1 Yang & Barron 1

Minimax Estimation:

$\theta \in S \subseteq \Theta$

"parameter", could be functions, densities, etc.

Loss function: $d(\theta, \hat{\theta})$

$\hat{\theta}$ = estimator of θ, $(x_1, \cdots, x_n) \mapsto \hat{\theta}$.

$r_n = \min_{\hat{\theta}} \max_{\theta \in S} \mathbb{E}[d^2(\theta, \hat{\theta})]$: function of S and n. $r_n = r_n(S)$.

Examples: (non-parametric regression)

$y_i = f(x_i) + \varepsilon_i$, $f: [0,1] \to \mathbb{R}$, $i = 1, \cdots, n$, $\varepsilon_i \sim N(0, \sigma^2)$

Data: (x_i, y_i), $i = 1, \cdots, n$

$\theta \equiv f$, $S^a = \{f \in C^a \mid f$ is a-times cts differentiable, $\|f\| \le K\}$ (sobolev-ball)

$r_n(S^a) \asymp \left(\frac{1}{n}\right)^{\frac{2a}{2a+1}}$ $d^2(f, \hat{f}) = \int_0^1 |f(t) - \hat{f}(t)|^2 dt$.

$S^k \subseteq S^{k-1} \subseteq \cdots S^1$

Need to measure complexity of parameter space.

ε-packing

Finite set $N_\varepsilon \subset S$ s.t. $\forall \theta, \theta' \in N_\varepsilon$, $\theta \neq \theta'$ $d(\theta, \theta') > \varepsilon$

packing ε-entropy $\log |N_\varepsilon|$

Examples:

(a) $\{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le 1\}$.

pack it with Euclidean balls. (last semester).

$M(\varepsilon) = \log |N_\varepsilon| = d \log\left(\frac{1}{\varepsilon}\right)$

(b) Sobolov-ball S^a (Birman, 1967)

d-dimension a = smoothness.

$\log |N_\varepsilon| \asymp \left(\frac{1}{\varepsilon}\right)^{d/a}$. (also for covering)

ε-covering:

A set $G_\varepsilon \subset S$ is an ε-cover of S if $\forall\, \theta \in S$, $\exists$ some $\theta' \in G_\varepsilon$ s.t. $d(\theta, \theta') < \varepsilon$

$\underbrace{\log |G_\varepsilon|}_{V(\varepsilon)}$ = covering entropy

Step 1: choose $\varepsilon_n \to 0$ s.t. $\varepsilon_n^2 n = V(\varepsilon_n)$

Step 2: choose $\varepsilon_{n,d} \to 0$ s.t.

$M(\varepsilon_{n,d}) = 4n\varepsilon_n^2 + 2\log 2$

Thm 1: $\min_{\hat\theta} \max_{\theta \in S} \mathbb{E}\left[d^2(\theta, \hat\theta)\right] \geq (\text{const})\, \varepsilon_{n,d}^2$

Example: $\varepsilon_n^2 n = \left(\frac{1}{\varepsilon_n}\right)^{d/\alpha} \Rightarrow \varepsilon_n^2 = \left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+d}}$

$d=1,\ \alpha=2 \to n^{-\frac{4}{5}}$

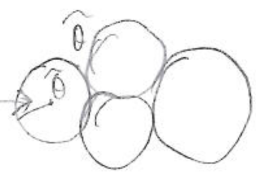
Proof: Construct $N_{\varepsilon_{n,d}}$ packing set (with $|N_{\varepsilon_{n,d}}|$ elements)

Let $\hat\theta$ be any estimator.

set $\tilde\theta = \arg\min_{\theta' \in N_{\varepsilon_{n,d}}} d(\theta', \hat\theta)$

claim: $\forall\, \theta \in N_{\varepsilon_{n,d}}, \theta \neq \tilde\theta$, $d(\theta, \hat\theta) \geq A\,\varepsilon_{n,d}/2$.

$\frac{\varepsilon_{n,d}}{2}$



$$\min_{\hat\theta} \max_{\theta \in S} \mathbb{P}_\theta\left(d(\theta, \hat\theta) \geq \frac{A}{2}\varepsilon_{n,d}\right)$$

$$\geq \min_{\hat\theta} \max_{\theta \in N_{\varepsilon_{n,d}}} \mathbb{P}_\theta\left(d(\theta, \hat\theta) \geq \frac{A}{2}\varepsilon_{n,d}\right)$$

$$= \min_{\tilde\theta} \max_{\theta \in N_{\varepsilon_{n,d}}} \mathbb{P}_\theta(\theta \neq \tilde\theta)$$

$\tilde\theta$ is implicitly defined through $\hat\theta$.

$$\geq \min_{\tilde\theta} \sum_{\theta \in N_{\varepsilon_{n,d}}} w(\theta)\, \mathbb{P}_\theta(\theta \neq \tilde\theta) = \min_{\tilde\theta} \mathbb{P}_w(\theta \neq \tilde\theta)$$

$w(\theta) \equiv$ uniform over $N_{\varepsilon n d}$ (parameter space)
"codewords" $\equiv$ ball centers $\theta \in N_{\varepsilon n d}$.
"channel" given $\theta \to (x_1, \cdots, x_n)$.

Fano's $\leq$: Estimate $\theta \in N_{\varepsilon n d}$ from X

$$P_{err} \geq \frac{H(\theta|X) - 1}{\log |N_{\varepsilon n d}|}.$$

our case: Θ is uniform over $N_{\varepsilon n d}$, $H(\Theta) = \log |N_{\varepsilon n d}|$.

$$I(\Theta : X) = H(\Theta) - H(\Theta|X).$$

$$\mathbb{P}_w(\theta \neq \hat{\theta}) \geq 1 - \frac{I(\Theta; X_1^n)}{\log |N_{\varepsilon n d}|} + o(1)$$

Remains to upper bound

(new): $I(\Theta : X_1^n) = D(P(\Theta, X_1^n) \,\|\, \mathbb{P}(\theta)\,\mathbb{P}(X_1^n))$

$$= \sum_\theta w(\theta) \underbrace{\int P(x_1^n|\theta) \log \frac{P(x_1^n|\theta)}{P_w(x_1^n)}\, dx}_{D(P(x_1^n|\theta) \,\|\, P_w(x_1^n))}$$

claim: $\leq \sum_\theta w(\theta)\, D(P(x^n|\theta) \,\|\, q(x^n))$

for any choice of $q(\cdot)$ over $\underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{n\text{-times}}$

$$\leq \max_{\theta \in N_{\varepsilon n d}} D(\mathbb{P}(x^n|\theta) \,|\, q(x^n)).$$

Last step: construct a ε-covering in KL-divergence.
G_{ε_n} For all $\theta \in S$, $\exists\, \bar{\theta} \in G_{\varepsilon_n}$, with

$$d_K^2(\theta, \bar{\theta}) \leq \varepsilon_n^2$$

choose $q(x^n) = \frac{1}{|G_{\varepsilon_n}|} \sum_{\theta \in G_{\varepsilon_n}} P(x^n|\theta)$.

$$\mathbb{P}_{err} \geq 1 - \frac{\overset{n\varepsilon_n^2}{\overset{\|}{V(\varepsilon_n)}} + n\varepsilon_n^2}{M(\varepsilon_n, d) = 4n\varepsilon_n^2}$$

estimation $\xrightarrow{\text{packing}}$ hypothesis testing

unified view of minimax