SDS8102: Advanced Mathematical Statistics

Week 13 – Lecture 1
November 24, 2026
Lecturer: Alexandre Tsybakov

Definition 4. *Let $x \in \mathbb{R}$ be fixed, and let $\mathcal{P}$ be a class of denisties on $\mathbb{R}$. A sequence $\{\psi_n\}$, $\psi_n > 0$, is called an* **optimal rate of convergence** *of MSE on the class $\mathcal{P}$ if the following two conditions are satisfied:*

(i) $\inf_{T_n} \sup_{p \in \mathcal{P}} \mathbb{E}_p[(T_n(x) - p(x))^2] \geq c\psi_n^2$, *where $c > 0$ is a constant independent of n.*

(ii) There exist an estimator $p_n(\cdot)$, and a constant $C > 0$ independent of n such that

$$\sup_{p \in \mathcal{P}} \mathbb{E}_p[(p_n(x) - p(x))^2] \leq C\psi_n^2.$$

If (i) and (ii) hold, p_n is called a **rate optimal estimator** *for MSE on the class $\mathcal{P}$.*

Corollary 2. *Let $\beta > 0$, $L > 0$. The kernel density estimator with bandwidth $h = \alpha n^{-\frac{1}{2\beta+1}}$, $\alpha > 0$, and kernel of order $l = \lfloor \beta \rfloor$ is rate optimal for MSE on the Hölder class $\mathcal{P}(\beta, L)$, and $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ is corresponding optimal rate.*

Remark: We only have established lower bound for pointwise risk. One can show that the rate is optimal also in L_2 sense (minimax $MISE$):

$$\inf_{T_n} \sup_{p \in \mathcal{P}(\beta,L)} \mathbb{E}_p \int (T_n(x) - p(x))^2 dx \geq cn^{-\frac{2\beta}{2\beta+1}}.$$

However, the proof of this lower bound is more complicated; the construction based on two densities will not be enough.

Summary. We have seen that the following issues play the key role in nonparametric estimation.

- *Bias-variance trade-off.* For nonparametric estimation, the bias is not negligible, which brings in the problem of optimal choice of the smoothing parameter. For the kernel density estimator, the smoothing parameter is the bandwidth.
- *Optimality in minimax sense.* Is the upper bound obtained from bias-variance trade-off indeed optimal? We need minimax lower bounds to answer this question.
- *Adaptation.* What is the optimal data-driven choice of smoothing parameter? Adaptive estimator is an estimator which is rate optimal on a large scale of classes without any knowledge about the parameters of the classes. Cross-validation is an example of successful adaptation procedure.

9 The problem of nonparametric regression

One can distinguish between two types of problems: regression with random design and regression with fixed design.

- Regression with random design.

  Let (X, Y) be a pair of random variables with $x \in \mathcal{X}$ and $Y \in \mathbb{R}$. Assume that the conditional expectation $f(x) = \mathbb{E}(Y|X = x)$ exists. The function $f : \mathcal{X} \to \mathbb{R}$ is the regression function. Suppose that we know that f is in a wide (nonparametric) class of functions $\mathcal{F}$. We observe a sample of i.i.d. pairs $(X_1, Y_1), ..., (X_n, Y_n)$ that have the same distibution as (X, Y). The problem is to estimate f. Note that we have $Y_i = f(X_i) + \xi_i$, where $\mathbb{E}(\xi_i) = 0$.

- Regression with fixed design.

  We observe independent pairs $(X_1, Y_1), ..., (X_n, Y_n)$, where $Y_i = f(X_i) + \xi_i$, the X_i are deterministic and $\mathbb{E}\xi_i = 0$. Suppose that we know that f is in a wide (nonparametric) class of functions $\mathcal{F}$. The problem is to estimate f.

In what follows, we discuss the following two nonparametric methods of estimation of f:

- Local polynomial estimators.
- Spline estimators (penalized least squares).

10 Local polynomial estimators

Suppose that the regression function $f : \mathbb{R} \to \mathbb{R}$ is in $\Sigma(\beta, L)$. Let $l = \lfloor \beta \rfloor$. We want to estimate f at point $x \in \mathbb{R}$. For z close to x ($|x - z| \le h$ and for small h), do a Taylor expansion of f around x:

$$f(z) \approx f(x) + f'(x)(z - x) + ... + \frac{f^{(l)}(x)}{l!}(z - x)^l = \theta(x)^T U\Big(\frac{z - x}{h}\Big),$$

where

$$\begin{aligned} \theta(x) &= (f(x), f'(x)h, ..., f^{(l)}(x)h^l)^T, \\ U(u) &= (1, u, u^2/2, ..., u^l/l!)^T. \end{aligned}$$

Estimating the $l + 1$ coefficients of $\theta(x)$ based on the data, we can get the estimators of $f(x)$ and of the derivatives of f at point x.

Given (X_i, Y_i)'s, we consider Least Squares estimation based on the data in a neighborhood of x, that is, we minimize

$$\sum_{i:|X_i - x| \le h} \Big(Y_i - \theta^T U\Big(\frac{X_i - x}{h}\Big)\Big)^2 \tag{1}$$

over all $\theta \in \mathbb{R}^{l+1}$. More generally, consider the estimator

$$\hat{\theta}_n(x) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^{n} \Big(Y_i - \theta^T U\Big(\frac{X_i - x}{h}\Big)\Big)^2 K\Big(\frac{X_i - x}{h}\Big), \tag{2}$$

where $K : \mathbb{R} \to \mathbb{R}$ is some kernel. The case in (1) corresponds to taking K as the rectangular kernel.

The estimator of $f(x)$ defined as

$$\hat{f}_n^{LP}(x) = U^T(0)\hat{\theta}_n(x)$$

is called the *local polynomial estimator of order* l of $f(x)$ (or shortly the LP(l) estimator). Notice that it is equal to the first coordinate of $\hat{\theta}_n(x)$.

A famous special case is the LP(0) estimator. Write for brevity $U_i = U\big(\frac{X_i - x}{h}\big), K_i = K\big(\frac{X_i - x}{h}\big)$. If $l = 0$ the problem of minimization in (2) can be written as

$$\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K_i.$$

Assume that $K_i \geq 0$. The minimizer θ must satisfy

$$\sum_{i=1}^{n} (Y_i - \theta) K_i = 0.$$

Solving this equation with respect to θ we obtain that the LP(0) estimator of $f(x)$ has the form

$$\hat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K_i}{\sum_{i=1}^n K_i} = \frac{\frac{1}{nh}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})}{\frac{1}{nh}\sum_{i=1}^n K(\frac{X_i - x}{h})}.$$

This estimator is called the *Nadaraya-Watson estimator* (proposed independently by Nadaraya (1964)[1] and Watson (1964)[2]). However, its performance is not completely satisfactory since one can show that it does not work efficiently for functions f with higher smoothness β.

Note that the denominator in the definition of $\hat{f}_n^{NW}$ is an estimator of the marginal density of X. If the X_i are random and uniformly distributed on $[0, 1]$, or $X_i = i/n, i = 1, ..., n$, then we can replace the denominator by a constant 1, which leads to the estimator

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

Definition 5. *An estimator $\hat{f}_n(x)$ of $f(x)$ is called linear if it has the form*

$$\hat{f}_n(x) = \sum_{i=1}^{n} W_{ni}(x, X_1, X_2, ..., X_n) Y_i.$$

In this definition, the weight W_{ni} depends only on $(x, X_1, X_2, ..., X_n)$. We will use a shorter writing $W_{ni}(x)$ instead of $W_{ni}(x, X_1, X_2, ..., X_n)$. Note that the Nadaraya-Watson estimator is linear with weights

$$W_{ni}^{NW}(x) = \frac{K_i}{\sum_{i=1}^n K_i}, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}^{NW}(x) = 1.$$

[1] Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and Its Applications*, 9, 141–142.
[2] Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā*, Series A, 26, 359–372.

Proposition 3. *Let the matrix*

$$\mathcal{B}_x = \frac{1}{nh}\sum_{i=1}^{n} U_i U_i^T K_i$$

be positive definite for given x. Then $\hat{f}_n^{LP}(x)$ is a linear estimator with weights

$$W_{ni}^{LP}(x) = \frac{1}{nh} U^T(0)\mathcal{B}_x^{-1} U_i K_i.$$

Proof.

$$\begin{aligned}\hat{\theta}_n(x) &= \underset{\theta\in\mathbb{R}^{l+1}}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - \theta^T U_i)^2 K_i \\ &= \underset{\theta\in\mathbb{R}^{l+1}}{\operatorname{argmin}} \Big[\sum_{i=1}^{n} Y_i^2 K_i - 2\theta^T \sum_{i=1}^{n} Y_i U_i K_i + \sum_{i=1}^{n}(\theta^T U_i)^2 K_i\Big].\end{aligned}$$

We now drop the term $\sum_{i=1}^{n} Y_i^2 K_i$ since it does not involve θ. Set $a = \frac{1}{nh}\sum_{i=1}^{n} Y_i U_i K_i$. Then we have

$$\hat{\theta}_n(x) = \underset{\theta\in\mathbb{R}^{l+1}}{\operatorname{argmin}} [-2\theta^T a + \theta^T \mathcal{B}_x \theta].$$

The minimum is achieved when $\mathcal{B}_x\theta = a$. Since $\mathcal{B}_x$ is positive definite,

$$\hat{\theta}_n(x) = \mathcal{B}_x^{-1} a.$$

Then, by definition, the LP(l) estimator of $f(x)$ is

$$\hat{f}_n^{LP}(x) = U^T(0)\hat{\theta}_n(x) = \frac{1}{nh} U^T(0)\mathcal{B}_x^{-1}\sum_{i=1}^{n} Y_i U_i K_i.$$

Defining the weights as $W_{ni}^{LP}(x) = \frac{1}{nh} U^T(0)\mathcal{B}_x^{-1} U_i K_i$ we obtain the proposition. □

Assume that there is no noise, and we observe $Y_i = Q(X_i), i = 1, ..., n$, where Q is a polynomial and $\deg(Q) \le l$. We have the following property of reproduction of the polynomial Q.

Proposition 4. *If the matrix $\mathcal{B}_x$ is positive definite then the LP(l) estimator $\hat{f}_n^{LP}(x)$ reproduces the polynomials of degree less than or equal to l, that is,*

$$\sum_{i=1}^{n} W_{ni}^{LP}(x) Q(X_i) = Q(x) \tag{3}$$

for any polynomial Q with $\deg(Q) \le l$.

Proof. Taylor expansion of Q gives

$$Y_i = Q(X_i) = Q(x) + Q'(x)(X_i - x) + \cdots + \frac{Q^{(l)}(x)}{l!}(X_i - x)^l.$$

Introduce the notation $q(x) = (Q(x), Q'(x), ..., Q^{(l)}(x))^T$. Then, by definition (2) we have

$$\hat{\theta}_n(x) \in \underset{\theta\in\mathbb{R}^{l+1}}{\operatorname{argmin}} \sum_{i=1}^{n}(Q(X_i) - \theta^T U_i)^2 K_i.$$

Here,

$$\sum_{i=1}^{n}(Q(X_i)-\theta^T U_i)^2 K_i = \sum_{i=1}^{n}[(q(x)-\theta)^T U_i]^2 K_i = (q(x)-\theta)^T \mathcal{B}_x (q(x)-\theta).$$

Since $\mathcal{B}_x$ is positive definite the unique minimizer over θ of this quadratic form is $q(x)$. It means that $\hat{\theta}_n(x) = q(x)$. By definition, $\hat{f}_n^{LP}(x)$ is equal to the first coordinate of $\hat{\theta}_n(x)$, so $\hat{f}_n^{LP}(x) = Q(x)$. Finally, we invoke Proposition 3 to obtain (3). □

Remark. Solving the minimization problem (2) also provides estimators of the derivatives of f. Thus, the LP(l) estimator of the jth derivative $f^{(j)}(x)$ has the form

$$\hat{f}_{n,j}^{LP}(x) = h^{-j} U^{(j)}(0)^T \hat{\theta}_n(x), \ j \le l.$$

Note that the jth coordinate of $U^{(j)}$ is 1 and all other coordinates are 0.

Corollary 3. *If the matrix $\mathcal{B}_x$ is positive definite then the weights of the LP(l) estimator satisfy:*

$$\sum_{i=1}^{n} W_{ni}^{LP}(x) = 1, \quad \sum_{i=1}^{n} W_{ni}^{LP}(x)(X_i - x)^j = 0, \quad j = 1, ..., l.$$

Proof. From Proposition 4 we have that, for any polynomial Q of degree at most l,

$$\sum_{i=1}^{n} W_{ni}^{LP}(x) Q(X_i) = Q(x).$$

Consider the polynomials $Q(t) \equiv 1$ and $Q(t) = (t-x)^j$ to obtain the result. □

11 Mean squared error of local polynomial estimator

In this section we study statistical properties of the LP(ℓ) estimator constructed from observations (X_i, Y_i), $i = 1, \dots, n$, such that

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \dots, n, \tag{4}$$

where ξ_i are independent zero mean random variables ($\mathbb{E}(\xi_i) = 0$), the X_i are deterministic values belonging to $[0,1]$, and f is a function from $[0,1]$ to $\mathbb{R}$.

In this section, we fix $x \in [0,1]$. Let $\hat{f}_n(x)$ be an LP(ℓ) estimator of $f(x)$ at point x. The bias and the variance of $\hat{f}_n(x)$ are given by the formulas

$$b(x) = \mathbb{E}\left[\hat{f}_n(x)\right] - f(x), \qquad \sigma^2(x) = \mathbb{E}\left[\hat{f}_n^2(x)\right] - \left(\mathbb{E}\left[\hat{f}_n(x)\right]\right)^2,$$

respectively, where $\mathbb{E}$ denotes the expectation with respect to the distribution of the random vector $(Y_1, \dots, Y_n)$ whose coordinates satisfy (4). The mean squared risk of the estimator $\hat{f}_n(x)$ is

$$\mathrm{MSE} = \mathrm{MSE}(x) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] = b^2(x) + \sigma^2(x).$$

We will study separately the bias and the variance terms in this representation of the risk. First, we introduce the following assumptions.

Assumptions (LP)

(LP1) *There exist a real number* $\lambda_0 > 0$ *and a positive integer* n_0 *such that the smallest eigenvalue* $\lambda_{\min}(\mathcal{B}_x)$ *of* $\mathcal{B}_x$ *satisfies*

$$\lambda_{\min}(\mathcal{B}_x) \geq \lambda_0$$

for all $n \geq n_0$.

(LP2) *There exists a real number* $a_0 > 0$ *such that for any interval* $A \subseteq [0,1]$ *and all* $n \geq 1$

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \leq a_0 \max(\mathrm{Leb}(A), 1/n)$$

where $\mathrm{Leb}(A)$ *denotes the Lebesgue measure of* A.

(LP3) *The support of kernel* K *is included in* $[-1,1]$ *and there exists a number* $K_{\max} < \infty$ *such that* $|K(u)| \leq K_{\max},\ \forall\, u \in \mathbb{R}$.

Assumption (LP1) is stronger than the condition that matrix $\mathcal{B}_x$ is positive definite introduced in the previous section since now we require uniformity with respect to n. This assumption is natural in the case where the matrix $\mathcal{B}_x$ converges to a limit as $n \to \infty$. Assumption (LP2) means that the points X_i are dense enough in the interval $[0,1]$. It holds for a sufficiently wide range of designs. An important example is given by the regular design: $X_i = i/n$, for which (LP2) is satisfied with $a_0 = 2$. Finally, assumption (LP3) is not restrictive since the choice of K belongs to the statistician.

Since the matrix $\mathcal{B}_x$ is symmetric, assumption (LP1) implies that, for all $n \geq n_0$ and $v \in \mathbb{R}^{\ell+1}$,

$$\|\mathcal{B}_x^{-1} v\| \leq \|v\|/\lambda_0, \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{\ell+1}$.

Lemma 4. *Let Assumptions (LP1) – (LP3) hold. Then, for all* $n \geq n_0$ *and* $h \geq 1/(2n)$, *the weights* $W_{ni}(x) = W_{ni}^{LP}(x)$ *of the* $\mathrm{LP}(\ell)$ *estimator are such that:*

(i) $\max_{1\leq i\leq n} |W_{ni}(x)| \leq \frac{C_*}{nh}$;

(ii) $\sum_{i=1}^{n} |W_{ni}(x)| \leq C_*$;

(iii) $W_{ni}(x) = 0 \quad \text{if} \quad |X_i - x| > h,$

where the constant C_* *depends only on* λ_0, a_0, *and* $K_{\max}$.

Proof. (i) Using (5) and the fact that $\|U(0)\| = 1$ we obtain:

$$\begin{aligned}
|W_{ni}(x)| &\leq \frac{1}{nh}\left\|\mathcal{B}_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right\| \\
&\leq \frac{1}{nh\lambda_0}\left\|U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right\| \\
&\leq \frac{K_{\max}}{nh\lambda_0}\left\|U\left(\frac{X_i - x}{h}\right)\right\| I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right) \\
&\leq \frac{K_{\max}}{nh\lambda_0}\sqrt{1 + 1 + \frac{1}{(2!)^2} + \cdots + \frac{1}{(\ell!)^2}} \leq \frac{2K_{\max}}{nh\lambda_0}.
\end{aligned}$$

(ii) In a similar way, by (LP2), we have

$$\begin{aligned}\sum_{i=1}^{n}|W_{ni}(x)| &\leq \frac{K_{\max}}{nh\lambda_0}\sum_{i=1}^{n}\left\|U\left(\frac{X_i-x}{h}\right)\right\| \; I\left(\left|\frac{X_i-x}{h}\right|\leq 1\right)\\ &\leq \frac{2K_{\max}}{nh\lambda_0}\sum_{i=1}^{n}I(x-h\leq X_i\leq x+h)\\ &\leq \frac{2K_{\max}}{nh\lambda_0}\max\left(2,\frac{1}{nh}\right)\leq\frac{4K_{\max}a_0}{\lambda_0}.\end{aligned}$$

To complete the proof, we take $C_* = \max\{2K_{\max}/\lambda_0, 4K_{\max}a_0/\lambda_0\}$ and observe that (iii) follows from the fact that the support of K is contained in $[-1,1]$.