Lecture 16  Consistency

Unless mentioned otherwise in the next few lectures, we assume that \( \{X_n\} \) is a sequence of iid samples from \( P_\theta, \theta \in \Theta \), and will present some basic results in asymptotic statistics, i.e. large sample theory.

**Definition 1** \( T_n \Delta = T(X^n) \) is said to be a (weakly) consistent estimator of \( \theta \) if \( T_n \) converges to \( \theta \) in probability as \( n \to \infty \), denoted by \( T_n \xrightarrow{P} \theta \). Moreover, \( T_n \) is said to be a strongly consistent estimator of \( \theta \) if with probability one, \( T_n \) converges to \( \theta \), or equivalently, \( T_n \) is said to converge to \( \theta \) almost surely, denoted by \( T_n \xrightarrow{a.s.} \theta \).

**Note:**
- Strong consistency implies weak consistency, but not vice versa.
- By definition, \( T_n \xrightarrow{P} \theta \) if and only if for every \( \epsilon > 0 \), \( P(\|T_n - \theta\| > \epsilon) \to 0 \) as \( n \to \infty \). However, \( T_n \xrightarrow{a.s.} \theta \) requires a certain rate for the convergence \( P(\|T_n - \theta\| > \epsilon) \to 0 \). A general sufficient condition for strong consistency of \( T_n \) is that for every \( \epsilon > 0, \sum_{n=1}^{\infty} P(\|T_n - \theta\| > \epsilon) < \infty \), called Borel-Cantelli Lemma.
- Here is a useful result of “continuous mapping”: Let \( g \) be a continuous function. Then \( T_n \xrightarrow{P} \theta \) implies \( g(T_n) \xrightarrow{P} g(\theta) \), and \( T_n \xrightarrow{a.s.} \theta \) implies \( g(T_n) \xrightarrow{a.s.} g(\theta) \).

The consistency of sample means and sample variances follows from the strong law of large numbers along with continuous mappings.

**Theorem 1** Let \( \mu = E_\theta X_1 \) and \( \sigma^2 = Var_\theta X_1 \). Then \( \bar{X} \xrightarrow{a.s.} \mu \) and \( \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \xrightarrow{a.s.} \sigma^2 \).

Consistency is a minimal requirement in asymptotic statistics. There are usually many consistent estimators for \( \theta \) in a problem, some of which are quite dubious and silly, e.g.

\[
T_n = \begin{cases} 
1, & \text{if } n \leq 1,000,000, \\
\bar{X}, & \text{if } n > 1,000,000,
\end{cases}
\]

is only one of many consistent estimators of population mean \( \mu \). Obviously, there should be some additional information about convergence rates attached whenever possible. See the next two examples for illustration.
Example 16.1 Let \( X_1 \sim U(0, \theta) \). Then the MLE \( \hat{\theta} = X_{(n)} \). Furthermore, for any \( \epsilon \in (0, \theta) \), \( P(|X_{(n)} - \theta| > \epsilon) = P(X_{(n)} < \theta - \epsilon) = \left( \frac{\theta - \epsilon}{\theta} \right)^n = e^{-bn} \) with \( b = \log \frac{\theta}{\theta - \epsilon} \) represents an exact rate of exponential convergence.

Example 16.2 Let \( X_1 \sim f(x|\theta) = \exp[\theta T(x) - \psi(\theta)], \theta \in \Theta^o \subset \mathbb{R} \). Write \( Y_i = T(X_i), i = 1, 2, \ldots, \) and \( U_n = \sum_{i=1}^n Y_i \). Clearly, \( \bar{Y} \) is a consistent estimator of \( \Delta \triangleq \frac{\Delta}{d\psi/d\theta} \). To get some information about the convergence rate, start with

\[
P(|\bar{Y} - \lambda| > \epsilon) \leq P(A_n) + P(B_n),
\]

where \( A_n = \{ \bar{Y} > \lambda + \epsilon \} \) and \( B_n = \{ \bar{Y} < \lambda - \epsilon \} \). For any \( t > 0 \) with \( \theta + t \in \Theta^o \), we have

\[
P(A_n) = P(e^{tU_n} > e^{tn(\lambda+\epsilon)}) \leq E\{\exp[t U_n - tn(\lambda + \epsilon)] | A_n\}
\leq e^{-tn(\lambda+\epsilon)} E[e^{tY_1}]^n
= e^{-tn(\lambda+\epsilon)} \exp\{n[\psi(\theta + t) - \psi(\theta)]\} \leq e^{-n\delta}
\]

for some \( \delta > 0 \) that depends on \( \theta \) and \( \epsilon \) because \( t^{-1}[\psi(\theta + t) - \psi(\theta)] \to \lambda \) as \( t \to 0 \). \( P(B_n) \) can be bounded similarly. It can also be shown (details skipped) that

\[
P(|\bar{Y} - \lambda| > \epsilon) \geq e^{-n\delta'}
\]

for some \( \delta' > 0 \) that depends on \( \theta \) and \( \epsilon \).

The following two examples concern consistency in nonparametric estimation. See Lecture 6 for the notation.

Example 16.3 Let \( X_1, X_2, \ldots \) be iid random variables with a common unknown cdf \( F \). Obviously, for every fixed \( x \), we have \( \bar{F}_n(x) \stackrel{a.s.}{\to} F(x) \). A stronger result (Glivenko-Cantelli Theorem) holds:

\[
\sup_{x \in \mathbb{R}} |\bar{F}_n(x) - F(x)| \stackrel{a.s.}{\to} 0.
\]

Example 16.4 Consider a discrete model with an unknown mass function \( p_\xi = P(X_1 = \xi), \xi \in D \) — a discrete set in \( \mathcal{X} \). First, \( \frac{S_\xi(n)}{n} \stackrel{a.s.}{\to} p_\xi \) for every fixed \( \xi \). Again, a stronger result holds:

\[
\sum_{\xi \in D} \left| \frac{S_\xi(n)}{n} - p_\xi \right| \stackrel{a.s.}{\to} 0. \tag{16.1}
\]

To verify (16.1), note the following facts:

(a) For every \( y \in \mathbb{R} \), we have \( y = y^+ - y^- \) and \( |y| = y^+ + y^- \).
(b) Since \( \sum_{\xi \in D} \left( p_\xi - \frac{S^{(\xi)}_n}{n} \right)^+ \leq \sum_{\xi \in D} p_\xi \) and \( \left( p_\xi - \frac{S^{(\xi)}_n}{n} \right)^+ \xrightarrow{a.s.} 0 \), it follows from Lebesgue Dominated Convergence Theorem that \( \sum_{\xi \in D} \left( p_\xi - \frac{S^{(\xi)}_n}{n} \right)^+ \xrightarrow{a.s.} 0 \).

(c) \( \sum_{\xi \in D} \left( p_\xi - \frac{S^{(\xi)}_n}{n} \right) = \sum_{\xi \in D} p_\xi - \sum_{\xi \in D} \frac{S^{(\xi)}_n}{n} = 1 - 1 = 0 \).

(d) (a), (b) and (c) together imply \( \sum_{\xi \in D} \left( p_\xi - \frac{S^{(\xi)}_n}{n} \right)^- \xrightarrow{a.s.} 0 \).

Finally, (16.1) follows from (a), (b) and (d). We should be careful with the null sets: Let \( B_\xi \) be the null set on which \( \frac{S^{(\xi)}_n}{n} \xrightarrow{a.s.} p_\xi \) does not hold, and let \( B \) be the null set on which (16.1) fails. Then \( B = \bigcup_{\xi \in D} B_\xi \). Hence \( P(B) \leq \sum_{\xi \in D} P(B_\xi) = 0 \).