STA 300 OF   Lecture 10

Nonparametric estimation

- Density estimation

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

Goal: estimate p^* (under some metric)

e.g. $|\hat{p}_n(x_0) - p^*(x_0)|$ for given x_0

$\|\hat{p}_n - p^*\|_{L^\infty}$, $d_{TV}(\hat{p}_n, p^*)$

$D_{KL}(\hat{p}_n \| p^*)$, $D_{KL}(p^* \| \hat{p}_n)$

$W_2(\hat{p}_n, p^*)$, ...

- Regression. $(X_i, Y_i)_{i=1}^n$ iid.

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

(random design).

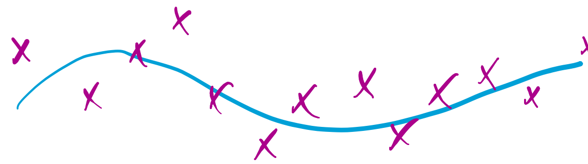
Simplification: fixed design

$x_1, x_2, \cdots, x_n$ deterministic

Try to recover f^* on $x_1, x_2, \cdots, x_n$

(n-dim vector).

e.g. denoising.



signal.

Fixed-design regression (i.e. denoising problem).

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

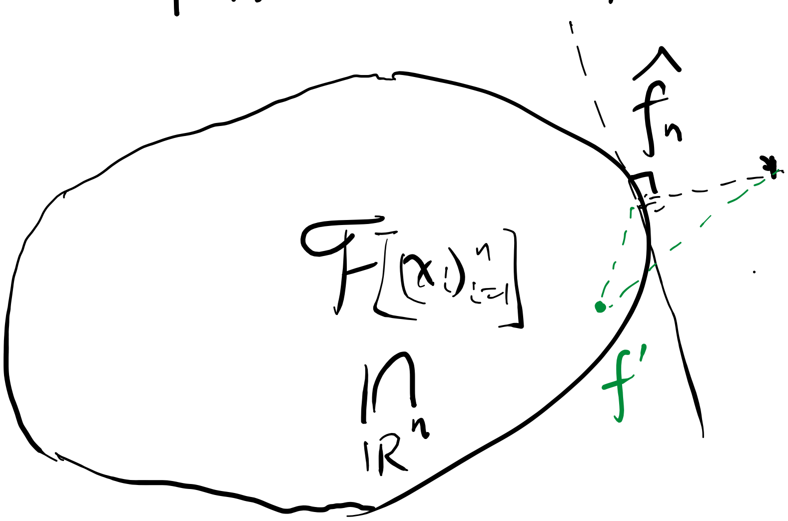
(x_i's deterministic).

MLE $\Longleftrightarrow$ least square.

$$\hat{f}_n := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(x_i) \right)^2 \right\}$$

(Efficiently computable when $\mathcal{F}$ is convex)

First-order optimality condition.



$Y = (Y_i)_{i=1}^n$

(in literature, convexity can be relaxed to star-convexity)

$$\sum_{i=1}^{n} \left(Y_i - \hat{f}_n(x_i)\right) \cdot \left(f'(x_i) - \hat{f}_n(x_i)\right) \leq 0$$

$$(\forall f' \in \mathcal{F}).$$ "Basic ineq".

Substitute with f^*, re-arranging.

$$\sum_{i=1}^{n} \left[(f^* - \hat{f}_n)(x_i)\right]^2 \leq \sum_{i=1}^{n} (\hat{f}_n - f^*)(x_i) \cdot \varepsilon_i$$

Denote $\hat{\Delta}_n := \hat{f}_n - f^*$, $\|f\|_n := \sqrt{\frac{1}{n}\sum_i f(x_i)^2}$

we have $\|\hat{\Delta}_n\|_n^2 \leq \langle \hat{\Delta}_n, \varepsilon \rangle_n$

(naive bound : $\|\hat{\Delta}_n\|_n^2 \leq \|\hat{\Delta}_n\|_n \cdot \|\varepsilon\|_n \Rightarrow \|\hat{\Delta}_n\|_n \leq \|\varepsilon\|_n$

this leads to useless bound, $\|\varepsilon\|_n = O(1)$).

We know $\hat{\Delta}_n \in \mathcal{F}^* = \{f - f^* : f \in \mathcal{F}\}$.

$$\langle \hat{\Delta}_n, \varepsilon \rangle_n \leq \sup_{\substack{\|h\|_n \leq \|\hat{\Delta}_n\|_n \\ h \in \mathcal{F}^*}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i).$$

Define : "Gaussian complexity"

$$\mathcal{G}_n(r) := \mathbb{E}\left[\sup_{\substack{\|h\|_n \leq r \\ h \in \mathcal{F}^*}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(x_i)\right|\right]$$

where $\varepsilon_i \overset{iid}{\sim} N(0,1)$.

We have $\|\hat{\Delta}_n\|_n^2 \le \langle \varepsilon, \hat{\Delta}_n \rangle_n \le G_n(\|\hat{\Delta}_n\|_n)$.

similar "chicken-egg problem"

Thm: Suppose $G_n(r) \le \phi_n(r)$

s.t. $\phi_n(cr) \le c^\alpha \phi_n(r)$ for some $\alpha < 2$. $(*)$

If δ_n solves $\delta_n^2 = \phi_n(\delta_n)$

then we have $\|\hat{f}_n - f^*\|_n \le C \cdot \delta_n$

w.h.p.

(Proof: see lecture 8).

Verifying growth condition on Gaussian complexity.

$(c > 1)$.

$\forall f \in \mathcal{F}^* \cap \mathbb{B}(cr)$, we have $\dfrac{f}{c} \in \mathcal{F}^* \cap \mathbb{B}(r)$.

By convexity, we have

So $\mathbb{E}\left[\sup_{f \in \mathcal{F}^* \cap \mathbb{B}(cr)} \langle \varepsilon, f \rangle_n\right] \le c\, \mathbb{E}\left[\sup_{f \in \mathcal{F}^* \cap \mathbb{B}(r)} \langle \varepsilon, f \rangle_n\right]$

So Gaussian complexities grow sub-linearly.

(*) is satisfied with $\alpha \le 1$.

Key observation:

MGF of Rade $\le$ MGF of $N(0,1)$

and previous proofs for Rademacher complexity are basically bounding it using MGF of $N(0,1)$.

Thm. $$G_n(r) \le \frac{c}{\sqrt{n}} \int_0^r \sqrt{\log N(\delta; \mathcal{F}^* \cap \mathbb{B}_n(r), \|\cdot\|_n)}\, d\delta$$

(Proof: see lecture 6).

---

e.g. $\mathcal{F} = \{ f : [0,1] \to [0,1],\ |f(x) - f(y)| \le |x-y|,\ \forall x, y \}$.

$$N(\delta; \mathcal{F}^* \cap \mathbb{B}_n(r), \|\cdot\|_n)$$

$$\le N(\delta; \mathcal{F}^*, \|\cdot\|_n) \le N(\delta; \tilde{\mathcal{F}}, \|\cdot\|_n)$$

$(\mathcal{F}^* \subseteq \tilde{\mathcal{F}} := \{ f : [0,1] \to [-1,1],\ |f(x) - f(y)| \le 2|x-y|\ \forall x, y \})$.

$$\underset{\text{(Rudelson-Vershynin)}}{\le} \exp\left(c_1 \operatorname{fat}_{c_2 \delta}(\tilde{\mathcal{F}})\right).$$

(From Lecture 8), $\text{fat}_{c_2\delta}(\tilde{\mathcal{F}}) \leq C'/\delta$.

Applying the results.

$$\mathcal{G}_n(r) \lesssim \frac{c}{\sqrt{n}} \int_0^r \sqrt{\log N(\delta; \tilde{\mathcal{F}}, \|\cdot\|_n)}\, d\delta$$

$$\lesssim \frac{c}{\sqrt{n}} \int_0^r \sqrt{\frac{c_1 C'}{\delta}}\, d\delta \lesssim c_3 \sqrt{\frac{r}{n}}.$$

Solving for fixed pt, $r^2 \lesssim c_3 \sqrt{\frac{r}{n}}$

$$r_n \asymp n^{-1/3}.$$

So $\|\hat{f}_n - f^*\|_n \lesssim n^{-1/3}$ w.h.p.

Final exam: Dec 15, noon - 16, 23:59

---

e.g. In general, β-Hölder class.

$\beta > 0$, $\beta = k + \gamma$, $k \in \mathbb{N}$, $\gamma \in [0,1)$.

$$\Sigma(\beta, L) := \left\{ f : [0,1]^d \to [0,1],\ \forall \alpha \text{ multi-index s.t. } |\alpha| = k,\ |\partial^\alpha f(x) - \partial^\alpha f(y)| \leq L |x-y|^\gamma\ \forall x, y \in [0,1]^d \right\}$$

(A multi-index is non-negative-integer-valued vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$, $|\alpha| := \sum_{i=1}^d \alpha_i$

$$\partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}.$$

Thm For any Q,

$$N\left(\varepsilon; \Sigma(\beta, 1), \|\cdot\|_{L^2(Q)}\right) \le c_1 \exp\left(\left(\frac{c_2}{\varepsilon}\right)^{d/\beta}\right).$$

(c_1, c_2 may depend on (d, β)).

Implication to Hölder regression:

$$G_n(r) \le \frac{c}{\sqrt{n}} \int_0^r \sqrt{\log N(\delta; \mathcal{F}, \|\cdot\|_n)}\, d\delta$$

$$\left(\le \frac{c'}{\sqrt{n}} \int_0^r \left(\frac{1}{\delta}\right)^{\frac{d}{2\beta}} d\delta \quad \text{may diverge}\right).$$

To avoid divergence, we note that

$$N(\delta; \mathcal{F}, \|\cdot\|_n) \le N\left(\delta; [0,1]^n, \|\cdot\|_n\right) \le \left(\frac{c}{\delta}\right)^n$$

$$\int_0^r \sqrt{\log N(\delta; \mathcal{F}, \|\cdot\|_n)}\, d\delta$$

$$= c\int_0^{r_0} \sqrt{n \log(1/\delta)}\, d\delta + c\int_{r_0}^r \left(\frac{1}{\delta}\right)^{\frac{d}{2\beta}} d\delta$$

$\beta > d/2$, pick $r_0 = 0$.

$$G_n(r) \le \frac{c}{\sqrt{n}} \cdot \int_0^r \delta^{-\frac{d}{2\beta}}\, d\delta \le \frac{c'}{\sqrt{n}} r^{1-\frac{d}{2\beta}}$$

Implication to rate: $r_n^2 = \dfrac{r_n^{1-\frac{d}{2\beta}}}{\sqrt{n}} \Rightarrow r_n \asymp n^{-\frac{\beta}{d+2\beta}}$

- $\beta < d/2$,

$$G_n(r) \le r_0 + \frac{c}{\sqrt{n}} \int_{r_0}^{r} \delta^{-\frac{d}{2\beta}} \, d\delta$$

$$\le r_0 + \frac{c}{\sqrt{n}} \cdot \left(r_0^{1-\frac{d}{2\beta}} - \cancel{r^{1-\frac{d}{2\beta}}} \right)$$

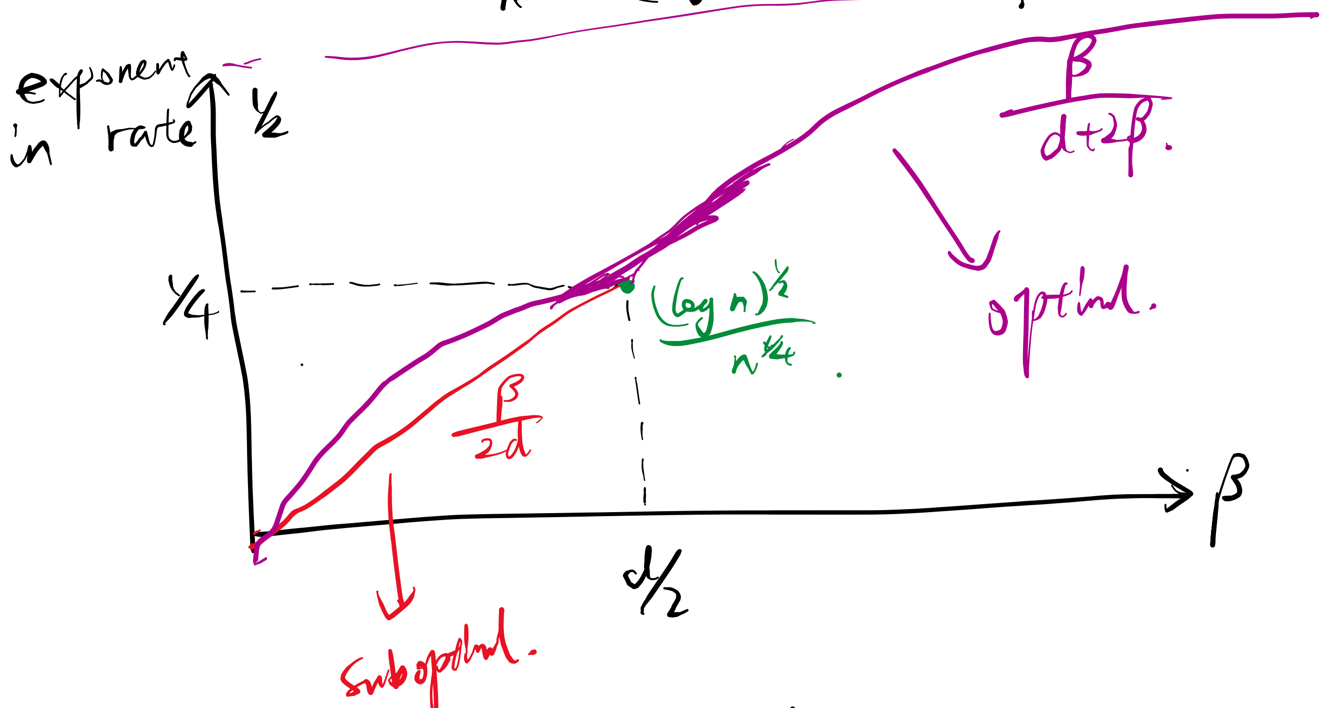
$$\le r_0 + \frac{c}{\sqrt{n}} r_0^{1-\frac{d}{2\beta}}$$

Choose $r_0 = n^{-\beta/d}$ $\quad G_n(r) \le c \cdot n^{-\beta/d}$.

Implication to rate: $r_n \asymp n^{-\frac{\beta}{2d}}$

- $\beta = d/2$, we can similarly get

$$r_n = (\log n)^{1/2} \cdot n^{-1/4}$$



- Suboptimality in $\beta \le \frac{d}{2}$ regime

(optimal rate for Hölder class is always $n^{-\frac{\beta}{d+2\beta}}$).

intrinsic for constrained least-square (regularized).

- For Hölder class, projection/local poly leads to optimal rates.

- In general, divergent integral bounds is known as "non-Donsker", usually strong indicator of suboptimality of least squares method.

(construction of optimal estimator can be ad-hoc).

e.g. multivariate convex functions.

$$\mathcal{F} = \{ f : [0,1]^d \to [0,1] \mid |f(x) - f(y)| \le |x-y|, \ f \text{ is convex} \}.$$

Optimal rate similar to $\Sigma(2,1)$

$n^{-\frac{2}{d+4}}$, achieved by LS when $d \le 3$.

LS becomes suboptimal for $d \ge 4$.

Projection estimators for Sobolev classes.

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

- $\beta \in \mathbb{N}^+$, $S(\beta, L) = \{ f : [0,1]^d \to [0,1],$

$$\sum_{|\alpha| \le \beta} \int |\partial^\alpha f(x)|^2 dx < L^2 \}$$

Under Fourier transform.

Let $\varphi_1, \varphi_2, \dots, \varphi_j, \dots$ be Fourier basis in $[0,1]^d$.

- $d=1$ $\quad \sum_{j=1}^{+\infty} \langle f, \varphi_j \rangle_{L^2}^2 \cdot j^{2\beta} \le L^2.$

- In general dim, φ_z for $z \in \mathbb{Z}^d$.

$$\sum_{z \in \mathbb{Z}^d} \langle f, \varphi_j \rangle_{L^2}^2 \cdot |z|^{2\beta} \le L^2.$$

Fixed design setting

in 1-D, $\quad x_i = i/n \quad$ (equispaced)

in general dimensions,

$$x_\alpha = \frac{\alpha}{m} \quad \text{where } \alpha = (\alpha_1, \alpha_2, \dots, \alpha_m) \in \mathbb{N}^d$$

$$m^d = n$$

$\varphi_1, \varphi_2, \cdots \varphi_n$ "Discrete Fourier basis".

Orthonormal basis of $(\mathbb{R}^n, \|\cdot\|_n)$.

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{n} \varphi_i \cdot c_i \;\middle|\; \sum_{j=1}^{n} c_j^2 \cdot j^{\frac{2\beta}{d}} \leq L^2 \right\}.$$

. We know that constrained LS is suboptimal for $\mathcal{F}$ (when $\beta < \frac{d}{2}$).

Alternative idea: truncate in Fourier domain.

. Observe $Y \in \mathbb{R}^n$

. $\hat{c}_j := \langle Y, \varphi_j \rangle_n$ (for $j = 1, 2, \cdots n$)

. $\hat{f}_n := \sum_{j=1}^{N} \hat{c}_j \varphi_j$ $(N < n)$.

Analysis:

$$\mathbb{E}\left[\|\hat{f}_n - f^*\|_n^2\right] = \frac{1}{n}\sum_{j=1}^{N} \mathbb{E}\left[|\hat{c}_j - c_j^*|^2\right] + \frac{1}{n}\sum_{j=N+1}^{n} (c_j^*)^2$$

$$= \frac{N}{n} + \sum_{j=N+1}^{n} (c_j^*)^2$$

$\leq L^2$

$$\leq \frac{N}{n} + \frac{1}{(N+1)^{2\beta/d}} \cdot \boxed{\sum_{j=N+1}^{n} j^{\frac{2\beta}{d}} \cdot (c_j^*)^2}$$

$$\leq \frac{N}{n} + \frac{L^2}{N^{2\beta/d}}$$

Choose $N = n^{\frac{d}{2\beta+d}} \Longrightarrow r_n = n^{-\frac{\beta}{2\beta+d}}$.