36-705: Intermediate Statistics **Fall 2019**

Lecture 18: October 9

Lecturer: Siva Balakrishnan

18.1 Asymptotic theory

This lecture and the next will focus on asymptotic theory for the MLE. We suppose that we obtain a sample $X_1, \ldots, X_n \sim p(X;\theta)$ and are interested in estimating θ.

Analogous to the asymptotic theory we developed for the average of i.i.d. random variables we will be interested in two questions:

1. **Consistency:** Does the MLE converge in probability to θ, i.e. does $\widehat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$? This is analogous to the LLN.
2. **Asymptotic distribution:** What can we say about the distribution of $\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta)$? This is analogous to the CLT.

We will begin with the question of consistency.

18.2 Consistency of the MLE

The main take-home from this section is that under somewhat mild conditions the MLE is a consistent estimator. We will try to develop the necessary conditions and build some intuition about the MLE and about what consistency entails.

18.2.1 MLE as Empirical Risk Minimization

We have discussed previously the idea of empirical risk minimization, where we construct an estimator by minimizing an empirical estimate of the risk. We looked at the particular case of classification with the 0/1 loss. The MLE can be viewed as a special case of ERM with a different loss function.

Suppose we define the risk function:

$$R_n(\widehat{\theta}, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i;\theta)}{p(X_i;\widehat{\theta})},$$

then we can observe that minimizing this risk function is identical to maximizing the likelihood. Notice that we introduced an extra $p(X_i;\theta)$ term but this does not affect anything. Of course, if this is the empirical risk it is natural to wonder what the associated population risk is. This is given as:

$$R(\widehat{\theta},\theta) = \mathbb{E}_\theta \log \frac{p(X;\theta)}{p(X;\widehat{\theta})},$$

which is known as the Kullback-Leibler divergence, i.e. the population risk is the KL divergence $\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta}))$.

Notice that, the empirical risk is a sum of i.i.d terms so by the LLN we have that for any fixed $\widetilde{\theta}$

$$R_n(\widetilde{\theta},\theta) \xrightarrow{p} R(\widetilde{\theta},\theta).$$

To analyze empirical risk minimization we needed a *uniform* LLN and we will need exactly this to show consistency.

An important property of the KL divergence is that it is zero iff $p(X;\theta) = p(X;\widehat{\theta})$ almost everywhere (i.e. they are equal except on sets of measure 0).

The main thing to remember is the connection between MLE and KL divergence.

18.2.2 Conditions for consistency

Condition 1: Identifiability: A basic requirement for constructing any consistent estimator is that the model be identifiable, i.e. if $\theta_1 \neq \theta_2$ then it must be the case that $p(X;\theta_1) \neq p(X;\theta_2)$.

We will in general require something slightly stronger than this:

Condition 2: Strong identifiability: We assume that for every $\epsilon > 0$

$$\inf_{\widetilde{\theta}:|\widetilde{\theta}-\theta|\geq\epsilon} \mathrm{KL}(p(X;\theta)\|p(X;\widetilde{\theta})) > 0.$$

This condition is essentially the same as Condition 1, except that it does not allow the difference between the two distributions to be vanishingly small. The two conditions are equivalent if θ is restricted to lie in a compact set.

Condition 3: Uniform LLN: Assume that,

$$\sup_{\widetilde{\theta}} |R_n(\widetilde{\theta},\theta) - R(\widetilde{\theta},\theta)| \xrightarrow{p} 0.$$

This condition is a uniform LLN. As we have seen before it holds for instance if the Rademacher complexity of the class of functions of the form: $f_{\widetilde{\theta}}(X) = \log p(X;\widetilde{\theta})/p(X;\theta)$ is not too large. In 36-702/36-708/10-716/... you will explore this idea further.

Theorem 18.1 *Suppose that Conditions 2 and 3 above hold, then the MLE is consistent.*

Proof: Fix an $\epsilon > 0$. Using the strong identifiability condition we see that for every $\epsilon > 0$, we have that there is an $\eta > 0$ such that,

$$\mathrm{KL}(p(X;\theta)\|p(X;\widetilde{\theta})) \geq \eta,$$

if $|\widetilde{\theta} - \theta| \geq \epsilon$. We will show that for the MLE $\widehat{\theta}$, we have that $\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) \leq \eta$, as $n \to \infty$ in probability. This in turn implies that $|\widehat{\theta} - \theta| \leq \epsilon$ which implies that $\widehat{\theta} \xrightarrow{p} \theta$.

If remains to show that $\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) \leq \eta$, as $n \to \infty$. Notice that,

$$\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) = R(\widehat{\theta},\theta) = R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) + R_n(\widehat{\theta},\theta) \overset{\text{(i)}}{\leq} R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) \xrightarrow{p} 0,$$

where the final convergence simply uses Condition 3. The inequality (i) follows since,

$$R_n(\widehat{\theta},\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i;\theta)}{p(X_i;\widehat{\theta})} \leq 0,$$

since $\widehat{\theta}$ is the MLE. ■

18.3 Inconsistency of the MLE

The MLE can fail to be consistent. When the model is not identifiable it is clear that we cannot have consistent estimators.

The other possible failure is the failure of the uniform law. This typically happens when the parameter space is too large. Here is a simple example:

Example: Suppose that we measure some outcome (say their blood sugar) for n individuals using a machine. We do it twice for every individual so that we can assess the variability of the machine, i.e. suppose we observe:

$$\begin{aligned} &Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2) \\ &\vdots \\ &Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2), \end{aligned}$$

and want to estimate σ^2. Even though we only want to estimate σ^2 the model has a growing number of parameters $\mu_1, \ldots, \mu_n, \sigma^2$ and the MLE for σ^2 will depend on estimating μ_i. Formally, we can see that the MLE for the means is:

$$\widehat{\mu}_i = \frac{Y_{i1} + Y_{i2}}{2}.$$

The log-likelihood for σ^2 can be written as:

$$\mathcal{LL}(\sigma^2, \mu) = -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(Y_{i1} - \mu_i)^2 + (Y_{i2} - \mu_i)^2 \right],$$

which is maximized when we take:

$$\widehat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} \left[(Y_{i1} - \widehat{\mu}_i)^2 + (Y_{i2} - \widehat{\mu}_i)^2 \right] = \frac{1}{4n} \sum_{i=1}^{n} (Y_{i1} - Y_{i2})^2.$$

Notice that,

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{\sigma^2}{2},$$

so by the LLN the MLE is inconsistent. One could easily fix this in this problem (by multiplying the MLE by 2) but more generally this could be tricky. We note that in this type of problem where the number of parameters is not fixed (and grows with the sample size) it is not even clear how to define convergence of the log-likelihood since its limit changes with the sample size.

18.4 MLE under misspecification

In statistical modeling we do not typically believe the model is correct, i.e. that the samples were in fact generated by some distribution in our model. Rather, we think of the model as a useful idealization or a simplification. In this (more realistic) case, one might wonder what the MLE converges to, or if it converges at all?

Concretely, suppose $X_1, \ldots, X_n \sim q$, and we estimate $\widehat{\theta}_{\text{MLE}}$, then what can we say about our estimate? To answer this, we can follow a similar argument to what we did in the beginning of the lecture and observe that at the population-level (i.e. with infinite samples) the MLE is:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathbb{E}_q \log p(X; \theta)$$

How do we interpret this statement? As before we can re-write it in terms of KL divergences and see that:

$$\text{KL}(q \| p_{\widehat{\theta}_{\text{MLE}}}) \leq \text{KL}(q \| p_\theta) \quad \text{for all } \theta \in \Theta.$$

So that at the population-level we can conclude that the MLE is estimating the KL projection of the data-generating distribution on our model, i.e. when q does not belong to our model the MLE is essentially estimating the KL projection of q onto our model. One can also impose similar conditions to what we had in the last section (uniform law + strong identifiability) to complete the consistency argument under model misspecification.