MATH888: High-dimensional probability and statistics Fall 2021

Lecture 31 — 17th November, 2021

Sebastien Roch, UW-Madison *Scribe: Nicholas Chelales*

1 Overview

Recall in the previous lectures where we defined the minimax risk:

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

We showed the special case where $M = 2$ where the minimax risk can be bounded as

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

In this lecture we will examine the case where $M > 2$ and develop and information theory based approach to bounding the mini-max risk using Fano's method. The proof of this method will be discussed in the upcoming lecture, with examples and results being shown today.

2 Mutual Information and the Minimax Risk

The mutual information between two variables is an information theory technique which helps numerically describe the relationship between two random variables. If one were to observe variable A, the mutual information would describe the amount of information gained regarding variable B with this single observation of variable A. The higher the mutual information, the more knowledge one gains regarding the non-observed variable by observing a particular variable. It follows that zero mutual information indicates independent random variables.

2.1 Mutual Information and KL Divergence

The mutual information can be described in terms of the Kullback Leibler divergence:

$$I(Z, J) = KL(\mathbb{Q}_{Z,J} || \mathbb{Q}_Z \mathbb{Q}_J) \tag{3}$$

where $\mathbb{Q}_{\mathbb{Z},\mathbb{J}}$ represent the joint distribution and $\mathbb{Q}_Z$ and $\mathbb{Q}_J$ represent the marginal distributions respectively.

Theorem 1. *Proposition 15.12 in W [1]*

$$\mathfrak{M} \geq \Phi(\delta)[1 - \frac{I(Z, J) + 1}{log_2(M)}] \tag{4}$$

It follows that the lower the mutual information (i.e. less information J reveals about Z, the larger $[1 - \frac{I(Z,J)+1}{log_2(M)}]$ becomes, thus increasing the lower bound for the minimax risk.

Lemma 2. *15.34 in W [1]*

$$I(Z;J) \leq \frac{1}{M^2} \sum_{i,j}^{M} KL(\mathbb{P}_{\theta^i} || \mathbb{P}_{\theta^j}) \tag{5}$$

The above lemma is useful to reduce the problem to two different distributions, which can be plugged into equation 4 to get a more descriptive bound. The proof of this theorem and lemma will be discussed in the next lecture.

3 Linear Regression Example

Recall the linear regression problem, where we have an unknown parameter $\theta^* \in \mathbb{R}^P$ of which we wish to estimate. We are given n observations of $y_i = f(\underline{x}_i) + \epsilon_i$ where ϵ is some form of noise, and f represents the true function, i.e. $f(\underline{x}) = {\theta^*}^T \underline{x}$

The Mean squared error (MSE) is defined as:

$$MSE(\hat{f}_n) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(\underline{x}_i) - f(\underline{x}_i))^2 \tag{6}$$

this can be simplified in terms of the matrix form of $\underline{x}$

where the rows of $\mathbb{X}$ contain $\underline{x_i}$

$$MSE(\hat{f}_n) = \frac{1}{n} ||\mathbb{X}\underline{\hat{\theta}} - \mathbb{X}\underline{\theta}^*||^2 \tag{7}$$

and

$$\hat{\theta}^{LS} \in argMin_{\underline{\hat{\theta}}} ||\underline{y} - \mathbb{X}\underline{\theta}||^2 \tag{8}$$

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

Recall we previously showed w.p. $1 - \delta$

$$MSE(\mathbb{X}\hat{\theta}^{LS}) \leq \frac{\sigma^2}{n}(rank(\mathbb{X}) + log(1/\delta)) \tag{9}$$

when ϵ has sub-guassian norm $||\underline{\epsilon}||_{\psi_2} \leq \sigma$

Suppose our ϵ_i are generated from a normal distribution $N(\underline{0}, \sigma^2 I_n)$

We wish to find as many possible $\theta^1 ... \theta^m$ such that:

$$\frac{1}{\sqrt{n}}||\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j||_2 > 2\delta \ \forall i \neq j \tag{10}$$

We wish to find the largest M number of θ since earlier we showed that the upper bound on the mutual information $I(Z; J)$ is inversely proportional to M (see Lemma 2). Thus the larger M is, the tighter upper bound we can get on the mutual information. since X is fixed, the above expression is really a combination of the columns of X, and thus we can obtain the list of θ's by finding all vectors γ^i in **the range of** $\mathbb{X}$ that fit the criteria:

$$||\underline{\gamma}^i - \underline{\gamma}^j||_2 \geq 2\delta\sqrt{n} \ \forall i \neq j \tag{11}$$

then there must exist a set of θ^i such that:

$$\underline{\gamma}^i = \mathbb{X}\underline{\theta}^i \tag{12}$$

Note that all possible γ values lie within the range of $\mathbb{X}$, and we wish the find the maximum possible set which fit the δ criteria as described above.

what remains is a probability distribution of the form:

$$\mathbb{P}_{\theta^j} = N(\mathbb{X}\underline{\theta}^j, \sigma^2 I_n) \tag{13}$$

of which we wish to bound the KL divergence:

$$KL(\mathbb{P}_{\underline{\theta}^i}||\mathbb{P}_{\underline{\theta}^j}) = \frac{1}{2\sigma^2}||\mathbb{X}\underline{\theta}^i - \mathbb{X}\underline{\theta}^j||_2^2 \tag{14}$$

to be continued in more detail the next lecture.

References

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