STA3000F Lecture 12

Nonparametric regression

- Least-square estimators (for any function class with covering/packing bounds)
- Fourier projection. (Sobolev class, requires basis functions)
- "Local" estimators. (Hölder class. pointwise bounds). (does not need basis functions).

Warmup: Nadaraya-Watson estimator.

Problem: x_i's deterministic,

$$Y_i = f^*(x_i) + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} P. \quad \begin{cases} \mathbb{E}[\varepsilon_i] = 0 \\ \mathrm{var}(\varepsilon_i) \le \sigma^2. \end{cases}$$

(Fixed design analysis)

(For random design, first condition on $(X_i)_{i=1}^n$).

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}.$$

(for simplicity, we define weight function).

$$W_{n,i}(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}.$$

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

Note: $\hat{f}_n$ is a linear functional of $(Y_i)_{i=1}^n$.

$W_{n,i}$ are deterministic.

(cf. Tsybakov Chapter 3. best linear estimator).

- Bias: $b(x_0) = \sum_{i=1}^{n} W_{n,i}(x_0) \cdot \left(f^*(x_i) - f^*(x_0) \right)$

- Variance: $\sigma^2(x_0) = \sigma^2 \sum_{i=1}^{n} W_{n,i}^2(x_0)$.

Assuming $f^* \in \Sigma(\beta, L)$ for $0 < \beta \leq 1$

$$|b(x_0)| \leq \sum_{i=1}^{n} |W_{n,i}(x_0)| \cdot L \cdot |x_i - x_0|^{\beta}$$

(Assume that K is supported on $[-1, 1]$

$W_{n,i}(x_0)$ is non-zero only when

$|x_i - x_0| \leq h$).

$$\leq \sum_{i=1}^{n} \left| W_{n,i}(x_0) \right| \cdot L h^{\beta}.$$

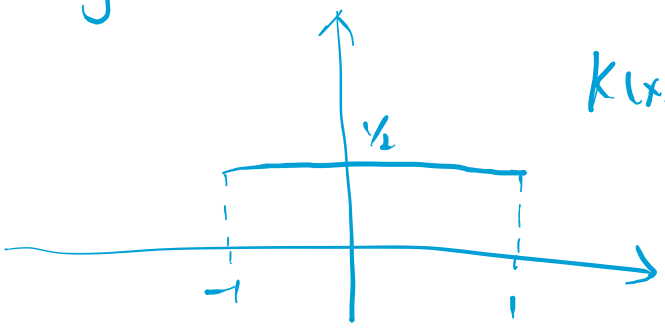
(Assuming $K \geq 0$).

$$= L \cdot h^{\beta}$$

$$\sigma^2(x_0) = \sigma^2 \cdot \sum_{i=1}^{n} W_{n,i}^2(x_0).$$

$$\leq \sigma^2 \cdot \underbrace{\left(\sum_{i=1}^{n} W_{n,i}(x_0) \right)}_{\geq 1} \cdot \underbrace{\max_i \left| W_{n,i}(x_0) \right|}_{\leq \frac{1}{\#j : |x_j - x_0| \leq h}}.$$

Taking box-shaped kernel



$$K(x) = \frac{1}{2} \mathbb{1}\{x \in [-1, 1]\}.$$

$$W_{n,i}(x_0) = \frac{\mathbb{1}\{x_i \in [x_0 - h, x_0 + h)\}}{\#\, j, \text{ s.t. } x_j \in [x_0 - h, x_0 + h]}$$

For equi-spaced design $x_i = i/n$.

(this can be relaxed, e.g. distance between adjacent design points).

$$\left| \{ j : |x_j - x_0| \le h \} \right| \ge \lfloor 2nh \rfloor \quad \text{for} \quad h > n^{-1}.$$

$$\sigma^2(x_0) \le \frac{\sigma^2}{nh}.$$

$$\therefore h_n = n^{\frac{-1}{2\beta+1}} \Longrightarrow MSE(x_0) = C \cdot n^{\frac{-2\beta}{2\beta+1}}.$$

How about $\beta > 1$? Local polynomial estimation.

Recall NW estimator

$$\hat{f}_n(x) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right).$$

(In general, we may use a polynomial instead of const)

$$f(x_i) \approx f(x) + f'(x)\cdot(x_i - x) + \frac{1}{2!} f''(x)(x_i - x)^2 + \cdots + \frac{f^{(l)}(x)}{l!}(x_i - x)^l$$

$$= \begin{bmatrix} 1 \\ x_i - x \\ (x_i - x)^2/2! \\ \vdots \\ (x_i - x)^l/l! \end{bmatrix}^T \cdot \begin{bmatrix} f(x) \\ f'(x) \\ \vdots \\ f^{(l)}(x) \end{bmatrix}.$$

Notation: $u(t) = \begin{bmatrix} 1 \\ t \\ t^2/2! \\ \vdots \\ t^l/l! \end{bmatrix} \in \mathbb{R}^{l+1}.$

$$\hat{\theta}_n(x) = \underset{\theta \in \mathbb{R}^{l+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \theta^T U\left(\frac{X_i - x}{h}\right)\right)^2 \cdot K\left(\frac{X_i - x}{h}\right)$$

(By taking K as box function, we're approximating f^* in $[x-h, x+h]$ using polynomials).

$$\hat{f}_n(x) = e_1^T \hat{\theta}_n(x) \qquad \text{(first coordinate)}$$

(We can also use $e_2^T \hat{\theta}_n(x)$ to estimate $f'(x)$, etc.)

Analysis. We define

$$B_{n,x} := \frac{1}{nh} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right)^T \cdot K\left(\frac{X_i - x}{h}\right)$$

(matrix in the quadratic form)

then we have

$$W_{n,i}(x) = \frac{1}{nh} e_1^T \cdot B_{n,x}^{-1} \cdot U\left(\frac{X_i - x}{h}\right) \cdot K\left(\frac{X_i - x}{h}\right)$$

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

Variance:

$$\sigma^2(x_0) = \sigma^2 \cdot \sum_{i=1}^{n} W_{n,i}(x_0)^2$$

$$\leq \sigma^2 \cdot \sum_{i=1}^{n} |W_{n,i}(x_0)| \cdot \max_{1 \leq i \leq n} |W_{n,i}(x_0)|.$$

$$\max_{1\le i\le n} |W_{n,i}(x_0)| \le \frac{1}{nh} \cdot \max_{i:\,|x_i-x_0|\le h} \left\| B_{n,x}^{-1} \, u\left(\frac{x_i-x_0}{h}\right)\right\|_2.$$

$$\le \frac{1}{nh} \cdot \|B_{n,x}^{-1}\|_{op} \cdot \max_{|x_i-x_0|\le h} \left\| u\left(\frac{x_i-x_0}{h}\right)\right\|_2.$$

$$\left\| u\left(\frac{x_i-x_0}{h}\right)\right\|_2^2 \le \sum_{k=0}^{l} \frac{1}{(k!)^2} \le 3.$$

(See Tsybakov). If using equispaced design,

$B_{n,x} \longrightarrow$ something positive definite (defined by integral)

$\exists n_0$, s.t. when $n > n_0$ $B_{n,x} \ge \lambda_0 I_{l+1}$.

(Assuming $X_1, X_2 \cdots X_n \overset{iid}{\sim} P$, with density bounds)

we can verify using empirical process tools

$$\left\| B_{n,x} - \frac{\mathbb{E}\left[u\left(\frac{X-x}{h}\right) u\left(\frac{X-x}{h}\right)^T K\left(\frac{X-x}{h}\right)\right]}{h} \right\|_{op} \le O\left(\frac{1}{\sqrt{nh}}\right)$$

w.h.p. then we can verify PD of $\mathbb{E}[\cdots]$.

For our analysis, we assume design points are good enough to satisfy $B_{n,x_0} \ge \lambda_0 I_{l+1}$.

$$\max_{|x_i - x_0| \le h} |W_{n,i}(x_0)| \le \frac{c}{\lambda_0 \, n h}.$$

$$\sum_{i=1}^{n} |W_{n,i}(x_0)| \le \frac{a_0 \cdot c}{\lambda_0}.$$

(Assuming that for any interval $A \subseteq [0,1]$

$$\frac{1}{n} |\{i : x_i \in A\}| \le a_0 \cdot \max\{|A|, \tfrac{1}{n}\}$$

This can be verified using prob assumptions).

So $$\sigma^2(x_0) \le \frac{a_0 \cdot c^2}{\lambda_0^2 \, n h}.$$

Bias. $$b(x_0) = \sum_{i=1}^{n} W_{n,i}(x_0) \cdot \left(f^*(x_i) - f^*(x_0) \right)$$

$f^*(x_i) - f^*(x_0)$ → indep of i.

$$= \sum_{i=1}^{n} W_{n,i}(x_0) \cdot \sum_{k=1}^{l} \frac{(f^*)^{(k)}(x_0) \cdot (x_i - x_0)^k}{k!}$$

$$+ \sum_{i=1}^{n} (x_i - x_0)^l W_{n,i}(x_0) \cdot \left(\frac{(f^*)^{(l)}(x_0 + \tau_i (x_i - x_0))}{l!} - \frac{(f^*)^{(l)}(x_0)}{l!} \right)$$

$$|\cdots| \le |x_i - x_0|^{\beta - l} \cdot L.$$

$$|\cdots| \le L \cdot h^{\beta}.$$

Hope: make the first term cancel out.

Lemma. For a degree-l polynomial Q.

we have $\sum_{i=1}^{n} W_{n,i}(x) \cdot Q(X_i) = Q(x)$.

(Proof. see Tsybakov).

Using this Lemma, we have

$$\sum_{i=1}^{n} W_{n,i}(x) \cdot (X_i - x)^k = 0. \quad \text{for } k = 1, 2, \cdots, l$$

$$\sum_{k=1}^{l} \frac{(f^*)^{(k)}(x_0)}{k!} \sum_{i=1}^{n} W_{n,i}(x_0) \cdot (X_i - x_0)^k = 0$$

Therefore, we can bound the bias as

$$|b(x_0)| \leq L \cdot h^{\beta}.$$

By choosing $h_n = n^{-\frac{1}{2\beta+1}}$, $MSE(x_0) \leq n^{-\frac{2\beta}{2\beta+1}}$.

Question: minimax optimality?

Recap: Le Cam's two-point method:

Construct $f_1, f_2 \in \mathcal{F}$.

$$\inf_{\hat{T}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[|\hat{T} - T(f)|^2\right] \geq \frac{1}{8}\left(T(f_1) - T(f_2)\right)^2 \cdot \left|1 - d_{TV}(P_1, P_2)\right|$$

Application to nonparametric regression.

(results for density estimation are similar).

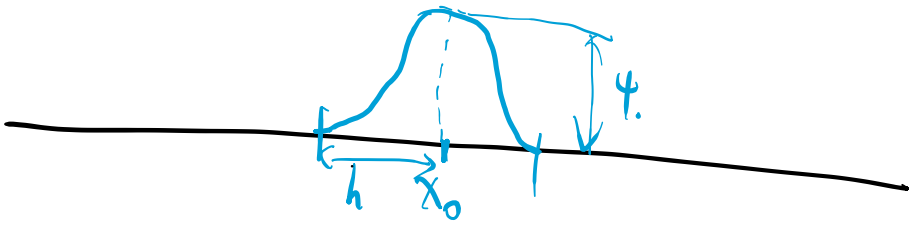
$x_1, x_2, \ldots, x_n$ deterministic, satisfying spacing assumptions (e.g. equispaced).

$$Y_i = f^*(x_i) + \varepsilon_i. \qquad \varepsilon_i \sim N(0,1).$$

$$f^* \in \mathcal{F} = \Sigma(\beta, L). \qquad T(f) = f(x_0).$$

WLOG $f_2(x) \equiv 0 \ (\forall x)$.

$$f_1(x) := \psi \cdot K\left(\frac{x - x_0}{h}\right).$$



Idea: we want large ψ and small h.

Choose K: C^∞ smooth, and bounded support.

$$K(u) = \exp\left(\frac{-1}{1-u^2}\right) \cdot 1_{\{|u| \le 1\}}.$$

Under this construction.

$$|f_1(x_0) - f_2(x_0)| = \psi.$$

$$d_{TV}(P_1, P_2) \le \sqrt{\frac{1}{2} D_{KL}(P_1 \| P_2)}$$

$$= \sqrt{\frac{1}{2} \sum_{i=1}^{n} D_{KL}(P_{1,i} \| P_{2,i})}$$

(where $P_{j,i}$ is the dist'n of i^{th} obs under P_j).

$$= \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(f_1(x_i) - f_2(x_i)\right)^2}$$

$$\le \sqrt{\frac{\psi^2}{2} \cdot \left|\{i : |x_i - x_0| \le h\}\right|} \le \sqrt{\frac{a_0 \psi^2 n h}{2}}.$$

Still need to verify $f_1 \in \Sigma(\beta, L)$.

$$f_1(x) = \psi \cdot K\left(\frac{x - x_0}{h}\right).$$

$$f^{(l)}(x) = \frac{\psi}{h^l} K^{(l)}\left(\frac{x - x_0}{h}\right)$$

$$\left|f^{(l)}(x) - f^{(l)}(y)\right| \le \frac{\psi}{h^l} \cdot C_l \cdot \frac{|x - y|}{h}$$

$$\le \frac{\psi}{h^l} \cdot C_l \cdot \frac{|x-y|^{\beta-l} \cdot (2h)^{1-(\beta-1)}}{h}$$

$$\le 2 C_l \cdot |x-y|^{\beta-l} \cdot \frac{\psi}{h^\beta} \quad \left(\text{need } \le L|x-y|^{\beta-l}\right)$$

Choose $\psi = \frac{L h^\beta}{2 C_L}$

Choose c_0 small enough with $h_n = c_0 \cdot n^{-\frac{1}{2\beta+1}}$.

$$d_{TV}(P_1, P_2) \le \sqrt{\frac{a_0}{2} \cdot n \cdot h_n \cdot \left(\frac{L h_n^\beta}{2C_L}\right)^2}$$

$$\sqrt{\text{const.} \cdot n \cdot (h_n)^{2\beta+1}} < \frac{1}{2}.$$

$$\inf_{\hat{T}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[|f(x_0) - \hat{T}|^2\right] \ge \frac{c_0^2}{16} \cdot c' \cdot n^{-\frac{2\beta}{2\beta+1}}.$$

- Remark: for MISE, we need something more "Fano's method"
  - Le Cam finds a pair of hard instances.
  - Fano deals with difficulties from a large collection of problem instances using mutual info arguments.

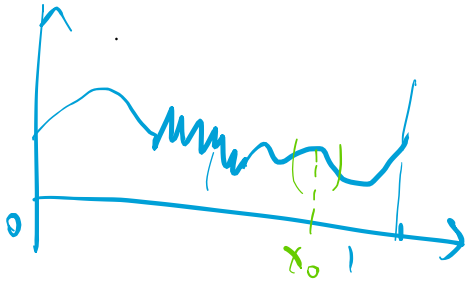
Adaptivity:

Question: Can we do well w/o knowing β?

Want to construct a single $\hat{f}_n$.

s.t. $\forall \beta,\ \forall f^* \in \Sigma(\beta, L),\quad MSE(x_0) \le n^{-\frac{2\beta}{2\beta+1}}$.

In general, f^* may have different smoothness at different places.



This also allows us to adapt to local smoothness around x_0.

Surprisingly, this is not possible. (w/o additional log factor)

Thm (Lepski, Lower bound).

Suppose $0 < \beta_1 < a < \beta_2 \le 1$.

$$r_1^2(n) = \left(\frac{\log n}{n}\right)^{\frac{2\beta_1}{2\beta_1+1}} \qquad r_2^2(n) = \left(\frac{1}{n}\right)^{\frac{2a}{2a+1}}$$

$$\inf_{\hat{f}} \sup_{i \in \{1,2\}} \sup_{f^* \in \Sigma(\beta_i, L)} \frac{\mathbb{E}\left[\left|\hat{f}_n(x_0) - f^*(x_0)\right|^2\right]}{r_i^2(n)} \ge C.$$

- Asymmetric Lower bound. Penalize more

for m/s- estimating the easier class.

- If we want to achieve anything faster than $n^{\frac{-2\beta_1}{2\beta_1+1}}$ under $\Sigma(\beta_2, L)$.

then this estimator must pay log factor under $\Sigma(\beta_1, L)$.

Proof: from HW1, asymmetric two-point Lemma.

$$f_2^* \in \Sigma(\beta_2, L), \quad f_1^* \in \Sigma(\beta_1, L).$$

$$\inf_{\hat{f}} \sup_{i \in \{1,2\}} \frac{\mathbb{E}\left[\left|\hat{f}_n(x_0) - f_i^*(x_0)\right|^2\right]}{r_i(n)^2} = \psi^2, \text{ under construction.}$$

$$\geq \frac{\left(f_1^*(x_0) - f_2^*(x_0)\right)^2}{8 \cdot r_1(n)^2} \left\{ 1 - \frac{r_2^2(n)}{r_1^2(n)} \left(1 + \chi^2\left(P_{f_1} \| P_{f_2}\right)\right) \right\}.$$

Construction: $f_2^* = 0$.

$$f_1^* = \psi \cdot K\left(\frac{x - x_0}{h}\right).$$

Need to ensure.

$$\frac{r_2^2(n)}{r_1^2(n)} \cdot \left\{1+\chi^2(P_{f_1} \| P_{f_2})\right\} < \frac{1}{2}.$$

$\asymp \frac{1}{\text{poly}(n)}.$

We only need

$$\chi^2(P_{f_1} \| P_{f_2}) \le \text{poly}(n)$$

(e.g. $n^{\frac{a-\beta_1}{2}}$).

(c.f. for symmetric version, we need $D_{KL}(P_{f_1} \| P_{f_2}) \le \frac{1}{2}$).

By choosing

$$\begin{cases} h = C_0 \cdot \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_1+1}} \\ \psi = c' \cdot h^{\beta_1} \end{cases}$$

We can ensure $\chi^2(\cdots) \le \text{poly}(n)$.

Substituting back, we prove the result.

How to achieve it? Lepski's method.

Key observation:

for each β, let h_β be optimal bandwidth.

Let $\hat{f}_{h_\beta}(x_0)$ be the estimator using h_β

Let β^* be the true smoothness.

If we under-smooth, $\beta < \beta^*$.

$$\left| \hat{f}_{h_\beta}(x_0) - \hat{f}_{h_{\beta^*}}(x_0) \right| \leq C \cdot h_\beta^\beta$$

(w.h.p.) for any $\forall \beta < \beta^*$.

and if we use this test for $\beta > \beta^*$,

this may not be satisfied.

Step I. discretize $\beta^* \in [\beta_{min}, \beta_{max}]$

$$\mathcal{B} = \{ \beta_{min} = \beta_1 < \beta_2 < \cdots < \beta_N = \beta_{max} \}$$

where $N = \log n$. $\quad \beta_j - \beta_{j-1} = \frac{1}{\log n}$.

For $\beta \in [\beta_{j-1}, \beta_j]$.

$$C \cdot n^{\frac{-2\beta_{j-1}}{2\beta_{j-1}+1}} \leq n^{\frac{-2\beta}{2\beta+1}} \leq n^{\frac{-2\beta_{j-1}}{2\beta_{j-1}+1}}$$

The rates within each interval are the same.

Step II.

$$\hat{\beta} = \max\left\{ \beta \in \mathcal{B} : \quad \forall \beta' \in \mathcal{B},\ \beta' < \beta,\ |\hat{f}_{h_{\beta'}}(x_0) - \hat{f}_{h_\beta}(x_0)| \le C_0 \cdot h_{\beta'}^{\beta'} \right\}$$

then estimate by

$$\hat{f}_{\text{Lepski}}(x_0) := \hat{f}_{h_{\hat{\beta}}}(x_0).$$

Thm:

$$\mathbb{E}\left[|\hat{f}_{\text{Lepski}}(x_0) - f^*(x_0)|^2 \right] \le C \cdot \left(\frac{\log n}{n}\right)^{\frac{2\beta_*}{2\beta_*+1}}.$$

Proof idea: Define event $\mathcal{E}_j = \{\hat{\beta} = \beta_j\}$.

$$\mathbb{E}\left[|\hat{f}(x_0) - f^*(x_0)|^2 \right] = \sum_{j=1}^{N} \mathbb{E}\left[|\hat{f}(x_0) - f^*(x_0)|^2 \cdot 1_{\mathcal{E}_j} \right].$$

(1) For $j \ge j^*$, on the event $\mathcal{E}_j$,

$$|\hat{f}_{h_{\beta^*}}(x_0) - \hat{f}_{h_{\beta_j}}(x_0)| \le C \cdot h_{\beta^*}^{\beta^*}.$$

even if we over estimate β^*, it doesn't matter (by our screening criteria).

$$\sum_{j=j^*}^{N} \mathbb{E}\left[|\hat{f}(x_0) - f^*(x_0)|^2 1_{\mathcal{E}_j} \right] \lesssim h_{\beta^*}^{2\beta^*} \text{ by NW analysis}$$

$$\leq 2 \cdot \sum_{j=j^*}^{N} \mathbb{E}\left[|\hat{f}_{h_{\beta^*}}(x_0) - f^*(x_0)|^2 \cdot 1_{\mathcal{E}_j} \right]$$

$$+ 2 \cdot \sum_{j=j^*}^{N} \mathbb{E}\left[|\hat{f}_{h_{\beta_j}}(x_0) - \hat{f}_{h_{\beta^*}}(x_0)|^2 \cdot 1_{\mathcal{E}_j} \right].$$

$\lesssim h_{\beta^*}^{2\beta^*}$ by def $\mathcal{E}_j$.

(!!) $j < j^*$. We can show using Gaussian tail bound.

$$\mathbb{P}(\mathcal{E}_j) \leq n^{-c'} \quad \text{(for } c' \text{ large enough)}.$$

this requires $h_{\beta_j} = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_j + 1}}$.