IEMS 402: Statistical Learning **Winter 2024-2025**

Lecture 5: Delta Methods and Asymptotic Normality

Lecturer: Yiping Lu *Scribes:*

- `https://web.stanford.edu/class/stats300b/ScribeNotes/2021/lecture-03.pdf`
- `https://web.stanford.edu/class/stats300b/ScribeNotes/2021/lecture-04.pdf`

5.1 "Reduction" for Asymptotics

General set-up: Suppose we know

$$r_n(T_n - \theta) \xrightarrow{d} A.$$

Here, θ is some parameter we want to estimate, T_n is some statistic of our data, r_n is a deterministic rate, and A is some random variable. What can we say about the law of $\phi(T_n) - \phi(\theta)$?

Example 1 (Population loss of estimated parameter) *Consider the following linear regression setting from Lecture 1 where (X_i, Y_i) with $Y_i = \langle X_i, w\rangle + \text{noise}$. If T_n estimates weight vector w from n samples, what is the law of the ℓ_2 loss*

$$f(T_n) = \mathbb{E}[(\langle X_i, T_n\rangle - Y_i)^2]?$$

The delta method lets us understand the law of $f(T_n) - f(\theta)$ purely in terms of the law of $T_n - \theta$ using Taylor Expansion, as long as f is nice enough.

5.2 First Order Delta Method

General set-up: Suppose $T_n, \theta \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at θ; that is, it has Jacobian $f'_\theta \in \mathbb{R}^{k\times d}$ satisfying

$$f(\theta + h) - f(\theta) = f'_\theta h + o(\|h\|) \text{ as } h \to 0$$

Remark: A sufficient condition is that f is continuously differentiable at θ.

Theorem 5.1 (Delta Method) *Let $r_n \to \infty$ and let f be differentiable at θ. If $r_n(T_n - \theta) \xrightarrow{d} A$, then (1) $r_n(f(T_n) - f(\theta)) \xrightarrow{d} f'_\theta A$; and (2) $r_n(f(T_n) - f(\theta)) - f'_\theta(r_n(T_n - \theta)) \xrightarrow{p} 0$*

Proof: By differentiability of f, $f(\theta + h) - f(\theta) = f'_\theta h + o(\|h\|)$ as $h \to 0$. Take $h = T_n - \theta$. Since $r_n(T_n - \theta) \xrightarrow{d} A$ and $r_n \to \infty$, then $T_n - \theta \xrightarrow{p} 0$ by Slutsky's since $1/r_n \to 0$. Swapping in $T_n - \theta$ for h (Lemma **??** applies since $T_n - \theta \xrightarrow{p} 0$) and then multiplying by r_n on both sides, the linear approximation yields

$$r_n(f(T_n) - f(\theta)) = r_n f'_\theta(T_n - \theta) + o_p(r_n\|T_n - \theta\|) \tag{5.1}$$

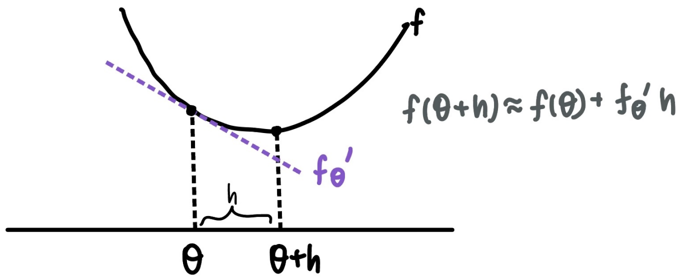

Figure 5.1: In 1 dimension, $f'_\theta h$ is linear approximation to $R(h) = f(\theta + h) - f(\theta)$

Matrix multiplication is continuous, so by Continuous Mapping Theorem, $r_n f'_\theta(T_n - \theta) \xrightarrow{d} f'_\theta(A)$. The sequence $r_n(T_n - \theta)$ is uniformly tight so $o_p(r_n\|T_n - \theta\|) \xrightarrow{p} 0$. Applying Slutsky's, $r_n(f(T_n) - f(\theta)) \xrightarrow{d} f'_\theta A$. To prove the second part, subtracting $r_n f'_\theta(T_n - \theta)$ from both sides of 5.1 gives the desired result:

$$r_n(f(T_n) - f(\theta)) - r_n f'_\theta(T_n - \theta) = o_p(r_n\|T_n - \theta\|) \xrightarrow{p} 0$$

■

Example 2 (Quadratic function) *Let* $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mathcal{D}$ *with* $\mathbb{E}X = \mu$, $\mathrm{Cov}(X) = \Sigma$, *take* $f(x) = \frac{1}{2}x^T Mx$ *for symmetric* M. *Then,*

$$\sqrt{n}(f(\overline{X}_n) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, \mu^T M\Sigma M\mu)$$

Why? By Central Limit Theorem, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. *The derivative of* f *is* $f'_\mu = \mu^T M$. *Then, by Delta Method (Theorem 5.1), applying the linear transformation* f'_μ *gives the desired result.*

Example 3 (Sample variance) *Let* $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mathcal{D}$ *with* $\mathbf{Var}(X_i) = \sigma^2$, $\mathbb{E}X_i^4 = \alpha_4 < \infty$ *and define* $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X}_n)^2$. *Then,* $\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[(X_i - \mu)^4] - \sigma^4)$. *Why? Denoting* $\overline{X^2}_n = \frac{1}{n}\sum_{i=1}^n X_i^2$, *we can define a function* $f(a, b) = b - a^2$ *such that* $S_n^2 = f(\overline{X}_n, \overline{X^2}_n)$. *By CLT,*

$$\sqrt{n}\begin{pmatrix}\overline{X}_n \\ \overline{X^2}_n\end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{bmatrix}\mathbb{E}X \\ \mathbb{E}X^2\end{bmatrix}, \begin{bmatrix}\mathbf{Var}(X) & \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 \\ \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 & \mathbf{Var}(X^2)\end{bmatrix}\Big)$$

Let $\mu = \mathbf{E}[X_i]$, *so* $\mathbf{E}[X^2] = \mu^2 + \sigma^2$. *Computing the first derivative,* $f'_{(a,b)} = [-2a\ 1]$, *and* $f'_{(\mu,\mu^2+\sigma^2)} = [-2\mu\ 1]$. *Then, applying the Delta Method (Theorem 5.1),*

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}\Big(0, [-2\mu\ 1]\begin{bmatrix}\mathbf{Var}(X) & \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 \\ \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 & \mathbf{Var}(X^2)\end{bmatrix}\begin{bmatrix}-2\mu \\ 1\end{bmatrix}\Big)$$

Expanding the computation for asymptotic variance gives

$$4\mu^2\sigma^2 - 4\mu(\mathbb{E}X^3 - \mu\mathbb{E}X^2) + \mathbb{E}X^4 - (\mathbb{E}X^2)^2 = 6\mu^2\sigma^2 + 3\mu^4 - 4\mu\mathbb{E}X^3 + \mathbb{E}X^4 - \sigma^4$$

which is exactly $\mathbb{E}(X - \mu)^4 - \sigma^4$. *Note that the derivation can be simplified by applying delta method to the centered variable* $Y_i = X_i - \mathbb{E}X$.

See this week's problem set (PS 1) for more examples.

5.3 Higher Order Delta Method

Sometimes the first-order Taylor expansion may not be informative, for example if $f'_\theta = 0$. In this case, $r_n(f(T_n) - f(\theta)) \xrightarrow{p} 0$, but we might want to also know the law of the fluctuations.

If f is twice continuously differentiable, $f(\theta + h) - f(\theta) = f'_\theta h + \frac{1}{2} f''_\theta (h \otimes h) + o(\|h\|^2)$ for $h \to 0$, where $f''_\theta : \mathbb{R}^{d^2} \to \mathbb{R}^k$. Here $\otimes$ denotes the tensor product so for $a \in \mathbb{R}^d, b \in \mathbb{R}^k$, $a \otimes b \in \mathbb{R}^{d \times k}$ and $(a \otimes b)_{ij} = a_i b_j$. Based on this heuristic, we have the following corollary to the first-order Delta Method (Theorem 5.1):

Corollary 5.2 (2nd order Delta Method) *If $r_n \to \infty$ and f is twice continuously differentiable at θ, $f'_\theta = 0$, and $r_n(T_n - \theta) \xrightarrow{d} A$, then*

$$r_n^2(f(T_n) - f(\theta)) \xrightarrow{d} \frac{1}{2} f''_\theta A \otimes A$$

Remark: Before proving the corollary, we first note that if $k = 1$, then $\frac{1}{2} f''_\theta A \otimes A = \frac{1}{2} A^T \nabla^2 f_\theta A$ (which we recognize as the Hessian) so the corollary gives us the following limiting law

$$r_n^2(f(T_n) - f(\theta)) \xrightarrow{d} \frac{1}{2} A^T \nabla^2 f_\theta A$$

Proof: The proof is very similar to that of Theorem 5.1, except we now use the second order Taylor expansion of f:

$$r_n^2(f(T_n) - f(\theta)) = \underbrace{r_n^2 \frac{1}{2} f''_\theta((\theta - T_n) \otimes (\theta - T_n))}_{I} + \underbrace{r_n^2 o(\|T_n - \theta\|^2)}_{II}$$

where the first order term vanishes since we assume $f'_\theta = 0$. By CMT, the first term I converges in distribution to $\frac{1}{2} f''_\theta A \otimes A$ since matrix multiplication is continuous. As $r_n^2 \|T_n - \theta\|^2$ is uniformly tight, the second term converges in probability to 0. Summing the two terms together via Slutsky's Theorem, we get the desired conclusion. ∎

Remark: By induction, this argument extends so we can derive other higher order Delta Methods. The generalization follows by taking higher-order Taylor expansions:

$$f(\theta + h) - f(\theta) = f'_\theta h + \frac{1}{2} f''_\theta (h \otimes h) + \frac{1}{6} f'''_\theta (h \otimes h \otimes h) + \dots \tag{5.2}$$

For any integer k, if f is k-times continuously differentiable at θ, and if $f_\theta^{(j)} = 0$ for $j < k$, then:

$$r_n^k(f(T_n) - f(\theta)) \xrightarrow{d} \frac{1}{k!} f_\theta^{(k)} (A \otimes \cdots \otimes A) \tag{5.3}$$

where $f_\theta^{(k)}$ is the k-th derivative tensor.

Example: Relative entropy and log likelihood

First, we define relative entropy (also referred to as Kullback-Liebler divergence, KL divergence).

Definition 5.3 *The **relative entropy** (aka Kullback-Liebler divergence) between distributions $\mathcal{P}$ and $\mathcal{Q}$ is defined as*

$$D(\mathcal{P} \parallel \mathcal{Q}) = \int d\mathcal{P} \log \frac{d\mathcal{P}}{d\mathcal{Q}} = \int p \log \frac{p}{q} \, d\mu$$

where p, q are defined as the densities corresponding to $\mathcal{P}, \mathcal{Q}$.

Example 4 (iid Bernoulli) *Let* $X_i \overset{i.i.d.}{\sim} \text{Bern}(\mu)$, $\mu \in (0,1)$. *Then, we claim*

$$n \cdot D(\text{Bern}(\overline{X}_n) \,\|\, \text{Bern}(\mu)) \xrightarrow{d} \frac{1}{2}Z^2 \text{ and}$$
$$n \cdot D(\text{Bern}(\mu) \,\|\, \text{Bern}(\overline{X}_n)) \xrightarrow{d} \frac{1}{2}Z^2$$

for $Z \sim \mathcal{N}(0,1)$ *and where the notation* $\text{Bern}(p)$ *is used to refer to the Bernoulli distribution with parameter* p. *First, we note that by Central Limit Theorem,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \sqrt{\mu(1-\mu)}Z$$

The above claim then follows from a direct application of the second order Delta Method (Corollary 5.2). Let $h(x,y) = D(\text{Bern}(x) \,\|\, \text{Bern}(y)) = x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ *calculated based on Definition 5.3. In this case, since* $k = 1$, *we are working with the Hessian so it suffices to compute first and second order partial derivatives:*

$$\frac{d}{dx}h(x,y) = \log\frac{x}{y} - \log\frac{1-x}{1-y} \qquad \left(\frac{d}{dx}\right)^2 h(x,y) = \frac{1}{x(1-x)}$$
$$\frac{d}{dy}h(x,y) = \frac{y-x}{y(1-y)} \qquad \left(\frac{d}{dy}\right)^2 h(x,y) = \frac{y^2-2xy+x}{y^2(1-y)^2}$$

Evaluating at $x = y = \mu$,

$$\frac{d}{dx}h(\mu,\mu) = \frac{d}{dy}h(\mu,\mu) = 0, \left(\frac{d}{dx}\right)^2 h(\mu,\mu) = \left(\frac{d}{dy}\right)^2 h(\mu,\mu) = \frac{1}{\mu(1-\mu)}$$

Applying Corollary 5.2, we conclude

$$n \cdot D(\text{Bern}(\overline{X}_n) \,\|\, \text{Bern}(\mu)) = n \cdot h(\overline{X}_n, \mu) \xrightarrow{d} \frac{1}{2}Z^2 \text{ and}$$
$$n \cdot D(\text{Bern}(\mu) \,\|\, \text{Bern}(\overline{X}_n)) = n \cdot h(\mu, \overline{X}_n) \xrightarrow{d} \frac{1}{2}Z^2$$

where the second line follows since the second order partial derivatives match, so we have symmetry in this particular example.

5.4 Asymptotic Normality

Often, we might be concerned with the limiting law of the MLE $\hat{\theta}_n$ (or more generally M- or Z- estimators). For example, a limiting distribution is useful for finding confidence intervals. In this section, we'll prove that under certain regularity conditions, that the limiting law of the MLE is normal. In particular, a model $\{P_\theta\}_{\theta\in\Theta}$ that is *"smooth/nice"* (as defined below) sets the stage nicely for asymptotic normality of the MLE.

Definition 5.4 *A model* $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^d}$ *with densities* p_θ *that is "smooth/nice" at* θ^* *if*

1. *The Hessian of the log-likelihood (from now on referred to simply as "the Hessian") exists and is Lipschitz near* θ^*: *i.e. there exists* $\varepsilon > 0$ *such that for all* θ_1, θ_2 *such that* $\|\theta^* - \theta_i\| \leqslant \varepsilon$,

$$\|\nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_2}(x)\|_{op} < M(x)\|\theta_1 - \theta_2\|_2,$$

where $\nabla^2 \ell_\theta = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta\right]_{i,j=1}^d \in \mathbb{R}^{d\times d} := \ddot{\ell}_\theta$ *is the Hessian and where* $P_{\theta^*}[M(X)] < \infty$, *and*

2. *The gradient is bounded in the sense $P_{\theta^*}\|\nabla \ell_{\theta^*}\|^2 < \infty$.*

Theorem 5.5 *Suppose $\{P_\theta\}_{\theta\in\Theta}$ is nice/smooth at θ^* and Θ is an open subset of $\mathbb{R}^d$. Suppose also that the Hessian has finite mean (or alternatively, that we can exchange the order of differentiation wrt θ and expectation X). Let $X_1, .., X_n \overset{i.i.d.}{\sim} P_{\theta^*}$, $\hat{\theta}_n = \underset{\theta\in\Theta}{arg\,max} P_n \ell_\theta$, and $\hat{\theta}_n$ is consistent (i.e. $\hat{\theta}_n \overset{p}{\to} \theta^*$). Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \overset{d}{\to} \mathcal{N}(0, \Sigma_{\theta^*}),$$

where

$$\Sigma_{\theta^*} := (P_{\theta^*}[\nabla^2 \ell_{\theta^*}])^{-1} \mathbf{Cov}_{\theta^*}(\nabla \ell_{\theta^*})(P_{\theta^*}[\nabla^2 \ell_{\theta^*}])^{-1} \text{ (More on this quantity next class...)}$$

Proof: Before diving in, let's first make a few observations:

- The existence of gradients and Hessians in the limiting distribution's variance (and the fact that we're working with asymptotics) indicates that Taylor expansions probably play a role in this proof.
- Since $P_n \nabla \ell_{\theta^*} = \frac{1}{n}\sum_{i=1}^n \nabla \ell_{\theta^*}(X_i)$, where $\nabla \ell_{\theta^*}(X_i)$ are IID with $P_{\theta^*}\nabla \ell_{\theta^*} = \mathbf{0}$ (by optimality of θ^*) and $P_{\theta^*}\|\nabla \ell_{\theta^*}\|_2^2 < \infty$ (from nice/smoothness criterion), we can apply CLT, which tells us

$$\sqrt{n}(P_n \nabla \ell_{\theta^*} - P\nabla \ell_{\theta^*}) \overset{d}{\to} \mathcal{N}(0, \mathbf{Cov}(\nabla \ell_{\theta^*})).$$

- By the definition of MLE, $\hat{\theta}_n$ satisfies $P_n \ell_{\hat{\theta}_n} = 0$.

Now, we want to use these facts/observations to show asymptotic normality. First, use a (0th order) Taylor expansion (i.e. the definition of differentiability for $\nabla \ell_{\theta_0}$) for any fixed $\theta_0 \in \Theta$ and $x \in \mathbb{R}$:

$$\nabla \ell_{\theta_0}(x) = \nabla \ell_{\theta^*} + \nabla^2 \ell_{\theta^*}(\theta_0 - \theta^*) + R(\theta_0 - \theta^*),$$

where the remainder term represents the error in the Hessian. Note that the mean-value theorem doesn't apply when the function is vector-valued, but can still be upperbounded (see Remark 1).

Averaging with respect to the empirical CDF and plugging in $\theta_0 = \hat{\theta}_n$ gives

$$\underbrace{P_n \nabla \ell_{\hat{\theta}_n}}_{=0 \text{ by optimality}} = \underbrace{P_n \nabla \ell_{\theta^*}}_{\text{normal after scaling}} + \underbrace{P_n \nabla^2 \ell_{\theta^*}}_{(A)} \underbrace{(\hat{\theta}_n - \theta^*)}_{\text{main object}} + \underbrace{P_n R(\hat{\theta}_n - \theta^*)}_{(B)}.$$

We now want to find limiting behaviors of terms (A) and (B).

(A). Since the mean of the Hessian is finite, by LLN, $P_n \nabla^2 \ell_{\theta^*} \overset{p}{\to} P_{\theta^*}\nabla^2 \ell_{\theta^*}$, so $P_n \nabla^2 \ell_{\theta^*} = P_{\theta^*}\nabla^2 \ell_{\theta^*} + o_p(1)$.

(B). $\|P_n R\|_{op} \leqslant P_n \|R\|_{op} \leqslant (P_n M)\|\hat{\theta}_n - \theta^*\|_2$, where the first inequality follows from triangle inequality and the second from the bound in Remark 1 and the niceness of the model. $P_n M \overset{a.s.}{\to} P_{\theta^*} M$ by SLLN and $\|\hat{\theta}_n - \theta^*\|_2 \overset{p}{\to} 0$ by consistency. Thus, by Slutsky's lemma, $(P_n M)\|\hat{\theta}_n - \theta^*\|_2 \overset{p}{\to} 0$. Therefore, $\|P_n R\|_{op} \overset{p}{\to} 0$, which implies $P_n R \overset{p}{\to} 0$, i.e. $P_n R = o_p(1)$.

Putting this all together and combining the last two terms in the Taylor expansion, we get

$$0 = P_n \nabla \ell_{\theta^*} + (P_{\theta^*}\nabla^2 \ell_{\theta^*} + o_p(1))(\hat{\theta}_n - \theta^*).$$

Rearranging and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = (P_{\theta^*}\nabla^2 \ell_{\theta^*} + o_p(1))^{-1} \underbrace{(-\sqrt{n}P_n \ell_{\theta^*})}_{\xrightarrow{d} \mathcal{N}(0, \mathbf{Cov}(\nabla \ell_{\theta^*}))} \xrightarrow{d} \mathcal{N}(0, \Sigma_{\theta^*}),$$

where the last line follows from Slutsky's Lemma. ■

Remark 1 (Taylor expansion bounds on the remainder of a vector valued function) *Suppose that* $f : \mathbb{R}^d \to \mathbb{R}^k$*, differentiable and* ∇f *is* M*-Lipschitz. Then, we have the following Taylor expansion with remainder term:*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + R(\boldsymbol{y} - \boldsymbol{x}),$$

where $\|R\|_{op} \leqslant M\|\boldsymbol{y} - \boldsymbol{x}\|_2$*.* **Proof:** *Define* $\phi_i(t) = f_i((1-t)x + ty)$*,* $\phi_i : [0,1] \to \mathbb{R}$*, and observe the following properties of* ϕ_i*: (1)* $\phi_i(0) = f_i(x)$ *and* $\phi_i(1) = f_i(y)$ *and (2)* $\phi_i' = (\nabla f_i((1-t)x + ty))^T(y - x)$*. Then*

$$Df((1-t)x + ty)(y - x) = \begin{bmatrix} \nabla^T f_1((1-t)x + ty) \\ \nabla^T f_2((1-t)x + ty) \\ \vdots \\ \nabla^T f_k((1-t)x + ty) \end{bmatrix} (y - x) = \begin{bmatrix} \phi_1'(t) \\ \phi_2'(t) \\ \vdots \\ \phi_k'(t) \end{bmatrix}.$$

Then, by the fundamental theorem of calculus (FTOC),

$$\begin{aligned} f(y) - f(x) - Df(x)(y - x) &= \int_0^1 Df((1-t)x + ty)(y - x)dt - Df(x)(y - x) \\ &= \int_0^1 (Df((1-t)x + ty) - Df(x))dt(y - x). \end{aligned}$$

Then, we can express the remainder term as

$$R = \int_0^1 (Df((1-t)x + ty) - Df(x))dt.$$

Now, for any $u \in S^{d-1}$*,*

$$\begin{aligned} \|Ru\|_2 &= \left\| \int_0^1 (Df((1-t)x + ty) - Df(x))udt \right\| \\ &\leqslant \int_0^1 \|Df((1-t)x + ty) - Df(x)\|_{op} \|u\|_2 dt \\ &\leqslant \int_0^1 M\|(1-t)x + ty - x\|dt \\ &= \frac{M}{2}\|y - x\|_2. \end{aligned}$$

Thus, taking the supremum over all such $\boldsymbol{u}$ *gives* $\|R\|_{op} \leqslant \frac{M}{2}\|y - x\|_2$*.*[1] ■

[1]Thanks to Etaash Katiyar for helping work this out!