STA3000F   Lecture 11

Fixed design: $x_1, x_2, \dots x_n$ deterministic

$$Y_i = f^*(x_i) + \varepsilon_i$$ (e.g. denoising)

Random design. $(X_i, Y_i)_{i=1}^n \overset{iid}{\sim} \mathbb{P}$

$$f^*(x) = \mathbb{E}[Y_i \mid X_i = x]$$

Goal: generate an $\hat{f}_n$ s.t. $\|\hat{f}_n - f^*\|_{L^2(\mathbb{P})}$ small.

(Operationally, given a new data point X predict Y, s.t. err is small)

($\mathbb{E}[Y \mid X=x]$

Constrained LS.

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(Y_i - f(X_i)\right)^2.$$

(Assuming $f^* \in \mathcal{F}$)

From first-order-condition

$$\frac{1}{n} \sum_{i=1}^n \left(\hat{f}_n - f^*\right)(X_i)^2 \leq \frac{1}{n} \sum_{i=1}^n \left(Y_i - f^*(X_i)\right) \cdot \left(\hat{f}_n - f^*\right)(X_i)$$

Question: relate $\frac{1}{n}\sum_i^n (\hat{f}_n - f^*)(x_i)^2$ to $\|\hat{f}_n - f^*\|_{L^2(\mathbb{P})}$?

Solution:

$$Z_n(r) := \mathbb{E}\left[\sup_{\substack{\|h\|_{L^2(\mathbb{P})} \le r \\ h \in \mathcal{F}^*}} \left|(\mathbb{P}_n - \mathbb{P})\, h^2\right|\right].$$

$$\left(= \mathbb{E}\left[\sup_{\dots} \left|\frac{1}{n}\sum_i^n h(x_i)^2 - \mathbb{E}[h(x)^2]\right|\right]\right).$$

$$\le 2 \cdot \mathbb{E}\left[\sup_{\dots} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, h(x_i)^2\right|\right].$$

Once we're able to bound this, we'll have

$$r_n^2 = \|\hat{f}_n - f^*\|_{L^2(\mathbb{P})} \le \frac{1}{n}\sum_i^n \zeta_i (\hat{f}_n - f^*)(x_i) + Z_n\left(\|\hat{f}_n - f^*\|_{L^2}\right).$$

$$\le \mathbb{E}\left[\sup_{\substack{h \in \mathcal{F} \\ \|h\|_{L^2} \le r_n}} \left|\frac{1}{n}\sum_i^n \zeta_i h(x_i)\right|\right] + Z_n(r_n)$$

(w.h.p.)

using the same argument in Lec 8.

From hw 2. Assuming $\forall f \in \mathcal{F}$, $\|f\|_\infty \le M$.

$x \mapsto x^2$ is a $2M$-Lipschitz mapping

when $x \in [-M, M]$.

So we have

$$\mathbb{E}\left[\sup_{h \in \mathcal{F}^*(r)} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h^2(X_i) \right| \right] \leq 2M \cdot \mathbb{E}\left[\sup_{h \in \mathcal{F}^*(r)} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(X_i) \right| \right].$$

Applying the arguments we have seen yields the bounds.

- To improve failure prob, use concentration ineq for sup of empirical process.
- The contraction-based bound here is not tight..

  eg. If we have $Y_i = f^*(X_i) + \varepsilon_i$

  $$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

  when σ is small (high signal, low noise).

  or even w/ $\sigma = 0$

  Faster rates are expected, but not achieved using above arguments.

  (c.f. Mendelson (2012), "small ball argument").

Kernel density estimation.

$$X_1, X_2, \cdots X_n \overset{iid}{\sim} p^* \in \mathcal{P}.$$

Possible methods:

- MLE $\quad \hat{p} = \underset{p \in \mathcal{P}}{\arg\max} \frac{1}{n} \sum_i^n \log p(X_i)$

(then use empirical process to study the error).

- "skeleton estimator" / "integral probability metric".

(see Hw 3)

(theoretically nice, but computationally infeasible).

- Local methods for smoothness classes.

Suppose we have $X_1, X_2, \cdots X_n \overset{iid}{\sim} p^*$ on $\mathbb{R}$

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

where K is a kernel function

satisfying $\int_{\mathbb{R}} K(x)\, dx = 1.$

e.g. $K = \frac{1}{2} \mathbb{1}_{[-1,1]}$. local averaging

Analysis: at the point x_0 (deterministic)

- $\mathrm{var}(\hat{p}_n(x_0)) = \frac{1}{h^2 n} \mathrm{var}\left(K\left(\frac{X-x_0}{h}\right)\right)$

$$\leq \frac{1}{h^2 n} \cdot \int K^2\left(\frac{y-x_0}{h}\right) \cdot p(y)\, dy$$

$$\leq \frac{p_{max}}{h\, n} \int K^2(x)\, dx.$$

const.

- Bias.

Assuming p is Hölder w/ exponent β.

$(0<\beta\leq 1)$.

$(\ |p(x) - p(y)| \leq L \cdot |x-y|^{\beta} \quad \forall x, y\ ).$

$$\mathbb{E}[\hat{p}_n(x_0)] - p(x_0)$$

$$= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{y-x_0}{h}\right) \cdot \left(p(y) - p(x_0)\right) dy.$$

$$\leq \frac{L}{h} \int_{\mathbb{R}} K\left(\frac{y-x_0}{h}\right) \cdot |y-x_0|^{\beta}\, dy$$

$$\leq h^{\beta} \cdot L \cdot \int K(u) \cdot |u|^{\beta} du$$

const.

(Assuming $\int K(u) \cdot |u|^{\beta} du < +\infty$).

Bias - var tradeoff.

$$\mathbb{E} |\hat{p}_n(x_0) - p^*(x_0)|^2 \leq C \cdot \frac{1}{nh} + C' \cdot h^{2\beta}$$

(optimal $h_n = C_1 \cdot n^{-\frac{1}{2\beta+1}}$)

$$\mathbb{E}\left[|\hat{p}_n(x_0) - p^*(x_0)|^2 \right] \leq C_2 n^{-\frac{2\beta}{2\beta+1}}.$$

What if we have more smoothness?

$$p(x_0+uh) - p(x_0)$$

$$= p'(x_0) \cdot uh + \frac{p''(x_0)}{2!} (uh)^2 + \cdots + \frac{p^{(l-1)}(x_0)}{(l-1)!} (uh)^{l-1}$$

$$+ \frac{p^{(l)}(x_0 + \tau_u uh)}{l!} \cdot (uh)^l.$$

(for some $\tau_u \in [0,1]$).

Here we assume $p \in$ Hölder(β, L)

where $\beta = l + \gamma$, $l \in \mathbb{N}$, $\gamma \in (0,1]$.

$$\mathrm{Bias}(x_0) = \int_{\mathbb{R}} K(u) \cdot \left[p'(x_0) \cdot uh + p''(x_0) \cdot \frac{(uh)^2}{2} + \cdots + p^{(\ell)}(x_0) \cdot \frac{(uh)^\ell}{\ell!} \right] du$$

If K is symmetric,

$\int K(u)\, uh\, du = 0.$

$$+ \boxed{\int_{\mathbb{R}} K(u) \frac{\left(p^{(\ell)}(x_0 + \tau_u uh) - p^{(\ell)}(x_0)\right)}{\ell!} (uh)^\ell du.}$$

$$|\cdots| \leq L \cdot \int_{\mathbb{R}} |K(u)| \cdot |\tau_u uh|^\gamma \cdot (uh)^\ell du$$

$$\leq L \cdot h^\beta \int_{\mathbb{R}} |K(u)| \cdot |u|^\beta du.$$

The entire sum can be made 0 by careful choice of K.

Note: to cancel (≥ 2) terms, we need to choose K that is negative somewhere.

Def. ℓ-th order kernel. K such that

$j = 1, 2, \ldots \ell$, $\quad \int_{\mathbb{R}} u^j K(u)\, du = 0.$

(e.g. box and Gaussian are both 1st order).

Construction via Legendre polynomials.

$(\varphi_k)_{k=0}^{+\infty}$ orthonormal basis on $L^2([-1,1])$.

$$K(u) = \sum_{m=0}^{l} \varphi_m(0)\,\varphi_m(u) \cdot 1\{u \in [-1,1]\}.$$

Easy to show

- $\int_{-1}^{1} K(u)\,du = 1.$

$$K = \sum_{m=0}^{l} \varphi_m(0)\,\varphi_m$$

$$u^j = \sum_{m=0}^{l} b_m \varphi_m$$

- $\int_{-1}^{1} K(u)\,u^j\,du = \langle K, u^j \rangle_{L^2}$

$$= \sum_{m=0}^{l} b_m \cdot \varphi_m(0) = u^j \big|_{u=0} = 0.$$

(for $j = 1, 2, \dots l$).

Easy to see, $\int_{-1}^{1} K(x)^2\,dx < +\infty$, $\int_{-1}^{1} |K(x)| \cdot |x|^\beta < +\infty$ $(\forall \beta)$

So we get

$$|\text{bias}(X_0)| \leq C \cdot h^\beta$$

For any $\beta > 0$, we always have

$$MSE(X_0) \leq C \cdot \left(\frac{1}{nh} + h^{2\beta} \right)$$

Choosing $h = n^{-\frac{1}{2\beta+1}}$, we get the rate $n^{-\frac{2\beta}{2\beta+1}}$.

- Not hard to extend to multivariate case

you'll get the rate $n^{-\frac{2\beta}{2\beta+d}}$.

$$MISE = \int_{\mathbb{R}} MSE(x)\, dx$$

(For Hölder class, just integrate the bound above).

$\beta \in \mathbb{N}_+$, Sobolev class

$$\int |p^{(\beta)}(x)|^2 dx \le L^2 < +\infty$$

· variance. $\displaystyle\int_{\mathbb{R}} \sigma^2(x) dx = \frac{1}{nh^2} \int \operatorname{var}\left(K\left(\frac{X-x}{h}\right)\right) dx$

$$\le \frac{1}{nh} \int K^2(x)\, dx.$$

· Bias (assuming $(\beta-1)$-th order kernel).

$$b(x) = \int K(u) \frac{(uh)^\beta}{\beta!} \int_0^1 (1-\tau)^{\beta-1} p^{(\beta)}(x+\tau uh)\, d\tau\, du$$

$$\int b(x)^2 dx$$

$$\le \int \left(\int |K(u)| \cdot \frac{|uh|^\beta}{\beta!} \int_0^1 (1-\tau)^{\beta-1} p^{(\beta)}(x+\tau uh)\, d\tau\, du\right)^2 dx$$

Detour: generalized Minkowski ineq.

· Minkowski ineq.

$$\| g(\cdot, u_1) + g(\cdot, u_2) + \cdots + g(\cdot, u_m) \|_{L^2} \le \sum_{j=1}^m \| g(\cdot, u_j) \|_{L^2}.$$

· Generalized Minkowski:

$$\left(\int \left| \int g(x,u)\,du \right|^2 dx\right)^{1/2} \le \int \left(\int |g(x,u)|^2 dx\right)^{1/2} du.$$

Using this ineq, we have (indep of x)

$$\int b(x)^2 dx \le h^{2\beta} \left[\int \left(\int K^2(u) \cdot |u|^{2\beta} \underbrace{\left(\int_0^1 p^{(\beta)}(x+\tau u h)\,d\tau\right)^2}_{\text{by Cauchy-Schwarz}} dx \right)^{1/2} du \right]^2$$

$$\left(\int_0^1 p^{(\beta)}(x+\tau u h)\,d\tau\right)^2 \le \int_0^1 p^{(\beta)}(x+\tau u h)^2\,d\tau$$

$$\le h^{2\beta} \left[\int |K(u)| \cdot |u|^\beta \underbrace{\left(\int_{\mathbb{R}} \int_0^1 p^{(\beta)}(x+\tau u h)^2\,d\tau\,dx \right)^{1/2}}_{\le L} du \right]^2$$

$$\le L^2 \cdot h^{2\beta} \cdot \left(\int |K(u)| \cdot |u|^\beta du\right)^2$$

Then apply the same trade off.

Question: asymptotic optimal bandwidth?

e.g. in parametric estimation.

$$\| \hat{\theta}_n - \theta^* \|_2 \le O_p(1/\sqrt{n}).$$

We can actually show that asymptotic MSE $\lim_{n\to+\infty} \mathbb{E}[n \cdot \|\hat{\theta}_n - \theta^*\|_2^2] \geq I(\theta^*)^{-1}$ achieved asymptotically by MLE.

- Operationally, we can choose estimators by comparing asymptotic MSE.

Can we do the same for KDE (e.g. choosing h, K).

Do we have sth. like

$$\lim_{n\to+\infty} \mathbb{E}\left[n^{\frac{2\beta}{2\beta+1}} \|\hat{p}_n - p^*\|_{L^2}^2 \right] \geq \text{sth. ?}$$

Lack of asymptotic optimality.

Let p be a density on $\mathbb{R}$, $\int (p''(x))^2 dx \leq L^2$.

optimal rate: $n^{-4/5}$ (MISE).

(achieved using KDE, 1st order kernel).

Thm: Suppose we use a second-order kernel, $\forall \varepsilon > 0$. take $h = \frac{n^{-1/5}}{\varepsilon} \int K(u)^2 du$

then we have

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

Rmk:

- This is for fixed p, $n\to+\infty$.
  If we take finite n minimax risk,
  we'll still get $C_0 n^{-4/5}$. where $C_0>0$ is a constant.
- Fixed n, the "hardest" problem in $S(2,L)$ at sample size n depends on n.
- For $S(2,L)$, every density is asymptotically easier than the average.
  (every density in this class is a little bit more smooth).

Proof ideas

- variance.

$$\int \sigma^2(x)\, dx = \frac{1}{nh}\int K(u)^2 du + o\left(\frac{1}{nh}\right)$$

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

by using second-order kernel.

So we trade off $\frac{1}{nh} + o(h^4)$.

Proof idea for the bias term:

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

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

We have

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

and

$$\int |b(x) - \tilde{b}(x)|^2 dx \longrightarrow 0.$$

Key step: $\|p''(\cdot + h) - p''\|_{L^2} \xrightarrow{h \to 0} 0$

(by density argument in L^2, approximation using "nice" functions).