bernstein bound.

Fix $x \in \mathcal{X}$ and define

$$Z_i(x) := \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Then

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n Z_i(x), \qquad p_h(x) = \mathbb{E}[Z_i(x)].$$

Since K is bounded,

$$|Z_i(x)| \leq \frac{\|K\|_\infty}{h^d}.$$

Also,

$$\mathrm{Var}(Z_i(x)) \le \mathbb{E}[Z_i(x)^2] = \frac{1}{h^{2d}} \int K^2\left(\frac{x-u}{h}\right) p(u)\, du.$$

With the change of variables $v = (x-u)/h$, so that $du = h^d\, dv$, we get

$$\mathbb{E}[Z_i(x)^2] = \frac{1}{h^d} \int K^2(v)\, p(x-hv)\, dv \le \frac{\|p\|_\infty}{h^d} \int K^2(v)\, dv.$$

Hence there exists a constant $C_0 > 0$ such that

$$\mathrm{Var}(Z_i(x)) \le \frac{C_0}{h^d} \qquad \text{for all } x \in \mathcal{X}.$$

Now set

$$Y_i(x) := Z_i(x) - \mathbb{E}[Z_i(x)].$$

Then

$$\hat{p}_h(x) - p_h(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x),$$

and

$$|Y_i(x)| \le |Z_i(x)| + |\mathbb{E}[Z_i(x)]| \le \frac{2\|K\|_\infty}{h^d},$$

since $|\mathbb{E}[Z_i(x)]| \le \mathbb{E}|Z_i(x)| \le \|K\|_\infty h^{-d}$. Moreover,

$$\mathrm{Var}(Y_i(x)) = \mathrm{Var}(Z_i(x)) \le \frac{C_0}{h^d}.$$

Applying Bernstein's inequality to the iid centered variables $Y_i(x)$ yields

$$P(|\hat{p}_h(x) - p_h(x)| > \epsilon) \le 2\exp\left(-\frac{n\epsilon^2}{2(C_0/h^d) + (4\|K\|_\infty \epsilon)/(3h^d)}\right).$$

Therefore, if $0 < \epsilon \le c_3 h^{-d}$ with $c_3 > 0$ sufficiently small, then the denominator is bounded by a constant multiple of h^{-d}, and so

$$P(|\hat{p}_h(x) - p_h(x)| > \epsilon) \le 2\exp(-c_2 n h^d \epsilon^2) \tag{4}$$

for some constant $c_2 > 0$ independent of x, h, n, and ϵ.

Step 2: Reduction to a finite net.

Because $\mathcal{X}$ is compact, for every $\eta > 0$ there exists an η-net $x_1, \dots, x_N \in \mathcal{X}$ such that

$$\mathcal{X} \subset \bigcup_{j=1}^{N} B(x_j, \eta), \qquad N \le C_{\mathcal{X}} \eta^{-d},$$

where $C_{\mathcal{X}} > 0$ depends only on $\mathcal{X}$ and d.

Since K is Lipschitz, there exists $L_K > 0$ such that

$$|K(u) - K(v)| \le L_K \|u - v\| \qquad \text{for all } u, v \in \mathbb{R}^d.$$

Hence for any $x, y \in \mathcal{X}$,

$$|\hat{p}_h(x) - \hat{p}_h(y)| \le \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} \left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{y - X_i}{h}\right) \right| \le \frac{L_K}{h^{d+1}} \|x - y\|.$$

Taking expectations gives the same bound for p_h:

$$|p_h(x) - p_h(y)| \le \frac{L_K}{h^{d+1}}\|x - y\|.$$

Therefore

$$|(\hat{p}_h - p_h)(x) - (\hat{p}_h - p_h)(y)| \le \frac{2L_K}{h^{d+1}}\|x - y\|.$$

Now choose

$$\eta := \frac{\epsilon h^{d+1}}{4L_K}.$$

Then for any $x \in \mathcal{X}$, if x_j satisfies $\|x - x_j\| \le \eta$, we have

$$|(\hat{p}_h - p_h)(x) - (\hat{p}_h - p_h)(x_j)| \le \frac{2L_K}{h^{d+1}}\eta = \frac{\epsilon}{2}.$$

Consequently,

$$\|\hat{p}_h - p_h\|_\infty > \epsilon \quad \Longrightarrow \quad \max_{1\le j\le N} |\hat{p}_h(x_j) - p_h(x_j)| > \frac{\epsilon}{2}.$$

Hence

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \le \sum_{j=1}^{N} P\Big(|\hat{p}_h(x_j) - p_h(x_j)| > \frac{\epsilon}{2}\Big). \tag{5}$$

Step 3: Union bound.

Applying the pointwise bound from Step 1 at each net point x_j gives

$$P\Big(|\hat{p}_h(x_j) - p_h(x_j)| > \frac{\epsilon}{2}\Big) \le 2\exp(-c_2' n h^d \epsilon^2)$$

for some constant $c_2' > 0$. Therefore

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \le 2N \exp(-c_2' n h^d \epsilon^2).$$

Since

$$N \le C_{\mathcal{X}} \eta^{-d} = C_{\mathcal{X}} \left(\frac{4L_K}{\epsilon h^{d+1}}\right)^d,$$

we obtain

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \le c_1 \left(\frac{1}{h^{d+1}\epsilon}\right)^d \exp(-c_2 n h^d \epsilon^2)$$

for suitable constants $c_1, c_2 > 0$. This proves the theorem. □

The theorem controls the stochastic fluctuation of the kernel density estimator uniformly over x. To translate this into a bound on $\|\hat{p}_h - p\|_\infty$, we combine it with the bias bound.

Corollary 2.9 (Uniform error bound)**.** *Assume in addition that $p \in \Sigma(\beta, L)$ and that the kernel satisfies the moment conditions used in the bias lemma. Then*

$$\|\hat{p}_h - p\|_\infty \le \|\hat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty \le \|\hat{p}_h - p_h\|_\infty + Ch^\beta,$$

where $C > 0$ depends only on β, L, K, and d. Consequently, if $h = h_n \to 0$ and

$$\frac{n h_n^d}{\log(1/h_n)} \to \infty,$$

then

$$\|\hat{p}_{h_n} - p\|_\infty = O_P\left(\sqrt{\frac{\log(1/h_n)}{nh_n^d}} + h_n^\beta\right).$$

Proof. The decomposition

$$\hat{p}_h - p = (\hat{p}_h - p_h) + (p_h - p)$$

implies

$$\|\hat{p}_h - p\|_\infty \le \|\hat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty.$$

The bias lemma gives

$$\|p_h - p\|_\infty \le Ch^\beta.$$

It therefore remains to bound $\|\hat{p}_h - p_h\|_\infty$.

Let

$$\epsilon_n := M\sqrt{\frac{\log(1/h_n)}{nh_n^d}},$$

where $M > 0$ is a constant to be chosen. Then

$$nh_n^d\epsilon_n^2 = M^2\log(1/h_n).$$

Applying the theorem,

$$P(\|\hat{p}_{h_n} - p_{h_n}\|_\infty > \epsilon_n) \le c_1\left(\frac{1}{h_n^{d+1}\epsilon_n}\right)^d \exp(-c_2M^2\log(1/h_n)).$$

Since

$$\exp(-c_2M^2\log(1/h_n)) = h_n^{c_2M^2},$$

the right-hand side is bounded by

$$c_1\left(\frac{1}{h_n^{d+1}\epsilon_n}\right)^d h_n^{c_2M^2}.$$

Also,

$$\frac{1}{h_n^{d+1}\epsilon_n} = \frac{\sqrt{nh_n^d}}{M\,h_n^{d+1}\sqrt{\log(1/h_n)}}.$$

By assumption,

$$\frac{nh_n^d}{\log(1/h_n)} \to \infty,$$

so the prefactor grows at most polynomially in $1/h_n$, whereas the factor $h_n^{c_2M^2}$ decays faster than any fixed polynomial if M is chosen large enough. Hence

$$P(\|\hat{p}_{h_n} - p_{h_n}\|_\infty > \epsilon_n) \to 0.$$

Thus

$$\|\hat{p}_{h_n} - p_{h_n}\|_\infty = O_P\left(\sqrt{\frac{\log(1/h_n)}{nh_n^d}}\right).$$

Combining this with the bias term yields

$$\|\hat{p}_{h_n} - p\|_\infty = O_P\left(\sqrt{\frac{\log(1/h_n)}{nh_n^d}} + h_n^\beta\right).$$

□

2.5 Giné-Guillou's Approach

An alternative approach is to replace Bernstein's inequality with a more sophisticated inequality due to Talagrand. We follow the analysis in Giné and Guillou (2002) [5]. Let

$$\mathcal{F} = \left\{ K\left(\frac{x-\cdot}{h}\right),\ x \in \mathbb{R}^d,\ h > 0 \right\}.$$

We assume there exist positive numbers A and v such that

$$\sup_P N(\mathcal{F}_h,\, L_2(P),\, \epsilon \|F\|_{L_2(P)}) \leq \left(\frac{A}{\epsilon}\right)^v, \tag{6}$$

where $N(T, d, \epsilon)$ denotes the ϵ-covering number of the metric space (T, d), F is the envelope function of $\mathcal{F}$ and the supremum is taken over the set of all probability measures on $\mathbb{R}^d$. The quantities A and v are called the VC characteristics of $\mathcal{F}_h$.

Theorem 2.10 (Giné and Guillou, 2002 [5])**.** *Assume that the kernel satisfies the above property.*

1. *Let $h > 0$ be fixed. Then, there exist constants $c_1 > 0$ and $c_2 > 0$ such that, for all small $\epsilon > 0$ and all large n,*

$$P\left(\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| > \epsilon \right) \leq c_1 \exp\left(-c_2 n h^d \epsilon^2\right). \tag{7}$$

2. *Let $h_n \to 0$ as $n \to \infty$ in such a way that $nh_n^d / |\log h_n^d| \to \infty$. Let*

$$\epsilon_n \geq \sqrt{\frac{|\log h_n|}{n h_n^d}}. \tag{8}$$

Then, for all n large enough, the tail bound in part (1) holds with h and ϵ replaced by h_n and ϵ_n, respectively.

We will use this result as a black box and do not prove it here; its proof requires empirical process theory.

Remark 2.11 (Idea of the proof)**.** The class of functions

$$\mathcal{F}_h = \left\{ u \mapsto \frac{1}{h^d} K\left(\frac{x-u}{h}\right) : x \in \mathbb{R}^d \right\}$$

is a translated kernel class. Under the stated regularity assumptions, this class has sufficiently small covering numbers uniformly in h. One then applies an exponential inequality for empirical process suprema over VC-type classes to obtain

$$\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| = \sup_{f \in \mathcal{F}_h} |(P_n - P)f|,$$

together with the tail bound

$$P\left(\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| > \epsilon \right) \leq c_1 \exp(-c_2 n h^d \epsilon^2).$$

The second part follows by substituting $h = h_n$ and $\epsilon = \epsilon_n$ and using the condition

$$\frac{nh_n^d}{|\log h_n|} \to \infty.$$

The above theorem imposes minimal assumptions on the kernel K and, more importantly, on the probability distribution P, whose density is not required to be bounded or smooth, and, in fact, may not even exist. Combining the above theorem with the bias bound gives the following result.

Theorem 2.12. *Suppose that $p \in \Sigma(\beta, L)$. Fix any $\delta > 0$. Then*

$$P\left(\sup_x |\hat{p}(x) - p(x)| > \sqrt{\frac{C \log n}{nh^d}} + ch^\beta\right) < \delta$$

for some constants C and c where C depends on δ. Choosing $h \asymp (\log n/n)^{1/(2\beta+d)}$ we have

$$P\left(\sup_x |\hat{p}_h(x) - p(x)|^2 > \frac{C \log n}{n^{2\beta/(2\beta+d)}}\right) < \delta.$$

Remark 2.13 (Boundary Bias)**.** We have ignored what happens near the boundary of the sample space. If x is $O(h)$ close to the boundary, the bias is $O(h)$ instead of $O(h^2)$. There are a variety of fixes including: data reflection, transformations, boundary kernels, local likelihood.

3 High Dimensions

The rate of convergence $n^{-2\beta/(2\beta+d)}$ is slow when the dimension d is large. In this case it is hopeless to try to estimate the true density p precisely in the L_2 norm (or any similar norm). We need to change our notion of what it means to estimate p in a high-dimensional problem. Instead of estimating p precisely we have to settle for finding an adequate approximation of p. Any estimator that finds the regions where p puts large amounts of mass should be considered an adequate approximation. Let us consider a few ways to implement this type of thinking.

Biased Density Estimation. Let

$$p_h(x) = \mathbb{E}[\hat{p}_h(x)] = \int \frac{1}{h^d} K\left(\frac{x-u}{h}\right) p(u)\, du.$$

Thus the mean of $\hat{p}_h$ can be viewed as a smoothed version of p. Let P_h denote the probability distribution with density p_h. Then

$$P_h = P \star K_h,$$

where $\star$ denotes convolution and K_h is the distribution with density

$$u \mapsto \frac{1}{h^d} K\left(\frac{u}{h}\right).$$

Equivalently, if $Y \sim P$ and $Z \sim K_h$ are independent, then $Y + Z \sim P_h$. In this sense, P_h is a blurred or smoothed version of P.

Rather than trying to estimate p itself, one may instead fix $h > 0$ and regard p_h as the target. This corresponds to accepting the smoothing bias and estimating the regularized density p_h. For fixed h, the stochastic error decays much faster than in the usual nonparametric regime in which $h \to 0$.

Theorem 3.1 (Fixed-bandwidth estimation of the smoothed density)**.** *Fix $h > 0$. Under the assumptions of Theorem 2.8, there exist constants $C_h, c_h > 0$, depending on h, R, d, K, and p, such that for all sufficiently small $\epsilon > 0$,*

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \leq C_h\, \epsilon^{-d} \exp(-c_h n \epsilon^2).$$

Consequently,

$$\|\hat{p}_h - p_h\|_\infty = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Thus, when the target is the smoothed density p_h with fixed bandwidth h, the stochastic error has a parametric-type rate in n. The dependence on dimension enters through the constants, not through the exponent of n.

Proof. This follows directly from Theorem 2.8. When $h > 0$ is fixed, all factors involving h can be absorbed into the constants, so there exist constants $C_h, c_h > 0$ such that

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon) \leq C_h\, \epsilon^{-d} \exp(-c_h n \epsilon^2)$$

for all sufficiently small $\epsilon > 0$.

Now let

$$\epsilon_n = M\sqrt{\frac{\log n}{n}},$$

where $M > 0$ is a constant. Then

$$n\epsilon_n^2 = M^2 \log n, \qquad \epsilon_n^{-d} = M^{-d}\left(\frac{n}{\log n}\right)^{d/2}.$$

Hence

$$P(\|\hat{p}_h - p_h\|_\infty > \epsilon_n) \leq C_h M^{-d}\left(\frac{n}{\log n}\right)^{d/2} n^{-c_h M^2}.$$

If M is chosen so that $c_h M^2 > d/2$, then the right-hand side converges to zero. Therefore

$$\|\hat{p}_h - p_h\|_\infty = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

□

The theorem shows that if one is willing to estimate the blurred density p_h rather than the original density p, then the stochastic difficulty is greatly reduced. The remaining question is how to choose the smoothing level h so that p_h still preserves the features of interest of p.

Independence Based Methods. If we can live with some bias, we can reduce the dimensionality by imposing some independence assumptions. The simplest example is to treat the components $(X_1, \ldots, X_d)$ as if they are independent. In that case

$$p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p_j(x_j)$$

and the problem is reduced to a set of one-dimensional density estimation problems.

4 Miscellanea

4.1 Robust Density Estimation under Wasserstein Contamination

An interesting problem concerns with robust density estimation, i.e., how well can we recover the true density based on some contaminated data. One way to formalize robustness is to allow the sample to be perturbed by a bounded Wasserstein contamination. Following the discussion in Section 2 of [4], suppose that we observe contaminated data $X_1', \dots, X_n' \sim \widetilde{p}$, while the target density is p. The contamination level is measured by a Wasserstein metric: for some $q \in [1, \infty]$, $r \geq 1$, and $\varepsilon > 0$, we assume that there is a coupling of $X_i \sim p$ and $X_i' \sim \widetilde{p}$ such that

$$\left(\mathbb{E}\|X_i' - X_i\|_q^r\right)^{1/r} \leq \varepsilon.$$

Equivalently, the contaminated distribution lies within Wasserstein radius ε of the true distribution. This models small but systematic perturbations of the whole sample, in contrast to the classical Huber contamination model where an arbitrary fraction of observations may be replaced by gross outliers.

If the target is pointwise estimation of $p(x_0)$ over an s-Hölder class in dimension d, then the minimax squared risk under Wasserstein contamination is the maximum of the classical nonparametric term and a contamination term.

Theorem 4.1 (Density estimation under Wasserstein contamination [2])**.** *Let $s > 0$, $L > 0$, $q \in [1, \infty]$, and $r \in [1, \infty)$. For pointwise estimation of an s-Hölder smooth density in dimension d from i.i.d. data under ε-bounded $W_{q,r}$ Wasserstein contamination, the minimax squared risk has rate*

$$\max\left\{n^{-2s/(2s+d)},\ \varepsilon^{2s/(s+1+d/r)}\right\}.$$

Moreover, this rate is achieved by kernel density estimation with a suitably smooth bounded kernel and bandwidth

$$h^\star \asymp \max\left\{n^{-1/(2s+d)},\ \varepsilon^{1/(s+1+d/r)}\right\}.$$

This result shows that Wasserstein contamination acts like an additional source of smoothing bias. When ε is below the threshold

$$\varepsilon \lesssim n^{-(s+1+d/r)/(2s+d)},$$

the classical rate $n^{-2s/(2s+d)}$ remains dominant. Above that threshold, the contamination term determines both the optimal risk and the optimal bandwidth. A notable feature is that kernel smoothing remains minimax optimal under this form of robust contamination, unlike in the classical Huber contamination model where median-type robustification is often needed.

4.2 Manifolds and Singularities

Sometimes a distribution is concentrated near a lower-dimensional set. This causes problems for density estimation. In fact the density, as we usually think of it, may not be defined. As a simple example, suppose P is supported on the unit circle in $\mathbb{R}^2$. The distribution P is singular with respect to Lebesgue measure μ. This means that there are sets A with $P(A) > 0$ even though $\mu(A) = 0$. Effectively, this means that the density is infinite. To see this, consider a point x on the circle. Let $B(x, \epsilon)$ be a ball of radius ϵ centered at x. Then

$$p(x) = \lim_{\epsilon \to 0} \frac{P(B(x,\epsilon))}{\mu(B(x,\epsilon))} \to \infty.$$

Note also that the L_2 loss does not make any sense. If you tried to use cross-validation, you would find that the estimated risk is minimized at $h = 0$.

A simple solution is to focus on estimating the smoothed density $p_h(x)$ which is well-defined for every $h > 0$. More sophisticated ideas are based on topological data analysis.

4.3 Choosing the Bandwidth

In practice, we need a data-based method for choosing the bandwidth h. To do this, we will need to estimate the risk of the estimator and minimize the estimated risk over h. Two cross-validation methods are leave one out and V-fold cross-validation (e.g., $V = 10$) [3].

Another method for selecting h which is sometimes used when p is thought to be very smooth is the plug-in method. The idea is to take the formula for the mean squared error (equation 31), insert a guess of p'' and then solve for the optimal bandwidth h. For example, if $d = 1$ and under the idealized assumption that p is a univariate Normal this yields $h^* = 1.06\sigma n^{-1/5}$. Usually, σ is estimated by $\min\{s, Q/1.34\}$ where s is the sample standard deviation and Q is the interquartile range.[1] This choice of h^* works well if the true density is very smooth and is called the Normal reference rule. Since we don't want to necessarily assume that p is very smooth, it is usually better to estimate h using cross-validation.

A generalization of the kernel method is to use adaptive kernels where one uses a different bandwidth $h(x)$ for each point x. One can also use a different bandwidth $h(x_i)$ for each data point. This makes the estimator more flexible and allows it to adapt to regions of varying smoothness. But now we have the very difficult task of choosing many bandwidths instead of just one.

References

[1] Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024.

[2] Patrick Chao and Edgar Dobriban. Statistical estimation under distribution shift: Wasserstein perturbations and minimax theory. *arXiv preprint arXiv:2308.01853*, 2023.

[3] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

[4] Edgar Dobriban. Solving a research problem in mathematical statistics with AI assistance. *arXiv preprint arXiv:2511.18828*, 2025.

[5] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

[6] Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.

[7] Alexandre B Tsybakov. Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pages 1–76. Springer, 2008.

[1] Recall that the interquartile range is the 75th percentile minus the 25th percentile. The reason for dividing by 1.34 is that $Q/1.34$ is a consistent estimate of σ if the data are from a $N(\mu, \sigma^2)$.