LECTURE 7 NOTES

1. Convergence of random variables. Before delving into the large sample properties of the MLE, we review some concepts from large sample theory.

1. *Convergence in probability:* $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ if, for any $\delta > 0$,

$$\mathbf{P}(|\mathbf{x}_n - \mathbf{x}| > \delta) \to 0$$

2. *Convergence in distribution:* $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ if

$$\mathbf{E}\left[g(\mathbf{x}_n)\right] \to \mathbf{E}\left[g(\mathbf{x})\right]$$

for all bounded, continuous functions $g : \mathcal{X} \to \mathbf{R}$. If $\mathbf{x}_n$ and $\mathbf{x}$ are real valued, $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ is equivalent to

$$F_n(t) \to F(t) \text{ at the continuity points of } F,$$

where F_n is the CDF of $\mathbf{x}_n$ and F is the CDF of $\mathbf{x}$.

Convergence in probability is the stronger notion of convergence. In particular, $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ implies $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$. There is a partial converse: if $\mathbf{x}_n \xrightarrow{d} c$, where c is a constant, then $\mathbf{x}_n \xrightarrow{p} c$.

For convenience, we say o_P and O_P for convergence and boundedness in probability. Recall for a non-negative sequence of scalars $\{a_n\}$,

1. $a_n \sim o(1)$ means $a_n \to 0$ as n grows.
2. $a_n \sim o(b_n)$ for a non-negative sequence $\{b_n\}$ means $\frac{a_n}{b_n} \sim o(1)$.

Further,

1. $a_n \sim O(1)$ means $\{a_n\}$ is bounded; i.e. there is some (large) $M > 0$ such that a_n
2. $a_n \sim O(b_n)$ (for a non-negative sequence $\{b_n\}$) means $\frac{a_n}{b_n} \sim O(1)$.

There are probabilistic analogues of the preceding statements. Let $\mathbf{x}_n$ be a sequence of non-negative real-valued random variables.

1. $\mathbf{x}_n \sim o_P(1)$ means $\mathbf{x}_n \xrightarrow{p} 0$ as n grows.
2. $\mathbf{x}_n \sim o_P(b_n)$ for a non-negative sequence $\{b_n\}$ means $\frac{\mathbf{x}_n}{b_n} \sim o_P(1)$.
3. $\mathbf{x}_n \sim o_P(\mathbf{y}_n)$ for a sequence of non-negative random variables $\{\mathbf{y}_n\}$ implies $\frac{\mathbf{x}_n}{\mathbf{y}_n} \sim o_P(1)$.

Further,

1. $\mathbf{x}_n \sim O_P(1)$ means $\{\mathbf{x}_n\}$ is bounded in probability; i.e. for any $\epsilon > 0$, there is $M > 0$ such that $\sup_n \mathbf{P}(\mathbf{x}_n > M) \leq \epsilon$.
2. $\mathbf{x}_n \sim O_P(b_n)$ means $\frac{\mathbf{x}_n}{b_n} \sim O_P(1)$.
3. $\mathbf{x}_n \sim O_P(\mathbf{y}_n)$ means $\frac{\mathbf{x}_n}{\mathbf{y}_n} \sim O_P(1)$.

The aforementioned convergence properties are preserved under transformations.

THEOREM 1.1 (Continuous mapping theorem). *For any continuous function g,*

1. $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ *implies* $g(\mathbf{x}_n) \xrightarrow{p} g(\mathbf{x})$.
2. $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ *implies* $g(\mathbf{x}_n) \xrightarrow{d} g(\mathbf{x})$.

THEOREM 1.2 (Slutsky's theorem). *Let $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ and $\mathbf{x}_n \xrightarrow{d} c$. We have*

1. $\mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x} + c$,
2. $\mathbf{x}_n \mathbf{y}_n \xrightarrow{d} c\mathbf{x}$.

More generally, $g(\mathbf{x}_n, \mathbf{y}_n) \xrightarrow{d} g(\mathbf{x}, c)$, for any continuous function g.

PROOF SKETCH. The key step in the proof of Slutsky's theorem shows that $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{d} c$ implies $(\mathbf{x}_n, \mathbf{y}_n) \xrightarrow{d} (\mathbf{x}, c)$. We appeal to the continuous mapping theorem to obtain the stated conclusion. □

We remark that $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{d} \mathbf{y}$ (convergence of the marginal distributions) generally does not imply $\mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x} + \mathbf{y}$; convergence of the joint distribution is crucial.

LEMMA 1.3 (delta method). *Let $a_n(\mathbf{x}_n - c) \xrightarrow{d} \mathbf{x}$, where $a_n \to \infty$ as n grows. For any differentiable g,*

$$a_n(g(\mathbf{x}_n) - g(c)) \xrightarrow{d} \nabla g(c)^T \mathbf{x}.$$

PROOF. By Taylor's theorem,

$$g(x_n) - g(c) = \nabla g(c)^T (x_n - c) + o(\|x_n - c\|_2)$$

for any sequence $\{x_n\} \to c$. Since $a_n(\mathbf{x}_n - c) \xrightarrow{d} \mathbf{x}$, by Slutsky's theorem, we deduce $\mathbf{x}_n - c \xrightarrow{d} 0$, which implies $\mathbf{x}_n - c \xrightarrow{p} 0$. Thus

$$g(\mathbf{x}_n) - g(c) = \nabla g(c)^T (\mathbf{x}_n - c) + o_P(\|\mathbf{x}_n - c\|_2),$$

which implies

$$\begin{aligned} a_n(g(\mathbf{x}_n) - g(c)) &= a_n \nabla g(c)^T (\mathbf{x}_n - c) + o_P(a_n \|\mathbf{x}_n - c\|_2) \\ &= a_n \nabla g(c)^T (\mathbf{x}_n - c) + o_P(1). \end{aligned}$$

Rearranging,

$$a_n(g(\mathbf{x}_n) - g(c)) - a_n \nabla g(c)^T (\mathbf{x}_n - c) \xrightarrow{p} 0.$$

By the continuous mapping theorem,

$$a_n \nabla g(c)^T (\mathbf{x}_n - c) \xrightarrow{d} \nabla g(c)^T \mathbf{x},$$

which allows us to conclude $a_n(g(\mathbf{x}_n) - g(c)) \xrightarrow{d} \nabla g(c)^T \mathbf{x}$. □

COROLLARY 1.4. *Let* $\sqrt{n}(\mathbf{x}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. *We have*

$$\sqrt{n}(g(\mathbf{x}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}\big(0, \nabla g(\theta)^T \Sigma \nabla g(\theta)\big).$$

Corollary 1.4 is by far the most common application of the delta method to the extent that some sources go as far as to call it the delta method.

2. Consistency. Most asymptotic statements concern a sequence of estimators $\{\hat{\theta}_n\}$ indexed by sample size n. In the usual asymptotic framework, we observe *i.i.d.* random variables $\{\mathbf{x}_i\}$ and consider the sequence of estimators

$$\hat{\theta}_n = \hat{\theta}\big(\{\mathbf{x}\}_{i \in [n]}\big)$$

that consists of the same estimation procedure performed on growing samples. For example, in the coin tossing investigation, the sequence of MLE's of p consists of the means of growing samples:

$$\hat{p}_n = \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_i.$$

In statistical parlance, statements about asymptotic properties of an estimator (instead of a sequence of estimators) are common. Such statements are actually about the underlying estimation procedure, and should be interpreted as such.

DEFINITION 2.1. *A sequence of point estimators* $\{\hat{\theta}_n\}$ *of* $\theta^* \in \Theta$ *is consistent if* $\hat{\theta}$ *converges in probability to* θ.

We remark that since convergence in mean squared implies convergence in probability, a sufficient condition for consistency is

$$\mathbf{bias}_{\theta^*}\big[\hat{\theta}_n\big]^2 + \mathbf{var}_{\theta^*}\big[\hat{\theta}_n\big] = \mathbf{E}_{\theta^*}\Big[\|\hat{\theta}_n - \theta^*\|_2^2\Big] \to 0.$$

Consistency is the basic notion of "correctness" for a point estimator. It ensures an estimator converges to its target as the sample size grows. For simple estimators, such as those that are smooth functions of the observations $\{\mathbf{x}_i\}_{i\in[n]}$, consistency is often a direct consequence of the law of large numbers (LLN). In the coin tossing investigation, the (weak) LLN implies

$$\hat{p}_n \xrightarrow{p} p^*.$$

Further, by the continuous mapping theorem, the estimator $g(\hat{\theta}_n)$ is a consistent estimator of $g(\theta^*)$ as long as $\hat{\theta}_n$ is a consistent estimator of θ^*.

EXAMPLE 2.2. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p^*)$. *Recall the MLE of* p^* *is* $\bar{\mathbf{x}}_n$. *By the law of large numbers,* $\bar{\mathbf{x}}_n \xrightarrow{p} p^*$, *which shows that the MLE is consistent.*

EXAMPLE 2.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{unif}(0, \theta^*)$. *Recall the MLE of* θ^* *is* $\mathbf{x}_{(n)}$. *It is possible to show the MLE is consistent by direct calculation. Indeed,*

$$\begin{aligned}\mathbf{P}_\theta\big(\theta^* - \mathbf{x}_{(n)} > \epsilon\big) &= \mathbf{P}_\theta\big(\max_{i\in[n]} \mathbf{x}_i < \theta^* - \epsilon\big) \\ &= \prod_{i\in[n]} \mathbf{P}_\theta\big(\mathbf{x}_i < \theta^* - \epsilon\big) \\ &= \prod_{i\in[n]} \frac{\theta^* - \epsilon}{\theta^*},\end{aligned}$$

which converges to zero as n *grows.*

We turn our attention to establishing the consistency of the MLE. The MLE is an example of an *extremum estimator*: it is the maximizer of the log-likelihood $\ell_{\mathbf{x}}(\theta)$. Since the log-likelihood depends on the observations $\mathbf{x}$, it is a random function (of θ). The basic idea is to show that

1. the log-likelihood converges *uniformly* to a population objective,
2. the true parameter θ^* is the maximizer of the population objective.

The first condition is known as a *uniform law of large numbers*.

DEFINITION 2.4. *A set of (real-valued) functions* $\mathcal{F}$ *satisfies a* uniform law of large numbers *(ULLN) if*

$$\sup_{f\in\mathcal{F}} \Big|\tfrac{1}{n}\textstyle\sum_{i\in[n]} f(\mathbf{x}_i) - \mathbf{E}\big[f(\mathbf{x}_1)\big]\Big| \xrightarrow{p} 0.$$

We remark that as long as $\mathbf{E}\big[f\big]$ is finite, the (weak) LLN ensures

$$\left|\tfrac{1}{n}\textstyle\sum_{i\in[n]} f(\mathbf{x}_i) - \mathbf{E}\big[f(\mathbf{x}_1)\big]\right| \xrightarrow{p} 0$$

for each $f \in \mathcal{F}$. A ULLN is a stronger statement that says the convergence is uniform over $\mathcal{F}$. Why is the stronger statement necessary? Consider the MLE $\hat{\theta}_n$ of θ in a parametric model. The LLN ensures

$$\tfrac{1}{n}\textstyle\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta) \xrightarrow{p} \mathbf{E}_\theta\big[\ell_{\mathbf{x}_1}(\theta)\big]$$

for any $\theta \in \Theta$. Unfortunately, pointwise convergence of the log-likelihood is not even enough to ensure the convergence of the maximum, much less the maximizer, which is required to conclude the consistency of the MLE.

THEOREM 2.5. *Let* $\{\mathbf{x}_i\}$ *be* i.i.d. *random variables, and* $f_{\theta^*}(x)$ *be their density for some* $\theta^* \in \Theta$. *Assume*

1. $f_\theta(x) \overset{a.e.}{=} f_{\theta^*}(x)$ *if and only if* $\theta = \theta^*$,
2. $\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta) - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]\right| \xrightarrow{p} 0$,

then θ^* *is the unique maximizer of* $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$ *on* Θ *and*

$$\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta^*)\big] - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\big] \xrightarrow{p} 0.$$

PROOF. The proof consists of two parts:

1. show that θ^* is the unique maximizer of $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$ on Θ.
2. show that the maximum of $\frac{1}{n}\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta))$ converges to the maximum of $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$.

We begin by showing

$$\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta^*)\big] - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big] > 0$$

for any $\theta \neq \theta^*$. By the linearity of the expectation,

$$\begin{aligned}
&\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta^*)\big] - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big] \\
&\quad= \mathbf{E}_{\theta^*}\big[\log f_{\theta^*}(\mathbf{x}_1)\big] - \mathbf{E}_{\theta^*}\big[\log f_\theta(\mathbf{x}_1)\big] \\
&\quad= \mathbf{E}_{\theta^*}\left[\log\left[\frac{f_{\theta^*}(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)}\right]\right],
\end{aligned}$$

which is the Kullback-Leibler divergence between F_{θ^*} and F_θ. The KL divergence is always non-negative. Indeed, by Jensen's inequality,

$$\begin{aligned}\mathbf{E}_{\theta^*}\left[\log\left[\frac{f_{\theta^*}(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)}\right]\right] &= \mathbf{E}_{\theta^*}\left[-\log\left[\frac{f_\theta(\mathbf{x}_1)}{f_{\theta^*}(\mathbf{x}_1)}\right]\right] \\ &\geq -\log\left(\int_{\mathcal{X}} f_\theta(x_1)dx_1\right) = -\log 1.\end{aligned}$$

Further, since $\log x$ is strictly concave, the inequality is strict unless

$$\frac{f_{\theta^*}(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)} \overset{a.s.}{=} \mathbf{E}_{\theta^*}\left[\frac{f_{\theta^*}(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)}\right],$$

which, by the first assumption, is not possible at any $\theta \neq \theta^*$. Thus θ^* is the unique maximizer of $\mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta)\right]$.

To complete the proof, we show that

$$\mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right] - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\right] \xrightarrow{p} 0. \tag{2.1}$$

We expand $\mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\right] - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right]$:

$$\begin{aligned}&\mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right] - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\right] \\ &\quad= \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right] - \tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta^*)) + \underbrace{\tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta^*) - \tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\hat{\theta}_n)}_{\leq 0 \text{ since } \hat{\theta}_n \text{ maximizes } \sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta)} \\ &\qquad+ \tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\hat{\theta}_n) - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\right] \\ &\quad\leq \left|\mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right] - \tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta^*))\right| + \left|\tfrac{1}{n}\textstyle\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\hat{\theta}_n) - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\right]\right|,\end{aligned}$$

The first term is $o_P(1)$ by the LLN, and the second term is also $o_P(1)$ by the uniform convergence assumption, which establishes (2.1). □

Theorem 2.5 captures the essence of most results on the consistency of the MLE. The key ingredients are identifiability and uniform convergence. The first assumption of Theorem 2.5 is an identifiability assumption. It ensures the maximizer of $\mathbf{E}_\theta\left[\ell_{\mathbf{x}_1}(\theta)\right]]$ is unique by disallowing two parameters θ_1, θ_2 to correspond to the same density. The second assumption is a ULLN for the set of log-likelihood functions $\{\log f_\theta(x) : \theta \in \Theta\}$. Here is an example of such a ULLN.

LEMMA 2.6. *Assume*

1. *$\ell_x(\theta)$ is a continuous function (of θ) for any $x \in \mathcal{X}$*

2. *there is an* envelope function b *such that* $\mathbf{E}\big[b(\mathbf{x})\big] < \infty$ *and* $|\ell_x(\theta)| \leq b(x)$ *for any* $\theta \in \Theta$.

If Θ *is compact, then*

$$\sup_{\theta\in\Theta} \left| \tfrac{1}{n} \textstyle\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta) - \mathbf{E}\big[\ell_{\mathbf{x}_1}(\theta)\big] \right| \xrightarrow{p} 0.$$

We remark that Theorem 2.5 establishes the *excess risk* of the MLE

$$\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta^*)\big] - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\hat{\theta}_n)\big]$$

vanishes. However, sans other assumptions, we cannot conclude $\hat{\theta}_n \xrightarrow{p} \theta^*$. Most results on the consistency of the MLE state a laundry list of assumptions, most of which are required to establish uniform convergence and to conclude $\hat{\theta}_n \xrightarrow{p} \theta^*$ from (2.1). For completeness, we state such a result.

COROLLARY 2.7. *Let* $\{\mathbf{x}_i\}$ *be* i.i.d. *random variables, and* $f_{\theta^*}(x)$ *be their density for some* θ^* *in a compact parameter space* Θ. *If*

1. $f_\theta(x) = f_{\theta^*}(x)$ *if and only if* $\theta = \theta^*$,
2. $\ell_x(\theta)$ *is continuous for any* $x \in \mathcal{X}$
3. *there is a function* $b : \mathcal{X} \to \mathbf{R}$ *such that* $|\ell_x(\theta)| \leq b(x)$ *for any* $\theta \in \Theta$ *and* $\mathbf{E}_{\theta^*}\big[b(\mathbf{x}_1)\big] < \infty$.

then the MLE is consistent; i.e. $\arg\max_{\theta\in\Theta} \sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta) \xrightarrow{p} \theta^*$.

PROOF. First, we show that $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$ is continuous under the stated assumptions. For any sequence $\{\theta_n\}$ converging to θ,

$$\{\ell_x(\theta_n)\} \to \ell_x(\theta)$$

by the continuity of ℓ_x. Since ℓ_x is dominated by b and $\mathbf{E}_{\theta^*}\big[b(\mathbf{x}_1)\big]$ is finite,

$$\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta_n)\big] \to \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$$

by the dominated convergence theorem.

To complete the proof, we show that $\hat{\theta}_n \xrightarrow{p} \theta^*$. For any $\delta > 0$, consider

$$\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2\geq\delta} \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big].$$

Since $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$ is continuous and $\{\theta \in \Theta : \|\theta - \theta^*\|_2 \geq \delta\}$ is compact, the sup is attained at some $\bar{\theta} \in \{\theta \in \Theta : \|\theta - \theta^*\|_2 \geq \delta\}$. By Theorem 2.5, θ^* is the unique maximizer of $\mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta)\big]$ on Θ. Thus

$$\epsilon := \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\theta^*)\big] - \mathbf{E}_{\theta^*}\big[\ell_{\mathbf{x}_1}(\bar{\theta})\big] > 0.$$

When $\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta) - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta)\right]\right| < \frac{\epsilon}{2}$,

$$\begin{aligned}
\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta) &\geq \frac{1}{n}\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta^*) \\
&> \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta^*)\right] - \frac{\epsilon}{2} \\
&= \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\bar{\theta})\right] + \frac{\epsilon}{2} \\
&\geq \sup_{\theta\in\Theta:\|\theta-\theta^*\|_2\geq\delta}\frac{1}{n}\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta)
\end{aligned}$$

which implies $\left\|\hat{\theta}_n - \theta^*\right\|_2 \leq \delta$. By Lemma 2.6,

$$\mathbf{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta) - \mathbf{E}_{\theta^*}\left[\ell_{\mathbf{x}_1}(\theta)\right]\right| < \frac{\epsilon}{2}\right) \to 1,$$

We conclude $\mathbf{P}_{\theta^*}\left(\left\|\hat{\theta}_n - \theta^*\right\|_2 \leq \delta\right) \to 1$. □

The takeaways from the lengthy preceding section on the consistency of the MLE are

1. as long as $\frac{1}{n}\sum_{i\in[n]}\ell_{\mathbf{x}_i}(\theta)$ converges to $\mathbf{E}\left[\ell_{\mathbf{x}_1}(\theta)\right]$ uniformly on $\theta\in\Theta$, the risk of the MLE converges to the risk of θ^*. Establishing uniform convergence is non-trivial, and we do not dwell on the topic.
2. to show consistency $\hat{\theta}_n \xrightarrow{p} \theta^*$ requires additional technical conditions on the parametric model. There are many approaches to establishing consistency: we followed Wald's original approach.

To wrap up, we consider the consequences of an incorrect model; i.e. the parametric model

$$\mathcal{F} := \{f_\theta(x) : \mathcal{X} \to \mathbf{R} : \theta\in\Theta\}$$

does not include the generative distribution of the observations. We know that the MLE converges to the maximizer of $\mathbf{E}_F\left[\log f_\theta(\mathbf{x})\right]$ or equivalently, the minimizer of

$$\mathbf{E}_F\left[\log f(\mathbf{x})\right] - \mathbf{E}_F\left[\log f_\theta(\mathbf{x})\right],$$

where F is the generative distribution and $f(x)$ is its density. We recognize the difference as the KL divergence between F and F_θ:

$$= \mathbf{E}_F\left[\log\left[\frac{f(\mathbf{x})}{f_\theta(\mathbf{x})}\right]\right].$$

Thus the MLE generally converges to a $\theta^*\in\Theta$ that is the best approximation of F in the parametric model. That is F_{θ^*} has the smallest KL divergence to F among the distributions in the parametric model.

Yuekai Sun
Berkeley, California
December 4, 2015