STA 3000 F Lecture 11

From last lecture: lack of asymptotic optimality

p density on $\mathbb{R}$ $\quad p \in \mathbb{H}^2$. i.e. $\int (p''(x))^2 dx \le L^2$.

Using 1st order kernel, $\quad MISE \asymp n^{-4/5}$.

Thm. If we use a second kernel $\quad \|K\|_2^2 < +\infty$.

Then $\forall \varepsilon > 0$, take

$$h = n^{-1/5} \varepsilon^{-1} \int K^2(u)\, du$$

We have

$$\limsup_{n \to +\infty} n^{4/5} \cdot \mathbb{E} \int \left(\hat{p}_n(x) - p(x) \right)^2 dx \le \varepsilon.$$

- Remark: Only for fixed density p, $n \to +\infty$.

  In practice, don't know when such an asymptotics kicks in (depends on p).

Rationale: For fixed n, the "hardest ~~problem~~" in class $\mathbb{H}^2$ depends on value of n.

Key step:

$$(*) \int b^2(x)\, dx = \cancel{\frac{h^4}{4} \left(\int u^2 K(u)\, du \right)^2 \cdot \int (p''(x))^2 dx} + o(h^4).$$

$$\int \sigma^2(x)\, dx \asymp \frac{1}{nh}$$

Trade-off.

Proof of (*).

$$b(x) = h^2 \int u^2 K(u) \left[\int_0^1 (1-\tau)\, \underbrace{p''(x+\tau u h)}_{\approx p''(x) \text{ on average}} d\tau \right] du.$$

(h small)

$$\tilde{b}(x) = h^2 \int u^2 K(u) \left[\int_0^1 (1-\tau)\, p''(x)\, d\tau \right] du$$

$$\int \tilde{b}(x)^2 dx = \frac{h^4}{4} \cdot \left(\int u^2 K(u)\, du \right)^2 \left(\int (p''(x))^2 dx \right).$$

It remains to bound approx. err. by $\tilde{b}$.

$$\left| \int \left(b^2(x) - \tilde{b}^2(x) \right) dx \right|$$

$$\leq \underbrace{\left(\int \left(b(x) + \tilde{b}(x) \right)^2 dx \right)^{1/2}}_{\text{Bdd. on } p \in H^2} \cdot \left(\int \left(b(x) - \tilde{b}(x) \right)^2 dx \right)^{1/2}$$

$$\int \left\{ \int u^2 K(u) \cdot \left[\int_0^1 (1-\tau) \left(p''(x+\tau u h) + p''(x) \right) d\tau \right] du \right\}^2 dx < +\infty$$

(By Generalized Minkowski).

$$\int \left(b(x) - \tilde{b}(x) \right)^2 dx$$

$$= \int \left\{ \int u^2 K(u) \cdot \left[\int_0^1 (1-\tau) \left(p''(x+\tau u h) - p''(x) \right) d\tau \right] du \right\}^2 dx$$

$$\leq \left(\int u^2 |K(u)| \cdot \left\{ \int \underbrace{\left[\int_0^1 \left| p''(x+\tau u h) - p''(x) \right| d\tau \right]^2}_{\leq \int_0^1 \int \left(p''(x+\tau u h) - p''(x)\right)^2 dx\, d\tau} dx \right\}^{1/2} du \right)^2 .$$

So we have

$$\sqrt{\int \left(b(x) - \tilde{b}(x) \right)^2 dx}$$

$$\leq \int u^2 |K(u)| \cdot \sup_{\tau \in [0,1]} \| p''(\cdot + \tau u h) - p'' \|_{L^2} \, du$$

$$\leq 2 \underbrace{\| p'' \|_{L^2}}_{\text{Bounded}} \cdot \underbrace{\int_{|u| \geq 1/\sqrt{h}} u^2 |K(u)| \, du}_{\to 0 \text{ as } h \to 0} + \boxed{\sup_{\substack{|u| < 1/\sqrt{h} \\ 0 \leq \tau \leq 1}} \| p''(\cdot + \tau u h) - p'' \|_{L^2}} \cdot \underbrace{\int u^2 |K(u)| \, du}_{\text{Bounded.}}$$

Bounding the key term.

$p \in H^2$, $\quad p'' \in L^2$

$\forall \varepsilon > 0$, $\exists g \in C_c$ $\quad \| p'' - g \|_{L^2} \leq \varepsilon$.

$$\| p''(\cdot + \tau u h) - p'' \|_{L^2}$$

$$\leq \underbrace{\| p''(\cdot + \tau u h) - g(\cdot + \tau u h) \|_{L^2}}_{\leq \varepsilon} + \underbrace{\| g(\cdot + \tau u h) - g \|_{L^2}}_{\to 0 \text{ as } \tau u h \to 0} + \underbrace{\| g - p'' \|_{L^2}}_{\leq \varepsilon}$$

Q.E.D.

Back to nonparametric regression.

So far,

— Constrained LS, general, sub-optimal in non-Donsker (entropy integral diverges)

— Projection (Fourier truncation), optimal L^2-rate in specialized case

Local polynomial fitting.

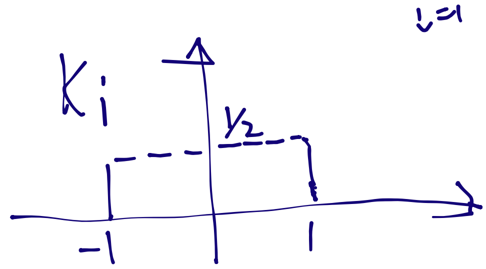
— Able to achieve pointwise bound.

— Works well without need to find orthonormal basis.

Warmup: Nadaraya-Watson estimator. (Data: $(X_i, Y_i)_{i=1}^n$). X_i not necessarily i/n, X_i deterministic

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

e.g.



$\hat{f}_n$ takes average of obs within $[x-h, \cancel{x+h}]$

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

$$W_{n,i}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}$$

weight of i-th data when estimating for x.

$$b(x_0) = \frac{\sum_{i=1}^{n} f^*(x_i)\cdot K\left(\frac{x_i-x_0}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i-x_0}{h}\right)} - f^*(x_0)$$

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

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

When $f \in \Sigma(\beta, L)$ for some $\beta \in (0,1]$.

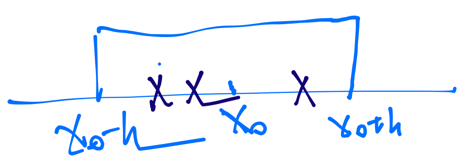
$$|f(x_1) - f(x_2)| \leq L\cdot|x_1 - x_2|^\beta.$$

Choice of K is unimportant, $K = \frac{1}{2}\mathbb{1}_{x\in[-1,1]}$

$$|b(x_0)| \leq \sum_{i=1}^{n} \underbrace{W_{n,i}(x_0)}_{\text{non zero only when } |x_i-x_0|\leq h}\cdot L\cdot|x_i - x_0|^\beta$$

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

$$= L\cdot h^\beta.$$



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

$$\leq \max_i |W_{n,i}(x_0)|\cdot\sum_{i=1}^{n}|W_{n,i}(x_0)|$$

$$\leq \left|\{j\in[n] : |x_j - x_0| \leq h\}\right|^{-1}.$$

If Equispaced design, for $h > n^{-1}$,

$$\left|\{j : |X_j - x_0| \le h\}\right| \ge \lfloor 2nh \rfloor.$$

Require weaker assumption than Equispace.

Only needs X_i's "dense enough".

$$\sigma(x_0)^2 \ge \frac{c}{nh}$$

$$h_n^* = n^{-\frac{1}{2\beta+1}} \qquad MSE(x_0) \le c' \cdot n^{-\frac{2\beta}{2\beta+1}}.$$

Dealing with $\beta > 1$?

Idea: use linear estimator, w/ better weights.

Alternative perspective on NW:

$$\hat{f}_n(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 \cdot K\left(\frac{X_i - x}{h}\right)$$

Zero-th order approximation to f around x.

How about higher-order expansion?

X_i Near x,

$$f(X_i) \approx f(x) + f'(x) \cdot (X_i - x) + \frac{f''(x)}{2!}(X_i - x)^2 + \cdots + \frac{f^{(\ell)}}{\ell!}(X_i - x)^\ell.$$

$$U(t) = \begin{bmatrix} 1 \\ t \\ t^2/2 \\ \vdots \\ t^\ell/\ell! \end{bmatrix}$$

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

$$\hat{f}_n(x) = e_1^T \hat{\theta}_n(x)$$

Still linear in $Y_1, Y_2, \ldots, Y_n$

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

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

(normalize by $\frac{1}{nh}$ since there are $O(nh)$ terms when K has compact support).

Key property of $\underline{W_{n,i}(\cdot)}$: When $B_{n,x} \succ 0$

Lemma. for any degree-l polynomial Q

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

(LP exactly reproduces polynomials, no bias).

Proof. $\sum_{i=1}^{n} Q(X_i) W_{n,i}(x)$ is LP estimator evaluated at x

by taking $(Q(X_i))_{1 \le i \le n}$ as inputs.

$$\underset{\theta \in \mathbb{R}^{l+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Q(X_i) - \theta^T u\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right)$$

min value ≥ 0.

Construct $\theta^* = \begin{bmatrix} Q(x) \\ Q'(x) \cdot h \\ \vdots \\ Q^{(\ell)}(x) h^\ell \end{bmatrix} \in \mathbb{R}^{\ell+1}$

(Obj function at θ^*) $= 0$.

$B_{n,x} \succ 0 \Rightarrow$ minimizer is unique.

So the LP outputs $Q(x)$

completing the proof of lemma.

Consequently, $\sum_{i=1}^n W_{n,i}(x) = 1$

for $k = 1, 2, \dots, \ell$ $\quad \sum_{i=1}^n (x_i - x)^k W_{n,i}(x) = 0$.

$\left(Q(z) := (z-x)^k \right)$

Analysis of LP. $\quad f^* \in \Sigma(\beta, L)$

$\ell :=$ largest int strictly less than β.

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

$$= \sum_{i=1}^n W_{n,i}(x_0) \left[\cancel{\sum_{k=1}^{\ell-1} \frac{f^{(k)}(x_0)(x_i - x_0)^k}{k!}} + \frac{f^{(\ell)}(x_0 + \tau_i(x_i - x_0))}{\ell!} (x_i - x_0)^\ell \right]$$

(By lemma)

Furthermore, $\sum_{i=1}^n W_{n,i}(x_0) \cdot (x_i - x_0)^\ell \cdot \frac{f^{(\ell)}(x_0)}{\ell!} = 0$.

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

$$|b(x_0)| \le \sum_{i=1}^{n} |W_{n,i}(x_0)|\cdot \frac{L}{l!}\cdot |\tau_i(x_i-x_0)|^{\beta-l}\cdot (x_i-x_0)^l \quad (\tau_i\in[0,1])$$

$$\le \sum_{i=1}^{n} |W_{n,i}(x_0)|\cdot L\cdot h^{\beta} \quad (K \text{ supported on } [-1,1]).$$

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

Holding true under

- K supported on $[-1,1]$
- $B_{n,x} \succ 0$.

Need to bound (1) $\max_{1\le i\le n} |W_{n,i}(x_0)|$ (2) $\sum_{i=1}^{n} |W_{n,i}(x_0)|$.

(1). $$|W_{n,i}(x_0)| = \frac{1}{nh}\left|e_1^T B_{n,x}^{-1}\, u\left(\frac{x_i-x_0}{h}\right)\cdot \underbrace{K\left(\frac{x_i-x_0}{h}\right)}_{|K|\le K_{max}}\right|$$

$$\le \frac{K_{max}}{nh}\cdot \cancel{\|e_1\|_2}\cdot \left\|u\left(\frac{x_i-x_0}{h}\right)\right\|_2\cdot |||B_{n,x}^{-1}|||_{op}$$

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

$$\le \sum_{k=0}^{l}\frac{1}{(k!)^2} \le 3.$$

$$|W_{n,i}(x_0)| \le \frac{2}{nh}K_{max}\cdot |||B_{n,x}^{-1}|||_{op}.$$

Assume that $B_{n,x} \succeq \lambda_0 I$ for some $\lambda_0>0$ independent of (n,h)

See Tsybakov, w/ equispaced design

$B_{n,x} \to$ something positive definite as $n \to +\infty$, $h \to 0$, $nh \to +\infty$.

(When n large enough, $B_{n,x} \succcurlyeq \lambda_0 I$).

e.g. When $X_i \overset{iid}{\sim} p$.

$$B_{n,x} = \frac{1}{nh} \sum_{i=1}^{n} u\left(\frac{X_i - x}{h}\right) \cdot u\left(\frac{X_i - x}{h}\right) \mathbb{1}\{X_i \in [x-h, x+h]\}$$

Using empirical process / LLN on $\mathbb{R}^{(l+1)\times(l+1)}$.

$$\left|\left|\left| B_{n,x} - \frac{\mathbb{E}_p\left[u\left(\frac{X-x}{h}\right) \cdot u\left(\frac{X-x}{h}\right)^T \mathbb{1}\{X \in [x-h, x+h]\}\right]}{h} \right|\right|\right|_{op} \leq \cdots$$

---

(2). $\sum_{i=1}^{n} |W_{n,i}(x)|$

$$\leq \frac{2K_{max}}{nh\lambda_0} \left|\{i : |X_i - x| \leq h\}\right| \leq \frac{4K_{max} a_0}{\lambda_0}.$$

Assumption. for any interval A

$$\frac{1}{n} \left|\{i : X_i \in A\}\right| \leq a_0 \cdot \max\left(|A|, \frac{1}{n}\right)$$

To conclude.

- $B_{n,x} \succcurlyeq \lambda_0 I_{l+1}$

- K supported on $[-1,1]$, $|K| \le K_{max}$
- $\frac{1}{n} \left| \{ i : x_i \in A \} \right| \le a_0 \cdot \max\left(|A|, \frac{1}{n} \right)$

We have

$$|b(x_0)| \le \frac{4 K_{max} a_0}{\lambda_0} \cdot L \cdot h^{\beta}$$

$$\sigma^2(x_0) \le \frac{8 K_{max} a_0}{\lambda_0^2} \cdot \frac{1}{nh}.$$

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

Information theoretically optimal?

Tools I introduced earlier.

- Le Cam two point
- Bayesian CR

(• Fano's method using mutual information).

Recall Thm (Le Cam) for any $f_0, f_1 \in \mathcal{F}$

$$\inf_{\hat{T}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[|\hat{T} - T(f)|^2 \right] \ge \inf_{\hat{T}} \sup_{f \in \{f_0, f_1\}} \mathbb{E}\left[|\hat{T} - T(f)|^2 \right]$$

(hw 1).

$$\ge \frac{1}{8} |T(f_1) - T(f_0)|^2 \cdot \{ 1 - d_{TV}(P_0, P_1) \}$$

Goal: Construct f_0, f_1 s.t. $\begin{cases} |T(f_1) - T(f_0)| \text{ large} \\ d_{TV} \leq \frac{1}{2}. \end{cases}$

Facts:

- $d_{TV} \leq \sqrt{\frac{1}{2} D_{KL}}$ (Pinsker)
- $D_{KL}\left(\bigotimes_{i=1}^{n} P_i \,\|\, \bigotimes_{i=1}^{n} Q_i\right) = \sum_{i=1}^{n} D_{KL}(P_i \| Q_i)$.

(Strategy applies to χ^2, Hellinger).

Application to nonpara regression.

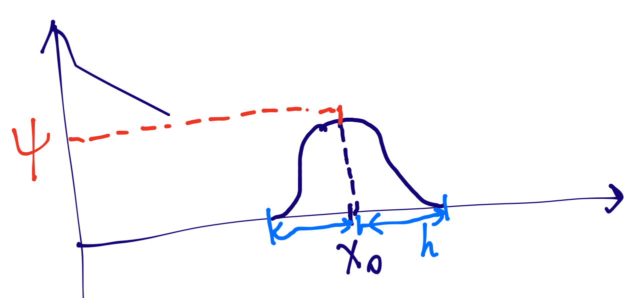
$$Y_i = f^*(x_i) + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0,1).$$

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

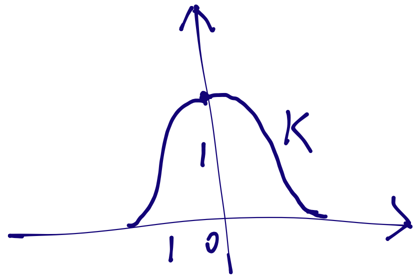
$$T(f) = f(x_0)$$

Idea of construction

- f_0 and f_1 differs only near x_0
- WLOG, assume $f_0 = 0$



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



K : C^∞ smooth, bounded support on $[-1,1]$.

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

Under construction.

- $$|f_1(x_0) - f_0(x)| = \psi.$$

- $$2 d_{TV}^2\left(P_1^{(n)}, P_0^{(n)}\right) \leq D_{KL}\left(P_1^{(n)} \| P_0^{(n)}\right)$$

$$= \sum_{i=1}^n D_{KL}\left(P_{1,i} \| P_{0,i}\right)$$

($P_{z,i}$ denotes the distribution of Y_i under f_z for $z \in \{0,1\}$).

$$= \frac{1}{2} \sum_{i=1}^n \left(f_1(x_i) - f_0(x_i)\right)^2$$

$$\leq \frac{\psi^2}{2} \cdot \left|\{i : |x_i - x_0| \leq h\}\right|.$$

(under same assumption as LP)

$$\leq \frac{\psi^2}{2} \cdot a_0 \cdot nh$$ (when $h > 1/n$).

$$\inf_{\hat{T}} \sup_{f \in \{f_0, f_1\}} \mathbb{E}\left[\underbrace{|\hat{T} - T(f)|^2}\right] \geq \frac{1}{8}\psi^2 \cdot \left\{1 - \frac{\psi}{2}\sqrt{a_0 nh}\right\}.$$

Need ~~to~~ make sure $f_1 \in \Sigma(\beta, L)$.

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

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

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

($K^{(l)}$ is a Lipschitz function)

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

$$\leq \text{sth} \cdot |x-y|^{\beta - l}$$

$$\text{sth} = 2 C_l \cdot \frac{\psi}{h^\beta} \leq L$$

- $\psi \leq \frac{L}{2C_l} \cdot h^\beta$
- $\psi \cdot \sqrt{a_0 nh} \leq 1$

$\rightarrow$ maximize ψ.

Take $h_n = c' \cdot n^{-\frac{1}{2\beta+1}}$

$\psi_n = c'' \cdot n^{-\frac{\beta}{2\beta+1}}$.

So $\inf_{\hat{f}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[|f(x_0) - \hat{f}|^2 \right] \gtrsim c \cdot n^{-\frac{2\beta}{2\beta+1}}$.