EE290 Mathematics of Data Science Lecture 18 and 19 - 10/31/2019

Lecture 18 and 19: Nonparametric Function Estimation

Lecturer: Jiantao Jiao *Scribe: Yanjun Han and Jiantao Jiao*

This course notes was adapted from lecture 14 of EE378A offered at Stanford University by Tsachy Weissman and Jiantao Jiao.

1 Introduction

We look at the following regression setting: we have n observations

$$y_i = f(x_i) + \sigma\xi_i, \qquad i = 1, \cdots, n \tag{1}$$

where $f(\cdot)$ is our target function on $[0,1]$ which we assume to lie in some prescribed function class $\mathcal{F}$, $x_i = \frac{i}{n}$ are equally spaced points on $[0,1]$, σ is the known noise level, and $\xi_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ are i.i.d Gaussian noise. Our target is to find some estimator $\hat{f}$ which comes close to minimizing the worst-case L_q risk:

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|_q \lesssim \inf_{\hat{f}^*} \sup_{f\in\mathcal{F}} \mathbb{E}_f \|\hat{f}^* - f\|_q. \tag{2}$$

Notations: for non-negative sequences $\{a_n\}, \{b_n\}$, we use $a_n \lesssim b_n$ to denote $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$, i.e., there exists some constant C not depending on n (but could depend on other parameters) such that $a_n \le Cb_n$ for any $n \in \mathbb{N}$. Similarly, we write $a_n \gtrsim b_n$ to denote $b_n \lesssim b_n$, and $a_n \asymp b_n$ to denote $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

References:

1. A. Nemirovski. "Topics in non-parametric Statistics." Ecole d'Ete de Probabilites de Saint-Flour 28 (2000): 85.

2. A.B. Tsybakov. "Introduction to Nonparametric Estimation." Springer, 2009.

2 The Bias–Variance Decomposition

Before looking at specific function class $\mathcal{F}$ and estimators $\hat{f}$, we take a look at how to evaluate the performance of any given estimator $\hat{f}$. We define the following two quantities:

1. Deterministic error: $b(x) \triangleq \mathbb{E}_f \hat{f}(x) - f(x)$; this term corresponds to the bias, and is a deterministic scalar.

2. Stochastic error: $s(x) \triangleq \hat{f}(x) - \mathbb{E}_f \hat{f}(x)$; this term corresponds to the variance, and is a random variable.

With the previous definition, for the L_q loss with $1 \le q < \infty$, we can write

$$\begin{aligned}
\mathbb{E}_f \|\hat{f} - f\|_q &= \mathbb{E}_f \|\hat{f} - \mathbb{E}_f\hat{f} + \mathbb{E}_f\hat{f} - f\|_q && (3)\\
&\le \mathbb{E}_f \|\hat{f} - \mathbb{E}_f\hat{f}\|_q + \mathbb{E}_f \|\mathbb{E}_f\hat{f} - f\|_q && (4)\\
&= \mathbb{E}_f \|s\|_q + \|b\|_q && (5)\\
&\le (\mathbb{E}_f \|s\|_q^q)^{\frac{1}{q}} + \|b\|_q && (6)
\end{aligned}$$

where we have used the triangle inequality and Jensen's inequality. As a result, we can decompose the L_q risk into the bias term $\|b\|_q$ and the variance term $(\mathbb{E}_f \|s\|_q^q)^{\frac{1}{q}}$ separately.

3 Function Estimation in Hölder Balls

First we assume that the function class $\mathcal{F}$ is the Hölder ball with smoothness parameter s, and construct a sound estimator $\hat{f}$ for f in this case.

3.1 Introduction to Hölder Balls

The Hölder ball $\mathcal{H}^s(L)$ consists of all functions which are Hölder smooth of order $s > 0$, i.e., for $s = m + \alpha$ with $m \in \mathbb{N}, \alpha \in (0, 1]$,

$$\mathcal{H}^s(L) = \left\{ f \in C[0,1] : \sup_{x \neq y \in [0,1]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x - y|^\alpha} \leq L \right\}. \tag{7}$$

In other words, the Hölder ball consists of functions which are "smooth" of a certain smoothness order.

Exercise 1. Show that for $s > 1$, if we use $m = 0, \alpha = s$ in the definition (7), then the only function which satisfies the new definition is the constant function.

3.2 Kernel–based Estimator

Now we consider how to find an estimator $\hat{f}$ of f. First suppose that there is no noise, i.e., $\sigma = 0$. In this case, we know perfectly the value of f on points x_i, but do not know the value of $x \in (x_i, x_{i+1})$ in between. However, since we know that the function f is smooth, it may be tempted to use interpolation to obtain $f(x)$, e.g.,

$$\hat{f}(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} f(x_i) + \frac{x - x_i}{x_{i+1} - x_i} f(x_{i+1}). \tag{8}$$

In general, we may write

$$\hat{f}(x) = \sum_{i: x_i \in I_h(x)} w(x, x_i) f(x_i) \tag{9}$$

for some weight function $w(x, x_i)$, where $I_h(x) \triangleq [x-h, x+h]$ is a small "window" around x with bandwidth h (to be specified later). Also in the noisy case where only y_i is available, we may construct

$$\hat{f}(x) = \sum_{i: x_i \in I_h(x)} w(x, x_i) y_i. \tag{10}$$

Note that for simplicity we have ignored the boundary effect when x is close to zero or one. In other words, one may view that x lies on a torus $\mathbb{T}$ with length 1.

Now we specify our choice of the weight function. A natural requirement is that $\sum_{i: x_i \in I_h(x)} w(x, x_i) = 1$ for any x, i.e., the weights should sum into one. This requirement can be generalized in the following way: when $f(x) = x^k$ for $k = 0, \cdots, m$, the estimator in the LHS of (9) should be equal to the true value $f(x) = x^k$. In other words, we require that

$$\sum_{i: x_i \in I_h(x)} w(x, x_i) x_i^k = x^k, \qquad \forall k = 0, 1, \cdots, m. \tag{11}$$

Note that as long as $nh \to \infty$, (11) can be satisfied since we have $2nh$ degrees of freedom while there are only $m + 1$ linear constraints. The following lemma (Lemma 1.3.1 in Nemirovski notes) shows some properties of the weights:

Lemma 2. *For any $x \in [0,1]$, there exists some weight function $w(x,\cdot)$ which satisfies* (11) *and*

$$\|w(x,\cdot)\|_1 \lesssim 1, \qquad \|w(x,\cdot)\|_2 \lesssim \frac{1}{\sqrt{nh}}. \tag{12}$$

As a sanity check, just consider the case where $m = 0$ and $w(x, x_i) = \frac{1}{2nh}$ for any $x_i \in I_h(x)$.

3.3 Performance Analysis

Now we analyze the performance of the estimator in (10). By the bias–variance decomposition, it suffices to deal with the variance and the bias separately.

For the variance (stochastic error), we have

$$\begin{aligned} s(x) &= \hat{f}(x) - \mathbb{E}_f \hat{f}(x) && (13)\\ &= \sum_{i:x_i \in I_h(x)} w(x,x_i) y_i - \sum_{i:x_i \in I_h(x)} w(x,x_i)\mathbb{E} y_i && (14)\\ &= \sum_{i:x_i \in I_h(x)} w(x,x_i)(f(x_i) + \sigma\xi_i) - \sum_{i:x_i \in I_h(x)} w(x,x_i) f(x_i) && (15)\\ &= \sigma \sum_{i:x_i \in I_h(x)} w(x,x_i)\xi_i && (16)\\ &\sim \mathcal{N}(0, \sigma^2 \|w(x,\cdot)\|_2^2). && (17) \end{aligned}$$

Since for $X \sim \mathcal{N}(0,\sigma^2)$, we have $\mathbb{E}|X|^q \lesssim \sigma^q$ for any $q \in [1,\infty)$, for the stochastic error we have $\mathbb{E}|s(x)|^q \lesssim \|w(x,\cdot)\|_2^q$ (suppressing the dependence on σ), and thus

$$(\mathbb{E}\|s\|_q^q)^{\frac{1}{q}} \lesssim \|w(x,\cdot)\|_2 \lesssim \frac{1}{\sqrt{nh}} \tag{18}$$

where in the last step we have used Lemma 2.

Now we take a look at the bias. By definition,

$$b(x) = \mathbb{E}_f \hat{f}(x) - f(x) = \sum_{i:x_i \in I_h(x)} w(x,x_i) f(x_i) - f(x). \tag{19}$$

By our choice of the weights, we know that for any polynomial $P(x)$ of degree at most m, we have

$$\begin{aligned} |b(x)| &= \left| \sum_{i:x_i \in I_h(x)} w(x,x_i)(f(x_i) - P(x_i)) - (f(x) - P(x)) \right| && (20)\\ &\le (\|w(x,\cdot)\|_1 + 1) \cdot \|f - P\|_{\infty, I_h(x)} && (21)\\ &\lesssim \|f - P\|_{\infty, I_h(x)} && (22) \end{aligned}$$

where $\|f\|_{\infty,I} \triangleq \operatorname{ess\,sup}\{|f(x)| : x \in I\}$. Since this inequality holds for any P, we further have

$$|b(x)| \lesssim \inf_{P \in \mathrm{Poly}_m} \|f - P\|_{\infty, I_h(x)} \tag{23}$$

where Poly_m denotes the class of all polynomials of degree at most m. Hence, to control the bias, it suffices to find a good polynomial approximation of f in the interval $I_h(x)$.

Due to the definition of the Hölder ball, a natural choice of the polynomial is the Taylor expansion polynomial, i.e.,

$$P(y) = \sum_{k=0}^{m} \frac{f^{(k)}(x)}{k!}(y - x)^k. \tag{24}$$

By Taylor's theorem and the Lagrange remainder term, for any $y \in I_h(x)$ we have

$$|f(y) - P(y)| = \left| f(y) - \sum_{k=0}^{m} \frac{f^{(k)}(x)}{k!}(y-x)^k \right| \tag{25}$$

$$= \frac{|f^{(m)}(\xi) - f^{(m)}(x)|}{m!} \cdot |y-x|^m \tag{26}$$

$$\leq \frac{L|\xi - x|^\alpha}{m!} \cdot |y-x|^m \tag{27}$$

$$\leq \frac{Lh^\alpha}{m!} \cdot h^m \tag{28}$$

$$\lesssim h^s \tag{29}$$

where again we suppress the dependence on L. As a result, we have $|b(x)| \lesssim h^s$ for any $x \in [0,1]$, and thus

$$\|b\|_q \lesssim h^s. \tag{30}$$

Combining the bias and variance together, we have

$$\sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \|\hat{f} - f\|_q \lesssim \frac{1}{\sqrt{nh}} + h^s. \tag{31}$$

Setting these two terms to be equal yields $h \asymp n^{-\frac{1}{2s+1}}$, and thus

$$\sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \|\hat{f} - f\|_q \lesssim n^{-\frac{s}{2s+1}}. \tag{32}$$

In fact, this bound turns out to be optimal:

Theorem 3. *For any $s > 0$ and $q \in [1, \infty)$, the minimax L_q risk in estimating f is given by*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \|\hat{f} - f\|_q \asymp n^{-\frac{s}{2s+1}}. \tag{33}$$

Moreover, the kernel-based estimator in (10) *with bandwidth $h \asymp n^{-\frac{1}{2s+1}}$ attains the minimax risk within multiplicative constants.*

3.4 Discussions

We briefly discuss the previous result. The most important point is on the bias–variance tradeoff: when the bandwidth h increases, the deterministic error $b(x)$ gets larger and the stochastic error $s(x)$ gets smaller. Specifically, we have

$$(\mathbb{E}|s(x)|^q)^{\frac{1}{q}} \lesssim \frac{1}{\sqrt{nh}}, \tag{34}$$

$$|b(x)| \lesssim \inf_{P \in \mathrm{Poly}_m} \|f - P\|_{\infty, I_h(x)}. \tag{35}$$

The reason is also straightforward intuitively: as the bandwidth h increases, the error incurred by using $f(x \pm h)$ to approximate $f(x)$ gets larger, i.e., the bias gets larger. In contrast, the number of the points to be averaged over (i.e., $2nh$) gets larger, which yields to a smaller variance. Finally, we need to choose a suitable bandwidth h^* to balance the bias and variance, as is shown in the following figure 1.

The second point to note is that the overall estimator $\hat{f}$ is linear in the observations $y_1, \cdots, y_n$. This is called a *linear* estimator. Theorem 3 shows that, for nonparametric function estimation over Hölder balls, linear estimator can attain the minimax risk within constants. Then a natural question arises: can linear

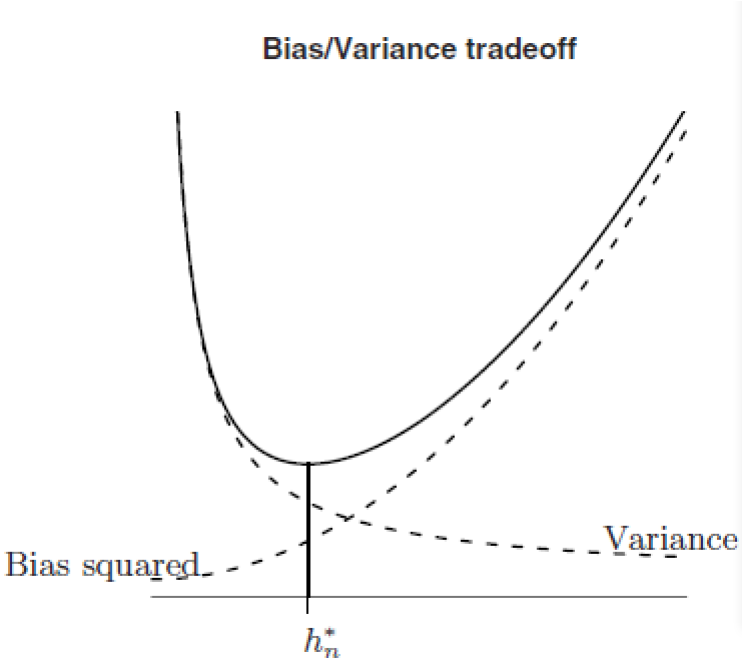

Figure 1: The bias-variance tradeoff.

approaches always attain the minimax risk in nonparametric function estimation for other natural function classes $\mathcal{F}$? We will show in the next section that the answer is *no* in general.

The third point is more technical and high-level: in choosing the weight function, we require that our weight should keep all polynomials of degree at most m. Some more thoughts lead to the following question: why are polynomials so special here? Can we choose other basis functions? To answer this question, we need to introduce the Kolmogorov n-width:

Definition 4 (Kolmogorov n-width)**.** *Let $(X, \|\cdot\|)$ be a normed vector space, and $K \subset X$ be a compact subset. The* Kolmogorov n-width *of K is defined by*

$$d_n(K) \triangleq d_n(K, \|\cdot\|) \triangleq \inf_{V \subset X} \sup_{x \in K} \inf_{y \in V} \|x - y\| \tag{36}$$

where the first infimum is taken over all linear subspaces $V \subset X$ of dimension at most n.

The following lemma shows the Kolmogorov n-width of the Hölder ball:

Lemma 5. *For the Hölder ball $\mathcal{H}^s(L)$ with smoothness parameter $s > 0$, its Kolmogorov n-width is*

$$d_n(\mathcal{H}^s(L)) \asymp Ln^{-s} \tag{37}$$

and is attained by the polynomial basis $V = span\{1, x, \cdots, x^{n-1}\}$.

Now we take a look at the bias (35), which shows that the bias is determined by the approximation error of the basis functions we choose. As a result, we should choose the basis functions which attain the Kolmogorov n-width, and by the previous lemma we know that polynomial basis is the right basis to use here. In general, we would like to remark another important point here: **the choice of basis is very important!** We will see more examples in the next lecture.

4 Function Estimation in Sobolev Balls

Next we consider the same regression problem in Sobolev balls, which is a generalization of the Hölder ball in the sense that some spatial inhomogeneity is allowed here. We will show that in certain scenarios, any linear approach fails to give the optimal minimax risk.

4.1 Introduction to Sobolev Balls

Sobolev space is another space which naturally measures the smoothness of a function, and plays an important role in approximation theory and partial differential equations. Specifically, the (1D) Sobolev ball $\mathcal{W}_1^{k,p}(L)$ consists of all functions f on $[0,1]$ such that

$$\|f^{(k)}\|_p \le L \tag{38}$$

where $k > 0$ is an integer, and $p \in [1, \infty]$. Technically, we remark that the derivative is defined in terms of distributions (not the classical derivative) and exists for any function. Moreover, to ensure a continuous embedding $\mathcal{W}_1^{k,p}(L) \subset C[0,1]$, we need an additional assumption $k > \frac{1}{p}$. To see a counterexample, consider $f(x) = \mathbb{1}(x \ge \frac{1}{2})$, we have $f'(x) = \delta(x - \frac{1}{2})$ with $\|f'\|_1 = 1$, so $f \in \mathcal{W}_1^{1,1}(1)$ but f is not continuous. We also note that the "effective" smoothness s of $\mathcal{W}_1^{k,p}(L)$ will become

$$s = k - \frac{1}{p}. \tag{39}$$

4.2 Performance Analysis of Linear Approaches

Let's consider the performance of the linear approach in (10), where the weight function keeps all polynomial of degree at most $k-1$. By the bias-variance tradeoff, it suffices to deal with these two quantities separately.

For the variance, (34) does not depend on the specific function class $\mathcal{F}$, so we still have

$$(\mathbb{E}\|s\|_q^q)^{\frac{1}{q}} \lesssim \frac{1}{\sqrt{nh}}. \tag{40}$$

For the bias, the polynomial approximation error incurred by the Taylor expansion polynomial will change for the Sobolev ball. In fact, by the same argument, for any $y \in I_h(x)$ we have

$$\begin{aligned}
|f(y) - P(y)| &\lesssim h^{k-1} \cdot |f^{(k-1)}(\xi) - f^{(k-1)}(x)| && (41)\\
&\le h^{k-1} \int_x^\xi |f^{(k)}(z)| dz && (42)\\
&\le h^{k-1} \left(\int_x^\xi |f^{(k)}(z)|^p dz \right)^{\frac{1}{p}} \left(\int_x^\xi 1 dz \right)^{1-\frac{1}{p}} && (43)\\
&\le h^{k-\frac{1}{p}} \cdot \left(\int_{I_h(x)} |f^{(k)}(z)|^p dz \right)^{\frac{1}{p}}. && (44)
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\|b\|_q^q &\lesssim \int_0^1 \left(h^{k-\frac{1}{p}} \cdot \left(\int_{I_h(x)} |f^{(k)}(z)|^p dz \right)^{\frac{1}{p}} \right)^q dx && (45)\\
&= h^{(k-\frac{1}{p})q} \cdot \int_0^1 \left(\int_{I_h(x)} |f^{(k)}(z)|^p dz \right)^{\frac{q}{p}} dx. && (46)
\end{aligned}$$

To upper bound this quantity, we need to employ the Sobolev ball condition. Note that without the exponent q/p, we have

$$\int_0^1 \left(\int_{I_h(x)} |f^{(k)}(z)|^p dz \right) dx = \iint_{|x-z|\le h} |f^{(k)}(z)|^p dx dz \le 2h \int_0^1 |f^{(k)}(z)|^p dz \le 2hL^p \lesssim h. \tag{47}$$

In other words, we would like to find the maximum of some integral with exponent q/p given that this integral without the exponent is a fixed constant. Define

$$g(x) = \int_{I_h(x)} |f^{(k)}(z)|^p dz,$$

then for every $x \in [0, 2h]$ we have

$$g(x) + g(x+2h) + \ldots + g(x+1-2h) = \int_0^1 |f^{(k)}(z)|^p dz \le L^p. \tag{48}$$

The rest follows once we recall the following fact:

Exercise 6. For non-negative reals $a_1, \cdots, a_n$, show that

$$a_1^r + a_2^r + \cdots + a_n^r \le \begin{cases} n^{1-r}(a_1 + a_2 + \cdots + a_n)^r & \text{if } r \in [0,1], \\ (a_1 + a_2 + \cdots + a_n)^r & \text{if } r > 1. \end{cases} \tag{49}$$

Using this fact (and after some analysis), we can obtain that

$$\|b\|_q^q \lesssim \begin{cases} h^{kq} & \text{if } q \le p, \\ h^{(k-\frac{1}{p}+\frac{1}{q})q} & \text{if } p < q < \infty. \end{cases} \tag{50}$$

Now upon choosing the optimal bandwidth h, we arrive at

$$\sup_{f \in \mathcal{W}_1^{k,p}(L)} \mathbb{E}_f \|\hat{f} - f\|_q \lesssim \begin{cases} n^{-\frac{k}{2k+1}} & \text{if } q \le p, \\ n^{-\frac{k-\frac{1}{p}+\frac{1}{q}}{2(k-\frac{1}{p}+\frac{1}{q})+1}} & \text{if } p < q < \infty. \end{cases} \tag{51}$$

It turns out that this is the best performance we can hope for using linear approaches:

Theorem 7. *For any $p \in [1, \infty], k > \frac{1}{p}$ and $q \in [1, \infty)$, the minimax* linear *risk in estimating f over Sobolev ball $\mathcal{W}_1^{k,p}(L)$ is given by*

$$\inf_{\hat{f}^{lin}} \sup_{f \in \mathcal{W}_1^{k,p}(L)} \mathbb{E}_f \|\hat{f}^{lin} - f\|_q \asymp \begin{cases} n^{-\frac{k}{2k+1}} & \text{if } q \le p, \\ n^{-\frac{k-\frac{1}{p}+\frac{1}{q}}{2(k-\frac{1}{p}+\frac{1}{q})+1}} & \text{if } p < q < \infty. \end{cases} \tag{52}$$

where the infimum is taken over all linear estimators.

4.3 Optimal Performance

One may wonder whether the performance of the best linear estimator in Theorem 7 is optimal even if we allow for non-linear estimators. The answer is yes when $q \le p$, and is no otherwise. This can be seen intuitively with the help of the previous exercise:

1. When $q \le p$, the bias-maximizing function f is non-zero everywhere in $[0,1]$. In this case, the function f is almost homogeneous, and it is fine to use the same bandwidth h everywhere. This is called the *"regular" regime*.

2. When $p < q < \infty$, the bias-maximizing function f is only supported on some interval of length h, and vanishes outside. As a result, if we knew this supporting interval, we can simply set the bandwidth outside this interval to be $\Theta(1)$ (otherwise we are accumulating noise but no signal!), and the variance becomes

$$(\mathbb{E}\|s\|_q^q)^{\frac{1}{q}} \lesssim \frac{1}{\sqrt{nh}} \cdot h^{\frac{1}{q}}, \tag{53}$$

which is smaller than the case where we use the same bandwidth everywhere. Combining the bias bound $\|b\|_q \lesssim h^{k-\frac{1}{p}+\frac{1}{q}}$, the optimal bandwidth will result in the error

$$\sup_{f \in \mathcal{W}_1^{k,p}(L)} \mathbb{E}_f \|\hat{f} - f\|_q \lesssim n^{-\frac{k-\frac{1}{p}+\frac{1}{q}}{2(k-\frac{1}{p})+1}}. \tag{54}$$

This is called the *"sparse" regime*.

The previous insights explain the reason why linear approaches cannot always obtain the minimax risk, and remark that different bandwidths should be used at different areas. The exact minimax risk is given in the following theorem:

Theorem 8. *For any $p \in [1,\infty], k > \frac{1}{p}$ and $q \in [1,\infty]$, the minimax risk in estimating f over Sobolev ball $\mathcal{W}_1^{k,p}(L)$ is given by*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}_1^{k,p}(L)} \mathbb{E}_f \|\hat{f} - f\|_q \asymp \begin{cases} n^{-\frac{k}{2k+1}} & \text{if } q \le (2k+1)p, \\ (\frac{n}{\log n})^{-\frac{k-\frac{1}{p}+\frac{1}{q}}{2(k-\frac{1}{p})+1}} & \text{if } (2k+1)p < q \le \infty. \end{cases} \tag{55}$$

where the infimum is taken over all possible estimators.

5 Adaptive Bandwidth: Lepski's Trick

As is shown in the previous section, we need an adaptive bandwidth when:

1. there is spatial homogeneity: the function f may behave differently at different points, and we should choose a suitable bandwidth to achieve the bias-variance tradeoff "locally";

2. some parameters, e.g., s, k, p, are unknown: previously our choice of the optimal bandwidth relies on the parameters s, k, p. If these parameters are unknown, we need to select a good bandwidth based on empirical data we have seen.

Let's first take a look at what we have already known. For any $x \in [0,1]$, consider the linear estimator $\hat{f}_h$ in (10) with bandwidth $h \equiv h(x)$ to estimate $f(x)$. Denote by $b_h(x), s_h(x)$ the corresponding bias and variance of this estimator with bandwidth h, then we would like to find some optimal bandwidth $h^* \equiv h^*(x)$ to balance the bias and variance[1]:

$$b_{h^*}(x) = s_{h^*}(x). \tag{56}$$

[1]strictly speaking, $b_h(x), s_h(x)$ are not the exact deterministic and stochastic errors, but suitable upper bounds (e.g., (35), (34)) of these quantities.

The problem is that we do not know anything about the LHS other than the monotonicity of $b_h(x)$ in h, and for the RHS we also only know that $|s_h(x)| \lesssim \sqrt{\frac{\log n}{nh}}$ with overwhelming probability. Let's just define

$$s_h(x) \triangleq \Theta\left(\sqrt{\frac{\log n}{nh}}\right). \tag{57}$$

5.1 Lepski's Selection Rule

Surprisingly, the monotonicity of the bias suffices to provide the knowledge to choose the bandwidth adaptively. Here is Lepski's selection rule: for any $x \in [0,1]$, define the set[2]

$$A(x) \triangleq \{h \geq 0 : |\hat{f}_h(x) - \hat{f}_{h'}(x)| \leq 4s_{h'}(x), \quad \forall h' \in (0,h)\} \tag{58}$$

consisting of "admissible" bandwidths, and then choose the bandwidth $\hat{h}(x)$ to be

$$\hat{h}(x) \triangleq \max A(x). \tag{59}$$

Finally, the overall estimator $\hat{f}^{\text{Lep}}$ is given by

$$\hat{f}^{\text{Lep}}(x) = \hat{f}_{\hat{h}(x)}(x). \tag{60}$$

Note that we only add a little bit "non-linearity" here: we still use a kernel-based estimator at any point, but the bandwidth differs spatially. Moreover, the overall estimator does not require the knowledge of any parameters (except for an *upper bound* on the smoothness parameter, for we need to fix a weight function (kernel) to keep all polynomials of degree up to the smoothness parameter).

5.2 Intuitions of Lepski's Trick

Let's explain intuitively why Lepski's trick will work, i.e., we will have

$$|\hat{f}_{\hat{h}}(x) - f(x)| \lesssim |\hat{f}_{h^*}(x) - f(x)|. \tag{61}$$

The analysis decomposes into two steps:

1. We first show that $h^* \in A(x)$. It suffices to check the definition: when $h < h^*$, we have

$$\begin{aligned}
|\hat{f}_{h^*}(x) - \hat{f}_h(x)| &\leq |\hat{f}_{h^*}(x) - f(x)| + |\hat{f}_h(x) - f(x)| && (62)\\
&\leq b_{h^*}(x) + s_{h^*}(x) + b_h(x) + s_h(x) && (63)\\
&\leq s_{h^*}(x) + s_{h^*}(x) + s_h(x) + s_h(x) && (64)\\
&\leq s_h(x) + s_h(x) + s_h(x) + s_h(x) && (65)\\
&= 4s_h(x). && (66)
\end{aligned}$$

   Hence, h^* satisfies the condition, and $h^* \in A(x)$.

2. Next we show that $\hat{h}$ indeed works. Since $\hat{h} = \max A(x)$ and $h^* \in A(x)$, we have $\hat{h} \geq h^*$, and by the definition of $A(x)$ we know that

$$\begin{aligned}
|\hat{f}_{\hat{h}}(x) - f(x)| &\leq |\hat{f}_{\hat{h}}(x) - \hat{f}_{h^*}(x)| + |\hat{f}_{h^*}(x) - f(x)| && (67)\\
&\leq 4s_{h^*}(x) + b_{h^*}(x) + s_{h^*}(x) && (68)\\
&\leq 6s_{h^*}(x) && (69)
\end{aligned}$$

   as desired.

[2] strictly speaking, we should require that h' take value in a finite fine grid of $[0,1]$ instead of the continuum.

5.3 Performance Analysis of Lepski's Trick

The performance of Lepski's estimator $\hat{f}^{\text{Lep}}$ is summarized in the following theorem:

Theorem 9. *For any $p \in [1, \infty], k > \frac{1}{p}$ and $q \in [1, \infty]$, the Lepski's estimator $\hat{f}^{\text{Lep}}$ satisfies*

$$\sup_{f \in \mathcal{W}_1^{k,p}(L)} \mathbb{E}_f \|\hat{f}^{\text{Lep}} - f\|_q \asymp \begin{cases} (\frac{n}{\log n})^{-\frac{k}{2k+1}} & \text{if } q \leq (2k+1)p, \\ (\frac{n}{\log n})^{-\frac{k-\frac{1}{p}+\frac{1}{q}}{2(k-\frac{1}{p})+1}} & \text{if } (2k+1)p < q \leq \infty. \end{cases} \tag{70}$$

Compared with Theorem 8, we see that this estimator is optimal in estimating function over Sobolev balls within logarithmic factors. In fact, this logarithmic factor is unavoidable if we want our estimator to be adaptive for any smoothness parameters and any not-too-small subset of $[0, 1]$, and is the price we need to pay for adaptivity.