**MATH888: High-dimensional probability and statistics** Fall 2021

Lecture 32 — 19th November, 2021

*Sebastien Roch, UW-Madison* *Scribe: Gokcan Tatli*

# 1 Overview

In the last lecture, we lower bounded the minimax risk $\mathfrak{M}$ using Proposition 15.12 from W [1]. Then, we used the relation between mutual information and Kullback Leibler divergence (15.34 in W [1]), to find the following result:

**Recall:** If $(\underset{\sim}{\theta}^1), \ldots, \underset{\sim}{\theta}^M)$ are $2\delta$ separated under P, then

$$\mathfrak{M} \geq \phi(\delta) \left( 1 - \frac{\frac{1}{\mu^2} \sum_{i \neq j} KL(P_{\underset{\sim}{\theta}^i} || P_{\underset{\sim}{\theta}^j}) + 1}{\log_2 M} \right)$$

.

In this lecture, we continue to linear regression example and prove the lower bound for minimax MSE for that example. After that, we cover some basics of information theory.

# 2 Example: Linear Regression (continued)

- Suppose $\underset{\sim}{\mathbb{Y}} \sim N(\mathbb{X}\underset{\sim}{\theta}^*), \sigma^2 I_n)$ when $(\underset{\sim}{\mathbb{Y}}, \mathbb{X})$ are observed and $\underset{\sim}{\theta}^*$ is unknown.
- We want a lower bound on minimax MSE $(\mathbb{X}\underset{\sim}{\theta}^*))$, where

$$\text{MSE}(\mathbb{X}\underset{\sim}{\theta}^*)) = \frac{1}{n} ||\mathbb{X}\underset{\sim}{\theta}^*) - \mathbb{X}\underset{\sim}{\hat{\theta}})||_2^2.$$

  That is why we need following claims.

- **Claim 1.** If $\underset{\sim}{\gamma}^1, \ldots, \underset{\sim}{\gamma}^M \in range(\mathbb{X})$ are such that

$$||\underset{\sim}{\gamma}^i - \underset{\sim}{\gamma}^j|| > 2\delta\sqrt{n} \quad \forall i \neq j,$$

  then $\exists \underset{\sim}{\theta}^1, \ldots, \underset{\sim}{\theta}^M$ such that

$$\rho(\underset{\sim}{\theta}^i, \underset{\sim}{\theta}^j) \frac{1}{\sqrt{n}} ||\mathbb{X}\underset{\sim}{\theta}^i - \mathbb{X}\underset{\sim}{\theta}^j||_2 > 2\delta$$

- **Claim 2.** If $\underset{\sim}{\gamma}^1, \ldots, \underset{\sim}{\gamma}^M \in range(\mathbb{X})$ are such that

$$||\underset{\sim}{\gamma}^i|| \leq 4\delta\sqrt{n} \quad \forall i,$$

  then

$$KL(P_{\underset{\sim}{\theta}^i} || P_{\underset{\sim}{\theta}^j}) \leq \frac{32n\delta^2}{\sigma^2}$$

**Proof of Claim 2:** We know that $P_{\underset{\sim}{\theta}^j} = N(\mathbb{X}\underset{\sim}{\theta}^j, \sigma^2 I_n)$. That is why,

$$KL(P_{\underset{\sim}{\theta}^i} || P_{\underset{\sim}{\theta}^j}) = \frac{1}{2\sigma^2} ||\mathbb{X}\underset{\sim}{\theta}^i) - \mathbb{X}\underset{\sim}{\theta}^j)|| \quad \text{(by explicit calculation)} \tag{1}$$

$$\leq \frac{1}{2\sigma^2} \left[ ||\underset{\sim}{\gamma}^i|| + ||\underset{\sim}{\gamma}^j|| \right]^2 \quad \text{(by triangle inequality)} \tag{2}$$

Then, we use the norm bound, i.e., $4\delta\sqrt{n}$, in Claim 2 to get the desired KL-divergence bound $\frac{32n\delta^2}{\sigma^2}$.

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

We proved that

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

where $N(K, \epsilon)$ is the covering number and $B_2^r$ is the unit norm-ball.

In our case (linear regression example), $K \subseteq \mathbb{R}^r$, rank($\mathbb{X}$)=r, $\epsilon = 2\delta\sqrt{n}$, $K =$ "$4\delta\sqrt{n}B_2^r$" (inside the range($\mathbb{X}$)). Therefore,

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

So, we can write

$$\mathfrak{M} \geq \delta^2 \left( 1 - \frac{\frac{32n\delta^2}{\sigma^2} + 1}{r} \right) \tag{3}$$

$$\approx \frac{\delta^2}{2} \tag{4}$$

where (3) comes from combining above bound for r and the bound we got from Claim 2, and plugging them in the lower bound of minimax risk given in Section 1. Then, we can write the lower bound in (4) by taking $\delta^2 = \frac{\sigma^2}{64}\frac{r}{n}$ when r is sufficiently large.

# 3 Quick Tour of Information Theory

- Note that in the interest of keeping things simple, we will do everything in discrete spaces

- Mostly from Chapter 2 of Cover-Thomas [2]

**Definition 1.** If X$\in \mathcal{X}$ is a discrete random variable with pmf p(x), the entropy of X is

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

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

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

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

with equality if and only if X is uniform over $\mathcal{X}$.

**Proof:** Suppose U is uniform over $\mathcal{X}$, then

$$\begin{aligned} H(U) &= -\sum_{u \in \mathcal{X}} p(u) \log p(u) && (5) \\ &= -\sum_{u \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \log \frac{1}{|\mathcal{X}|} && (6) \\ &= \log |\mathcal{X}| && (7) \end{aligned}$$

On the other hand,

$$\begin{aligned} 0 &\leq KL(X||U) && (8) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{1/|\mathcal{X}|} && (9) \\ &= \log |\mathcal{X}| - (-\sum_{x \in \mathcal{X}} p(x) \log p(x)) && (10) \\ &= \log |\mathcal{X}| - H(X) && (11) \end{aligned}$$

which completes the proof.

**Definition 3.** For a pair of random variables (X, Y), the conditional entropy of X|Y is

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

where $p(x, y)$ can be also written as $p(y)p(x|y)$. This can be considered as expectation of conditional distribution x|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.