MATH888: High-dimensional probability and statistics	Fall 2021

Lecture 32 — 19th November, 2021

Sebastien Roch, UW-Madison *Scribe: Yu Sun, Sebastien Roch*

Recall from last time, under the hypothesis testing problem setting:

- $\underline{\theta}^1, \dots, \underline{\theta}^M$ are 2δ-separated under ρ over the space $\Theta(\mathcal{P})$.
- Pick J uniformly at random on $[M]$. Given $J = j$, pick $Z \sim \mathbb{P}_{\theta^j}$ where $\theta(\mathbb{P}_{\theta^j}) = \theta^j$.
- $\mathbb{Q}_{Z,J}$ is the joint distribution of (Z, J).

Fano's method can be stated as (Proposition 15.12 and equation (15.34) in [1]; we will give a proof in the next lectures)

$$\begin{aligned}\mathfrak{M} \triangleq \mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) &= \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\hat{\theta}, \theta(\mathbb{P}))] \\ &\geq \Phi(\delta) \cdot \left\{ 1 - \frac{\frac{1}{M^2}\sum_{i\neq j}^{M} KL(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta^j}) + 1}{\log_2 M} \right\}. \end{aligned} \tag{1}$$

Today we finish the linear regression example and show that the least-squares estimate achieves the minimax risk (under the mean-squared error) up to constant.

1 Example: Linear Regression (15.14 in [1]): Continued

Recall the assumption: $\underline{Y} \sim \mathcal{N}(\mathbb{X}\underline{\theta}^*, \sigma^2 I_n)$, where $\underline{Y}$ and $\mathbb{X}$ are observed, and $\underline{\theta}^*$ is unknown. We want a lower bound for minimax of MSE, i.e.

$$\mathrm{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}\|\mathbb{X}\underline{\theta}^* - \mathbb{X}\hat{\theta}\|_2^2, \tag{2}$$

we have 2 claims as follow:

- **Claim 1.** (proved at the end of the last lecture): If $\underline{\gamma}^1, \dots, \underline{\gamma}^M \in range(\mathbb{X})$, s.t.

  $$\|\underline{\gamma}^i - \underline{\gamma}^j\|_2 > 2\delta\sqrt{n} \quad \forall i \neq j, \tag{3}$$

  then there $\exists \underline{\theta}^1, \dots, \underline{\theta}^M$, s.t.

  $$\rho(\underline{\theta}^i, \underline{\theta}^j) \triangleq \frac{1}{\sqrt{n}}\|\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j\|_2 > 2\delta \quad \forall i \neq j. \tag{4}$$

- **Claim 2.** : If $\underline{\gamma}^1, \dots, \underline{\gamma}^M \in range(\mathbb{X})$, s.t.

  $$\|\underline{\gamma}^i\|_2 < 4\delta\sqrt{n} \quad \forall i, \tag{5}$$

then

$$KL(\mathbb{P}_{\underline{\theta}^i} \| \mathbb{P}_{\underline{\theta}^j}) \leq 32n\frac{\delta^2}{\sigma^2}. \tag{6}$$

Proof. (**Claim 2.**) Since $\mathbb{P}_{\underline{\theta}} = \mathcal{N}(\mathbb{X}\underline{\theta}, \sigma^2 I_n)$,

$$\begin{aligned} KL(\mathbb{P}_{\underline{\theta}^i} \| \mathbb{P}_{\underline{\theta}^j}) &= \frac{1}{2\sigma^2}\|\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j\|_2^2 \\ &\leq \frac{1}{2\sigma^2} \cdot \left(\|\underline{\gamma}^i\|_2 + \|\underline{\gamma}^j\|_2\right)^2 \text{(triangular inequality of norm)} \\ &= 32n\frac{\delta^2}{\sigma^2} \end{aligned} \tag{7}$$

□

Recall the *packing number*: the largest cardinality of an ϵ-separated set in a subset K of a metric space$(\mathcal{T}, d)$ is called the *packing number* of K, i.e. $\mathcal{P}(K, \epsilon)$. We proved,

$$\frac{Vol(K)}{Vol(\epsilon B_2^r)} \leq \mathcal{N}(K, \epsilon) \leq \mathcal{P}(K, \epsilon) \tag{8}$$

with $(\mathcal{T}, d)$ as $\mathbb{R}^r$ with l_2-norm, and B_2^r is the unit ball under Euclidian distance.

We adapt it into our linear regression problem, $K \subset \mathbb{R}^r$, with $r = rank(\mathbb{X})$, chose $\epsilon = 2\delta\sqrt{n}$ and $K = 4\delta\sqrt{n}B_2^r$ (technically, inside the range of $\mathbb{X}$), we have

$$M \geq \left(\frac{4\delta\sqrt{n}}{2\delta\sqrt{n}}\right)^r = 2^r \iff \log_2 M \geq r. \tag{9}$$

Combine **Claim 2** and the result from Fano's method at the beginning,

$$\begin{aligned} \mathfrak{M} &\geq \delta^2\left(1 - \frac{\frac{32n\delta^2}{\sigma^2} + 1}{r}\right) \qquad \left(\text{since all KL's are bounded by either 0 or } \frac{32n\delta^2}{\sigma^2}\right) \\ &\approx \Omega(\delta^2) \end{aligned} \tag{10}$$

by taking $\delta^2 = \frac{\sigma^2 r}{64n}$ (when r is sufficiently large).

2 Quick Tour for Information Theory

Remark 2.1. *In the interest of keeping things simple, we will derive everything on discrete spaces.*

Most of contents are covered in Chapter 2 of [2].

Definition 1. *If $X \in \mathcal{X}$ is a discrete r.v. with probability mass function $p(x)$, the textitentropy of X is*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{11}$$

with the criteria that $0\log 0 = 0$, and $0\log\frac{0}{0} = 0$.

Claim 2. *If $\mathcal{X}$ is finite, then*

$$H(X) \leq \log(|\mathcal{X}|) \tag{12}$$

with equality iff X is uniform on $\mathcal{X}$.

Proof.

$\Leftarrow$ Suppose U is uniform on $\mathcal{X}$, then

$$\begin{aligned} H(U) &= -\sum_{u \in \mathcal{X}} p(u) \log p(u) \\ &= \sum_{u \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \log |\mathcal{X}| \\ &= \log |\mathcal{X}| \end{aligned} \tag{13}$$

$\Rightarrow$

$$\begin{aligned} 0 \leq KL(X|U) &= \sum_{u \in \mathcal{X}} p(x) \log \frac{p(x)}{1/|\mathcal{X}|} \\ &= \sum_{u \in \mathcal{X}} p(u) \log p(u) - \log |\mathcal{X}| \\ &= -H(X) - \log |\mathcal{X}| \end{aligned} \tag{14}$$

□

Definition 3. *For a pair of r.v.s (X, Y), the conditional entropy $X|Y$ is*

$$H(X|Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x|y) \tag{15}$$

Remark 2.2. *By using $p(x, y) = p(y)p(x|y)$, we know above conditional entropy is the entropy of X given y, averaged over the distribution of $\mathcal{Y}$.*

References

[1] Wainwright, Martin J., *High-Dimensional Statistics: A non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, 2019

[2] Cover, Thomas M., *Elements of information theory*, John Wiley & Sons, 1999.