18.650: Fundamentals of Statistics December 1, 2023

Lecture 29 — Nonparametric curve estimation (Ch 20)

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Overview

Nonparametric curve estimation typically pertains to estimating either a density or a regression function. To give an example from regression, suppose we observe pairs (X_i, Y_i) as in Figure 1. We can see that the relationship between X and Y is not linear (the purple line is a bad approximation). So how should we fit a curve to the data? Should we choose the green curve in the lefthand plot or the green curve in the righthand plot? The lefthand curve is better because it's *smooth*. The guiding principle in nonparametric curve estimation is that the true underlying curve is smooth: the value of the function near a point x should be close to $f(x)$. There is

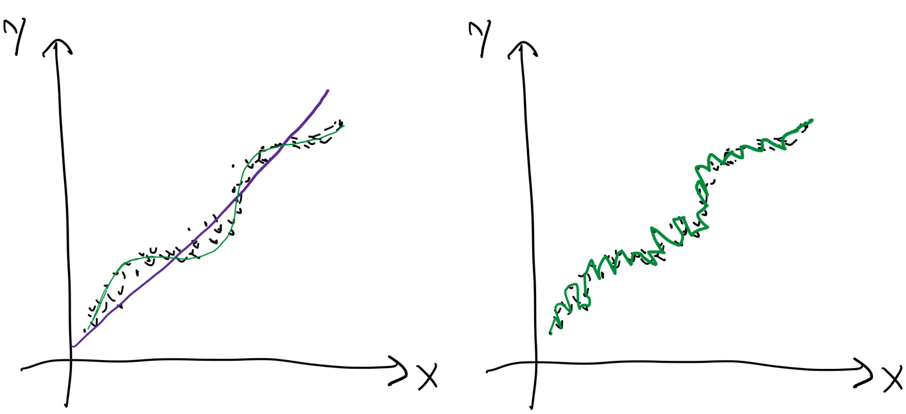

Figure 1: Curve estimation. We assume the true curve is smooth, so the green curve on the left is a better approximation then the green curve on the right. The purple line is clearly a bad fit.

still a question of *how* smooth the function should be. We will modulate how smooth the curve is by playing with the bias-variance tradeoff (intuitively, a smoother curve leads to lower variance but higher bias).

We will output a whole family of estimators that give different values of the bias and variance. Let's review these two quantities, now in the context of curve estimation. Suppose $g(\cdot) \rightsquigarrow \hat{g}(\cdot)$ ("g is estimated by $\hat{g}$"), which means $g(x) \rightsquigarrow \hat{g}(x)$

for all x. Then we define

$$\text{bias: } b(x) = \mathbb{E}[\hat{g}(x)] - g(x)$$
$$\text{variance: } v(x) = \mathbb{V}[\hat{g}(x)]$$
$$\text{mean squared error : } \text{MSE}(\hat{g}(x)) = b^2(x) + v(x).$$

Since there is a MSE for each x, we can get an overall error by integrating:

Definition 1.1: MISE

The mean integrated squared error (MISE) is defined to be

$$R(\hat{g}, g) = \int \text{MSE}(\hat{g}(x))dx = \underbrace{\int b(x)^2 dx}_{\text{bias term}} + \underbrace{\int v(x)dx}_{\text{variance term}}.$$

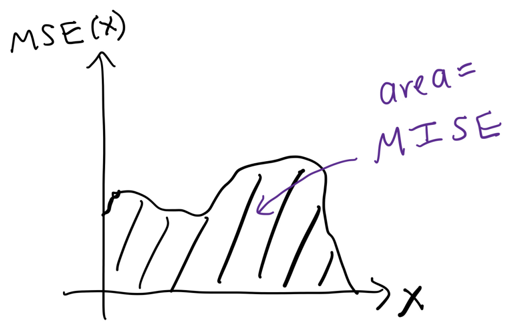

Figure 2: The MISE is the integral of the MSE over the domain.

2 Density estimation with histograms

Suppose we observe samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, where p is a density on the unit interval (the arguments below can easily be generalized to other intervals). To construct a histogram estimator $\hat{p}$ of p, create m equally spaced bins B_j of width $h = 1/m$, and define

$$n_j = \#\{i \ : \ X_i \in B_j\},$$
$$\hat{p}_j = \frac{n_j}{n} \text{ (proportion of observations in bin } B_j)$$

Definition 2.1: Histogram estimator

The histogram estimator $\hat{p}$ is the function which takes the value $\hat{p}(x) = \hat{p}_j/h$ when $x \in B_j$ for $j = 1, \ldots, m$. This can be written concisely as

$$\hat{p}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j).$$

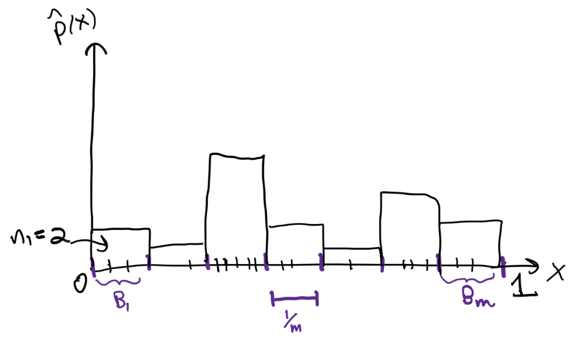

Figure 3: A visualization of the histogram estimator $\hat{p}(x)$ from Definition 4.2.

Let's show $\hat{p}$ is a true density, meaning $\hat{p}(x) \geq 0$ for all $x \geq 0$ and $\int \hat{p}(x)dx = 1$. The first criterion is clearly satisfied. To check the second criterion, we compute

$$\begin{aligned}\int \hat{p}(x)dx &= \sum_{j=1}^{m} \int \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j)dx = \sum_{j=1}^{m} \int_{B_j} \frac{\hat{p}_j}{h} dx \\ &= \sum_{j=1}^{m} h \cdot \frac{\hat{p}_j}{h} = \sum_{j=1}^{m} \hat{p}_j = \sum_{j=1}^{m} \frac{n_j}{n} = 1.\end{aligned}$$

Remark.

> The above calculation shows that the purpose of the $1/h$ normalization is to ensure $\hat{p}$ integrates to 1. If we use $\hat{p}$ only to visualize the data as a histogram, then this normalization is not important. But if we need $\hat{p}$ to do any quantitative estimation, then the $1/h$ is important.

2.1 Bias of $\hat{p}$

Let's compute the bias $b(x)$ and integrated bias. Fix an x in the unit interval, and find j such that x is in bin B_j. We then have

$$b(x) = \mathbb{E}[\hat{p}(x)] - p(x) = \mathbb{E}\left[\frac{\hat{p}_j}{h}\right] - p(x).$$

Now, recall that $\hat{p}_j = n_j/n$, where n_j is the number of samples in bin B_j. This implies

$$n_j \sim \mathrm{Bin}(n, \mathbb{P}(X \in B_j)) = \mathrm{Bin}\left(n, \int_{B_j} p(y)dy\right).$$

Using the fact that the expectation of $\mathrm{Bin}(n,p)$ is np, we get that

$$b(x) = \mathbb{E}\left[\frac{\hat{p}_j}{h}\right] - p(x) = \frac{1}{h}\int_{B_j} p(y)dy - p(x) \to 0 \quad \text{as} \quad h \to 0.$$

The reason that $\frac{1}{h}\int_{B_j} p(y)dy - p(x)$ goes to zero as $h \to 0$ is that $\frac{1}{h}\int_{B_j} p(y)dy$ is the average of p in B_j, and by smoothness, this average value is close to any value $p(x)$ for $x \in B_j$. See Figure 4 for a visualization of this.

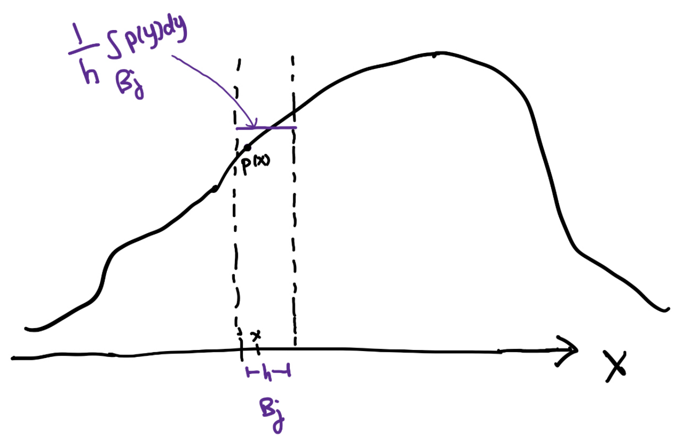

Figure 4: The $(1/h)$-normalized integral of p over B_j is just the average of the values of $p(y)$ over y's in bin B_j. By smoothness, this average is close to $p(x)$ for any $x \in B_j$.

One can show that the integrated bias also goes to zero as $h \to 0$:

$$\int b(x)dx \to 0 \quad \text{as} \quad h \to 0.$$

We conclude that *the smaller the bin width h, the smaller the bias.*

2.2 Variance of $\hat{p}$

Now let's compute the variance of $\hat{p}$. Fix an x in bin B_j and recall that $\hat{p}(x) = \hat{p}_j/h$. We again use that $n_j \sim \text{Bin}(n, \int_{B_j} p(y)dy)$, and the fact that the variance of $\text{Bin}(n,p)$ is $np(1-p)$. Therefore,

$$\begin{aligned} v(x) &= \mathbb{V}\left[\frac{\hat{p}_j}{h}\right] = \frac{1}{h^2}\mathbb{V}[\hat{p}_j] = \frac{1}{h^2}\mathbb{V}\left[\frac{n_j}{n}\right] = \frac{1}{h^2n^2}\mathbb{V}[n_j] \\ &= \frac{n}{h^2n^2}\int_{B_j} p(y)dy\left(1 - \int_{B_j} p(y)dy\right) = \frac{1}{nh}\underbrace{\left(\frac{1}{h}\int_{B_j} p(y)dy\right)}_{\to p(x) \text{ as } h \to 0}\underbrace{\left(1 - \int_{B_j} p(y)dy\right)}_{\to 1 \text{ as } h\to 0} \end{aligned}$$

We conclude that

$$v(x) \approx \frac{1}{nh}p(x) \quad \text{when } h \text{ is small}$$

Therefore, *the smaller the bin width h, the larger the variance.*

2.3 Bias-variance tradeoff

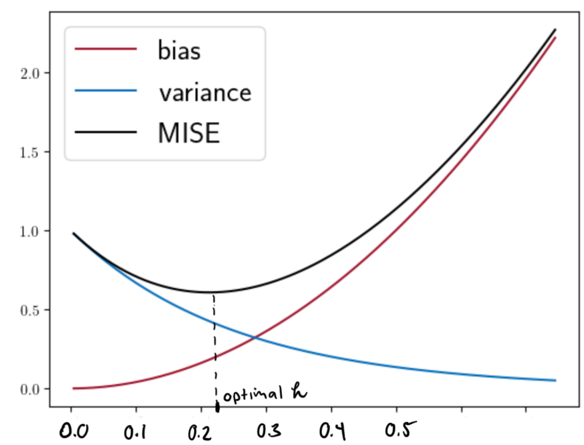

Figure 5: The MISE is minimized at an intermediate point, where the bias and variance are not too small and not too large.

The bias decreases as h gets smaller, and the variance decreases. Therefore, the optimal value of h which minimizes $\text{MISE}(\hat{p})$ is somewhere in between; see Figure 5. However, this optimal value of h depends on the true, unknown density p. To find the best choice of h without knowing p, we can use a method called *cross validation*; see Chapter 20 for a discussion of this.

3 Kernel density estimators

Definition 3.1: Kernel

A kernel K is a function satisfying $K(x) \geq 0$, $\int K(x)dx = 1$, and $\int xK(x)dx = 0$.

Any pdf symmetric about the origin satisfies these properties! For example, the Gaussian pdf

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \phi(x)$$

is a valid kernel.

Definition 3.2: Kernel density estimator

A kernel density estimator (KDE) of p with kernel K is the estimator

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

Figure 6 depicts several commonly used kernels.

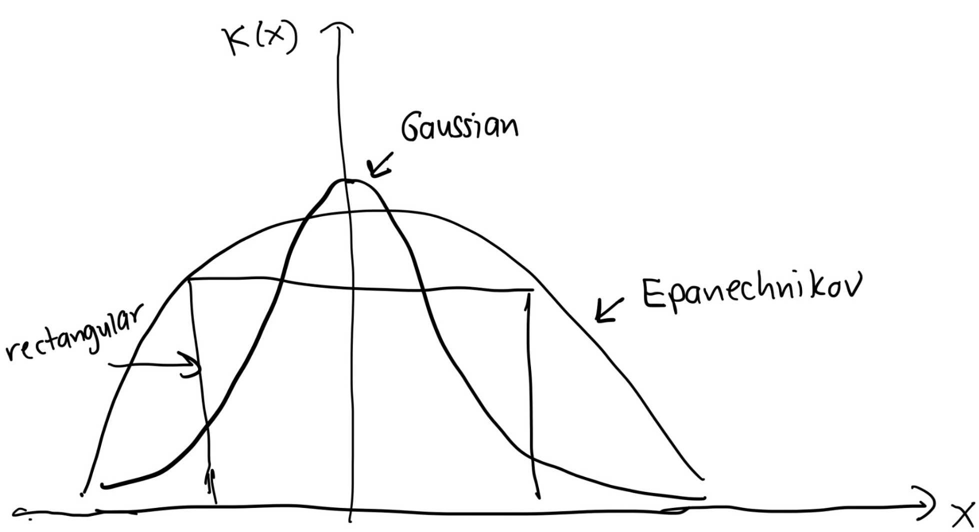

Figure 6: Common choices of kernel

Remark.

> There is freedom in how to choose h. In the context of KDEs, h is called the *bandwidth*.

The histogram estimator we constructed above is similar to the KDE with kernel

$$K(x) = \begin{cases} 1, & -\frac{1}{2} < x < \frac{1}{2}, \\ 0, & \text{otherwise} \end{cases}$$

For each x, let $n(x)$ be the number of samples which fall within the interval $[x - h/2, x + h/2]$. The KDE $\hat{p}(x)$ is then given by $\hat{p}(x) = n(x)/nh$. This is the same as the histogram, except now we use "sliding" bins.

4 Nonparametric Regression

Suppose we observe pairs (X_i, Y_i), $i = 1, \dots, n$. We want to estimate the regression function (which has the best prediction property), defined as

$$f(x) = \mathbb{E}[Y|X = x] = \int yp(y|x)dy.$$

One way to do this is the following: break up the x-axis into bins as before. Now, for each x, find the bin B_j it belongs to. Then

$$\mathbb{E}[Y|X = x] \approx \mathbb{E}[Y|X \in B_j] \approx \frac{\sum_{i:X_i \in B_j} Y_i}{\#\{i : X_i \in B_j\}} = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in B_j)}{\sum_{i=1}^n \mathbb{1}(X_i \in B_j)}$$

This is known as the *regressogram*.

Definition 4.1: Regressogram

The regressogram $\hat{f}$ is the estimator

$$\hat{f}(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in B_j)}{\sum_{i=1}^n \mathbb{1}(X_i \in B_j)}, \quad \text{for all } x \in B_j.$$

The regressogram is a step function, which is constant over each bin B_j.

The regressogram $\hat{f}$ is conceptually similar to the histogram estimator we used for density estimation. There is also an analogue to the KDE: you can estimate $p(x, y)$ using a two-dimensional KDE, estimate $p(x)$ using another KDE, and take the ratio of the two to get an estimate of $p(y|x)$. You can then plug this estimate into the formula $f(x) = \mathbb{E}[Y|X = x] = \int yp(y|x)dx$. After simplifying the resulting expression, you get the following estimator $\hat{f}$:

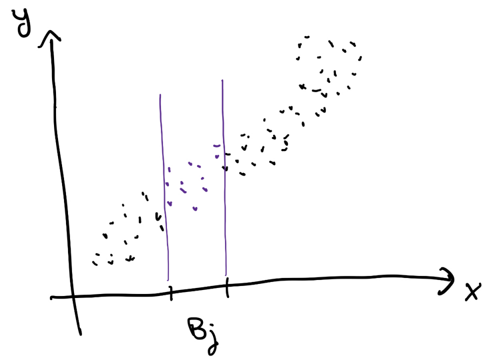

Figure 7: To compute the regressogram estimator $\hat{f}(x)$ for $x \in B_j$, simply average the Y_i's whose corresponding X_i lies in B_j (the purple points in this figure).

Definition 4.2: Nadaraya-Watson estimator

The Nadaraya-Watson estimator $\hat{f}$ is the estimator

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

Nonparametric curve estimation in higher dimensions: Nonparametric curve estimation can generalize to higher dimensions k of the variable x. However, the MISE of nonparametric estimators typically scales as $(1/h)^k/n$. This means the number of samples n we need to get a accurate estimate grows grows exponentially with dimension k.