Minimax bounds II

Kyoung Hee Kim

Sungshin University

1 Recap

2 Information theory

3 Fano, 1961

4 Yang & Barron, 1999

5 Devroye & Lugosi

Setting

- Family of probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$ on a sigma field $\mathcal{A}$
- Estimator $\hat{\theta}$: measurable map from Ω to Θ
- $L(\hat{\theta}, \theta)$: loss function
- Maximum risk $R(\Theta, \hat{\theta}) := \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\hat{\theta}, \theta)$
- Minimax risk with a minimax estimator $\hat{\theta}_{mm}$,

$$R(\Theta) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} E_\theta L(\tilde{\theta}, \theta) = \sup_{\theta \in \Theta} E_\theta L(\hat{\theta}_{mm}, \theta).$$

Minimax optimality

- Usually difficult to find minimax risk and minimax estimator.
- Typically satisfied if we find a 'good' lower bound $\ell(n)$ on $R(\Theta)$ and if we find a 'good' upper bound $u(n)$ by using a specific estimator $\tilde{\theta}$, where

$$\ell(n) \leq R(\Theta) \leq \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\tilde{\theta}, \theta) \leq u(n) \tag{1}$$

$$\lim_{n \to \infty} \frac{u(n)}{\ell(n)} \leq C.$$

- If $\tilde{\theta}$ satisfies (1), then $\tilde{\theta}$ is called a **minimax optimal** estimator.

Various distances

P, Q: two probability measures with densities p, q w.r.t ν.

- **Total variation** distance

$$V(P,Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \sup_{A \in \mathcal{A}} \left| \int_A (p-q) d\nu \right|$$

- Squared **Hellinger** distance

$$h^2(P,Q) = \int (p^{1/2} - q^{1/2})^2 d\nu.$$

- **Kullback–Leibler** (KL) divergence

$$KL(P,Q) = \int p \log \frac{p}{q} d\nu.$$

- **Chi-squared** χ^2 distance

$$\chi^2(P,Q) = \int \frac{p^2}{q} d\nu - 1.$$

Relation between distances

P, Q: two probability measures with densities p, q w.r.t ν.[1]

1. $\frac{1}{2}h^2(P,Q) \le V(P,Q) \le h(P,Q)\sqrt{1 - \frac{h^2(P,Q)}{4}}$
2. $V(P,Q) \le h(P,Q) \le \sqrt{KL(P,Q)} \le \sqrt{\chi^2(P,Q)}$
3. (Pinsker) $V(P,Q) \le \sqrt{\frac{KL(P,Q)}{2}}$ and $V(P,Q) \le 1 - \frac{1}{2}\exp(-KL(P,Q))$.
4. $h^2(P^n, Q^n) \le n h^2(P,Q)$

[1] For the proof of each statement and more details, see Chapter 2 of Tsybakov (2003).

Le Cam's Lemma

Construct

$$A := \{\theta_0, \theta_1\} \subseteq \Theta$$

such that

1. $\inf_{\xi \in \Theta} \left(L(\xi, \theta_0) + L(\xi, \theta_1) \right) \geq \delta$,
2. $||\mathbb{P}_{\theta_0} \wedge \mathbb{P}_{\theta_1}||_1 \geq c > 0$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\hat{\theta}, \theta) \geq \frac{c\delta}{4}.$$

Assouad's Lemma

Construct

$$A := \{\theta_\alpha,\ \alpha \in \{0,1\}^m\} \subseteq \Theta$$

such that

1. $\inf_{\xi\in\Theta} \left(L(\xi,\theta_\alpha) + L(\xi,\theta_\beta)\right) \geq \delta\|\alpha-\beta\|_0\ \forall\ \alpha,\beta \in \{0,1\}^m$,
2. $||\mathbb{P}_{\theta_\alpha} \wedge \mathbb{P}_{\theta_\beta}||_1 \geq c > 0$ if $\|\alpha-\beta\|_0 = 1$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta L(\theta,\hat{\theta}) \geq \frac{c\delta}{4} m.$$

Notation: $||\alpha-\beta||_0 = \sum_{k=1}^m \mathbb{1}_{\{\alpha_k\neq\beta_k\}}$.

Idea: *Metric entropy* structure of a class would determine the minimax rate of convergence of estimators.

Proposition (Yang & Barron, 1999)

For rich[2] *class* Θ *below, let* $\epsilon_n^2 = \frac{\log N_p(\epsilon, \Theta, d)}{n}$. *Then*

$$\min_{\hat{\theta}} \max_{\theta \in \Theta} \mathbb{E}_\theta d^2(\theta, \hat{\theta}) \sim \epsilon_n^2.$$

1. Θ*: class of densities s.t.* $0 < c \leq \theta \leq C$ *with* d^2 *is integrated squared* L_2*, sqaured Hellinger, or KL.*
2. Θ*: convex class of densities with* $\theta \leq C$ *and there exist one density in* Θ *bounded away from zero and* d *is* L_2*.*
3. Θ*: regression functions* θ *s.t.* $|\theta| \leq C$ *where* $Y = \theta(X) + \epsilon$*,* X *and* $\epsilon \sim N(0, \sigma^2)$ *are ind., and* d *is* $L_2(P_X)$ *norm.*

---

[2] $\liminf_{\epsilon \to 0} \frac{\log N_p(\epsilon/2, \Theta, d)}{\log N_p(\epsilon, \Theta, d)} > 1$.

Entropy

Let Y and Z be discrete random variables.

1. Entropy of Y:

$$H(Y) = -\sum_{y} p(y) \log p(y) = -\mathbb{E}(\log p(Y)).$$

2. Conditional entropy of Y given Z:

$$H(Y|Z) = -\sum_{y,z} p(y,z) \log p(y|z) = -\mathbb{E}(\log p(Y|Z)).$$

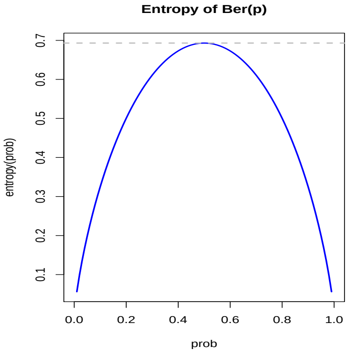
   where $p(y|z)$ is the conditional pmf of Y given $Z = z$.

3. Joint entropy of Y, Z:

$$H(Y,Z) = -\sum_{y,z} p(y,z) \log p(y,z).$$

Properties of entropy

- If $X \sim Ber(p)$, then
  $H(X) = -p \log p - (1-p)\log(1-p) \leq \log 2 = H(Ber(1/2))$.



- If X is discrete random variables on $\{1, \ldots, M\}$,
  $H(X) \leq \log(M)$.

Properties of entropy

- $H(Y, Z) = H(Y) + H(Z|Y) = H(Z) + H(Y|Z)$.
- $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$.
- $H(Y|Z) \le H(Y)$.
- $H(X_1, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i)$.
- For any function g, $H(g(Y)|Y) = 0$.

Information

Information between two random variables Y and Z is defined by

$$I(Y,Z) = KL(P_{Y,Z}, P_Y \times P_Z).$$

Also,

$$\begin{aligned}
0 \le I(Y,Z) &= \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)} \\
&= \sum_{y,z} p(y,z) \log \frac{p(y|z)}{p(y)} \\
&= -\sum_{y,z} p(z|y)p(y) \log(p(y)) + \sum_{y,z} p(y,z) \log p(y|z) \\
&= H(Y) - H(Y|Z),
\end{aligned}$$

giving $H(Y) \ge H(Y|Z)$.

Properties of information

- $I(Y, Z) \geq 0$ with equality iff Y and Z are independent.
- $I(X, (Y, Z)) = I(X, Y) + I(X, Z|Y) = I(X, Z) + I(X, Y|Z)$.
  If the following terms can be defined,

$$\begin{aligned} I(X, (Y, Z)) - I(X, Y) &= \sum p(x, y, z) \log \frac{p(x, y, z)}{p(z|y)p(x, y)} \\ &= \sum p(x, y, z) \log \frac{p(z, x|y)}{p(x|y)p(z|y)} \\ &=: I(X, Z|Y). \end{aligned}$$

- For any function g, $I(X, g(Y)) \leq I(X, Y)$.

Lemma (Fano's inequality)

Let Z, Y be discrete random variables on $\{1, \dots, M\}$. Then

$$\mathbb{P}(Z \neq Y) \geq \frac{H(Y|Z) - \log(2)}{\log M}.$$

Proof.

Let $E = \mathbb{1}_{\{Z \neq Y\}}$. By the definition of conditional pmf,

$$H(E, Y|Z) = H(Y|Z) + H(E|Y, Z) = H(Y|Z).$$

On the other hand,

$$\begin{aligned} H(E, Y|Z) &= H(E|Z) + H(Y|E, Z) \\ &\leq H(E) + \mathbb{P}(E = 0)H(Y|E = 0, Z) + \mathbb{P}(E = 1)H(Y|E = 1, Z) \\ &\leq \log(2) + \mathbb{P}(E = 1)H(Y) \\ &\leq \log(2) + \mathbb{P}(Z \neq Y)\log M. \end{aligned}$$

Lemma (Fano's Lemma)

Construct

$$F := \{\theta_j, j \in J\} \subseteq \Theta$$

where $|F| = M$ *satisfying the following: suppose* $\forall\ \theta_j, \theta_{j'} \in F$,

1. *(loss cond.)* $L(\theta_j, \theta_{j'}) \geq \delta$
2. *(testing cond.)* $KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_{j'}}) \leq \epsilon$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \geq \frac{\delta}{2}\left(1 - \frac{\epsilon + \log 2}{\log M}\right).$$

Define

1. Y: uniform random variable on $\{1, \ldots, M\} =: J$
2. X: random variable with a conditional distribution $\mathbb{P}_j := \mathbb{P}_{\theta_j}$ given $Y = j$.
3. $Z = \text{argmin}_{j \in J} L(\hat{\theta}, \theta_j)$.

- joint distribution of (X, Y): $\mathbb{P}(X \in A, Y = j) = \mathbb{P}(X \in A | Y = j)\mathbb{P}(Y = j) = \frac{\mathbb{P}_{\theta_j}(A)}{M}$.
- By bounding the supremum by the maximum followed by Markov inequality,

$$R(\Theta, \hat{\theta}) \geq \max_{j \in J} \mathbb{E}_{\theta_j} L(\theta_j, \hat{\theta}) \geq \frac{\delta}{2} \max_{j \in J} \mathbb{P}\left(L(\theta_j, \hat{\theta}) \geq \frac{\delta}{2} | Y = j \right).$$

- If $Z \neq j$, then $L(\theta_j, \hat{\theta}) \geq \frac{\delta}{2}$.

$$
\begin{aligned}
R(\Theta,\hat{\theta}) &\geq \frac{\delta}{2}\max_{j\in J}\mathbb{P}\left(Z\neq j|Y=j\right)\\
&\geq \frac{\delta}{2}\frac{1}{M}\sum_{j=1}^{M}\mathbb{P}\left(Z\neq j|Y=j\right)=\frac{\delta}{2}\mathbb{P}\left(Z\neq Y\right)\\
&\geq \frac{\delta}{2}\left(\frac{H(Y|Z)-\log(2)}{\log M}\right) \quad \text{by Fano's inequality}\\
&\geq \frac{\delta}{2}\left(\frac{H(Y)-I(X,Y)-\log(2)}{\log M}\right) \quad H(Y|Z)\geq H(Y)-I(X,Y)\\
&= \frac{\delta}{2}\left(1-\frac{I(X,Y)+\log(2)}{\log M}\right) \quad H(Y)=\log M.
\end{aligned}
$$

Note that for any function g, $I(Y,g(X))\leq I(Y,X)$ and $H(Y|Z)=H(Y)-I(Y,Z)\geq H(Y)-I(Y,X)$.

It suffices to bound $I(X,Y)$ from above. Let $p(x) = \frac{1}{M}\sum_j p_j(x)$.

$$I(X,Y) = \int \sum_j \frac{p_j(x)}{M} \log \frac{p_j(x)/M}{p(x)/M} d\lambda$$
$$= \frac{1}{M}\sum_j \int p_j \log \frac{p_j}{p} d\lambda = \frac{1}{M}\sum_j KL(\mathbb{P}_j, \bar{\mathbb{P}}),$$

where $\bar{\mathbb{P}} = \frac{1}{M}\sum_j \mathbb{P}_j$. By log-sum inequality[3],

$$KL(\mathbb{P}_j, \bar{\mathbb{P}}) \leq \frac{1}{M}\sum_i \int p_j \log \frac{p_j}{p_i} = \frac{1}{M}\sum_i KL(\mathbb{P}_j, \mathbb{P}_i).$$

Plug these into Fano's inequality,

$$R(\Theta, \hat{\theta}) \geq \frac{1}{\delta}\left(1 - \frac{\frac{1}{M^2}\sum_{i,j} KL(\mathbb{P}_j, \mathbb{P}_i) + \log(2)}{\log M}\right).$$

[3] Let $a_i, b_i > 0$ and $\sum_i a_i = a, \sum_i b_i = b$, then $a \log \frac{a}{b} \leq \sum_i a_i \log \frac{a_i}{b_i}$.

Example

(1) High-dimensional linear regression[4]

Suppose we have $\{(x_i, y_i)\}_{i=1}^n$ from $y_i = x_i^T\theta + w_i$ where $x_i \in \mathbb{R}^p$ and $w_i \sim N(0, \sigma^2)$ is i.i.d., and $\theta \in \mathbb{R}^p$. Assume $p > n$ and let

$$\Theta_s = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\}.$$

Also we let $\gamma_{2s} = \sup_{\theta \in \{\|\theta\|_0 \leq 2s\}} \frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2}$. Then

$$\sup_{\theta \in \Theta_s} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2 \geq \frac{C\sigma}{\gamma_{2s}} \sqrt{\frac{s}{n} \log\left(\frac{p - s/2}{s}\right)}.$$

[4] Raskutti, Wainwright, and Yu (2011)

- $\mathbb{P}_\theta = \prod_{i=1}^n P_{\theta,i}$ and $P_{\theta,i} = N(x_i^T\theta, \sigma^2)$ with a loss function $L(\hat\theta, \theta) = \|\hat\theta - \theta\|_2$.
- We need to construct as many parameters $\{\theta_1, \ldots, \theta_M\}$ as possible satisfying the following:
  1. (loss cond.) $\min_{j \neq j'} L(\theta_j, \theta_{j'}) \geq_{(?)} \delta$
  2. (testing cond.) $\max_{j \neq j'} KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_{j'}}) \leq_{(?)} \epsilon$
- Note that

$$KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_{j'}}) = \int^n \prod_{i=1}^n \phi_{\sigma^2}(u_i - x_i^T\theta_j) \log \frac{\prod_{i=1}^n \phi_{\sigma^2}(u_i - x_i^T\theta_j)}{\prod_{i=1}^n \phi_{\sigma^2}(u_i - x_i^T\theta_{j'})} du_1 \ldots du_n$$

$$= \int^n \prod_{i=1}^n \phi_{\sigma^2}(u_i - x_i^T\theta_j) \sum_{i=1}^n \left(u_i \frac{x_i^T(\theta_j - \theta_{j'})}{\sigma^2} - \frac{(x_i^T\theta_j)^2 - (x_i^T\theta_{j'})^2}{2\sigma^2} \right) du_1^n$$

$$= \sum_{i=1}^n \left(\frac{(x_i^T\theta_j)(x_i^T\theta_j - x_i^T\theta_{j'})}{\sigma^2} - \frac{(x_i^T\theta_j)^2 - (x_i^T\theta_{j'})^2}{2\sigma^2} \right) = \sum_{i=1}^n \frac{(x_i^T\theta_j - x_i^T\theta_{j'})^2}{2\sigma^2}$$

$$= \frac{\|X(\theta_j - \theta_{j'})\|_2^2}{2\sigma^2}.$$

Let $\theta_j, \theta_{j'} \in \Theta_s$, then $\|\theta_j - \theta_{j'}\|_0 \le 2s$. Let $\theta_{jj'} := \theta_j - \theta_{j'}$.

1. (testing cond.)

$$KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_{j'}}) = \frac{\|X\theta_{jj'}\|_2^2}{2\sigma^2} = \frac{n\|\theta_{jj'}\|_2^2}{2\sigma^2}\Big(\frac{\|X\theta_{jj'}\|_2}{\sqrt{n}\|\theta_{jj'}\|_2}\Big)^2$$
$$\le \frac{n\|\theta_{jj'}\|_2^2}{2\sigma^2}\gamma_{2s}^2 \le_{(?)} \epsilon$$

2. (loss cond.) $\|\theta_{jj'}\|_2 \ge_{(?)} \delta$

Since these two conditions need opposite direction for $\|\theta_{jj'}\|_2$, it would be good if we can construct $\{\theta_1, \dots, \theta_M\} \in \Theta_s$ so that for $j \ne j'$, $c_1\delta \le \|\theta_j - \theta_{j'}\|_2 \le C_1\delta$.

Lemma (Kühn (2001), Raskutti, et al.(2011))

There exists a subset $\Theta_0 \subseteq \Theta_s$ *such that* $\delta \leq \|\theta_j - \theta_{j'}\|_2 \leq 2\delta\sqrt{2}$ *for all* $1 \leq j < j' \leq M$ *and* $\log M \geq \frac{s}{2} \log\left(\frac{p-s/2}{s}\right)$.

Sketch of the proof:

- Define
$$\mathcal{H} = \mathcal{H}(s) := \{z \in \{-1, 0, 1\}^p, \|z\|_0 = s\}.$$
- For p, s (even) and $s < 2p/3$, there exists $\tilde{\mathcal{H}} \subset \mathcal{H}$ with $|\tilde{\mathcal{H}}| \geq \exp\left(\frac{s}{2} \log \frac{p-s/2}{s}\right)$ s.t. $\|z - z'\|_0 \geq s/2$ for all $z, z' \in \tilde{\mathcal{H}}$.
- Then use rescaled version $\sqrt{2/s}\delta\tilde{\mathcal{H}}$.

Sketch of the proof of the claim:

- $|\mathcal{H}| = \binom{p}{s}2^s$ and $\|z - z'\|_0 \leq 2s$ for all $z, z' \in \mathcal{H}$.
- For a fixed $z \in \mathcal{H}$, $|\{z' \in \mathcal{H} : H(z, z') \leq s/2\}| \leq \binom{p}{s/2}3^{s/2}$.
- Consider $\mathcal{H}_0 \subset \mathcal{H}$ with cardinality at most $|\mathcal{H}_0| \leq M := \frac{\binom{p}{s}}{\binom{p}{s/2}}$.
- The set of $z \in \mathcal{H}$ within Hamming distance $s/2$ of some element of $\mathcal{H}_0$ has cardinality at most $|\mathcal{H}_0|\binom{p}{s/2}3^{s/2} < |\mathcal{H}|$. Thus, for any such set with cardinality $\leq M$, there exists $z \in \mathcal{H}$ st. $H(z, z') > s/2$ for all $z' \in \mathcal{H}_0$. Adding this element inductively at each round, we can create a set $\mathcal{H}_0 \subset \mathcal{H}$ with $|\mathcal{H}_0| > M$ s.t. $H(z, z') > s/2$.
- Bound $M \geq \left(\frac{p - s/2}{s}\right)^{s/2}$.

- Using the above lemma,

$$KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_{j'}}) \leq \frac{4n\delta^2\gamma_{2s}^2}{\sigma^2}.$$

- Testing condition implies that we can take $\epsilon = c' \log M$, yielding

$$\delta^2 = \frac{c'\sigma^2}{4n}\gamma_{2s}^{-2} \log M \geq \frac{c'\sigma^2}{8}\gamma_{2s}^{-2}\frac{s}{n}\log(\frac{p-s/2}{s}).$$

- Fano's lemma gives

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \geq \tilde{c}\frac{\sigma}{\gamma_{2s}}\sqrt{\frac{s}{n}\log\left(\frac{p-s/2}{s}\right)}.$$

Relation to Assouad's lemma (density estimation)

For convenience, we assume L is a pseudo metric, i.e. $L(\xi, \theta_\alpha) + L(\xi, \theta_\beta) \geq c_0 L(\theta_\alpha, \theta_\beta)$. Suppose we construct

$$A := \{\theta_\alpha, \ \alpha \in \{0,1\}^m\} \subseteq \Theta$$

such that

1. (loss cond.) $L(\theta_\alpha, \theta_\beta) \sim \delta \|\alpha - \beta\|_0 \ \forall \ \alpha, \beta \in \{0,1\}^m$,
2. (testing cond.) $KL(\mathbb{P}_{\theta_\alpha}, \mathbb{P}_{\theta_\beta}) = nKL(\theta_\alpha, \theta_\beta) \leq 1 - 2c$ for $\|\alpha - \beta\|_0 = 1$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \gtrsim \delta m.$$

If $KL \sim L$, then $\delta \sim 1/n$, which implies the lower bound m/n.

Lemma (Varshamov–Gilbert)

Let $m \geq 8$. Let $w \in \{0,1\}^m$. Then there exists a subset $J_0 := \{w^{(0)}, \ldots, w^{(M)}\}$ of $\{0,1\}^m$ such that $w^{(0)} = (0, \ldots, 0)$,

$$\|w^{(j)} - w^{(k)}\|_0 \geq \frac{m}{8}, \quad \forall\ 0 \leq j < k \leq M,$$

and $M \geq 2^{m/8}$.

Sketch[5] of the proof: $|W| = 2^m$.

- Take $w^{(0)} = (0, \dots, 0)$ and exclude all w s.t. $\|w - w^{(0)}\|_0 \leq D := \lfloor m/8 \rfloor$.
- Set $W_1 = \{w \in W : \|w - w^{(0)}\|_0 > D\}$. Take $w^{(1)}$ an arbitrary element of W_1. Then exclude all $w \in W_1$ s.t. $\|w - w^{(1)}\|_0 \leq D$.
- Recurrently define W_j of W: $W_j = \{w \in W_{j-1} : \|w - w^{(j-1)}\|_0 > D\}$ for $j = 1, \dots, \mathsf{M}$ where M is the smallest integer s.t. $W_{\mathsf{M}+1} = \emptyset$.
- Let $A_j = \{w \in W_j : \|w - w^{(j)}\|_0 \leq D$ for $j = 0, \dots, \mathsf{M}$, then $|A_j| =: n_j \leq \sum_{i=1}^{D} \binom{m}{i}$ for $j = 0, \dots, \mathsf{M}$.
- $\|w^{(j)} - w^{(k)}\|_0 \geq D + 1 \geq m/8$ when $j \neq k$, and $2^m \leq \sum_{j=0}^{\mathsf{M}} n_j$, hence $(M+1) \geq \frac{2^m}{\sum_{i=0}^{D} \binom{m}{i}} = \frac{1}{P(Bi(m, 1/2) \leq \lfloor m/8 \rfloor)} \geq 2^{m/4}$ (via Hoeffding).

[5] For the detail, see Lemma 2.9 of Tsybakov(2003).

- Let δ be the same used in Assouad's method.
- Assume we constructed A in Assouad's method. Then by Varshamov–Gilbert Lemma, there exists more than $2^{m/8}$ indices s.t. the hamming dist. between any two is at least $m/8$. Let us take these as the index set J_0; then $L(\theta_j, \theta_{j'}) \sim m\delta$ for $j \neq j' \in J_0$.
- If $KL \sim L$, then testing cond. in Assouad's lemma implies $\delta \sim 1/n$ and $KL(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_j'}) \sim nL(\theta_j, \theta_{j'}) \sim m$.
- Thus Fano's lemma gives $\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \gtrsim \frac{m}{n}$.

Remarks

- When $KL(\theta,\theta') \sim L(\theta,\theta')$? If we assume that $c \le \theta \le C$, $KL \sim \chi^2 \sim L_2^2 \sim h^2$.
  - $$KL(\theta,\theta') \le \int \frac{(\theta-\theta')^2}{\theta'} \le \frac{1}{c} L_2^2(\theta,\theta').$$
  - $$L_2^2(\theta,\theta') = \int \left((\sqrt{\theta}-\sqrt{\theta'})(\sqrt{\theta}+\sqrt{\theta'})\right)^2 \le 2Ch^2(\theta,\theta').$$
- When $\mathbb{P}_\theta = N^n(\theta,\sigma^2)$, then $KL(\mathbb{P}_{\theta_j},\mathbb{P}_{\theta j'}) = \frac{n\|\theta_j-\theta_{j'}\|_2^2}{2\sigma^2}$. Let $L = L_2^2$, then $\epsilon \sim n\delta$. Hence we get the lower bound δ if we can construct $F = \{\theta_j, j \in J\}$ so that $\log|F|/n \sim \delta$ and $L_2^2(\theta_j,\theta_{j'}) \sim \delta$ for all $j \neq j' \in J$.

Metric entropy

- (Packing number) $N_p(\epsilon, \Theta, d) := \max\{N : \{\theta_1, \ldots, \theta_N\} \subseteq \Theta$ such that $d(\theta_j, \theta_{j'}) \geq \epsilon$ for all $i \neq j\}$. These set $\{\theta_1, \ldots, \theta_N\}$ is called an ϵ packing set.
- (Covering number) Let $\mathcal{P}_\Theta := \{\mathbb{P}_\theta : \theta \in \Theta\}$. $N_c(\epsilon, \mathcal{P}_\Theta, d) := \min\{N : \{\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_N}\} \subseteq \mathcal{P}_\Theta$ such that there exists j for which $d(\mathbb{P}_\theta, \mathbb{P}_{\theta_j}) \leq \epsilon$ for any $\theta \in \Theta\}$. These set $\{\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_N}\}$ is called an ϵ cover (net).

Lemma (Yang and Barron, 1999)

Construct

$$F := \{\theta_j, j \in J\} \subseteq \Theta$$

where $|F| = M$ *satisfying the following: suppose* $\forall\ \theta_j, \theta_{j'} \in F$,

1. $L(\theta_j, \theta_{j'}) \geq \delta$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \geq \frac{\delta}{2}\left(1 - \frac{\log N_c(\epsilon^2, \mathcal{P}_\Theta, KL) + \epsilon^2 + \log 2}{\log M}\right).$$

Following the proof of Fano's Lemma, it suffices to show $I(X,Y) = \frac{1}{M}\sum_j KL(\mathbb{P}_j, \bar{\mathbb{P}}) \leq \log N_c(\epsilon^2, \mathcal{P}_\Theta, KL) + \epsilon^2$.

- Consider an ϵ^2 net $\{\mathbb{P}_{\tilde{\theta}_j}, j = 1, \ldots, N_c\}$ under KL divergence. Then for any $\mathbb{P}_\theta \in \mathcal{P}_\Theta$, there exists $\mathbb{P}_{\tilde{\theta}_{j'}}$ such that $KL(\mathbb{P}_\theta, \mathbb{P}_{\tilde{\theta}_{j'}}) \leq \epsilon^2$.
- For any $\mathbb{Q}$,

$$\frac{1}{M}\sum_{j=1}^{M} KL(\mathbb{P}_{\theta_j}, \mathbb{Q}) - \frac{1}{M}\sum_{j=1}^{M} KL(\mathbb{P}_{\theta_j}, \bar{\mathbb{P}}) = KL(\bar{\mathbb{P}}, \mathbb{Q}) \geq 0.$$

- Use $\mathbb{Q} := \frac{1}{N_c}\sum_{j=1}^{N_c} \mathbb{P}_{\tilde{\theta}_j}$, then

$$\begin{aligned}
\frac{1}{M}\sum_{j} KL(\mathbb{P}_j, \bar{\mathbb{P}}) &\leq \frac{1}{M}\sum_{j=1}^{M}\int p_{\theta_j} \log \frac{p_{\theta_j}}{\frac{1}{N_c}\sum_{j=1}^{N_c} p_{\tilde{\theta}_j}} \\
&\leq \frac{1}{M}\sum_{j=1}^{M}\int p_{\theta_j} \log \frac{p_{\theta_j}}{\frac{1}{N_c} p_{\tilde{\theta}_{j'(j)}}} \\
&\leq \log N_c(\epsilon^2, \mathcal{P}_\Theta, KL) + \epsilon^2.
\end{aligned}$$

Example

(2) nonparametric regression

Let $Y_i = \theta(X_i) + w_i$ where $X_i \sim Uni[0,1]$, $w_i \sim N(0,1)$ and $X_i \perp w_i$. Assume $\theta \in \Theta_s$ where Θ_s satisfied

1. θ is differentiable $s-1$ times on $(0,1)$,
2. $\sup_{0 \le x \le 1} |\theta^{(k)}(x)| \le 1$ for all $k = 0, 1, \ldots, s-1$ where $\theta^{(0)} := \theta(x)$
3. $\theta^{(s-1)}$ is 1-Lipschitz on $(0,1)$.

Then for any estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta_s} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \ge c' n^{-\frac{2s}{2s+1}}.$$

- Given $X_i = x_i, i = 1, \ldots, n$, $\mathbb{P}_\theta = \prod_{i=1}^n P_{\theta,i} = N(\theta(x_i), 1)$.
- $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$.
- For $\mathbb{P}_\theta, \mathbb{P}_{\theta'} \in \mathcal{P}_{\Theta_s}$,

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \int^{2n} \prod_{i=1}^n \phi(u_i - \theta(x_i)) \log \frac{\prod \phi(u_i - \theta(x_i))}{\prod \phi(u_i - \theta'(x_i))} du_1^n dx_1^n$$

$$= \sum_{i=1}^n \left(\int^2 \phi(u_i - \theta(x_i)) \log \frac{\phi(u_i - \theta(x_i))}{\phi(u_i - \theta'(x_i))} du_i dx_i \right)$$

$$= \sum_{i=1}^n \left(\int (\theta(x_i) - \theta'(x_i))^2 dx_i \right) = \frac{n\|\theta - \theta'\|_2^2}{2}.$$

- Known:

$$c\epsilon^{-1/s} \le \log N_c(\epsilon, \Theta_s, L_2) \le C\epsilon^{-1/s}.$$

$\Rightarrow$ $\log N_p(\tilde{\delta}^2, \Theta_s, L_2^2) \sim \log N_c(\tilde{\delta}, \Theta_s, L_2) \ge c\tilde{\delta}^{-1/s}$.

$\Rightarrow$ $\log N_c(\epsilon^2, \mathcal{P}_{\Theta_s}, KL) \sim \log N_c(\sqrt{\frac{2}{n}}\epsilon, \Theta_s, L_2) \le C(2/n)^{-1/(2s)}\epsilon^{-1/s}$

- By Yang and Barron(YB),

$$\sup_{\theta\in\Theta_s} \mathbb{E}_\theta L(\theta, \hat{\theta}) \ge \frac{\tilde{\delta}^2}{2}\left(1 - \frac{C(2/n)^{-1/(2s)}\epsilon^{-1/s} + \epsilon^2 + \log 2}{\tilde{\delta}^{-1/s}}\right)$$
$$\ge c' n^{-2s/(1+2s)}$$

by taking $\epsilon \sim n^{\frac{1}{2(1+2s)}}$ and $\tilde{\delta}^2 \sim \epsilon^{-4s} \sim n^{-\frac{2s}{1+2s}}$.

Density estimation problem

Lemma (LB, YB)

Construct

$$F := \{\theta_j, j \in J\} \subseteq \Theta$$

where $|F| = M$ *satisfying the following: suppose* $\forall\ \theta_j, \theta_{j'} \in F$,

1. $L(\theta_j, \theta_{j'}) \geq \delta$.

Then, for every estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \geq \frac{\delta}{2}\left(1 - \frac{\log N_c(\epsilon^2, \Theta, KL) + n\epsilon^2 + \log 2}{\log M}\right).$$

Remark: If $\log N_p(\epsilon^2, \Theta, L) \sim \log N_c(\epsilon^2, \Theta, KL) \sim n\epsilon^2$, then $\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}) \geq \epsilon^2$.

Density estimation problem

Take the following (Bayes predictive) estimator

$$\bar{p}(x) = \frac{1}{n} \sum_{i=0}^{n-1} \hat{p}_i(x)$$

where $\hat{p}_i(x) = p_{X_{i+1}|X_1,\ldots,X_i}(x|X_1,\ldots,X_i)$ for $i > 0$ and $\hat{p}_0(x) = \frac{1}{N_c}\sum_{j=1}^{N_c} p_{\tilde{\theta}_j}(x) =: p(x)$.

Theorem (UB, YB)

Let i.i.d. sample $X_1, \ldots, X_n$ from a density $\theta \in \Theta$. Assume $\log N_c(\epsilon^2, \Theta, KL) \le n\epsilon^2$. Use Bayes predictive estimator $\hat{\theta}$. Then

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta KL(\theta, \hat{\theta}) \le C\epsilon^2.$$

$$
\begin{aligned}
\mathbb{E}_\theta KL(p_\theta, \bar{p}) &= \mathbb{E}_\theta \int p_\theta(x) \log \frac{p_\theta(x)}{\frac{1}{n}\sum_{i=0}^{n-1} \hat{p}_i(x)} dx \\
&\leq \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_\theta \int p_\theta(x) \log \frac{p_\theta(x)}{\hat{p}_i(x)} dx \\
&= \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{E}_\theta \int p_\theta(x) \log \frac{p_\theta(x)}{p(x|x_1, \dots, x_i)} dx + \frac{1}{n} \int p_\theta(x) \log \frac{p_\theta(x)}{p(x)} dx \\
&= \frac{1}{n} \int p_\theta(x_1, \dots, x_n) \log \left(\frac{p_\theta(x_2) \dots p_\theta(x_n)}{p(x_2|x_1) p(x_3|x_1, x_2) \dots p(x_n|x_1, \dots, x_{n-1})} \frac{p_\theta(x_1)}{p(x_1)} \right) \\
&= \frac{1}{n} \int p_\theta(x_1, \dots, x_n) \log \frac{p_\theta(x_1, \dots, x_n)}{p(x_1, \dots, x_n)} \\
&\leq \frac{1}{n} \left(\log N_c(\epsilon^2, \Theta, KL) + nKL(P_\theta, P_{\tilde{\theta}_j}) \right) \leq C\epsilon^2,
\end{aligned}
$$

using the same argument as before.

Example

Let $\Phi = \{\phi_1 = 1, \phi_2, \ldots, \phi_k, \ldots\}$ be a fundamental sequence (linear combinations are dense) in $L^2[0,1]^d$. Consider $\Theta = \{\theta \in L_2[0,1]^d : \min_{a_i} \|\theta - \sum_{i=1}^k a_i\phi_i\|_2 \leq k^{-\alpha}, k = 0, 1, \ldots, \}$.

(3) linear approxmation

In a regression setting, $Y_i = \theta(X_i) + \epsilon_i$,

$$\min_{\hat{\theta}} \max_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \asymp n^{-2\alpha/(1+2\alpha)}.$$

Known result: $\log N_c(\epsilon, \Theta, L_2) \sim k_\epsilon = \inf\{k : k^{-\alpha} \leq \epsilon\}$. By equating $\epsilon^{-1/\alpha} \sim n\epsilon^2$, we have $\epsilon^2 \sim n^{-2\alpha/(2\alpha+1)}$.

Skeleton estimates

- Let $\Theta_\epsilon = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_{N_c}\}$ the ϵ-cover of Θ (class of densities) where $N_c = N_c(\epsilon, \Theta, L_1)$.
- $\mathcal{A}_\epsilon = \left\{ \{x : \tilde{\theta}_j(x) > \tilde{\theta}_{j'}(x)\} : \tilde{\theta}_j, \tilde{\theta}_{j'} \in \Theta_\epsilon \right\}$.
- Define $\hat{\theta} = \operatorname{argmin}_{\tilde{\theta} \in \Theta_\epsilon} \sup_{A \in \mathcal{A}_\epsilon} \left| P_{\tilde{\theta}}(A) - \mathbb{P}_n(A) \right|$ where $\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}}$ is the empirical measure of A.

Theorem (Devroye & Lugosi (2001))

Assume that $\theta \in \Theta$. *Then*

$$\mathbb{E}_\theta L_1(\hat{\theta}, \theta) \leq 3\epsilon + \sqrt{\frac{8 \log(2N_c^2)}{n}}.$$

We need to equate $\log N_c(\epsilon, \Theta, L_1) \sim n\epsilon^2$ to obtain ϵ rate using L_1.

Theorem 6.3 (Devroye & Lugosi (2001))

For any density θ,

$$\int |\hat{\theta} - \theta| \leq 3 \min_{\tilde{\theta} \in \Theta_\epsilon} \int |\tilde{\theta} - \theta| + 4 \sup_{A \in \mathcal{A}} |P_\theta(A) - \mathbb{P}_n(A)| .$$

Let $\hat{\theta} = \tilde{\theta}_i$ and $\tilde{\theta}_j$ be any density minimising $\int |\tilde{\theta}_\ell - \theta|$ over all ℓ.
Assuming $j \neq i$, $\int |\hat{\theta} - \theta| \leq \int |\tilde{\theta}_j - \theta| + \int |\tilde{\theta}_i - \tilde{\theta}_j|$. W.l.o.g., let $i < j$.

$$\begin{aligned}
\int |\tilde{\theta}_i - \tilde{\theta}_j| &= 2 \sup_{A \in \mathcal{A}} \left| \int_A \tilde{\theta}_i - \int_A \tilde{\theta}_j \right| \\
&\leq 2 \sup_{A \in \mathcal{A}} \left| \int_A \tilde{\theta}_i - \mathbb{P}_n(A) \right| + \left| \int_A \tilde{\theta}_j - \mathbb{P}_n(A) \right| \leq 4 \sup_{A \in \mathcal{A}} \left| \int_A \tilde{\theta}_j - \mathbb{P}_n(A) \right| \\
&\leq 4 \sup_{B \in \mathcal{B}} \left| \int_B \tilde{\theta}_j - \int_B \theta \right| + 4 \sup_{A \in \mathcal{A}} \left| \int_A \theta - \mathbb{P}_n(A) \right| \\
&= 2 \int \left| \tilde{\theta}_j - \theta \right| + 4 \sup_{A \in \mathcal{A}} \left| \int_A \theta - \mathbb{P}_n(A) \right| .
\end{aligned}$$

In order to bound $\sup_{A \in \mathcal{A}} |P_\theta(A) - \mathbb{P}_n(A)|$, let $Y_i = \mathbb{1}_{\{X_i \in A\}} - P(A)$.

Lemma (Hoeffding)

Let Y_i be independent random variable with $\mathbb{E}Y_i = 0$, $-1 \leq Y_i \leq 1$ w.p. 1. Then for $s > 0$, $\mathbb{E}(e^{sY_i}) \leq e^{s^2/2}$. Thus

$$\mathbb{E}\left(e^{s\frac{1}{n}\sum_{i=1}^n Y_i}\right) = \left(\mathbb{E}e^{\frac{s}{n}Y_i}\right)^n \leq e^{s^2/(2n)}.$$

Lemma

Let $\sigma > 0$, $N \geq 2$, and let $Z_1, \ldots, Z_N$ be real-valued random variables such that for all $s > 0$ and $1 \leq j \leq N$, $E(e^{sZ_j}) \leq e^{s^2\sigma^2/2}$ and $E(e^{s(-Z_j)}) \leq e^{s^2\sigma^2/2}$. hen

$$E\left(\max_{j \leq N} Z_j\right) \leq \sigma\sqrt{2\log(2N)}.$$

References

Assouad, P. (1983). Deux remarques sur l'estimation. *Comptes Rendus de l' Academie des Sciences, Paris, Ser. I Math 296*, 1021–1024.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. Wiley.

Devroye, L. and G. Lugosi (2001). *Combinatorial Methods in Density Estimation*. Springer.

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communication*. Mass. and New York: M.I.T. Press and Cambridge and Wiley.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics 1*, 38–53.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation (Springer Texts in Statistics)* (2nd ed.). Springer.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory 57*(10), 6976–6994.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer-Verlag.

Yang, Y. and A. R. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Yu, B. (1997). Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang (Eds.), *A Festschrift for Lucien Le Cam*, pp. 423–435. New York: Springer-Verlag.