

A note on the bias-variance tradeoff

In this note we give a derivation of the bias-variance tradeoff and curse of dimensionality for kernel density estimation. The analysis for kernel smoothing is similar.

The kernel density estimator is

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

for a kernel $K(u) \geq 0$ satisfying $\int K(u) du = 1$ and $\int uK(u) du = 0$. In p -dimensions we can take the product kernel $K(u) \equiv \prod_{j=1}^p K(u_j)$ where $K(\cdot)$ is a 1-dimensional kernel.

To estimate the bias we calculate $\mathbb{E}\hat{f}(x)$ as

$$\mathbb{E}\hat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^n \mathbb{E}K\left(\frac{X_i - x}{h}\right) \quad (2)$$

$$= \frac{1}{h^p} \int K\left(\frac{u - x}{h}\right) f(u) du \quad (3)$$

since $X_i \sim F$ are iid with density f . Now make the change of variables $v = \frac{u-x}{h}$ so that $u = x + hv$ and $du = h^p dv$. Letting $h \rightarrow 0$ and assuming f is bounded and twice differentiable with bounded derivatives this gives

$$\mathbb{E}\hat{f}(x) = \int K(v) f(x + hv) dv \quad (4)$$

$$= \int K(v) \left(f(x) + hv^T \nabla f(x) + \frac{1}{2} h^2 v^T \nabla^2 f(x) v + o(h^2) \right) dv \quad (5)$$

$$= f(x) + C(x)h^2 + o(h^2) \quad (6)$$

where we use the assumptions $\int K(u) du = 1$ and $\int uK(u) du = 0$ on the kernel, and that

$$C(x) = \frac{1}{2} \int K(u) u^T \nabla^2 f(x) u du \leq C_1 h^2 \quad (7)$$

for some constant C_1 . This shows that the squared bias is of order h^4 :

$$\left(\mathbb{E}\hat{f}(x) - f(x) \right)^2 \leq C_1^2 h^4 + o(h^4). \quad (8)$$

Now we bound the variance. We use $\text{Var}(x) \leq \mathbb{E}\hat{f}(x)^2$ and calculate

$$\mathbb{E}\hat{f}(x) \leq C_2 \frac{1}{n^2 h^{2p}} \sum_{i=1}^n \mathbb{E} K \left(\frac{X_i - x}{h} \right)^2 \quad (9)$$

$$= C_2 \frac{1}{n h^{2p}} \int K \left(\frac{u - x}{h} \right)^2 f(u) du \quad (10)$$

$$= C_2 \frac{f(x)}{n h^p} \int K(v)^2 dv + o \left(\frac{1}{n h^p} \right) \quad (11)$$

$$= C_2 \frac{f(x)}{n h^p} + o \left(\frac{1}{n h^p} \right) \quad (12)$$

using another Taylor approximation, where we assume that $n h^p \rightarrow \infty$, with the constant C_2 changing from line to line.

To summarize, these calculations tell us that

$$\text{bias}^2(x) \approx h^4 \quad (13)$$

$$\text{var}(x) \approx \frac{1}{n h^p}. \quad (14)$$

Choosing h so that the squared bias and the variance are of the same order gives that

$$h \approx n^{\frac{-1}{4+p}} \quad (15)$$

$$\mathbb{E}(\hat{f}(x) - f(x))^2 = O \left(n^{\frac{-4}{4+p}} \right). \quad (16)$$

Lower bounds tell us that this is the fastest rate the risk can decrease, under the given assumptions on the density f . Flipping this risk around, we see that the sample size n must increase exponentially in p in order to drive the risk down to a fixed level ϵ . This is the “curse of dimensionality.”

Finally, note that we can think of the kernel smoothing estimator as what we get by plugging in the kernel density estimator for x and y . That is, using the “plug-in” estimate of $\hat{f}(y | x)$ we have that

$$\hat{m}(x) = \int y \hat{f}(y | x) dy \quad (17)$$

$$= \int y \frac{\hat{f}(x, y)}{\hat{f}(x)} dy \quad (18)$$

$$= \frac{\int \sum_{i=1}^n \frac{1}{n h^{p+1}} K \left(\frac{Y_i - y}{h} \right) K \left(\frac{X_i - x}{h} \right) dy}{\sum_{i=1}^n \frac{1}{n h^p} K \left(\frac{X_i - x}{h} \right)} \quad (19)$$

$$= \frac{\sum_{i=1}^n Y_i K \left(\frac{X_i - x}{h} \right)}{\sum_{i=1}^n K \left(\frac{X_i - x}{h} \right)} \quad (20)$$

using the fact that

$$\int y \frac{1}{h} K\left(\frac{Y_i - y}{h}\right) dy = Y_i \quad (21)$$

from our assumptions on the kernel. This is the usual kernel smoothing estimator. This suggests that the same bias-variance decomposition holds for kernel smoothing, which is indeed the case.