IEMS 402: Statistical Learning **Winter 2024-2025**

Lecture 3: Nonparametric Regression and Density Estimation

Lecturer: Yiping Lu *Scribes: Baicen Liu*

3.1 Introduction

3.1.1 Basic setup

For a random pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, the function

$$m_0(x) = \mathbb{E}(Y \mid X = x)$$

is known as the regression function of Y given X. The main objective in nonparametric regression is to construct a predictor for Y based on X, which amounts to estimating m_0 using i.i.d. samples $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$. When a new input X is observed, we use $\hat{m}(X)$ to predict Y. The variable X is often referred to as the input, feature, or predictor, while Y may be called the output, response, or outcome.

For i.i.d. observations (X_i, Y_i), $i = 1, \ldots, n$, we can always express the model as

$$Y_i = m_0(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the noise terms ε_i are i.i.d. random variables with mean zero. This allows us to think of the data as independent draws from a joint distribution over (X_i, ε_i), with $\mathbb{E}(\varepsilon_i) = 0$, and the corresponding Y_i generated according to the model above. Unlike the parametric setting, where the regression function m_0 is fully specified by a finite number of parameters, nonparametric models provide greater flexibility in capturing the shape of the regression function.

We define an empirical norm $\|\cdot\|_n$ using the training inputs X_i, $i = 1, \ldots, n$, applied to functions $m : \mathbb{R}^d \to \mathbb{R}$, as follows:

$$\|m\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} m^2(X_i).$$

This definition is valid whether the inputs X_i are considered fixed or random; in the latter case, the norm itself becomes a random quantity.

When we treat the inputs as random variables, we denote the distribution of X by P_X, and define the L_2 norm $\|\cdot\|_2$ with respect to this distribution. For a function $m : \mathbb{R}^d \to \mathbb{R}$, the L_2 norm is given by:

$$\|m\|_2^2 = \mathbb{E}[m^2(X)] = \int m^2(x)\, dP_X(x).$$

Thus, whenever the notation $\|\cdot\|_2$ appears, it indicates that X is being viewed as a random variable.

An important quantity to assess is the squared error of an estimator $\hat{m}$ for m_0, which can be measured under either norm:

$$\|\hat{m} - m_0\|_n^2 \quad \text{or} \quad \|\hat{m} - m_0\|_2^2.$$

In both scenarios, the error is random because $\hat{m}$ depends on the data. Our focus will be on analyzing such errors either through their expected value or through probabilistic bounds. The expected value of these squared errors—particularly when using the L_2 norm—is formally referred to as the *risk*. However, in practice, we may use the term "error" somewhat informally to refer to this quantity.

3.1.2 Bias-Variance Tradeoff

For a new observation (X, Y), the expected squared prediction error satisfies

$$\mathbb{E}(Y - \hat{m}(X))^2 = \int b_n^2(x)\, dP(x) + \int v(x)\, dP(x) + \tau^2 = \|\hat{m} - m_0\|_2^2 + \tau^2,$$

where $b_n(x) = \mathbb{E}[\hat{m}(x)] - m_0(x)$ represents the bias, $v(x) = \mathrm{Var}(\hat{m}(x))$ denotes the variance, and $\tau^2 = \mathbb{E}(Y - m_0(X))^2$ accounts for the irreducible error. In practice, selecting tuning parameters requires careful consideration to properly balance the trade-off between bias and variance.

3.1.3 What does "nonparametric" mean?

A key feature of nonparametric regression is that we do not impose a specific parametric structure on the function m_0. This does not preclude us from approximating m_0 using, for example, a linear combination of spline basis functions, as in

$$\hat{m}(x) = \sum_{j=1}^{p} \hat{\beta}_j g_j(x).$$

This often leads to a natural question: since the coefficients $\beta_1, \ldots, \beta_p$ are parameters, doesn't this make the method parametric? The crucial distinction is that *we are not assuming that m_0 has a fixed parametric form*; in other words, we are not presuming that m_0 is exactly expressible as a linear combination of the basis functions $g_1, \ldots, g_p$. The flexibility in the choice and number of basis functions allows the model to adapt to the structure of the data, which is the essence of nonparametric modeling.

3.2 k-nearest-neighbors regression

Here is a simple and widely used technique to begin with: *k-nearest neighbors (k-NN) regression*. We fix an integer $k \geq 1$, and define the estimator as

$$\hat{m}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} Y_i, \tag{1}$$

where $\mathcal{N}_k(x)$ denotes the set of indices corresponding to the k closest training inputs $X_1, \ldots, X_n$ to the point x.

The k-NN method is a classic example of what is known as *instance-based learning*, also referred to as *memory-based* or *lazy* learning. These methods do not involve fitting a fixed model to the data; instead, they retain the training data in full. To make a prediction at a new point x, they search for the k most similar instances and average their corresponding output values.

This approach is widely used across various domains, largely due to its simplicity. By adjusting the number of neighbors k, we can control the flexibility of the estimator $\hat{m}$. A smaller k typically yields a more flexible (but potentially noisy) estimate, while a larger k results in smoother, less flexible estimates. One clear

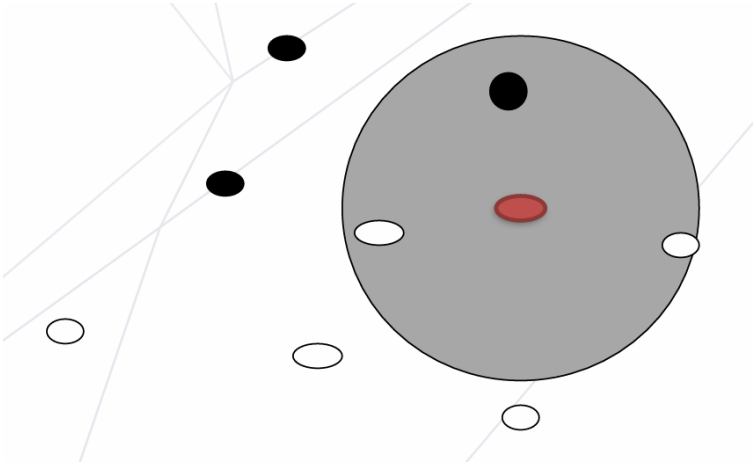

Figure 3.1: Illustration of 3-nearest neighbors.

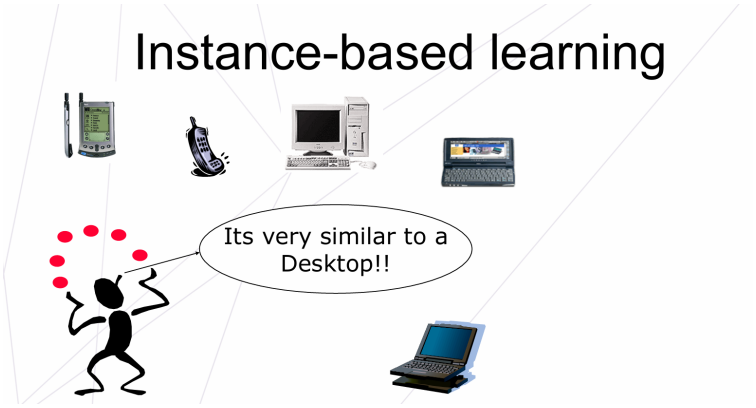

Figure 3.2: Real applications with instance-based learning.

limitation of the method is that the resulting function $\hat{m}$ often appears jagged, particularly when k is small or moderate. To understand why, it's helpful to rewrite the estimator as

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i, \tag{2}$$

where the weights $w_i(x)$, $i = 1, \ldots, n$, are defined as

$$w_i(x) = \begin{cases} 1/k & \text{if } X_i \text{ is among the } k \text{ nearest neighbors of } x, \\ 0 & \text{otherwise.} \end{cases}$$

Because each weight function $w_i(x)$ changes abruptly as x moves, the function $\hat{m}(x)$ is inherently discontinuous.

This form (2) also shows that the k-NN estimator belongs to a broader class known as *linear smoothers*. Specifically, if we define $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ as the vector of training responses, and $\hat{\mu} = (\hat{m}(X_1), \ldots, \hat{m}(X_n)) \in \mathbb{R}^n$ as the vector of fitted values at the training points, then we can write

$$\hat{\mu} = SY,$$

for some matrix S. Note that this linearity is in the responses Y_i, for fixed inputs X_i, and does not imply that $\hat{m}(x)$ is a linear function of x. Many estimators fall into the linear smoother framework, as we will see in later sections.

Importantly, the k-nearest-neighbors method is *universally consistent*, meaning that the expected squared L_2 error vanishes asymptotically:

$$\mathbb{E}\|\hat{m} - m_0\|_2^2 \to 0 \quad \text{as } n \to \infty,$$

assuming only that $\mathbb{E}(Y^2) < \infty$, provided that $k = k_n$ satisfies $k_n \to \infty$ and $k_n/n \to 0$. For instance, choosing $k = \sqrt{n}$ is sufficient. See Chapter 6.2 of [1] for more details.

Furthermore, assuming the underlying regression function m_0 is Lipschitz continuous, the k-nearest-neighbors estimate with $k \asymp n^{2/(2+d)}$ satisfies

$$\mathbb{E}\|\hat{m} - m_0\|_2^2 \lesssim n^{-2/(2+d)}. \tag{3}$$

Proof: Assume that $\mathrm{Var}(Y \mid X = x) = \sigma^2$, a constant, for simplicity, and fix (condition on) the training points. Using the bias-variance tradeoff,

$$\mathbb{E}[(\hat{m}(x) - m_0(x))^2] = \underbrace{(\mathbb{E}[\hat{m}(x)] - m_0(x))^2}_{\mathrm{Bias}^2(\hat{m}(x))} + \underbrace{\mathbb{E}[(\hat{m}(x) - \mathbb{E}[\hat{m}(x)])^2]}_{\mathrm{Var}(\hat{m}(x))},$$

we have

$$\mathbb{E}[(\hat{m}(x) - m_0(x))^2] = \left(\frac{1}{k}\sum_{i \in \mathcal{N}_k(x)} (m_0(X_i) - m_0(x))\right)^2 + \frac{\sigma^2}{k}.$$

Using the Lipschitz property $|m_0(x) - m_0(z)| \leq L\|x - z\|_2$ for some constant $L > 0$, we get

$$\mathbb{E}[(\hat{m}(x) - m_0(x))^2] \leq \left(\frac{L}{k}\sum_{i \in \mathcal{N}_k(x)} \|X_i - x\|_2\right)^2 + \frac{\sigma^2}{k}.$$

For "most" of the points we'll have $\|X_i - x\|_2 \leq C(k/n)^{1/d}$, for a constant C (think of having input points X_i, $i = 1, \ldots, n$, spaced equally over (say) $[0,1]^d$). Then our bias-variance upper bound becomes

$$(CL)^2 \left(\frac{k}{n}\right)^{\frac{2}{d}} + \frac{\sigma^2}{k}.$$

We can minimize this by balancing the two terms so that they are equal, giving

$$k^{1+\frac{2}{d}} \asymp n^{\frac{2}{d}},$$

i.e.,

$$k \asymp n^{\frac{2}{2+d}}$$

as claimed. Plugging this in gives the error bound of

$$n^{-\frac{2}{2+d}},$$

as claimed. ■

In short, as k increases, we would expect higher bias (since points in a valid neighbor would become less similar) and lower variance (since more data would average out noises).

3.2.1 Curse of dimensionality

It is important to observe that the error rate $n^{-\frac{2}{2+d}}$ demonstrates a significant sensitivity to the dimensionality d. To better understand this, suppose we fix a desired error level ε, and ask how large the sample size n must be to ensure that

$$n^{-\frac{2}{2+d}} \leq \varepsilon.$$

Rearranging this inequality gives

$$n \geq \varepsilon^{-\frac{2+d}{2}}.$$

This expression highlights that as the dimension d increases, the number of samples required to achieve an error below ε grows exponentially. Figure 3.3 illustrates this effect for $\varepsilon = 0.1$.

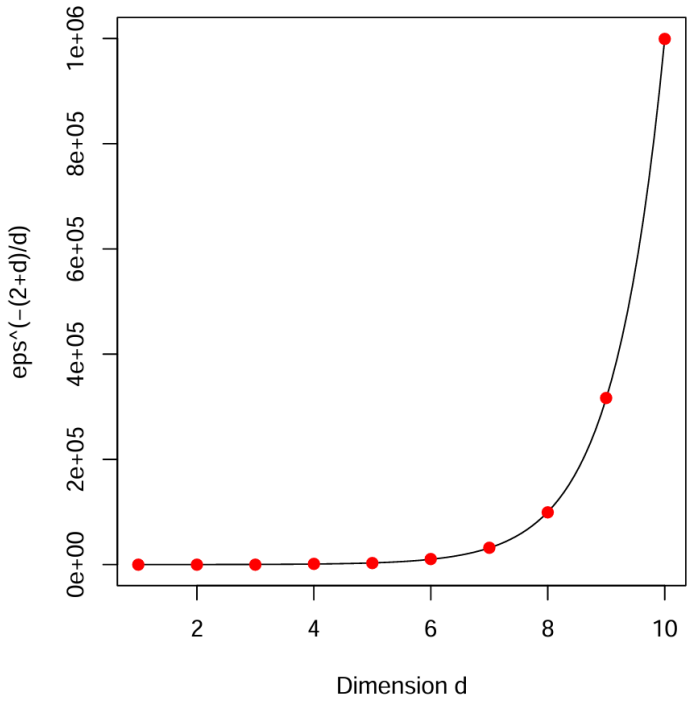

Figure 3.3: Illustration of the curse of dimensionality with $\varepsilon = 0.1$

This phenomenon is not unique to k-nearest neighbors; rather, it exemplifies the broader issue known as the *curse of dimensionality*. As the number of dimensions increases, the difficulty of estimation rises dramatically, often at an exponential rate. This challenge is formalized through minimax theory, which shows that the rate $n^{-\frac{2}{2+d}}$ is essentially the best achievable in a worst-case sense over certain function classes.

Specifically, for the class $\mathcal{H}_d(1, L)$ of L-Lipschitz functions on $\mathbb{R}^d$, with $L > 0$, one can prove that

$$\inf_{\hat{m}} \sup_{m \in \mathcal{H}_d(1,L)} \mathbb{E}[\|\hat{m} - m_0\|^2] \gtrsim n^{-\frac{2}{2+d}}, \tag{4}$$

where the infimum is taken over all possible estimators $\hat{m}$. This result confirms that no estimator can consistently outperform this rate across the entire Lipschitz function class.

3.3 Kernel smoothing

Kernel smoothing begins with a kernel function $K : \mathbb{R} \to \mathbb{R}$, satisfying

$$\int K(t)\,dt \;=\; 1, \quad \int t\,K(t)\,dt \;=\; 0, \quad 0 \;<\; \int t^2\,K(t)\,dt \;<\; \infty.$$

Three common examples are the box-car kernel:

$$K(t) \;=\; \begin{cases} 1, & \text{if } |t| \leq \dfrac{1}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

the *Gaussian* kernel:

$$K(t) \;=\; \frac{1}{\sqrt{2\pi}}\,\exp(-\frac{t^2}{2}),$$

and the Epanechnikov kernel:

$$K(t) \;=\; \begin{cases} \dfrac{3}{4}(1-t^2), & \text{if } |t| \leq\; 1, \\ 0, & \text{else.} \end{cases}$$

A reasonable approximation to the regression curve $m(x)$ will be the mean of response variables near a point x. This *local averaging procedure* can be defined as

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_{hi}(x) Y_i,$$

where W_{hi} are *local averaging weights* that depend on the distance via:

$$W_{hi}(x) = \frac{K_h(x - X_i)}{n^{-1} \sum_{i=1}^{n} K_h(x - X_i)},$$

where $K_h(u) = h^{-d} K(u/h)$. Note that h, called bandwidth in this context, controls the size of the neighborhood. Figure 3.4 illustrates how the bandwidth in kernel smoothing affects the resulting estimate. Specifically, the top-left panel shows the "True Density," a multi-modal distribution; the top-right panel ("Undersmoothed") has a very small bandwidth, so the estimate is extremely jagged, reflecting random noise rather than the true shape; the bottom-left ("Just Right") uses a moderate bandwidth that captures the main modes without introducing too much noise or too much smoothing; the bottom-right ("Oversmoothed") uses a large bandwidth, merging all the peaks into a single wide bump and losing important structure.

Given a bandwidth $h > 0$, the Nadaraya-Watson estimator is defined as:

$$\hat{m}_h(x) = \frac{n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i}{n^{-1} \sum_{i=1}^{n} K_h(x - X_i)} = \frac{\sum_{i=1}^{n} K(\frac{\|x - X_i\|_2}{h})\, Y_i}{\sum_{i=1}^{n} K(\frac{\|x - X_i\|_2}{h})}$$

In comparison to the k-nearest-neighbors estimator, which can be thought of as a raw (discontinuous) moving average of noisy responses, the Nadaraya-Watson estimator is a smooth moving average of responses. See Figure 3.5 for an example with $d = 1$.

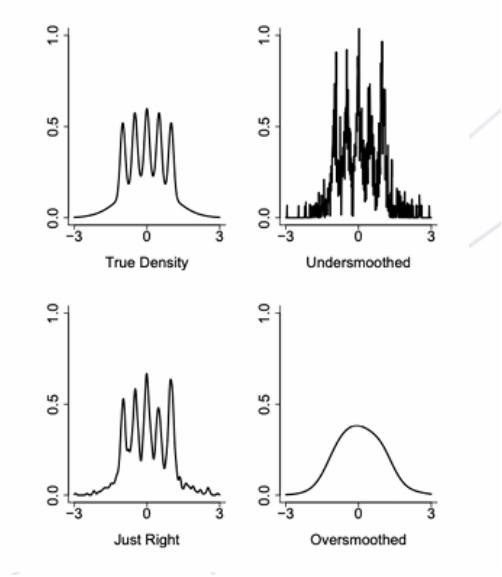

Figure 3.4: Effect of bandwidth in kernel smoothing.

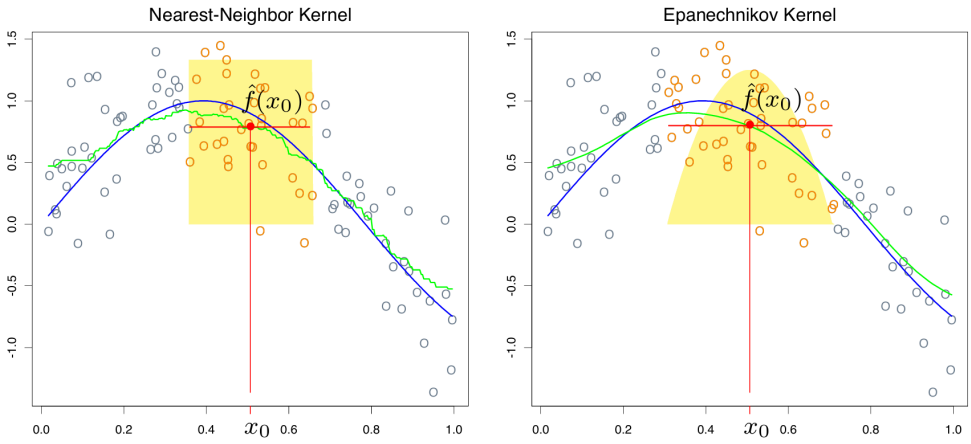

Figure 3.5: Comparing k-nearest-neighbor and Epanechnikov kernels when $d = 1$.

3.4 Error analysis

Theorem 3.1 *Suppose that the distribution of X has compact support and that* $\mathrm{Var}(Y \mid X = x) \leq \sigma^2 < \infty$ *for all x. Then*

$$\sup_{P \in \mathcal{H}_d(1,L)} \mathbb{E}[\|\hat{m} - m\|_P^2] \;\leq\; c_1\, h^2 \;+\; \frac{c_2}{n\, h^d}.$$

Hence, if $h \asymp n^{-\frac{1}{d+2}}$, then

$$\sup_{P \in \mathcal{H}_d(1,L)} \mathbb{E}[\|\hat{m} - m\|_P^2] \;\leq\; \frac{c}{n^{\frac{2}{d+2}}}.$$

Note that the rate $n^{-\frac{2}{d+2}}$ is slower than the pointwise rate $n^{-\frac{4}{d+2}}$ because we have made weaker assumptions.

Recall from (4) we saw that this was the minimax optimal rate over $H_d(1, L)$. More generally, the minimax rate over $H_d(\alpha, L)$, for a constant $L > 0$, is

$$\inf_{\hat{m}} \sup_{m_0 \in H_d(\alpha,L)} \mathbb{E}\|\hat{m} - m_0\|_2^2 \gtrsim n^{-2\alpha/(2\alpha+d)}. \tag{5}$$

Previously, we saw that with additional assumptions, a faster rate of $n^{-4/(4+d)}$ is attainable. This rate is in fact minimax for the class $H_d(2, L)$, and we will see later that it can be achieved under weaker assumptions using local linear regression techniques.

Now suppose that the distribution of X is supported on a smooth manifold with intrinsic dimension $r < d$. In this case, the expression

$$\int \frac{dP(x)}{nP(B(x,h))}$$

scales as $1/(nh^r)$ instead of the usual $1/(nh^d)$. This reduced intrinsic dimension leads to a better convergence rate of $n^{-2/(r+2)}$.

3.5 Density Estimation via Histograms

Perhaps the simplest density estimators are histograms. For convenience, assume that the data $X_1, \ldots, X_n$, are contained in the unit cube $\mathcal{X} = [0, 1]^d$ (although this assumption is not crucial). Divide $\mathcal{X}$ into bins, or sub-cubes, of size h. We discuss methods for choosing h later. There are $N \approx (1/h)^d$ such bins and each has volume h^d. Denote the bins by $B_1, \ldots, B_N$. The histogram density estimator is

$$\hat{p}_h(x) = \sum_{j=1}^{N} \frac{\hat{\theta}_j}{h^d} I(x \in B_j) \tag{3.1}$$

where

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in B_j) \tag{3.2}$$

is the fraction of data points in bin B_j. Now we bound the bias and variance of $\hat{p}_h$. We will assume that $p \in \mathcal{P}(L)$ where

$$\mathcal{P}(L) = \{p : |p(x) - p(y)| \leq L\|x - y\|, \text{ for all } x, y\}. \tag{3.3}$$

First, we bound the bias. Let $\theta_j = P(X \in B_j) = \int_{B_j} p(u)\, du$. For any $x \in B_j$,

$$p_h(x) \equiv \mathbb{E}(\hat{p}_h(x)) = \frac{\theta_j}{h^d} \tag{3.4}$$

and hence

$$p(x) - p_h(x) = p(x) - \frac{1}{h^d} \int_{B_j} p(u)\, du. \tag{3.5}$$

Thus,

$$|p(x) - p_h(x)| \leq \frac{1}{h^d} \int_{B_j} |p(x) - p(u)|\, du \leq \frac{1}{h^d} Lh\sqrt{d} \int_{B_j} du = Lh\sqrt{d}, \tag{3.6}$$

where we used the fact that if $x, u \in B_j$ then $\|x - u\| \leq \sqrt{d}h$.

Now we bound the variance. Since p is Lipschitz on a compact set, it is bounded. Hence,

$$\theta_j = \int_{B_j} p(u)\,du \leq Ch^d \text{ for some } C. \tag{3.7}$$

Thus, the variance is

$$\text{Var}(\hat{p}_h(x)) = \frac{1}{h^{2d}}\text{Var}(\hat{\theta}_j) = \frac{\theta_j(1-\theta_j)}{nh^{2d}} \leq \frac{\theta_j}{nh^{2d}} \leq \frac{C}{nh^d}. \tag{3.8}$$

We conclude that the L_2 risk is bounded by

$$\sup_{p\in\mathcal{P}(L)} R(p,\hat{p}) = \int \left(\mathbb{E}(\hat{p}_h(x)) - p(x)\right)^2\,dx \leq L^2h^2d + \frac{C}{nh^d}. \tag{3.9}$$

The upper bound is minimized by choosing

$$h = \left(\frac{C}{L^2nd}\right)^{1/(d+2)}. \tag{3.10}$$

With this choice,

$$\sup_{p\in\mathcal{P}(L)} R(p,\hat{p}) \leq C_0\left(\frac{1}{n}\right)^{2/(d+2)}, \tag{3.11}$$

where $C_0 = L^2d(C/(L^2d))^{2/(d+2)}$.

Later, we will prove the following theorem, which shows that this upper bound is tight. Specifically:

Theorem 2 There exists a constant $C > 0$ such that

$$\inf_{\hat{p}}\sup_{p\in\mathcal{P}(L)} \mathbb{E}\int (\hat{p}_h(x) - p(x))^2\,dx \geq C\left(\frac{1}{n}\right)^{2/(d+2)}. \tag{3.12}$$

3.5.1 Concentration Analysis for Histograms

Let us now derive a concentration result for $\hat{p}_h$. We will bound

$$\sup_p P^n(\|\hat{p}_h - p\|_\infty > \epsilon).$$

Assume that $\epsilon \leq 1$. Using Bernstein's inequality and bounding terms, we find:

$$\|\hat{p}_h - p\|_\infty \leq \sqrt{\frac{1}{nh^d}\log\left(\frac{2}{\delta h^d}\right)} + L\sqrt{d}h. \tag{3.13}$$

Choosing $h = (c_2/n)^{1/(2+d)}$, we conclude that

$$\|\hat{p}_h - p\|_\infty = O\left(\left(\frac{\log n}{n}\right)^{1/(2+d)}\right).$$

3.6 Kernel Density Estimation

A one-dimensional smoothing kernel is any smooth function K such that $\int K(x)\,\mathrm{d}x = 1$, $\int xK(x)\,\mathrm{d}x = 0$, and $\sigma_K^2 \equiv \int x^2 K(x)\,\mathrm{d}x > 0$. Smoothing kernels should not be confused with Mercer kernels, which we discuss later. Some commonly used kernels are the following:

$$\begin{aligned}
\text{Boxcar:} \quad & K(x) = \frac{1}{2} I(x) \\
\text{Gaussian:} \quad & K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\
\text{Epanechnikov:} \quad & K(x) = \frac{3}{4}(1 - x^2) I(x) \\
\text{Tricube:} \quad & K(x) = \frac{70}{81}(1 - |x|^3)^3 I(x)
\end{aligned}$$

where $I(x) = 1$ if $|x| \leq 1$ and $I(x) = 0$ otherwise. These kernels are plotted in Figure 2. Two commonly used multivariate kernels are $\prod_{j=1}^{d} K(x_j)$ and $K(\|x\|)$.

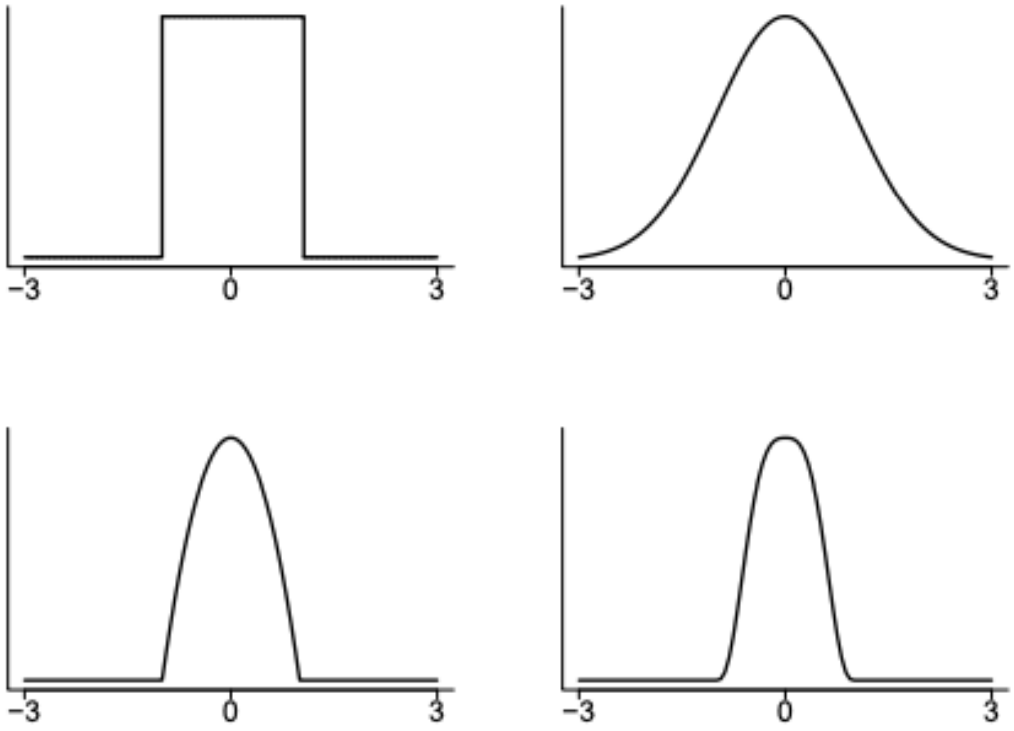

Figure 3.6: Examples of smoothing kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

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

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h} \right). \tag{3.14}$$

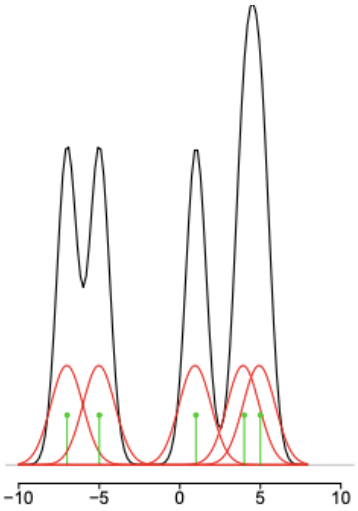

Figure 3.7: A kernel density estimator $\hat{p}$. At each point x, $\hat{p}(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars. The kernels are not drawn to scale.

More generally, we define

$$\hat{p}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i), \tag{3.15}$$

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.

Sometimes we write the estimator as $\hat{p}_h$ to emphasize the dependence on h. In the multivariate case, the coordinates of X_i should be standardized so that each has the same variance, since the norm $\|x - X_i\|$ treats all coordinates as if they are on the same scale.

The kernel estimator places a smoothed-out lump of mass of size $1/n$ over each data point X_i; see Figure 3. 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.

3.6.1 Additional Comments on the Kernel Density Estimation

Our bias/variance trade off depends on the size of the neighborhood. A smaller n will have a smaller bias.

In regards to the linear estimator, this is just a linear re-weighting, where we see the estimated label as a combination of Y. This is equivalent to local polynomial regression. If you do a local regression, you can write the estimator as a linear combination of the label Y. If you want to do a quadratic regression instead of only linear, you will need to change your input X via feature extraction. This will give you a matrix now to work with. Plus - a quadratic functions form a (linear) vector space! So you do not need the function to be linear, just the function space.

3.6.2 Risk Analysis

In this section, we examine the accuracy of kernel density estimation. We will first need a few definitions.

Assume that $X_i \in \mathcal{X} \subset \mathbb{R}^d$ where $\mathcal{X}$ is compact. Let β and L be positive numbers. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$, and

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}. \tag{3.16}$$

Let β be a positive integer. Define the Hölder class

$$\Sigma(\beta, L) = \{g : |D^s g(x) - D^s g(y)| \leq L\|x - y\|, \text{ for all } s \text{ such that } |s| = \beta - 1, \text{ and all } x, y\}. \tag{3.17}$$

For example, if $d = 1$ and $\beta = 2$, this means that

$$|g'(x) - g'(y)| \leq L|x - y|, \quad \text{for all } x, y. \tag{3.18}$$

The most common case is $\beta = 2$; roughly speaking, this means that the functions have bounded second derivatives.

If $g \in \Sigma(\beta, L)$, then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L\|u - x\|^\beta, \tag{3.19}$$

where

$$g_{x,\beta}(u) = \sum_{|s| \leq \beta} \frac{(u - x)^s}{s!} D^s g(x). \tag{3.20}$$

In the common case of $\beta = 2$, this means that

$$\left|p(u) - \left[p(x) + (x - u)^T \nabla p(x)\right]\right| \leq L\|x - u\|^2. \tag{3.21}$$

Assume now that the kernel K has the form $K(x) = G(x_1) \cdots G(x_d)$ where G has support on $[-1, 1]$, $\int G = 1$, $\int |G|^p < \infty$ for any $p \geq 1$, $\int |t|^\beta |K(t)| \, \mathrm{d}t < \infty$, and $\int t^s K(t) \, \mathrm{d}t = 0$ for $s \leq \beta$.

An example of a kernel that satisfies these conditions for $\beta = 2$ is $G(x) = (3/4)(1 - x^2)$ for $|x| \leq 1$. Constructing a kernel that satisfies $\int t^s K(t) \, \mathrm{d}t = 0$ for $\beta > 2$ requires using kernels that can take negative values.

Let

$$p_h(x) = \mathbb{E}[\hat{p}_h(x)]. \tag{3.22}$$

The next lemma provides a bound on the bias $p_h(x) - p(x)$.

Proposition 3.2 (Bias of higher-order kernel estimator) *Let $p \in \mathcal{H}^\beta(L)$ and let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel of order β, i.e.,*

1. $\int K(v) \, dv = 1$,
2. $\int v^\alpha K(v) \, dv = 0$ *for all multi-indices* α *with* $1 \leq |\alpha| < \beta$,
3. $\int \|v\|^\beta |K(v)| \, dv < \infty$.

Then the kernel smoothed density $p_h(x) := \int \frac{1}{h^d} K\left(\frac{u - x}{h}\right) p(u) \, du$ *satisfies* $\sup_x |p_h(x) - p(x)| = O(h^\beta)$.

Proof: By the change of variables $v = (u - x)/h$, we write

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

Let $m = \lfloor \beta \rfloor$. By Taylor's theorem, for each x,

$$p(x + hv) = p(x) + \sum_{1 \le |\alpha| \le m} \frac{(hv)^\alpha}{\alpha!} D^\alpha p(x) + R(x, hv),$$

where the remainder satisfies $|R(x, hv)| \le C\|hv\|^\beta = Ch^\beta \|v\|^\beta$.

Substituting into the bias expression yields

$$p_h(x) - p(x) = \sum_{1 \le |\alpha| \le m} \frac{h^{|\alpha|}}{\alpha!} D^\alpha p(x) \int v^\alpha K(v)\, dv + \int K(v) R(x, hv)\, dv.$$

Since K is a kernel of order β, we have

$$\int v^\alpha K(v)\, dv = 0 \quad \text{for all } |\alpha| < \beta,$$

so all polynomial terms vanish. Therefore,

$$|p_h(x) - p(x)| \le \int |K(v)|\, |R(x, hv)|\, dv \le Ch^\beta \int \|v\|^\beta |K(v)|\, dv.$$

Because $\int \|v\|^\beta |K(v)|\, dv < \infty$, we conclude that

$$\sup_x |p_h(x) - p(x)| \le C' h^\beta.$$

■

Lemma 4 The variance of $\hat{p}_h$ satisfies:

$$\sup_{p \in \Sigma(\beta, L)} \mathrm{Var}(\hat{p}_h(x)) \le \frac{c}{nh^d} \tag{3.23}$$

for some $c > 0$.

Proof. We can write $\hat{p}_h(x) = n^{-1} \sum_{i=1}^n Z_i$ where $Z_i = \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right)$. Then,

$$\mathrm{Var}(Z_i) \le \frac{\sup_p p(x)}{h^d} \int K^2(\|v\|) \mathrm{d}v \le \frac{c}{h^d},$$

for some c since the densities in $\Sigma(\beta, L)$ are uniformly bounded. The result follows. □

Theorem 5 The L_2 risk is bounded above, uniformly over $\Sigma(\beta, L)$, by $h^{4\beta} + \frac{1}{nh^d}$ (up to constants). If $h \asymp n^{-1/(2\beta+d)}$ then

$$\sup_{p \in \Sigma(\beta, L)} \mathbb{E}\left[\int \left(\hat{p}_h(x) - p(x)\right)^2 \mathrm{d}x\right] \asymp \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{3.24}$$

When $\beta = 2$ and $h \asymp n^{-1/(4+d)}$ we get the rate $n^{-4/(4+d)}$.

3.6.3 Minimax Bound

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. We will prove this later in the course.

Theorem 6 There exists C depending only on β and L such that

$$\inf_{\hat{p}} \sup_{p \in \Sigma(\beta,L)} \mathbb{E}_p \int (\hat{p}(x) - p(x))^2 \mathrm{d}x \geq C \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{3.25}$$

Theorem 6 together with (3.24) imply that kernel estimators are rate minimax.

3.6.4 Concentration Analysis of Kernel Density Estimator

Now we state a result which says how fast $\hat{p}(x)$ concentrates around $p(x)$. First, recall Bernstein's inequality: Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with mean μ, $\mathrm{Var}(Y_i) \leq \sigma^2$ and $|Y_i| \leq M$. Then

$$\mathbb{P}(|\bar{Y} - \mu| > \epsilon) \leq 2 \exp\left\{ -\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3} \right\}. \tag{3.26}$$

Theorem 7 For all small $\epsilon > 0$,

$$\mathbb{P}(|\hat{p}(x) - p_h(x)| > \epsilon) \leq 2 \exp\left\{ -cnh^d \epsilon^2 \right\}. \tag{3.27}$$

Hence, for any $\delta > 0$,

$$\sup_{p \in \Sigma(\beta,L)} \mathbb{P}\left(|\hat{p}(x) - p(x)| > \sqrt{\frac{\log(2/\delta)}{nh^d} + ch^\beta} \right) < \delta \tag{3.28}$$

for some constants C and c. If $h \asymp n^{-1/(2\beta+d)}$ then

$$\sup_{p \in \Sigma(\beta,L)} \mathbb{P}\left(|\hat{p}(x) - p(x)|^2 > \frac{C \log n}{n^{2\beta/(2\beta+d)}} \right) < \delta. \tag{3.29}$$

3.6.5 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, and local likelihood.

3.6.6 Confidence Bands and the CLT

Consider first a single point x. Let $s_n(x) = \sqrt{\mathrm{Var}(\hat{p}_h(x))}$. The CLT implies that

$$Z_n(x) \equiv \frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} \xrightarrow{d} N(0, \tau^2(x)) \tag{3.30}$$

for some $\tau(x)$. This is true even if h_n is decreasing. Specifically, suppose that $h_n \to 0$ and $nh_n \to \infty$. Note that $Z_n(x) = \sum_{i=1}^n L_{ni}$ say. According to Lyapunov's CLT, $\sum_{i=1}^n L_{ni} \xrightarrow{d} N(0,1)$ as long as

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}[L_{ni}^{2+\delta}] = 0 \tag{3.31}$$

for some $\delta > 0$. But this does not yield a confidence interval for $p(x)$. To see why, let us write

$$\frac{\hat{p}_h(x) - p(x)}{s_n(x)} = \frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{p_h(x) - p(x)}{s_n(x)} = Z_n(x) + \frac{\text{bias}}{\sqrt{\text{Var}(x)}}.$$

Assuming that we optimize the risk by balancing the bias and the variance, the second term is some constant c. So

$$\frac{\hat{p}_h(x) - p(x)}{s_n(x)} \xrightarrow{d} N(c, \tau^2(x)). \tag{3.32}$$

This means that the usual confidence interval $\hat{p}_h(x) \pm z_\alpha s(x)$ will not cover $p(x)$ with probability tending to $1-\alpha$. One fix for this is to undersmooth the estimator. (We sacrifice risk for coverage.) An easier approach is just to interpret $\hat{p}_h(x) \pm z_\alpha s(x)$ as a confidence interval for the smoothed density $p_h(x)$ instead of $p(x)$.

But this only gives an interval at one point. To get a confidence band we use the bootstrap. Let P_n be the empirical distribution of $X_1, \ldots, X_n$. The idea is to estimate the distribution

$$F_n(t) = \mathbb{P}\left(\sqrt{nh^d}\|\hat{p}_h(x) - p_h(x)\|_\infty \leq t\right)$$

with the bootstrap estimator

$$\hat{F}_n(t) = \mathbb{P}\left(\sqrt{nh^d}\|\hat{p}_h(x) - \hat{p}_h^*(x)\|_\infty \leq t \mid X_1, \ldots, X_n\right),$$

where $\hat{p}_h^*$ is constructed from the bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Later in the course, we will show that

$$\sup_t |F_n(t) - \hat{F}_n(t)| \xrightarrow{p} 0. \tag{3.33}$$

Algorithm:

1. Let P_n be the empirical distribution that puts mass $1/n$ at each data point X_i.
2. Draw $X_1^*, \ldots, X_n^* \sim P_n$. This is called a bootstrap sample.
3. Compute the density estimator $\hat{p}_h^*$ based on the bootstrap sample.
4. Compute $R = \sup_x \sqrt{nh^d}\|\hat{p}_h^* - \hat{p}_h\|_\infty$.
5. Repeat steps 2-4 B times. This gives $R_1, \ldots, R_B$.
6. Let z_α be the upper α quantile of the R_js. Thus
$$\frac{1}{B}\sum_{j=1}^B I(R_j > z_\alpha) \approx \alpha.$$
7. Let
$$\ell_n(x) = \hat{p}_h(x) - \frac{z_\alpha}{\sqrt{nh^d}}, \quad u_n(x) = \hat{p}_h(x) + \frac{z_\alpha}{\sqrt{nh^d}}.$$

Theorem 10 Under appropriate (very weak) conditions, we have

$$\liminf_{n\to\infty} \mathbb{P}\left(\ell_n(x) \le p_h(x) \le u_n(x) \text{ for all } x\right) \ge 1-\alpha.$$

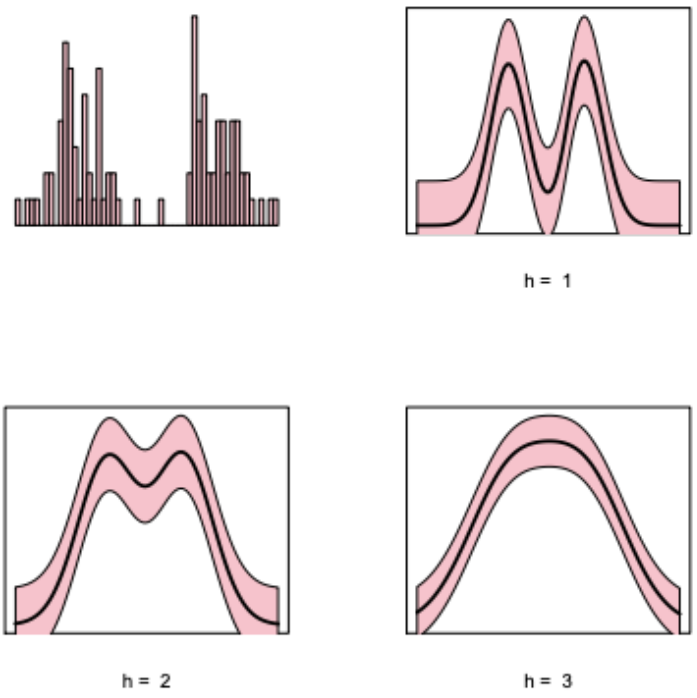

Figure 3.8: 95 percent bootstrap confidence bands using various bandwidths. The top-left panel shows the histogram of the data. The top-right, bottom-left, and bottom-right panels show the bootstrap confidence bands for bandwidths $h = 1$, $h = 2$, and $h = 3$, respectively.

See Figure 4.

If you want a confidence band for p you need to reduce the bias (undersmooth). A simple way to do this is with *twicing*. Suppose that $\beta = 2$ and that we use the kernel estimator $\hat{p}_h$. Note that

$$\mathbb{E}[\hat{p}_h(x)] = p(x) + C(x)h^2 + o(h^2),$$

$$\mathbb{E}[\hat{p}_{2h}(x)] = p(x) + C(x)4h^2 + o(h^2).$$

For some $C(x)$. That is, the leading term of the bias is $b(x) = C(x)h^2$. So if we define

$$\hat{b}(x) = \frac{\hat{p}_{2h}(x) - \hat{p}_h(x)}{3}$$

then $\mathbb{E}[\hat{b}(x)] = b(x)$.

We define the bias-reduced estimator

$$\tilde{p}_h(x) = \hat{p}_h(x) - \hat{b}(x) = \frac{4}{3}\left(\hat{p}_h(x) - \frac{1}{4}\hat{p}_{2h}\right).$$

A confidence set centered at $\tilde{p}_h$ will be asymptotically valid but will not be an optimal estimator. This is a fundamental conflict between estimation and inference.

References

[1] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.