MATH888: High-dimensional probability and statistics Fall 2021

Lecture 26 – 5th November, 2021

Sebastien Roch, UW-Madison *Scribe: Shubham Kumar Bharti*

1 Overview

In the last lecture we went through some standard results in information theory. Today, we start a new topic which focuses on providing information theoretic lower bounds on parameter estimation problems. In this lecture, we will give a lower bound proof for mean estimation of Gaussian distribution. This lower bound is tight and is achieved by the maximum likelihood estimator. The analysis follows the proof style in chapter 15 of Wainwright's book [1].

2 Parameter estimation problem

Consider the problem of parameter estimation under the following setup where we have

1. $\mathcal{P}$ – a family of distributions.
2. Θ – the parameter space of the family, every $\mathbb{P} \in \mathcal{P}$ is parameterized by a $\theta \in \Theta$.
3. $\rho : \Theta \times \Theta \to \mathbb{R}^+$ – a semi-metric loss function.
4. $\Phi : \mathbb{R} \to \mathbb{R}$ – a positive increasing function.

Given n i.i.d samples from a fixed distribution in the family, we are interested in estimating a parameter of that distribution. For any estimator $\hat{\theta} : \{X_i\}_{i\in[n]} \to \Theta$, we define its risk with respect to the true parameter $\theta(\mathbb{P})$ under the loss function $\Phi \circ \rho$ as follows:

$$\mathfrak{R}(\hat{\theta}, \theta(\mathbb{P}), \Phi \circ \rho) := \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}^n}\left[\Phi(\rho(\hat{\theta}, \theta(\mathbb{P})))\right] \tag{1}$$

The minimax risk of the estimation problem is then given as :

$$\begin{aligned} \mathfrak{M}(\theta(\mathbb{P}), \Phi \circ \rho) &= \inf_{\hat{\theta}} \mathfrak{R}(\hat{\theta}, \theta(\mathbb{P}), \Phi \circ \rho) &(2)\\ &= \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}^n}\left[\Phi(\rho(\hat{\theta}, \theta(\mathbb{P})))\right] &(3) \end{aligned}$$

2.1 Mean estimation of the Gaussian distribution

As our first example, consider the problem of estimating mean of the Gaussian distribution with a fixed variance σ^2. The distribution family,

$$\mathcal{P} := \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$$

is parameterized by the parameter $\theta \in \Theta := \mathbb{R}$. Since, we are interested in estimating the mean of the distribution, the parameter of interest is the mean itself which we denote by $\theta(\mathbb{P}_\mu) = \mu$.

Given n i.i.d samples drawn from a distribution $\mathbb{P}_\theta$, the **maximum likelihood estimator(MLE)** of the mean is given by :

$$\hat{\theta}(\{X_i\}_{i\in[n]}) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Suppose we are interested in the mean squared error of the estimator with respect to the true parameter. This can be captured by considering the semi-metric loss $\rho(\theta, \theta') = |\theta - \theta'|$ and loss transformation function $\Phi(z) = z^2$. The expected loss of the estimator is given as :

$$\mathbb{E}_{\mathbb{P}_\theta^n}\left[\Phi \circ \rho(\hat{\theta}, \theta)\right] = \mathbb{E}_{\mathbb{P}_\theta^n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right|^2\right] \tag{4}$$

$$= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n} \tag{5}$$

and the worst case error of the MLE estimator is $\sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\theta^n}\left[\Phi \circ \rho(\hat{\theta}, \theta)\right] = \frac{\sigma^2}{n}$.

Next, the question arises whether MLE estimator is optimal with respect to the minimax risk defined above in (3) or can we do better than $\frac{\sigma^2}{n}$ in the worst case? We will show that MLE is an optimal estimator for this problem in the next section.

3 MLE is optimal for Gaussian mean estimation

A standard recipe for providing a lower bound for a parameter estimation problem is to relate it to a hypothesis testing problem and then show that for a fixed sample size one can do no better in identifying the correct hypothesis. We outline this method next.

Let $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}\} \subseteq \Theta$ be a 2δ separated set i.e. $\forall i \neq j \in [M]^2, \rho(\theta^{(i)}, \theta^{(j)}) > \delta$. For now we assume that such a set is given but one has to always construct a problem specific set. Note that $\theta^{(i)} = \theta(\mathbb{P}_{\theta^{(i)}})$. We define a sampling procedure as follows:

1. Pick $J \sim Unif(M)$.
2. Sample $Z \sim \mathbb{P}_{\theta^{(J)}}$, Z is n i.i.d sample $\{X_i\}_{i\in[n]}$ s.t. $Z \sim \mathbb{P}^n_{\theta^{(J)}}$.

Note that Z comes from a mixture distribution $\bar{Q} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}_{\theta^{(J)}}$ and we denote the joint distribution as Q. Let $\psi : Z \to [M]$ denote an M-ary testing function. Given a 2δ separated set, we have the following theorem that relate the minimax risk of the parameter estimation problem to the hypothesis testing problem.

Theorem 1 (From estimation to a testing problem)**.** *For any semi-metric loss function ρ, increasing function Φ and a choice of 2δ-separated set Z, the minimax risk of the parameter estimation problem is lower bounded as*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} Q[\psi(Z) \neq J] \tag{6}$$

where infinimum is taken over all measurable testing functions.

Proof. Refer to Proposition 15.1 in [1] for a detailed proof. We will use this theorem to give an lower bound proof for the mean estimation problem. □

Theorem 2 (Minimax lower bound for Gaussian mean estimation)**.** *For Gaussian mean estimation problem defined above, the minimax risk satisfies*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{\sigma^2}{8n}$$

Proof. Consider two hypothesis defined by $\theta^{(1)} = 0, \theta^{(2)} = \frac{\sigma}{\sqrt{n}}$ and $\delta = \frac{1}{2} \cdot \frac{\sigma}{\sqrt{n}}$. Note that $\theta^{(1)}, \theta^{(2)}$ are 2δ separated. For any testing function ψ, we have,

$$\begin{align}
Q\left(\psi(Z) \neq J\right) &= \frac{1}{2}\left(\mathbb{P}^n_{\theta^{(1)}}[\psi(Z) \neq 1] + \mathbb{P}^n_{\theta^{(2)}}[\psi(Z) \neq 2]\right) \tag{7}\\
&= \frac{1}{2}\left(1 - ||\mathbb{P}^n_{\theta^{(1)}} - \mathbb{P}^n_{\theta^{(2)}}||_{TV}\right) && \text{(Prop. 4.4 in Rigollet's notes)} \tag{8}\\
&\geq \frac{1}{2}\left(1 - \sqrt{\mathrm{KL}\left(\mathbb{P}^n_{\theta^{(1)}}, \mathbb{P}^n_{\theta^{(2)}}\right)}\right) && \text{(by Pinskers inequality)} \tag{9}\\
&= \frac{1}{2}\left(1 - \sqrt{n \cdot \mathrm{KL}\left(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}\right)}\right) && \text{(KL factorizes for product distributions)} \tag{10}\\
&= \frac{1}{2}\left(1 - \sqrt{n \cdot \frac{\sigma^2/n}{2\sigma^2}}\right) \geq \frac{1}{4} && (\mathrm{KL}(\mathcal{N}(\theta_1, \sigma^2), \mathcal{N}(\theta_2, \sigma^2)) = \frac{||\theta_1 - \theta_2||^2}{2\sigma^2}) \tag{11}
\end{align}$$

Thus, $\inf_\psi Q\left(\psi(Z) \neq J\right) \geq \frac{1}{4}$. Plugging it back into 6, we get that for Gaussian mean estimation problem,

$$\begin{align}
\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) &\geq \Phi(\delta) \inf_\psi Q[\psi(Z) \neq J] \tag{12}\\
&\geq \frac{\sigma^2}{2n} \cdot \frac{1}{4} \tag{13}
\end{align}$$

which is of the same order as achieved by MLE estimator and hence MLE is an optimal minimax estimator for estimating mean of the Gaussian distribution. □

References

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