Lecture Notes 15
36-705

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. We are 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.

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.

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 loss function:

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

Observe that minimizing this loss 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

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \log \frac{p(X;\theta)}{p(X;\widehat{\theta})} = \int p(x;\theta) \log \left(\frac{p(x;\theta)}{p(x;\widehat{\theta})} \right) dx$$

which is 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.

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.

Theorem 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. ■

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.

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?

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.

5 Limiting Distribution of the MLE

Now we will address the question of what is the asymptotic distribution of the MLE. This is analogous to the CLT which gave the asymptotic distribution of averages. In some cases, we

can do this directly. For instance, if $X_1, \ldots, X_n \sim \text{Ber}(p)$ then the MLE is just the average:

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and so we know by the CLT:

$$\sqrt{n} \frac{\widehat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1),$$

which tells us the asymptotic distribution of the MLE.

More generally, however the MLE need not be a simple average of i.i.d. terms, but the main take-away is that asymptotically it often behaves like one.

Recall that the score function is

$$s(\theta) = \sum_{i=1}^{n} \nabla \log(p(X_i; \theta)),$$

which is the gradient of the log-likelihood, and the Fisher Information,

$$I(\theta) = \mathbb{E}[s(\theta) s(\theta)^T].$$

We showed that $s(\theta)$ has mean 0, so $I(\theta) = \text{Var}(s(\theta))$. The Fisher information is alternatively the expected Hessian of the log-likelihood:

$$I_n(\theta) = \mathbb{E}\left[\sum_{i=1}^{n} \nabla^2 \log p(X_i; \theta) \right].$$

It is worth remembering that the score is data-dependent, while the Fisher Information is not (it is an expectation over the data so does not depend on the values of $X_1, \ldots, X_n$).

Let $\widehat{\theta}$ denote the MLE. Our goal to show that (under enough regularity conditions),

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, [I_1(\theta)]^{-1}).$$

6 Counterexample

The usual counterexample to the above convergence in distribution is the MLE for the uniform distribution. For the uniform distribution most regularity conditions fail. Formally, we observe $X_1, \ldots, X_n \sim U[0, \theta]$ and want to estimate θ. The log-likelihood:

$$\ell(\theta) = \log \left[\frac{1}{\theta^n} \mathbb{I}(\theta \geq \max_i X_i) \right].$$

The MLE is $\widehat{\theta} = \max_{i=1}^{n} X_i$. Observe that the log-likelihood is not differentiable at the MLE, so the Fisher information is not defined at the MLE.

Another thing that we used frequently in defining the equivalent forms of the Fisher information was to exchange derivatives (with respect to θ) and integrals (with respect to X). This in general does not work if the domain of integration depends on the parameter with respect to which we are taking the derivative. For the uniform distribution the domain of density depends on the parameter.

On the other hand, things are usually nice for exponential families. They will automatically satisfy all the regularity conditions (provided it is identifiable, i.e. say full-rank and minimal) and the MLE is extremely well-behaved in such models.

Returning to the uniform case, we can directly analyze the distribution of the MLE. In a previous lecture we showed that

$$n(\widehat{\theta} - \theta) \xrightarrow{d} -\text{Exp}(1/\theta)$$

(we did this when $\theta = 1$ but you can work out the general case). It follows that $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \delta_0$, where δ_0 is a point mass at 0 and it does not have a Gaussian limit.

7 MLE asymptotics

We will only attempt a heuristic calculation here. If you are curious to see a rigorous proof with minimal regularity assumptions you should look at Van der Vaart's book on Asymptotic Statistics. Here is a list of some sufficient regularity conditions:

1. The dimension of the parameter space does not change with n, i.e. $\theta \in \mathbb{R}^d$ and d is fixed. We have seen that if d grows the MLE need not even be consistent.
2. $p(x;\theta)$ is a smooth (thrice differentiable) function of θ,
3. We can interchange differentiation with respect to θ and integration over X. This in turn requires that the range of X does not depend on θ, and some integrability conditions on $p(x;\theta)$.
4. The parameter θ is identifiable.
5. If the parameter space is restricted, i.e. $\theta \in \Theta$ for some set Θ then θ is in the interior of the set Θ (i.e. cannot be on its boundary).

We will focus on the case when the parameter is one-dimensional, although everything carries over almost exactly in the general (fixed) d case.

Theorem 2 *Under the regularity conditions above,*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, 1/I(\theta)).$$

We note that under the conditions of the theorem one can verify that the MLE is consistent, i.e. that $\widehat{\theta} \xrightarrow{p} \theta$. The basic idea is to verify that under the differentiability assumptions on the density, we can effectively treat the parameter space as compact, then derive a uniform law of large numbers, and then apply the proof from the previous lecture notes. This is a complicated technical proof but you can look it up by searching for Wald's proof of the consistency of the MLE.

The proof will use all the facts about scores and the Fisher information that we derived earlier.

Proof: To begin with let us note the following fact: if $\widehat{\theta} \xrightarrow{p} \theta$, then

$$\mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \widehat{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \theta)] = I(\theta).$$

Since $\widehat{\theta}$ maximizes the log-likelihood we know that the derivative of the log-likelihood at $\widehat{\theta}$ must be 0, i.e.

$$\ell'(\widehat{\theta}) = 0.$$

Formally you need to know that $\widehat{\theta}$ is not on the boundary of the parameter space. To prove this you will need to use the fact that θ is not on the boundary and that $\widehat{\theta} \xrightarrow{p} \theta$.

By a Taylor expansion of the derivative of the log-likelihood we obtain that,

$$0 = \ell'(\widehat{\theta}) = \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\widetilde{\theta}),$$

where $\widetilde{\theta}$ is some point in between $\widehat{\theta}$ and θ. This in turn gives us that,

$$(\widehat{\theta} - \theta) = \frac{\ell'(\theta)}{-\ell''(\widetilde{\theta})},$$

so that,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{\ell'(\theta)}{\sqrt{n}}}{-\frac{\ell''(\widetilde{\theta})}{n}}.$$

We will look at the numerator and denominator separately. The denominator is:

$$-\frac{\ell''(\widetilde{\theta})}{n} = \frac{1}{n}\sum_{i=1}^{n} -\nabla^2_\theta \log p(X_i; \widetilde{\theta}) \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \widetilde{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \theta)] = I(\theta)$$

where the last step uses the fact that $\widetilde{\theta} \xrightarrow{p} \theta$.

The numerator is just the score function, i.e.

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log p(X_i;\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} [\nabla_\theta \log p(X_i;\theta) - \mathbb{E}[\nabla_\theta \log p(X;\theta)]]$$
$$\xrightarrow{d} N(0, \mathrm{Var}(\nabla_\theta \log p(X;\theta))) \xrightarrow{d} N(0, I(\theta)),$$

where we used the facts that the score has mean 0, that the variance of the score is the Fisher information and that by the CLT $\sqrt{n}$ times an average of i.i.d. terms minus its expectation converges in distribution to a normal.

Putting the pieces together via Slutsky's theorem we obtain that,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \frac{1}{I(\theta)} N(0, I(\theta)) \xrightarrow{d} N(0, 1/I(\theta)),$$

which is what we wanted to prove. ■

Example: Suppose that $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$, then the log-likelihood,

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} X_i.$$

The score function:

$$s(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i,$$

and the Fisher information,

$$I(\theta) = \frac{n}{\theta^2}.$$

The MLE is $\widehat{\theta} = \frac{1}{\overline{X}}$. So we can use the above result to conclude that,

$$\widehat{\theta} - \theta \xrightarrow{d} N\left(0, \frac{\theta^2}{n}\right).$$

8 Influence Functions and Regular Asymptotically Linear Estimators

We could have followed a similar proof as above to conclude that the MLE can be written as:

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_\theta \log p(X_i;\theta)}{I(\theta)} + \text{Remainder},$$

where the remainder is small (roughly proportional to the previous term multiplied by $[I(\widetilde{\theta}) - I(\theta)] \to 0$). The term,

$$\psi(x) = \frac{\nabla_\theta \log p(x;\theta)}{I(\theta)},$$

is called the *influence function*.

Thinking of a complex predictor like a deep neural network, one can try to obtain some information about the predictor by trying to compute the influence function of training images on the final predictor. A paper that did this (and quite a bit more) won ICML's best paper a few years ago.

Returning to the expression:

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i).$$

Estimators that satisfy this type of expansion are called asymptotically linear estimators (many non-MLE estimators also satisfy expansions of this form). There is a classical result due to Le Cam that any sufficiently well-behaved (regular) estimator is asymptotically linear. It is not easy to prove (see Van Der Vaart's book). This together with the Cramér-Rao lower bound implies that the MLE is the "best regular asymptotically linear estimator".

9 Asymptotic Relative Efficiency

Once you restrict attention to asymptotically Normal estimators, comparing estimators in terms of their MSE boils down to comparing their variances. Specifically, if

$$\begin{aligned} \sqrt{n}(W_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_W^2) \\ \sqrt{n}(V_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_V^2) \end{aligned}$$

then the *asymptotic relative efficiency (ARE)* is

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 3 *Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. The mle of λ is $\overline{X}$. Let*

$$\tau = \mathbb{P}(X_i = 0).$$

So $\tau = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Another estimator is the mle

$$V_n = e^{-\widehat{\lambda}}.$$

The delta method gives

$$\text{Var}(V_n) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

We have

$$\begin{aligned} \sqrt{n}(W_n - \tau) &\rightsquigarrow N(0, e^{-\lambda}(1 - e^{-\lambda})) \\ \sqrt{n}(V_n - \tau) &\rightsquigarrow N(0, \lambda e^{-2\lambda}). \end{aligned}$$

So

$$\text{ARE}(W_n, V_n) = \frac{\lambda}{e^{\lambda} - 1} \leq 1. \quad \square$$

Since the mle is efficient, we know that, in general, $\text{ARE}(W_n, \text{mle}) \leq 1$.

10 Multivariate Case

Now let $\theta = (\theta_1, \ldots, \theta_k)$. In this case we have

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, I^{-1}(\theta))$$

where $I^{-1}(\theta)$ is the inverse of the Fisher information matrix. The approximate standard error of $\widehat{\theta}_j$ is $\sqrt{I_{jj}^{-1}/n}$. If $\tau = g(\theta)$ with $g : \mathbb{R}^k \to \mathbb{R}$ then by the delta method,

$$\sqrt{n}(\widehat{\tau} - \tau) \rightsquigarrow N(0, (g')^T I^{-1} g')$$

where g' is the gradient of g evaluated at θ.