Lecture 5

Review * Bias estimation. $n(\delta) \equiv n_{1/3}(\delta) = \Theta\left(\frac{1}{\delta^2}\right)$

Lower bounds

- Le Cam's 2-point method

$$R_n \geq \frac{|p-q|^2}{8}\left(1 - \sqrt{n D(p\|q)}\right)$$

- Fano's method

$$R_n \geq \delta^2 \left(1 - \frac{n \max_{1 \le i,j \le M} D(p_i \| p_j) + 1}{\log M}\right)$$

where $\min_{i \neq j} |p_i - p_j| \geq 2\delta$

* Gaussian mean estimation $R_n \leq \frac{d\delta^2}{n}$

Agenda * Lower bound for Gaussian mean estimation

* A generic estimation procedure (motivation for Scheffe Estimator)

A Gaussian mean estimation : Lower bounds

$$R_n = \min_{\hat{\mu}} \max_{\mu \in \mathbb{R}^d} \mathbb{E}_\mu \left[\|\mu - \hat{\mu}\|_2^2 \right]$$

Le Cam's 2 point method

Consider μ, ν s.t. $\|\mu - \nu\|_2 = 2\delta$. Then,

$$R_n \geq \frac{\delta^2}{2}\left(1 - \sqrt{\frac{\ln 2}{2} n D(P_\mu \| P_\nu)}\right)$$

Note that $D(P_\mu \| P_\nu)$

$$= \frac{1}{2\sigma^2} \mathbb{E}_\mu \left[\sum_{i=1}^{d} (X_i - \nu_i)^2 - (X_i - \mu_i)^2 \right] \frac{1}{\ln 2}$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{d} \mathbb{E}_\mu \left[\nu_i^2 - \mu_i^2 - 2 X_i (\nu_i - \mu_i) \right] \frac{1}{\ln 2}$$

$$= \frac{1}{2\ln 2\,\sigma^2} \sum_{i=1}^{d} \nu_i^2 + \mu_i^2 - 2\mu_i\nu_i = \frac{\|\mu-\nu\|_2^2}{2\ln 2\,\sigma^2} = \frac{2\delta^2}{\ln 2\,\sigma^2}$$

Thus, we obtain $R_n \geq \frac{\delta^2}{2}\left(1 - \sqrt{\frac{n\delta^2}{\sigma^2}}\right)$

Choose $\frac{n\delta^2}{\sigma^2} = \frac{1}{4} \Rightarrow \delta^2 = \frac{\sigma^2}{4n}$

$$\Rightarrow R_n \geq \frac{\sigma^2}{16n}$$!! Doesn't grow with d !!

Fano's method

Consider $\mu_1, \ldots, \mu_M$ s.t.

(i) $\min_{i \neq j} \|\mu_i - \mu_j\|_2 \geq 2\delta_{min}$

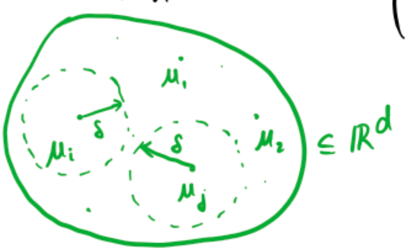
(ii) $\max_{i,j} \|\mu_i - \mu_j\|_2 \leq 2\delta_{max}$

Then, by Fano's inequality,

$$R_n \geq \delta_{min}^2 \left(1 - \frac{n \max_{i,j} D(P_{\mu_i} \| P_{\mu_j}) + 1}{\log M}\right)$$

where $D(P_{\mu_i} \| P_{\mu_j}) = \frac{\|\mu_i - \mu_j\|_2^2}{2\ln 2\,\sigma^2} \leq \frac{2\delta_{max}^2}{\ln 2\,\sigma^2}$.

Therefore, $R_n \geq \delta_{min}^2 \left(1 - \frac{2n\delta_{max}^2}{\ln 2\,\sigma^2 \log M} - \frac{1}{\log M}\right)$



Define

$$M(d, \delta_{min}, \delta_{max}) = \max\{ M : \exists \mu_1, \dots, \mu_M \in \mathbb{R}^d \text{ with } \min_{i \neq j} \|\mu_i - \mu_j\|_2 \geq 2\delta_{min} \text{ s.t. } \|\mu_i\|_2 \leq \delta_{max} \}$$

Then,
$$R_n \geq \delta_{min}^2 \left(1 - \frac{2n\delta_{max}^2/\sigma^2 + 1}{\log M(d, \delta_{min}, \delta_{max})} \right)$$

<u>Lemma</u> (Packing) For $c \leq 1/4$,
$$\log M(d, \delta\sqrt{c}, \delta) \geq d(1 - h(2c))$$

<u>Proof.</u> The proof uses the Gilbert-Varshamov construction from coding theory (or the Feinstein's maximal code construction) Instead of constructing a packing for the δ-ball in $\mathbb{R}^d$, we construct a packing for the binary hypercube $\{-1, 1\}^d$ with Hamming distance.

Let $u_1, \dots, u_M \in \{-1, 1\}^d$ be s.t. $\min_{i \neq j} d_H(u_i, u_j) \geq 2l$.

Let $\mu_i = \frac{\delta}{\sqrt{d}} u_i$. Then, $\|\mu_i\|_2 = \delta$ and
$$\min_{i \neq j} \|\mu_i - \nu_j\|_2 \geq \delta \sqrt{\frac{4l}{d}}$$

The GV construction is simple, we choose a point $u \in S_0 = \{-1, 1\}^n$ and remove all the points within a Hamming distance $2l$ of u to get

$S_1 = S_0 \setminus B_H(u, 2\ell)$ and continue till we can find a point.

Note that $|B_H(u,2\ell)| = |B_H(0,2\ell)| \leq 2^{d\,h(2\ell/d)}$

Thus, we can find at least $2^{d(1-h(2\ell/d))}$ points.

The proof is completed on choosing $\ell = dc$ ■

Therefore,

$$R_n \geq \delta^2 c \left(1 - \frac{2n\delta^2}{\sigma^2 d(1-h(2c))} - \frac{1}{d(1-h(2c))}\right)$$

Choose $\delta^2 = \frac{\sigma^2 d(1-h(2c))}{4n}$ to get

$$R_n \geq \frac{d\sigma^2}{n} \cdot \left(\frac{1-h(2c)}{4}\right) . c . \left(\frac{1}{2} - \frac{1}{d(1-h(2c))}\right)$$

Note that by Le Cam, we already have

$$R_n \geq \frac{\sigma^2}{16n}.$$

Therefore, $R_n \geq \frac{d\sigma^2}{n} \cdot \frac{(1-h(2c)).c}{8} - 4R_n c$

i.e., $R_n \geq \frac{d\sigma^2}{n} \cdot \frac{(1-h(2c)).c}{8(1+4c)}$

We can choose $c = 1/8$ to get

$$R_n \geq \frac{d\sigma^2}{1000n} \Rightarrow R_n = \Theta\left(\frac{d\sigma^2}{n}\right).$$

B A Generic Estimation Procedure

Consider $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$.

→ Estimating θ is tantamount to finding a $P \in \mathcal{P}$ that generated the observed samples

→ Our proof of lower bound, in essence, reduces this problem to that of solving an M-ary hypothesis testing problem for $P_1, \ldots, P_M \in \mathcal{P}$ where P_i's form a packing of $\mathcal{P}$ in an appropriate metric.

→ Perhaps inspired by these observations, a generic recipe for estimating $P \in \mathcal{P}$ by observing samples $X_1, \ldots, X_n$ from P was proposed by Devroye and Lugosi.

(See Chapter 6 of "Combinatorial methods in density estimation" by Luc Devroye and Gábor Lugosi)

Suppose our goal is to find a $\hat{P}$ s.t. $d(\hat{P}, P)$ is small.

DL Procedure

(1) Find a set $\{P_1, \ldots, P_M\} \subseteq \mathcal{P}$ s.t. $\forall\, P \in \mathcal{P}\ \exists\, P_i$ satisfying $d(P_i, P) \leq \delta$.

(2) Use samples $(X_1, \ldots, X_n)$ to find an i s.t.

$$d(P_i, P) \leq c \cdot \delta \text{ for some constant } c.$$

⑥

Note that the set $\{P_1, \dots, P_M\}$ itself can be found using the samples $(X_1, \dots, X_n)$.

The second step will be resolved using the so-called Scheffe Selector.

Problem: Given samples $X_1, \dots, X_n$ from P, find i s.t.

$$d(P_i, P) \le c \cdot \min_j d(P_j, P).$$

Let $P_1, \dots, P_M$ and P have densities $f_1, \dots, f_M$ and f w.r.t. ν.

Consider $M=2$ first.

Denote by μ_n the empirical measure

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}.$$

The Scheffe selector is given by

$$\hat{f}(X^n) = \underset{g \in \{f_1, f_2\}}{\operatorname{argmin}} \left| \int_{A^*} g(x)\, d\nu - \mu_n(A^*) \right|,$$

where

$$A^* \equiv A^*(f_1, f_2) = \{x : f_1(x) \ge f_2(x)\}.$$

Theorem

$$d(\hat{P}, P) \le 3 \min\{d(P_1, P), d(P_2, P)\} + 4 \left| P^n(A^*) - \mu_n(A^*) \right|.$$