MATH888: High-dimensional probability and statistics Fall 2021

Lecture 31 — 17th November, 2021

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

1 Overview

In previous lectures we have introduced the minimax risk, which is defined as

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) = \inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\hat{\theta}, \theta(\mathbb{P})]. \tag{1}$$

Recall the hypothesis testing problem setting:

- $\theta^1, \ldots, \theta^M$ 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).

We have shown that the minimax risk is lower bounded as

$$\mathfrak{M} \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}_{Z,J}[\psi(Z) \neq J] \tag{2}$$

where $\Phi(\delta)$ is an increasing function. We have also shown the special case when $M = 2$, where

$$\mathfrak{M} \geq \Phi(\delta) \cdot \frac{1}{2}\{1 - \|\mathbb{P}_{\theta^1} - \mathbb{P}_{\theta^2}\|_{TV}\}. \tag{3}$$

In this lecture we will examine a more general case when $M > 2$ and derive a lower bound of the minimax risk using Fano's method in an information theory perspective. We will show the theorem and an example today and give the proof in the next lecture.

2 Fano's method

The mutual information between Z and J is defined as:

$$I(Z, J) = KL(\mathbb{Q}_{Z,J} \| \mathbb{Q}_Z \mathbb{Q}_J) \tag{4}$$

where $\mathbb{Q}_Z, \mathbb{Q}_J$ are marginals of Z and J respectively.

The mutual information describes the amount of information gained on Z, if we only observe J. The higher the mutual information, the more knowledge one gain regarding the unobserved variable using the observed variable.

Theorem (Proposition 15.12 in [1]). Under the setting we describe in the beginning, we have

$$\mathfrak{M} \geq \Phi(\delta) \cdot \left\{1 - \frac{I(Z,J)+1}{\log_2 M}\right\}. \tag{5}$$

We will use the following lemma to reduce the problem into two different distributions,

Lemma (15.34 in [1]). Under the same setting,

$$I(Z,J) \leq \frac{1}{M^2}\sum_{i,j}^{M} KL(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta^j}) \tag{6}$$

We are going to give the proof of the theorem and the lemma next time.

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

3.1 Problem setting

- we have an unknown parameter $\underline{\theta}^* \in \mathbb{R}^P$
- given n observations $y_i = f(\underline{x}_i) + \epsilon_i$ where $f(\underline{x}) = \underline{\theta}^{*T}\underline{x}$
- Goal: given $(\underline{x}_i, y_i)$, we would want to estimate $\underline{\theta}^*$.

We use the Mean squared error (MSE) to measure our estimator $\hat{\underline{\theta}}^T$, MSE is defined as

$$MSE(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(\underline{x}_i) - f(\underline{x}_i))^2 \tag{7}$$

where $\hat{f}_n(\underline{x}) = \hat{\underline{\theta}}^T\underline{x}$. We would like to represent it in matrix form. Let $\mathbb{X}$ has the $\underline{x}_i$'s as rows,

$$MSE(\hat{f}_n) = \frac{1}{n}\|\mathbb{X}\hat{\underline{\theta}} - \mathbb{X}\underline{\theta}^*\|^2. \tag{8}$$

The least square estimator is

$$\hat{\theta}^{LS} \in \arg\min_{\underline{\theta}} \|\underline{Y} - \mathbb{X}\underline{\theta}^*\|^2 \tag{9}$$

where $\underline{Y} = (y_1, \ldots, y_n)$.

Note that $\hat{\theta}^{LS} = \mathbb{X}^T\underline{Y}$, we have shown that with probability $1-\delta$,

$$MSE(\mathbb{X}\hat{\theta}^{LS}) \lesssim \frac{\delta^2(rank(\mathbb{X}) + \log(1/\delta))}{n} \tag{10}$$

where $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ has $\|\underline{\epsilon}\|_{\psi_2} \leq \delta$.

3.2 Lower bound

Assume $\underline{\epsilon} \sim \mathcal{N}(\underline{\theta}, \sigma^2 I_n)$, $\rho(\underline{\hat{\theta}}, \underline{\theta}^*) = \frac{1}{\sqrt{n}}\|\mathbb{X}\underline{\hat{\theta}} - \mathbb{X}\underline{\theta}^*\|_2$, and $\Phi(\delta) = \delta^2$. We want to find $\theta_1, \ldots, \theta_M$ such that

$$\frac{1}{\sqrt{n}}\|\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j\|_2 > 2\delta \quad \forall i \neq j \quad (*) \tag{11}$$

We want to find the largest M since in the lemma we have shown that the upper bound of the mutual information is inversely proportional to M. The larger M is, the tighter the upper bound of the mutual information is.

Let $\underline{\gamma}^1, \ldots, \underline{\gamma}^M \in range(\mathbb{X})$ such that

$$\|\underline{\gamma}^i - \underline{\gamma}^j\|_2 > 2\delta\sqrt{n} \quad \forall i \neq j \quad (**) \tag{12}$$

then there exists $\theta_1, \ldots, \theta_M$ such that $\underline{\gamma}^i = \mathbb{X}\underline{\gamma}^i$ and (*) is satisfied.

Then let $\mathbb{P}_{\underline{\theta}^j} = \mathcal{N}(\mathbb{X}\underline{\theta}^j, \sigma^2 I_n)$, we need a bound of the KL divergence

$$KL(\mathbb{P}_{\underline{\theta}^i} \| \mathbb{P}_{\underline{\theta}^j}) = \frac{1}{2\sigma^2}\|\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j\|_2^2. \tag{13}$$

If we can prove $\forall i \neq j$, $KL(\mathbb{P}_{\underline{\theta}^i} \| \mathbb{P}_{\underline{\theta}^j}) \leq$ some upper bound Δ. Then $I(Z, J) \leq \Delta$. This will be discussed more in the next lecture.

References

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