Warning: these notes may contain factual errors

Reading: VDV Chapter 5.1-5.2, 5.5; TPE Chapter 2.5

Outline of Lecture 2:

1. Basic consistency and identifiability
2. Asymptotic Normality
   (a) Taylor expansions & Fisher Information
   (b) Moment method (not covered)
   (c) Exponential Family models (not covered)

Recap: Last lecture, we discussed the Delta Method (aka Taylor expansions). The basic idea was as follows:

Claim 1. If \( r_n(T_n - \theta) \overset{d}{\to} T \), then if \( \phi'(\theta) \) exists, \( r_n(\phi(T_n) - \phi(\theta)) = r_n(\phi'(\theta)(T_n - \theta)) + o_p(r_n(T_n - \theta)) \). Note that \( r_n(T_n - \theta) \overset{p}{\to} 0 \), i.e. \( o_p(1) \), since the sequence is uniformly tight. By applying the Continuous Mapping Theorem and Slutsky’s Lemma, we conclude that \( r_n(\phi(T_n) - \phi(\theta)) \overset{d}{\to} \phi'(\theta)T \).

Notation:

Definition 0.1. Given distribution \( P \) on \( \mathcal{X} \), function \( f : \mathcal{X} \to \mathbb{R}^d \),

\[
Pf := \int f dP = \int_{\mathcal{X}} f(x) dP(x) = \mathbb{E}_P[f(x)]
\]

If \( X_i, i = 1, \ldots, n \), are observations, use \( P_n \) to denote the empirical distribution:

\[
P_n f := \frac{1}{n} \sum_{i=1}^{n} f(x_i)
\]

for any function \( f \)

Example 1: \( P_n(A) = \frac{1}{n} \text{card } \{i \in [n] : x_i \in A\} \)

Setting: We have some model family \( \{P_\theta\}_{\theta \in \Theta} \) of distributions on \( \mathcal{X} \), where \( \Theta \subseteq \mathbb{R}^d \). Also, assume all \( P_\theta \) have density \( p_\theta \) with respect to base measure \( \mu \) on \( \mathcal{X} \), i.e. \( p_\theta = \frac{dP_\theta}{d\mu} \).
Idea: Our big idea for the day is that if we get enough observations from the same $P_{\theta}$, we should be able to:

1. Identify $\theta$

2. If $\ell_{\theta}(x) = \log p_{\theta}(x)$ is “smooth” enough, get explicit asymptotic normality results.

Definition 0.2.

$$\nabla \ell_{\theta}(x) := \left[ \frac{\partial}{\partial \theta_j} \log p_{\theta}(x) \right]_{j=1}^d \in \mathbb{R}^d$$

$$\nabla^2 \ell_{\theta}(x) := \left[ \frac{\partial^2}{\partial \theta_i \theta_j} \log p_{\theta}(x) \right]_{i,j=1}^{d \times d} \in \mathbb{R}^{d \times d}$$

Note: $\dot{\ell}_{\theta} \equiv \nabla \ell_{\theta}(x)$ and $\ddot{\ell}_{\theta} \equiv \nabla^2 \ell_{\theta}(x)$.

The gradient of the log-likelihood is often called the “score function.” We will use this term to refer to $\nabla \ell_{\theta}(x)$ throughout future lectures.

Estimating $\theta_0$: Observe $X_i \overset{iid}{\sim} P_{\theta_0}$ but $\theta_0$ is unknown. Our goal is to estimate $\theta_0$.

A standard estimator is to choose $\hat{\theta}_n$ to maximize the “likelihood,” i.e. the probability of the data.

$$\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} P_n \ell_{\theta}(x)$$

Example 2 (Gaussian mean): Let $X_i \overset{iid}{\sim} \text{N}(\theta, I)$ and $p_{\theta}(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2} \|x-\theta\|^2}$. Then, $P_n \ell_n(\theta) = -\frac{1}{2n} \sum_{i=1}^n \|x_i - \theta\|^2 + \frac{d}{2} \log(2\pi)$. Note that this function is concave, so it has a unique global maximum.

Since $\nabla P_n \ell_n(\theta) = -\frac{1}{n} \sum_{i=1}^n (\theta - x_i) = 0$, $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n x_i$. And by the CLT, we know that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{N}(0, I)$.

Example 3 (2-sided exponential): Let $X_i \overset{iid}{\sim} \text{Laplace}(\theta_0)$ and $p_{\theta}(x) = \frac{1}{2} e^{-|\theta - x|}$, i.e. $\log p_{\theta}(x) = -|\theta - x| - \log 2$. Then, $\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^n |\theta - x_i|$, i.e. $\hat{\theta}_n = \text{Median}(x_1, \ldots, x_n)$.

Note that even though the likelihood function is not smooth (the absolute value function has a discontinuity at 0), we are still able to derive asymptotic normality results for this estimator. These are, however, not presented in this lecture.

In general, there are two relevant questions to ask about Maximum Likelihood and other estimators of $\theta_0$:

1. Consistency: Does $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$?

2. What is the asymptotic distribution and the rate of convergence of $\hat{\theta}_n$ to $\theta_0$, i.e. for what $r_n \to \infty$, does $r_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$ and what is $Z$?
Consistency:

**Definition 0.3** (Identifiability). A model \( \{P_\theta \}_{\theta \in \Theta} \) is **identifiable** if \( P_{\theta_1} \neq P_{\theta_2} \) for all \( \theta_1, \theta_2 \in \Theta \) \((\theta_1 \neq \theta_2)\).

Equivalently, \( D_{\text{kl}}(P_{\theta_1} \| P_{\theta_2}) > 0 \) when \( \theta_1 \neq \theta_2 \). Recall that \( D_{\text{kl}}(P_{\theta_1} \| P_{\theta_2}) = \int \log \frac{dP_{\theta_1}}{dP_{\theta_2}} dP_{\theta_1} \).

Note that \( P_{\theta_1} \neq P_{\theta_2} \) means that \( \exists \text{ set } A \subseteq \mathcal{X} \) such that \( P_{\theta_1}(A) \neq P_{\theta_2}(A) \).

Now that we have established what both identifiability and consistency mean, we can prove a basic result regarding the finite consistency of the Maximum Likelihood estimator (MLE).

**Proposition 2** (Finite \( \Theta \) consistency of MLE). Suppose \( \{P_\theta \}_{\theta \in \Theta} \) is identifiable and \( \text{card} \Theta < \infty \). Then, if \( \hat{\theta}_n := \arg\max_{\theta \in \Theta} P_n \ell_\theta(x) \), \( \hat{\theta}_n \overset{p}{\to} \theta_0 \) when \( x_i \overset{iid}{\sim} P_{\theta_0} \).

**Proof of Proposition** We know by the Strong Law of Large Numbers that \( P_n \ell_\theta(x) \overset{a.s.}{\to} P_{\theta_0} \ell_\theta(x) \) when \( x_i \overset{iid}{\sim} P_{\theta_0} \).

\[
P_{\theta_0} \ell_\theta_0(x) - P_{\theta_0} \ell_\theta(x) = \mathbb{E}_{\theta_0} \left[ \log \frac{P_{\theta_0}(x)}{P_{\theta}(x)} \right] = D_{\text{kl}}(P_{\theta_0} \| P_{\theta})
\]

We know that \( D_{\text{kl}}(P_{\theta_0} \| P_{\theta}) > 0 \) unless \( \theta = \theta_0 \). So, for large enough \( n \) and finite \( \Theta \), we have that:

\[
P_n \ell_{\theta_0}(x) - P_n \ell_\theta(x) > 0 \quad \forall \theta \neq \theta_0
\]

Therefore, \( \hat{\theta}_n = \theta_0 \) “eventually.” \( \square \)

**Remark** The above result can fail for \( \Theta \) infinite even if \( \Theta \) is countable.

One sufficient condition often used for consistency results is a **uniform law**, i.e. for \( x_i \overset{iid}{\sim} P \), we have that \( \sup_{\theta \in \Theta} |P_n \ell_\theta - P \ell_\theta| \overset{p}{\to} 0 \).

**Example 4** (Uniform laws): Suppose \( \Theta = \mathbb{R}^d, x_i \overset{iid}{\sim} P \). Consider \( P_n \langle \theta, x \rangle = \frac{1}{n} \sum_{i=1}^{n} \langle \theta, x_i \rangle \). Then, for each \( \theta \in \Theta \), \( P_n \langle \theta, x \rangle \overset{a.s.}{\to} \langle \theta, \mathbb{E}[x] \rangle \).

Does the uniform law, i.e. \( \sup_{\theta \in \Theta} |\langle \theta, P_n x \rangle - \langle \theta, P x \rangle| \to 0 \), hold? If \( \text{Cov} x \neq 0 \), no. However, if \( \Theta \) is compact, we will be able to prove this result.

Alternatively, \( \hat{\theta}_n = \arg\max_{\theta \in \Theta} P_n (x, \theta) \) does satisfy \( \langle \hat{\theta}_n, P x \rangle \to \sup_{\theta \in \Theta} \langle \theta, P x \rangle \). ♠

Now, that we have established some basic definitions and results regarding the consistency of estimators, we turn our attention to understanding their asymptotic behavior.

**Asymptotic Normality via Taylor Expansions:**

**Definition 0.4** (Operator norm). \( \|A\|_{\text{op}} := \sup_{\|u\|_2 \leq 1} \|Au\|_2 \).

Note: \( A \in \mathbb{R}^{k \times d}, u \in \mathbb{R}^d \) and \( \|Ax\|_2 \leq \|A\|_{\text{op}} \|x\|_2 \).

Before we do anything, we have to make several assumptions.

1. We have a “nice, smooth” model, i.e. the Hessian is Lipschitz-continuous. To be rigorous, the following must hold:

\[
\|\nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_2}(x)\|_{\text{op}} \leq M(x) \|\theta_1 - \theta_2\|_2 \quad \mathbb{E}_\theta[M^2(x)] < \infty
\]
2. The MLE, \( \hat{\theta}_n \in \arg\max_{\theta \in \Theta} P_n \ell_\theta(x) \), is consistent, i.e. \( \hat{\theta}_n \overset{p}{\to} \theta_0 \) under \( P_{\theta_0} \).

3. \( \nabla P_n \ell_{\hat{\theta}_n} = 0 \).

4. \( \Theta \) is a convex set.

**Theorem 3.** Let \( x_i \overset{\text{iid}}{\sim} P_{\theta_0} \) and assume the conditions stated above. Then, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, (P_{\theta_0} \nabla^2 \ell_{\theta_0})^{-1} P_{\theta_0} \nabla \ell_{\theta_0} \nabla^T \ell_{\theta_0} (P_{\theta_0} \nabla^2 \ell_{\theta_0})^{-1}) \).

**Remark** We define the Fisher Information as \( I_{\theta} := \mathbb{E}_\theta[\nabla \ell_\theta(x) \nabla \ell_\theta(x)^T] = \text{Cov}_\theta \nabla \ell_\theta. \)

The final equality holds because \( \mathbb{E}_\theta[\nabla \ell_\theta(x)] = 0 (\theta \text{ maximizes } \mathbb{E}_\theta[\ell_\theta(x)]) \).

To show this, assume that we can swap \( \nabla, \mathbb{E} \). Then, \( \nabla \ell_\theta(x) = \nabla \log p_\theta(x) = \nabla p_\theta(x) p_\theta(x) \).

Using that result, we see that:

\[
\mathbb{E}_\theta[\nabla \ell_\theta] = \mathbb{E}\left[ \frac{\nabla p_\theta}{p_\theta} \right] = \int \frac{\nabla p_\theta}{p_\theta} p_\theta d\mu = \int \nabla p_\theta d\mu = \nabla(1) = 0.
\]

Similarly, given that \( \nabla^2 \ell_\theta = \nabla \left( \frac{\nabla p_\theta}{p_\theta} \right) = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla^T p_\theta}{p_\theta^2} \),

\[
\mathbb{E}_\theta \left[ \frac{\nabla^2 p_\theta}{p_\theta} \right] = \int \frac{\nabla^2 p_\theta}{p_\theta} p_\theta d\mu = \int \nabla^2 p_\theta d\mu = \nabla^2 \int p_\theta d\mu = 0
\]

As a result:

\[
\mathbb{E}_\theta[\nabla^2 \ell_\theta] = -\mathbb{E}_\theta \left[ \left( \frac{\nabla p_\theta}{p_\theta} \right) \left( \frac{\nabla p_\theta}{p_\theta} \right)^T \right] = -\text{Cov}_\theta(\nabla \ell_\theta(x)) = -I_{\theta}
\]

Using what we have shown about the Fisher information, we now have a more compact representation of the asymptotic distribution described in the Theorem above.

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, I_{\theta_0}^{-1} I_{\theta_0} I_{\theta_0}^{-1}) = N(0, I_{\theta_0}^{-1})
\] (3)

Consider \( I_{\theta} = -\nabla^2 \mathbb{E}[\ell_\theta(x)] \). If the magnitude of the second derivative is “large,” that implies that the log-likelihood is steep around the global maximum (making it “easy” to find). Alternatively, if the magnitude of \( -\nabla^2 \mathbb{E}[\ell_\theta(x)] \) is “small,” we do not have sufficient curvature to find the optimal \( \theta \).