Lecture 20: Minimax Lower Bounds - Global Fano's Method

Lecturer: Yudong Chen *Scribe: Xumei Xi*

1 Recap: Local Fano's Method

Recall the two key theorems from previous lectures.

Theorem 1 (Estimation to testing)**.** *If $\{\theta_1, \theta_2, \dots, \theta_M\}$ is a 2δ-packing of parameter space Θ in $\rho(\cdot,\cdot)$, then*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\theta}\left[\rho(\hat{\theta}(X),\theta)\right] \geq \delta \inf_{\psi} \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}\left[\psi(X)\neq j \middle| \mathbb{P}_{\theta_j}\right]. \tag{1}$$

Theorem 2 (Fano's inequality)**.** *For any testing procedure ψ, we have*

$$\mathbb{P}\left[\psi(X)\neq J\right] \geq 1 - \frac{I(X;J)+\log 2}{\log M}. \tag{2}$$

In the local Fano's method, we upper bound the mutual information by the maximum of pairwise KL-divergence,

$$I(X;J) \leq \max_{i,j} D(\mathbb{P}_{\theta_i}\|\mathbb{P}_{\theta_j}) \leq g(\delta).$$

This upper bound, based on convexity of KL, is relatively crude and not tight under some settings, especially for non-parametric problems. In particular, with this upper bound we only makes use of a *local* packing of Θ. In order to capture the full capacity of the entire parameter space, we will develop the so-called global Fano's method.

2 Global Fano's method

The global Fano's method is based on the following better upper bound on the mutual information.

2.1 Yang-Barron's upper bound

Lemma 1 (Yang-Barron's upper bound)**.** *Let $N_{\mathrm{KL}}(\varepsilon,\Theta)$ denote the ε-covering number of Θ in the pseudo-distance $\rho_{\mathrm{KL}}(\theta,\theta') := \sqrt{D(\mathbb{P}_\theta\|\mathbb{P}_{\theta'})}$. Then we have*

$$I(X;J) \leq \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon,\Theta), \qquad \forall\epsilon>0. \tag{3}$$

Proof Recall the notation $Q_X := \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}_{\theta_i}$ for the marginal distribution of X. Then for any distribution of X, denoted by Q', we have

$$\begin{aligned}
I(X;J) &= \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta_j}\|Q_X) && \text{definition of KL}\\
&\leq \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta_j}\|Q'), && \text{the mixture distribution } Q_X \text{ minimizes the average KL}\\
&\leq \max_{j=1,\dots,M} D(\mathbb{P}_{\theta_j}\|Q').
\end{aligned} \tag{4}$$

Note that the second step above is an analogue of the fact that

$$\arg\min_{\theta'} \left\{ \frac{1}{M} \sum_{j=1}^{M} \|\theta_j - \theta'\|_2^2 \right\} = \frac{1}{M} \sum_{j=1}^{M} \theta_j.$$

Let $\{\beta_1, \beta_2, \ldots, \beta_N\}$ be a minimal ε-covering of Θ w.r.t. ρ_{KL}, where $N = N_{\mathrm{KL}}(\varepsilon, \Theta)$. Note that we are free to choose any Q'. Here we set $Q' = \frac{1}{N} \sum_{\ell=1}^{N} \mathbb{P}_{\beta_\ell}$.

Fix an arbitrary index $j \in \{1, 2, \ldots, M\}$. By definition of ε-covering, there exists some β_i such that $\rho_{\mathrm{KL}}(\theta_j, \beta_i) \le \varepsilon$. We therefore have

$$\begin{aligned}
D(\mathbb{P}_{\theta_j} \| Q') &= \mathbb{E}_{\mathbb{P}_{\theta_j}} \left[\log \frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\frac{1}{N} \sum_{l=1}^{N} \mathrm{d}\mathbb{P}_{\beta_l}} \right], && \mathrm{d}\mathbb{P}_{\theta_j} \text{ denotes the density of } \mathbb{P}_{\theta_j} \\
&\le \mathbb{E}_{\mathbb{P}_{\theta_j}} \left[\log \frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\frac{1}{N} \mathrm{d}\mathbb{P}_{\beta_i}} \right] && \text{sum of } N \text{ terms is lower bounded by any one term} \\
&= D(\mathbb{P}_{\theta_j} \| \mathbb{P}_{\beta_i}) + \log N \\
&\le \varepsilon^2 + \log N.
\end{aligned}$$

Combining this upper bound (valid for all $j = 1, 2, \ldots, M$) with equation (4) gives the desired inequality $I(X; J) \le \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon, \Theta)$. □

Remark Note that we have two sets of points here:

- $\{\theta_1, \theta_2, \ldots, \theta_M\} \subset \Theta$ is a δ-packing in ρ, with cardinality $M = M(\delta, \Theta, \rho)$.
- $\{\beta_1, \beta_2, \ldots, \beta_N\} \subset \Theta$ is an ε-packing in ρ_{KL}, with cardinality $N = N(\varepsilon, \Theta, \rho_{\mathrm{KL}}) = N_{\mathrm{KL}}(\varepsilon, \Theta)$.

2.2 Procedure for using Yang-Barron's upper bound

We can employ the following two steps when choosing the parameters ϵ and δ.

1. Choose $\varepsilon > 0$ such that
$$\varepsilon^2 \ge \log N_{\mathrm{KL}}(\varepsilon, \Theta). \tag{5}$$
2. Choose the largest $\delta > 0$ such that
$$\log M(\delta, \Theta, \rho) \ge 4\varepsilon^2 + 2 \log 2. \tag{6}$$

With the above choice, we have the following lower bound the minimax error:

$$\begin{aligned}
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\left[\rho(\hat{\theta}, \theta) \right] &\ge \delta \left(1 - \frac{I(X; J) + \log 2}{\log M} \right), && \text{by combining (1) and (2)} \\
&\ge \delta \left(1 - \frac{\varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon, \Theta) + \log 2}{\log M} \right), && \text{by Lemma 1} \\
&\ge \delta \left(1 - \frac{2\varepsilon^2 + \log 2}{4\varepsilon^2 + 2 \log 2} \right), && \text{by what we just did in (5) and (6)} \\
&= \frac{1}{2} \delta. && (7)
\end{aligned}$$

3 Application: Lipschitz regression

For application, we consider a non-parametric regression problem over Lipschitz functions. We observe

$$y_i = f(x_i) + e_i, \quad i = 1, 2, \ldots, n,$$

where x_i's are fixed, $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for all i, and f is an unknown function from the function class

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R} \mid f(0) = 0, f \text{ is 1-Lipchitz}\}.$$

Here $\mathcal{F}$ acts as the parameter space Θ in the non-parametric setting.

We have proved in Homework 1 that

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log M(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta}.$$

So we can find a desired δ-packing of $\mathcal{F}$ in $\|\cdot\|_\infty$.

Next we need an ε-covering of $\mathcal{F}$ in ρ_{KL}. Observe that the distribution of the data $(y_i)_{i=1}^n$ is

$$\begin{aligned}\mathbb{P}_f &= \mathcal{N}(f(x_1), \sigma^2) \times \cdots \times \mathcal{N}(f(x_n), \sigma^2) \\ &= \mathcal{N}(f(x_1^n), \sigma^2 I_n),\end{aligned}$$

which represents an n-dimensional Gaussian distribution with mean vector $f(x_1^n) = (f(x_1), \ldots, f(x_n))$. Then we can calculate the pairwise KL-divergence:

$$\begin{aligned}D(\mathbb{P}_f \| \mathbb{P}_g) &= \frac{1}{2\sigma^2} \left\|f(x_1^n) - g(x_1^n)\right\|_2^2 \\ &\leq \frac{n}{2\sigma^2} \left\|f - g\right\|_\infty^2 .\end{aligned}$$

Hence the metric entropy of $\mathcal{F}$ in KL-divergence is proportional to that in $\|\cdot\|_\infty$:

$$\begin{aligned}\log N_{\mathrm{KL}}(\varepsilon, \mathcal{F}) &\asymp \log N\left(\sqrt{\frac{2\sigma^2}{n}}\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right) \\ &\asymp \frac{\sqrt{n}}{\varepsilon\sigma}.\end{aligned}$$

Now we are ready to set the parameters ε and δ according to our two-step procedure:

1. Choose $\varepsilon = \left(\frac{\sqrt{n}}{\sigma}\right)^{\frac{1}{3}}$ so that $\varepsilon^2 \geq \frac{\sqrt{n}}{\sigma\varepsilon} \gtrsim \log N_{\mathrm{KL}}(\varepsilon, \mathcal{F})$.
2. Choose $\delta \asymp \frac{1}{\varepsilon^2}$ so that $\log M(\delta, \Theta, \|\cdot\|_\infty) \gtrsim \frac{1}{\delta} \geq 4\varepsilon^2 + 2\log 2$.

Thus we satisfy the requirements in (5) and (6), and the minimax lower bound (7) holds. In particular, we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\infty\right] \geq \frac{1}{2}\delta \asymp \left(\frac{\sigma^2}{n}\right)^{\frac{1}{3}}.$$

Note that this tight lower bound cannot be achieved using local Fano's method instead.

We may compare this lower bound with the upper bound we derived in Lectures 14-15:

$$\left\|\hat{f} - f\right\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f(x_i)\right)^2} \lesssim \left(\frac{1}{n}\right)^{\frac{1}{3}}.$$

We can do some extra work to match the norms. In this case, the upper and lower bound match.