CS861: Theoretical Foundations of Machine Learning Lecture 14 - 10/06/2023
University of Wisconsin–Madison, Fall 2023

Lecture 14: Nonparametric Regression and Density Estimation

Lecturer: Kirthe[...] *Scribed by: Haoran Xiong, Zhihao Zhao*

Disclaimer: *These n[...]ected to the usual scrutiny reserved for formal publications. They may be distribute[...] with the permission of the instructor.*

In this lecture, we will develop upper and lower bounds for nonparametric regression and show that the minimax rate is $\tilde{\Theta}(n^{-2/3})$. We will also briefly introduce nonparametric density estimation.

1 Nonparametric Regression

Assume that we observe dataset $S = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ i.i.d drawn from some distribution $P_{XY} \in \mathcal{P}$, where

$$\begin{aligned}\mathcal{P} = \{P_{XY};\ & 0 < \alpha_0 \leq p(x) \leq \alpha_1 < \infty, \\ & f(x) = \mathbb{E}[Y|X = x] \text{ is L-Lipschitz}, \\ & \mathrm{Var}(Y|X = x) \leq \sigma^2\},\end{aligned}$$

in which $p(x)$ is the marginal density of X.

Our target regression function will be estimated via the following loss:

$$\ell(P_{XY}, g) \triangleq \int (f(x) - g(x))^2 p(x) dx,$$

where $p(x)$ and $f(x) = \mathbb{E}[Y|X = x]$ are defined above.

Then the minimax risk is defined as follows:

$$\begin{aligned} R_n^* &= \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{S \sim P_{XY}} \left[\ell(P_{XY}, \hat{f}) \right] \\ &= \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{S \sim P_{XY}} \left[\int (f(x) - \hat{f}(x))^2 p(x) dx \right] \end{aligned}$$

We want to show that the R_n^* is $\Theta(n^{-2/3})$ in two steps:

1. Establish a lower bound with Fano's method
2. Get an upper bound by using Nadaraya-Watson Estimation

1.1 Lower bound

By noticing that $\ell(P_{XY}, g)$ defined above cannot be written into the form of $\ell = \Phi \circ \rho$, which means we cannot utilize theorems and lemmas learnt in previous lectures, we circumvent this problem by constructing a sub-class $\mathcal{P}'$ of $\mathcal{P}$ as follows:

$$\mathcal{P}'' = \{P_{XY} \in \mathcal{P};\ p(x) = 1\}$$ [1]

Then $R_n^* \geq \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}''} \mathbb{E}_{S \sim P_{XY}} \left[\int (f(x) - \hat{f}(x))^2 dx \right]$, and now we can write $\Phi \circ \rho(f_1, f_2) = \|f_1 - f_2\|_2^2$.

[1] here we use the uniform density $p(x) = 1$ for convenience, but any fixed density $p(x)$ will still induce a metric.

1.1.1 Constructing alternatives

Define $\psi(x) = \begin{cases} x + \frac{1}{2} & \text{if } x \in [-\frac{1}{2}, 0), \\ -x + \frac{1}{2} & \text{if } x \in [0, \frac{1}{2}], \\ 0 & \text{o.w.} \end{cases}$ then we can easily note that ψ is 1-Lipschitz, and $\int \psi^2(x)dx = 1/12$.

Now let $h > 0$ (we'll [illegible]) and let $m = \frac{1}{h}$, we construct a new function class

$$[illegible] (\cdot) = \sum_{j=1}^{m} \omega_j \phi_j(\cdot), \omega \in \Omega_m \right\}$$

where Ω_m is the Varsha[illegible]percube of $\{0,1\}^m$, and $\phi_j(x) = Lh \cdot \psi\left(\frac{x-(j-1/2)h}{h}\right)$. Since $|\phi_j^{'}(x)| \leq L$, we know t[illegible] We can now define our alternatives:

$$\mathcal{P}^{'} = \left\{ P_{XY};\ p(x) \text{ uniform}, f(x) = \mathbb{E}[Y|X=x] \in \mathcal{F}^{'},\ Y|X = x \sim \mathcal{N}(f(x), \sigma^2) \right\}.$$

We see that $\mathcal{P}^{'} \subset \mathcal{P}'' \subset \mathcal{P}$.

1.1.2 Lower bound on $\|f_\omega - f_{\omega'}\}$

To better organize our result, we first calculate,

$$\int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2(x)dx = \int_{\frac{j-1}{m}}^{\frac{j}{m}} L^2h^2 \cdot \psi^2\left(\frac{x-(j-1/2)h}{h}\right)dx = \int_{-1/2}^{1/2} L^2h^3\psi^2(u)du = \frac{L^2h^3}{12}.$$

We then have,

$$\begin{aligned}
\rho^2(f_\omega, f_{\omega'}) &= \int_0^1 (f_\omega - f_{\omega'})^2 dx \\
&= \sum_{j=1}^{m} \int_{\frac{j-1}{m}}^{\frac{j}{m}} (\omega_j\phi_j(x) - \omega_j'\phi_j(x))^2 dx \\
&= \sum_{j=1}^{m} \mathbf{1}\{\omega_j \neq \omega_j'\} \int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2(x)dx \\
&= \frac{L^2h^3}{12} \sum_{j=1}^{m} \mathbf{1}\{\omega_j \neq \omega_j'\} = \frac{L^2h^3}{12} \cdot H(\omega_j, \omega_j')
\end{aligned}$$

where $H(\cdot,\cdot)$ is the Hamming distance and the last equation holds because of the definition of it.

Since $\omega, \omega' \in \Omega_m$, by Varshamov-Gilbert lemma, $H(\omega_j, \omega_j') \geq \frac{m}{8} = \frac{1}{8h}$. Then we have

$$\min_{\omega_j, \omega_j'} \rho(f_\omega, f_{\omega'}) \geq \frac{Lh}{\sqrt{96}} \triangleq \delta,$$

where δ is called the separation between hypotheses.

1.1.3 Upper bound KL

Next, we will upper bound the maximum KL divergence between our alternatives. Let $P_\omega, P_{\omega'} \in'$. Then,

$$
\begin{aligned}
KL(P_\omega, P_{\omega'}) &= \int p_\omega \log \frac{p_\omega}{p_{\omega'}} \\
&= \int\int p_\omega(x)p_\omega(y|x) \log \frac{p_\omega(x)p_\omega(y|x)}{p_{\omega'}(x)p_{\omega'}(y|x)} dydx \\
&= \int\int p_\omega(y|x) \log \frac{p_\omega(y|x)}{p_{\omega'}(y|x)} dydx \qquad (\text{as } p_\omega(x) = p_{\omega'}(x) = 1) \\
&= \int_0^1 KL\left(\mathcal{N}(f_\omega(x), \sigma^2), \mathcal{N}(f_{\omega'}(x), \sigma^2)\right) dx \qquad (\text{as } Y|X = x \sim \mathcal{N}(f(x), \sigma^2)) \\
&= \frac{1}{2\sigma^2} \int_0^1 \left(f_\omega(x) - f_{\omega'}(x)\right)^2 dx \\
&= \frac{1}{2\sigma^2} \rho^2(f_\omega, f_{\omega'}) = \frac{L^2h^3 \cdot H(\omega, \omega')}{24\sigma^2}.
\end{aligned}
$$

Then since $\max_{\omega,\omega'} H(\omega, \omega') \le m = 1/h$,

$$
\max_{\omega,\omega'} KL(P_\omega, P_{\omega'}) = \frac{L^2h^3}{24\sigma^2} \max_{\omega,\omega'} H(\omega, \omega') \le \frac{L^2h^2}{24\sigma^2}.
$$

1.1.4 Apply local Fano's method

In order to apply Fano's method, we need to satisfy $\max_{\omega,\omega'} KL(P_\omega, P_{\omega'}) \le \frac{\log|\mathcal{P}'|}{4n}$. Recall that by the Varshamov-Gilbert lemma, $|\mathcal{P}'| \ge 2^{m/8}$, so it is sufficient if we have,

$$
\frac{L^2h^2}{24\sigma^2} \le \frac{\log(2^{m/8})}{4n} = \frac{m \log 2}{32n} = \frac{\log 2}{32nh}.
$$

This suggests that we could choose $h = \left(\frac{3\log 2}{4}\right)^{\frac{1}{3}} \frac{\sigma^{2/3}}{n^{1/3}L^{2/3}}$.

Thus the separation between hypotheses $\delta = C_1 \frac{L^{1/3}\sigma^{2/3}}{n^{1/3}}$, where C_1 is some constant, and then by local Fano's method,

$$
R_n^* \ge \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{8}\delta^2 = C_2 \frac{L^{2/3}\sigma^{4/3}}{n^{2/3}}.
$$

Remark: In order to apply above local Fano's method, it's required that $|\mathcal{P}'| \ge 16$. It's sufficient to have $|\mathcal{P}'| \ge 2^{m/8} \ge 16$, i.e. $m = 1/h \ge 32$, which means the following must hold:

$$
h = \left(\frac{3\log 2}{4}\right)^{\frac{1}{3}} \frac{\sigma^{2/3}}{n^{1/3}L^{2/3}} \le \frac{1}{32} \implies n \ge C_3 \frac{\sigma^2}{L^2} \text{ for some constant } C_3.
$$

1.2 Upper Bound

To upper bound the minimax risk we introduce the following estimator. Later we will introduce the Nadaraya-Watson estimator, and show that our current estimator is a special case of the Nadaraya-Watson estimator.

Our estimator $\hat{f}(t)$ is defiend as follows: Let $N(t) = \sum_{i=1}^n \mathbf{1}\{X_i \in [t-h, t+h]\}$. Then define,

$$
\hat{f}(t) = \begin{cases} \text{clip}\left(\frac{1}{N(t)} \sum_{i=1}^n Y_i \mathbf{1}\{X_i \in [t-h, t+h]\}, 0, 1\right) & \text{if } N(t) > 0 \\ 0 & \text{if } N(t) = 0 \end{cases}
$$

where $\operatorname{clip}(x,0,1)$ means that

$$\operatorname{clip}(x,0,1) = \begin{cases} x, & 0 \leqslant x \leqslant 1 \\ 0, & x < 0 \\ 1, & x > 1. \end{cases}$$

By definition,

$$\begin{aligned} R(P_{XY}, \hat{f}) &= \mathbb{E}_S \left[\int (\hat{f}(x) - f(x))^2 p(x) dx \right] \\ &\leqslant \alpha_1 \mathbb{E}_S \left[\int (\hat{f}(x) - f(x))^2 dx \right] \\ &= \alpha_1 \int_0^1 \underbrace{\mathbb{E}_S \left[(\hat{f}(t) - f(t))^2 \right]}_{\operatorname{err}(t)} dt. \end{aligned}$$

We will next provide a pointwise bound on $\operatorname{err}(t)$ which will translate to an integrated bound. The calculations for the pointwise bound are very similar to an example we saw previously so we will only provide an overview and highlight the differences.

Let $G_t = \{N(t) \geqslant \alpha_0 nh\}$ denote the good event that there were a sufficient number of samples in a $2h$ neighborhood of t. We have,

$$\begin{aligned} \mathbb{P}(G_t^c) &= \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}\{X_i \in [t-h, t+h]\} < \alpha_0 nh \right) \\ &= \mathbb{P}\left(\sum_{i=1}^n (\mathbf{1}\{X_i \in [t-h, t+h]\} - \mathbb{P}([t-h, t+h])) < \alpha_0 nh - n\mathbb{P}([t-h, t+h]) \right), \end{aligned}$$

where $\mathbb{P}([t-h,t+h]) = \int_{t-h}^{t+h} p(x)dx \geqslant 2\alpha_0 h$. Thus we have $\alpha_0 nh - n\mathbb{P}([t-h,t+h]) \leqslant -\alpha_0 nh$. By Hoeffding's inequality, we have $\mathbb{P}(G_t^c) \leqslant \exp(-2\alpha_0^2 nh^2)$. By following the calculations from our previous example, we can show

$$\mathbb{E}_S \left[(\hat{f}(t) - f(t))^2 \right] \leqslant L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0^2 nh^2}.$$

Therefore,

$$R(P_{XY}, \hat{f}) \leqslant \alpha_1 \int_0^1 \mathbb{E}_S \left[(\hat{f}(t) - f(t))^2 \right] dt \leqslant \alpha_1 \left(L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0^2 nh^2} \right).$$

Now we choose $h = \sigma^{2/3} L^{-2/3} n^{-1/3}$, which implies that

$$R(P_{XY}, \hat{f}) \leqslant 2\alpha_1 \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} + \alpha_1 \exp\left(-2\alpha_0^2 \frac{\sigma^{4/3} n^{1/3}}{L^{4/3}} \right).$$

Remark: On an ancillary note, had we used the multiplication Chernoff bound instead of Hoeffding's inequality, we will have had the following bounds:

$$\begin{aligned} &\mathbb{P}(G^c) \leqslant e^{-\alpha_0 nh/8}, \\ &R(P_{XY}, \hat{f}) \leqslant 2\alpha_1 \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} + \alpha_1 \exp\left(-\frac{\alpha_0}{4} \frac{\sigma^{2/3} n^{2/3}}{L^{2/3}} \right). \end{aligned}$$

For i.i.d Bernoulli random variables with success probability close to 0 or 1, the multiplicative Chernoff bound can provide a tighter bound than Hoeffding's inequality. This does not significantly alter our conclusions in this example, but it may be significant in other use cases.

1.3 Nadaraya-Watson Estimator

An Nadaraya-Watson Estimator (also known as the kernel estimator), is defined to be

$$\hat{f}(t) = \sum^{n} y_i w_i(t), \; w_i(t) = \frac{K((t - X_i)/n)}{\sum_{j=1}^{n} K((t - X_j)/n)},$$

where K is called a sm[illegible]ample, in our previous case, the smoothing kernel is $K(t) = \mathbf{1}\{t \leqslant 1/2\}$.

Other kernel choices [illegible] under stronger smoothness assumptions. On such assumption is the Hölder class in $\mathbb{R}^d$ [illegible] (β, L) and defined to be the set of all functions whose $(\beta - 1)$th order partial derivatives [illegible] minimax rate in this class is $\Theta(n^{-2\beta/(2\beta+d)})$. To achieve this rate, we will need to d[illegible] the Nadaraya-Watson estimator. The same rates hold for density estimation in th[illegible]

2 Density Estimation

We will briefly introduce lower and upper bounds for density estimation. Let $\mathcal{F}$ be the class of bounded Lipschitz functions, i.e.

$$\mathcal{F} = \{f : [0,1] \to [0, B] \; : \; |f(x_1) - f(x_2)| \leqslant L|x_1 - x_2|, \forall x_1, x_2 \in [0,1]\}.$$

The corresponding nonparametric family of densities is then defined to be

$$\mathcal{P} = \{P : \text{ The p.d.f. } p \text{ of } P \text{ is in } \mathcal{F}\}.$$

Suppose we observe $S = (X_1, \ldots, X_n)$ drawn i.i.d from some distribution $P \in \mathcal{P}$. We wish to estimate the p.d.f. under the L_2 loss, i.e.

$$\Phi \circ \rho(P_1, P_2) = \int (p_1(t) - p_2(t))^2 dt.$$

By definition, the minimax risk is

$$R_n^* = \inf_{\widehat{p}} \sup_{p \in \mathcal{F}} \mathbb{E}_S \left[||\widehat{p} - p||_2^2 \right].$$

2.1 Lower bound

The first step is to construct alternatives. For this, define

$$\psi(x) = \begin{cases} x + \frac{1}{2}, & x \in \left[-\frac{1}{2}, -\frac{1}{4}\right] \\ -x, & x \in \left[-\frac{1}{4}, \frac{1}{4}\right] \\ x - \frac{1}{2}, & x \in \left[\frac{1}{4}, \frac{1}{2}\right]. \end{cases}$$

Note that ψ is 1-Lipschitz and always in $[-1/4, 1/4]$. Moreover, $\int \psi(t)dt = 0$, $\int \psi^2(t)dt = 1/48$, and $|\psi(t)| \leq 1/4$.