S&DS 411 (Fall 2019) Selected Topics in Statistical Decision Theory **Lecture: 2**

Nonparametric Estimation

Lecturer: Harrison Zhou *Scribe: Cynthia Yue, Sep 4, 2019*

1 Nonparametric Estimation

For nonparametric estimation, let us consider the following model, where $Z_i \sim \mathcal{N}(0,1)$ and $i = 1, 2, \ldots$:

$$Y_i = \theta_i + \frac{1}{\sqrt{n}} Z_i.$$

The parameter space is a Sobolev ball with smoothness α:

$$\theta \in \Theta = \{\theta : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq M\}.$$

Our goal is to estimate θ, or, in other words, to find $\hat{\theta}$ to minimize the worst risk $\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta} - \theta\|$.

Remark. We can think of the risk as $\mathbb{E}\int(\hat{f} - f)^2 dx = \mathbb{E}\sum_{i=1}^{\infty}(\hat{\theta}_i - \theta_i)^2$. This is often called the minimax criterion. The idea is that if we can estimate θ, we can estimate f well.

At a first glance, we may think that this is a ridiculous model and parameter space, so let us first look at the motivations for using this model and parameter space.

2 Motivations

2.1 Fourier Basis

The following motivates the parameter space.

We often have the following assumption for our function f, where often $\alpha = 1$ or 2:

$$f \in \int_0^1 (f^{(\alpha)}(x))^2 \leq M'.$$

Fourier basis:

$$\begin{aligned}
\varphi_1(x) &= 1_{[0,1]}(x) \\
\varphi_{2i}(x) &= \cos(2\pi i x) \quad i = 1, 2, \ldots \\
\varphi_{2i+1}(x) &= \sin(2\pi i x).
\end{aligned}$$

We can determine these are orthonormal bases, so we can expand our function $f(x)$ into this Fourier basis in order to have

$$f(x) = < f, \varphi_1 > \varphi_1(x) + < f, \varphi_2 > \varphi_2(x) + < f, \varphi_3 > \varphi_3(x) + \dots$$

where we can think of $< f, \varphi_1 >$ as θ_1, $< f, \varphi_2 >$ as θ_2, $< f, \varphi_3 >$ as θ_3, and so on and so forth.

Let us consider the simple case when $\alpha = 2$:

$$f^{(2)}(x) = < f, \varphi_2 > (2\pi i)^2(-\varphi_2(x)) + < f, \varphi_3 > (2\pi i)^2(-\varphi_3(x)) + \dots$$

$$\begin{aligned} \int_0^1 (f^{(2)}(x))^2 dx &= \sum_{i=1}^{\infty} (\theta_{2i}^2 + \theta_{2i+1}^2)(2\pi i)^4 \\ &\leq M'. \end{aligned}$$

2.2 Nonparametric Regression

With nonparametric regression, we have

$$Y_i = f\left(\frac{1}{n}\right) + Z_i \quad \text{where } Z_i \sim \mathcal{N}(0,1) \text{ and } i = 1, \dots, n.$$

How do we recover this function f? We can write Y_i as a vector and multiply both sides by $\frac{1}{\sqrt{n}}$:

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ Y_i \\ \vdots \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ f(\frac{i}{n}) \\ \vdots \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ Z_i \\ \vdots \end{pmatrix}$$

where $Z_i \sim \mathcal{N}(0,1)$ i.i.d. and $i = 1, \dots n$.

Remark.

$$\sum_{i=1}^{n} \left(\frac{1}{\sqrt{n}} f\left(\frac{i}{n}\right)\right)^2 = \frac{1}{n} \sum_{i=1}^{n} f^2\left(\frac{i}{n}\right) \approx f^2(x)dx.$$

Then, multiply both sides by an isonormal matrix (discrete Fourier transformation):

$$\begin{aligned} A_{n\times n} \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ Y_i \\ \vdots \end{pmatrix} &= A_{n\times n} \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ f(\frac{i}{n}) \\ \vdots \end{pmatrix} + A_{n\times n} \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ Z_i \\ \vdots \end{pmatrix} \\ A \frac{1}{\sqrt{n}} \begin{pmatrix} f(\frac{1}{n}) \\ \vdots \\ f(\frac{n}{n}) \end{pmatrix} &= \begin{pmatrix} \widetilde{\theta_1} \\ \vdots \\ \widetilde{\theta_n} \end{pmatrix} \approx \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}. \end{aligned}$$

2.3 Nonparametric Density Estimation

We often assume

$$Y_1, \dots, Y_n \text{ i.i.d. density } f.$$

We also assume f belongs to the space

$$f \in \{f : \int_0^1 (f^{(\alpha)}(x))^2 dx \leq M\}.$$

Our goal is to estimate f. We expand f by the Fourier basis:

$$f(x) = \sum_i^\infty < f, \varphi_i > \varphi_i(x)$$

where $< f, \varphi_i >= \theta_i = \int_0^1 f(x)\varphi(x)dx$.

$$\begin{aligned}
\theta_i &= \int_0^1 f(x)\varphi_i(x)dx \\
&= \mathbb{E}\varphi_i(Y_1) \\
\hat{\theta}_i &= \frac{1}{n}\sum_{j=1}^n \varphi_i(Y_j) \\
&\approx \mathcal{N}(\theta_i, \frac{1}{n}C_i) \qquad \text{where } C_i = \mathrm{Var}(\varphi_i(Y_i)).
\end{aligned}$$

3 Estimation

3.1 Minimax Rate

3.1.1 Upper bound

One way to estimate is to truncate somewhere:

$$\hat{\theta}_i = \begin{cases} Y_i & i \leq I \\ 0 & i > I. \end{cases}$$

$$\begin{aligned}
\mathbb{E}\|\hat{\theta} - \theta\|^2 &= \mathbb{E}\sum_{i=1}^I (Y_i - \theta_i)^2 + \mathbb{E}\sum_{i \geq I+1} (0 - \theta_i)^2 \\
&= \frac{I}{n} + \sum_{i \geq I+1} \theta_i^2 \qquad \text{Note: From the parameter space, we have } \sum_{i \geq I+1} i^{2\alpha}\theta_i^2 \leq M, \\
&\qquad\qquad \text{so } (I+1)^{2\alpha} \sum_{i \geq I+1} \theta_i^2 \leq M. \\
&\leq \frac{I}{n} + \frac{M}{(I+1)^2} \\
&\leq \frac{I}{n} + \frac{M}{I^{2\alpha}}.
\end{aligned}$$

Here, $\frac{I}{n}$ is variance, and $\frac{M}{I^{2\alpha}}$ is bias. If we set $\frac{I}{n}$ equal to $\frac{M}{I^{2\alpha}}$, we have

$$\begin{aligned}\frac{I}{n} &= \frac{M}{I^{2\alpha}}\\ I^{1+2\alpha} &= nM\\ I &= (nM)^{\frac{1}{1+2\alpha}}.\end{aligned}$$

Returning to our above statement, we then have

$$\begin{aligned}\mathbb{E}\|\hat{\theta}-\theta\|^2 &\leq \frac{I}{n}+\frac{M}{I^{2\alpha}}\\ &= 2\frac{(nM)^{\frac{1}{1+2\alpha}}}{n}\\ &= 2M^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}.\end{aligned}$$

This is our rate of convergence for the risk. This rate appears in literature very often.

Some questions to start considering are:
Q: Can we get a better rate?
Q: What if we don't know α and M.

3.1.2 Lower bound

For the lower bound, we have Le Cam's two point argument.

We want to show:

$$\inf_{\hat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}\|\hat{\theta}-\theta\|^2 \geq cM^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}.$$

For any $\hat{\theta}$, the worst risk is always bounded by the average:

$$\begin{aligned}\sup_{\theta}\mathbb{E}_{x|\theta}\|\hat{\theta}-\theta\|^2 &\geq \mathbb{E}_{\theta}\mathbb{E}_{x|\theta}\|\hat{\theta}-\theta\|^2 \quad \text{where } \theta\sim\pi \text{ supported on } \Theta\\ &\geq \mathbb{E}_{\theta}\mathbb{E}_{x|\theta}\|\hat{\theta}_{Bayes}-\theta\|^2 \\ &\geq cM^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}.\end{aligned}$$

Getting a good prior is key to getting a good lower bound:

Construction of prior π:

- θ_i i.i.d. $= \begin{cases}\frac{1}{\sqrt(n)} & \text{with probability } \frac{1}{2} \qquad i\leq I\\ 0 & \frac{1}{2}.\end{cases}$
- $\theta_i = 0 \qquad i\geq I+1.$

Is the prior supported on the parameter space? To do this, we need to check the following:

$$\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq M \impliedby \sum_{i=1}^{I} i^{2\alpha}(\frac{1}{\sqrt{n}})^2 \leq M.$$

Note that

$$\begin{aligned} i &\le I \\ I^{2\alpha}\frac{I}{n} &= \frac{I^{1+2\alpha}}{n} \\ &\le M. \end{aligned}$$

.

Now, let us follow the Bayes procedure:

$$\begin{aligned}
\hat{\theta}_{Bayes,i} &= \mathbb{E}(\theta_i|Y) = \mathbb{E}(\theta_i|Y_i) \\
\mathbb{E}_\theta \mathbb{E}_{Y|\theta}\|\hat{\theta}_{Bayes} - \theta\|^2 &= \sum_{i=1}^{\infty} \mathbb{E}_{\theta_i}\mathbb{E}_{Y_i|\theta_i}(\hat{\theta}_{Bayes,i}(Y_i) - \theta_i)^2 \\
&\ge \sum_{i=1}^{I}\left\{\frac{1}{2}\mathbb{E}_{Y_i|\theta_i=\frac{1}{\sqrt{n}}}\left(\hat{\theta}_{Bayes}(Y_i) - \frac{1}{\sqrt{n}}\right)^2 + \frac{1}{2}\mathbb{E}_{Y_i|\theta_i=0}\left(\hat{\theta}_{Bayes}(Y_i) - 0\right)^2\right\} \\
&= \sum_{i=1}^{I}\left\{\frac{1}{2}\int\left(\hat{\theta}_{Bayes,i}(Y_i) - \frac{1}{\sqrt{n}}\right)^2 \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}}\mathrm{e}^{-\frac{(Y_i - \frac{1}{\sqrt{n}})^2}{2\frac{1}{n}}}\right. \\
&\qquad \left. + \frac{1}{2}\int\left(\hat{\theta}_{Bayes,i}(Y_i) - 0\right)^2 \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{Y_i^2}{2\frac{1}{n}}}\right\} \\
&\ge \sum_{i=1}^{I}\left\{\frac{1}{2}\int\left[\left(\hat{\theta}_{Bayes,i}(Y_i) - \frac{1}{\sqrt{n}}\right)^2 + \left(\hat{\theta}_{Bayes,i}(Y_i) - 0\right)^2\right]\right. \\
&\qquad \left. \min\left\{\frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}}\mathrm{e}^{-\frac{(Y_i - \frac{1}{\sqrt{n}})^2}{2\frac{1}{n}}}, \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}}\mathrm{e}^{-\frac{Y_i^2}{2\frac{1}{n}}}\right\}\right\} \\
&\ge \frac{I}{4n}\underbrace{\int \min\left\{\frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}}\mathrm{e}^{-\frac{(x - \frac{1}{\sqrt{n}})^2}{2\frac{1}{n}}}, \frac{1}{\sqrt{2n}\frac{1}{\sqrt{n}}}\mathrm{e}^{-\frac{x^2 n}{2\frac{1}{n}}}\right\}dx}_{\ge c' \text{ (this is bounded by some constant c')}} \quad \text{where} \quad \underbrace{x = \frac{1}{\sqrt{n}}Y}_{\substack{\text{change of variables} \\ \text{so that this becomes} \\ \text{standard normal}}} \\
&= \frac{1}{4}M^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}.
\end{aligned}$$

We now consider the questions:
Q: We don't know α and M, so can we adapt the procedure to pretend we know M and α. Can we find a procedure that is adaptive to M and α?
Q: Going beyond weights $\{0, 1\}$

3.2 Best Linear Estimator

We consider the model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}Z_i \quad \text{where } i = 1, 2, \ldots$$

Claim.

$$\inf_{c_i \in [0,1]} \sup_{\theta \in \Theta} \mathbb{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 \leq (1 + O(1)) P_\alpha M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{2\alpha+1}} \quad \text{where } P_\alpha \text{ is the Pinsker constant}$$

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \sum_{i=1}^{\infty} (\hat{\theta}_i - \theta_i)^2 \geq (1 + O(1)) P_\alpha M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{2\alpha+1}}.$$

If we use the truncated Gaussian, we can show the lower bound. We can use the linear procedure to get the prior.

3.3 Adaptive Estimation

We have James-Stein estimation:

$$X = \theta + Z \quad \text{where } Z \sim \mathcal{N}(0, \sigma^2 I_d).$$

$$\mathbb{E}\|X - \theta\|^2 = d\sigma^2. \qquad \mathbb{E}\|X\|^2 = d + \|\theta\|^2.$$

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{\|X\|^2}\right) X.$$

Claim.

$$\mathbb{E}\|\hat{\theta}_{JS} - \theta\|^2 \leq 2 + \inf_{c \in [0,1]} \mathbb{E}\|cY - \theta\|^2.$$

For our adaptive procedure, we have blockwise James-Stein estimation $\hat{\theta}_{BJS}$:

$$\theta_i, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \ldots, \theta_{14}$$

$$\underbrace{Y_i, Y_2}_{2}, \underbrace{Y_3, Y_4, Y_5, Y_6}_{4}, \underbrace{Y_7, \ldots, Y_{14}}_{8}$$

When $\hat{\theta}_i = 0$ and $i \geq n$, we have

$$\sum_{i \geq n}^{\infty} \mathbb{E}(0 - \theta_i)^2 = \sum_{i \geq n} \theta_i^2 \leq \frac{M}{n^{2\alpha}}$$

$$= O(n^{-\frac{2\alpha}{2\alpha+1}}).$$

We will show that

$$\mathbb{E}\|\hat{\theta}_{BJS} - \theta\|^2 \leq c_1 M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{2\alpha+1}}.$$