6.S951 Modern Mathematical Statistics Fall 2024

Lecture 24 — December 10, 2024

Prof. Stephen Bates

Scribe: Theo X. Olausson

1 Outline

Agenda:

1. Optimality of maximum likelihood
   (a) Fisher information & MLE
   (b) Hardness bounds
2. Bonus: Course summary

Last time:

1. Asymptotic normality of M-estimators
   (a) Twice-differentiable (with proof) and general Lipschitz (without proof) results
   (b) Examples
      i. Linear regression
      ii. Robust standard errors

Recap (Asymptotics): So far, our discussion of asymptotic statistics has featured:

1. Using the central limit theorem, Slutsky's lemma, and the delta method to construct and manipulate asymptotic, approximate distributions of estimators
2. Moment estimators
3. M-estimators, with a focus on proving consistency as well as asymptotic normality

Today, we turn our focus to the Maximum Likelihood Estimator (MLE) as a particularly important instance of an M-estimator.

2 Fisher Information & MLE

Setting: Consider observations $X_i \overset{\text{i.i.d.}}{\sim} P$, $X_i \in \mathcal{X}$, for $i = 1, \ldots, n$. Suppose that we have a model class $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$, which may or may not contain the true distribution P. For brevity, let $\ell_\theta(X) \triangleq \log p_\theta(X)$. The MLE is then defined as:

$$\hat{\theta}^{MLE} \triangleq n^{-1} \sum_{i=1}^{n} \ell_\theta(X_i)$$

That is, seen as a special instance of our general framework for M-estimation, we can set

$$M_\theta(X_i) = \ell_\theta(X_i)$$

from which we obtain $\hat{\theta}^{MLE}$ by maximizing $M_n(\theta) = n^{-1} \sum_{i=1}^{n} M_\theta(X_i)$, as usual.

Target of MLE: The first question we would like to answer is what the *target* of $\hat{\theta}^{MLE}$ is, i.e., what the estimator converges to as $n \to \infty$. To answer this question, recall first that $M(\theta) \triangleq \mathbb{E}_p M_\theta(X_i)$. Then, plugging in the setting above and subtracting $\mathbb{E}_p \log p(X_i)$, we have that

$$\begin{aligned} M(\theta) &= \mathbb{E}_p \log p_\theta(X_i) \\ &= \mathbb{E}_P \log \frac{p_\theta(X_i)}{p(X_i)} + \mathbb{E}_p \log p(X_i) \\ &= -KL(P \| P_\theta) + C \end{aligned}$$

This observation immediately yields the following two results.

Proposition 1. *If P is in the model class (i.e., there is some $\theta_0 \in \Theta$ such that $P_{\theta_0} = P$), then $M(\theta)$ attains its maximum uniquely at θ_0.*

Remark 2. *One way to see this is to compare the statement above to Gibb's Inequality from information theory, which states that $KL(Q\|P) \geq KL(P\|P) = 0$ for all distributions Q, P. Thus, maximizing $M(\theta)$ is per the above equivalent to minimizing $KL(P\|P_\theta)$, which–per Gibb's Inequality–is in fact minimized precisely when $P_\theta = P$. What remains to show is uniqueness of the maximum, which follows from the convexity of KL in its first argument.*

Proposition 3. *If P is not in the model class, then $M(\theta)$ attains its maximum at a $\theta_0 \in \Theta$ for which $KL(P\|P_{\theta_0})$ is minimized.*

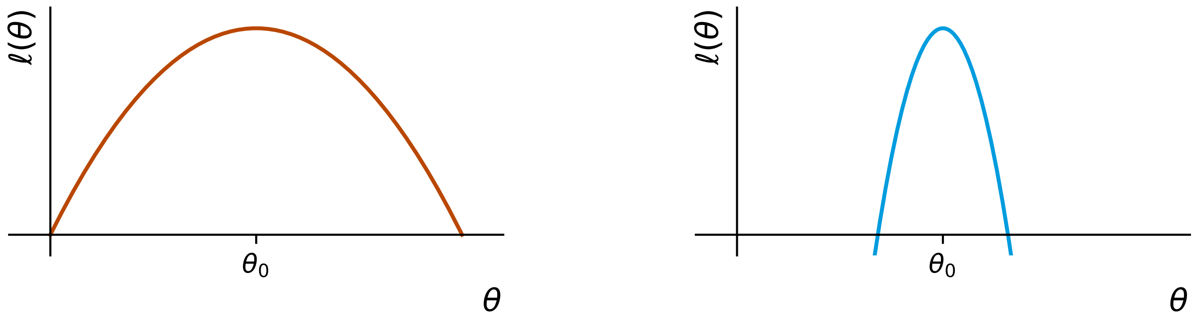
Asymptotic normality & Fisher Information: Our generic results from last time showed that M-estimators are asymptotically normal, and that we have an expression for their variance, under some suitable regularity assumptions. Applying these results in this setting, we have that

$$\sqrt{n}(\hat{\theta}^{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma)$$

where

$$\Gamma = (\mathbb{E}\ddot{\ell}_{\theta_0} X)^{-1} (\mathbb{E}\dot{\ell}_{\theta_0}(X) \dot{\ell}_{\theta_0}(X)^T)(\mathbb{E}\ddot{\ell}_{\theta_0}(X)^T)^{-1}$$

Note that this holds even for misspecified P, i.e., when P is not in the model class.



(a) Low Fisher information at θ_0.

(b) High Fisher information at θ_0.

Figure 1: Illuminating examples of instances where one would obtain high or low values of the Fisher information I_{θ_0}. The more pronounced the "peak" at θ_0 is, the higher the Fisher information.

Definition 4 ((Expected) Fisher Information)**.** *The (expected) fisher information given a distribution P_θ is defined as*

$$I_\theta = \mathbb{E}_\theta \dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^T$$

Intuitively, the Fisher Information encodes the information each X_i has about θ.

Remark 5. *Note that the Fisher Information is precisely the middle term of Γ above, when the expectation is taken over θ.*

We are now ready to prove our first result relating the Fisher Information to $\hat{\theta}^{MLE}$.

Proposition 6. *Under sufficient regularity (see vdV 5.39), if there is some $\theta_0 \in \Theta$ such that $P = P_{\theta_0}$, then*

$$\mathbb{E}\left[-\ddot{\ell}_{\theta_0}(X_i)\right] = \mathbb{E}\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i)^T = I_{\theta_0}$$

which implies that $\Gamma = I_{\theta_0}^{-1}$ in the above; i.e, $\hat{\theta}^{MLE} \overset{\cdot}{\sim} \mathcal{N}\left(\theta_0, I_{\theta_0}^{-1}/n\right)$.

See Figure 1 for a visualization of the intuition behind this proposition.

Informal proof. Suppose that we have enough regularity to exchange integrals and derivatives.[1] Then

$$\left(\int \dot{p}_\theta(x)\right)_j = \frac{\partial}{\partial \theta_j}\int p_\theta(x) = \frac{\partial}{\partial \theta_j} 1 = 0$$

for all θ. Similarly,

$$\left(\int \ddot{p}_\theta(x)\right)_{j,k} = \frac{\partial}{\partial \theta_k}\left(\int \dot{p}_\theta(x)\right)_j = 0$$

[1] This holds, for example, when the Dominated or Monotone Convergence Theorems apply; however, while this is sufficient for the property in question to hold it is not strictly necessary.

per the above. Taking these two observations together, it is easy to verify that

$$\begin{aligned}
\mathbb{E}_{\theta_0}\dot{\ell}_{\theta_0}(X_i) &= \mathbb{E}_{\theta_0}\frac{\dot{p}_{\theta_0}(X_i)}{p_{\theta_0}(X_i)} \\
&= \int \dot{p}_{\theta_0}(X_i) \\
&= 0 \\
\mathbb{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_i) &= \mathbb{E}_{\theta_0}\frac{\ddot{p}_{\theta_0}(X_i)}{p_{\theta_0}(X_i)} - \left(\frac{\dot{p}_{\theta_0}(X_i)}{p_{\theta_0}(X_i)}\right)\left(\frac{\dot{p}_{\theta_0}(X_i)}{p_{\theta_0}(X_i)}\right)^T \\
&= \left(\int \ddot{p}_{\theta_0}(X_i)\right) - \mathbb{E}_{\theta_0}\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i) \\
&= -\mathbb{E}_{\theta_0}\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i) \\
&= -I_{\theta_0}
\end{aligned}$$

from which we thus conclude that $\hat{\theta}^{MLE} \overset{\cdot}{\sim} \mathcal{N}\left(\theta_0, I_{\theta_0}^{-1}/n\right)$. □

3 Optimality of MLE

Anecdote: a lot of this lower bound theory was developed at Berkeley by Le Cam, although the MLE has been studied for much longer than that. These are surprisingly strong results and should thus be considered to be of significant importance.

Definition 7 (Asymptotic relative efficiency). *We say that an estimator $\hat{\theta}_1$ is* asymptotically more efficient *than another estimator $\hat{\theta}_2$ if $\Gamma_1 < \Gamma_2$ in the positive semi-definite ordering*[2] *and*

$$\sqrt{n}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma_1)$$

$$\sqrt{n}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma_2)$$

Theorem 8 (The almost-everywhere convolution theorem; vdV 8.9). *Suppose that the family $\{P_\theta : \theta \in \Theta\}$ is QMD (see vdV section 7.1, reproduced here as Definition 11 in the Appendix). Suppose also that $\hat{\theta}$ is an estimator such that for all $\theta_0 \in \Theta_0$, there is a distribution L_{θ_0} such that*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} L_{\theta_0}$$

Then, for almost all $\theta_0 \in \Theta_0$,

$$L_{\theta_0} = \mathcal{N}\left(0, I_{\theta_0}^{-1}\right) * D_{\theta_0}$$

for some distribution D_{θ_0}. That is,

$$\hat{\theta} \sim \mathcal{N}\left(\theta_0, I_{\theta_0}^{-1}/n\right) + Z/\sqrt{n}$$

where $Z \sim D_{\theta_0}$ (independently of the Gaussian).

[2] i.e., $x^T\Gamma_1 x < x^T\Gamma_2 x$ for all x.

Note: proving this result requires tools that are more advanced than we have had time to cover in this class. See section 8.9 of vdV for a full treatment of the theorem.

The conclusion of this theorem is rather strong, but then it also required us to make a rather strong assumption about the limiting behavior of the estimator sequence. Fortunately, there is a spiritually similar, complementary theorem, which yields a weaker conclusion but also requires a weakener set of assumptions:

Theorem 9 (Local asymptotic minimax theorem; vdV 8.11)**.** *Suppose as before that the family in question is QMD. Suppose also that at θ_0, the Fisher information I_{θ_0} exists. Let then $\hat{\theta}$ be any estimator, and take $L : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ to be a convex, bowl-shaped loss. Then*

$$\sup_{I \subset \mathbb{R}^D \,:\, |I| < \infty} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{(\theta_0 + h/\sqrt{n})} L(\sqrt{n}(\hat{\theta} - \theta_0 - h/\sqrt{n})) \geq \mathbb{E}L(Z)$$

where $Z \sim \mathcal{N}\left(0, I_{\theta_0}^{-1}\right)$.

It is a helpful exercise to compare the formal statement to the name of the theorem: the "local" part refers to the shifted target, $\theta_0 + h/\sqrt{n}$; the "asymptotic" part is of course the limit,; and "minimax" is due to taking the infimum (i.e., asserting that *no* other estimator could do better). That leaves us with the two suprema; the intuition is that one may construct pathological values of h for which the bound does not hold, but given a large enough set of such choices it will certainly hold for some h in that set.

Remark 10. *Spiritually, by combining these two theorems we obtain a pretty strong result that says that we cannot estimate θ_0 better than up to Gaussian noise with covariance equal to the inverse Fisher information (assuming our estimator satisfies some regularity conditions). Given that this is precisely what $\hat{\theta}^{MLE}$ does (by Proposition 6) these results thus prove the optimality of the MLE.*

4 Bonus: Quick Course Summary

This course has consisted of four chunks:

1. Statistical decision theory: as a general framework, with a handful of precise results within that. The goal was to develop a language to talk about abstract statistical problems (testing, estimation, prediction) precisely, so that we can prove results about them. We touched on sufficiency of statistics, as well as minimax and Bayes optimality; optimality of least squares (in the Gaussian model); and optimal testing (in particular, the Neyman-Pearson lemma).

2. Statistical inference procedures: in particular, confidence intervals (via the CLT, concentration inequalities like Hoeffding's, and the bootstrap); confidence bands for CDFs (the DKW inequality); the relationship between testing $\iff$ CIs (i.e., that CIs contain exactly those parameters which are consistent with the data); $p-$values (as measures of how extreme an observation is compared to the reference distribution); and permutation tests as a way of testing whether two distributions are the same without making specific assumptions about the distributions at hand.

3. Topics in advanced inference, in particular: predictive inference, explicitly in the linear model case and using conformal prediction as a black-box/generic method in the more general case; risk control as a variant of conformal prediction with a different criterium; as well as multiple hypothesis testing (Bonferroni to control the family-wise error rate–i.e., the probability of making any false discoveries–as well as the Benjamini-Hoeffberg procedure, which controls the false discovery rate, i.e., the proportion of discoveries in the final rejection set which are actually false).

4. Asymptotics (last five lectures): here the goal was to give approximate distributions of estimators; questions of importance then include optimality of the approximation as well as how to use these approximate distributions to give confidence intervals (e.g., giving a CI for a parameter by plugging in an estimate of the variance of the limiting distribution).

Appendix

Definition 11 (Differentiability in Quadratic Mean; vdW 7.1)**.** *We say that the model class* $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ *is* differentiable in Quadratic Mean *(or QMD, for short) at* θ *if there is a vector of measurable functions* $\dot{\ell}_\theta = (\dot{\ell}_{\theta,1}, \ldots, \dot{\ell}_{\theta,d})$ *such that as* $h \to 0$,

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \sqrt{p_\theta} \right]^2 \, d\mu = o(\|h\|^2)$$

(where p_θ *is the density of* P_θ *with respect to the measure* μ*).*