11/4/2021

Outline

1) Maximum Likelihood Estimator
2) Asymptotic Distribution of MLE
3) Consistency of MLE

Maximum Likelihood Estimation

For a generic dominated family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with densities p_θ, a simple estimator for θ is

$$\hat{\theta}_{MLE}(X) = \underset{\theta \in \Theta}{\arg\max}\ p_\theta(x)$$
$$= \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

Remark 1: argmax may not exist, be unique, or be computable

Remark 2: doesn't depend on parameterization or base measure, MLE for $g(\theta)$ is $g(\hat{\theta}_{MLE})$

Ex $p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$

$$\ell(\eta; x) = \eta' T(x) - A(\eta) + \log h(x)$$

$$\nabla \ell(\eta; X) = T(x) - \mathbb{E}_\eta T(X)$$

$\Rightarrow \hat{\eta}_{MLE}$ solves $T = \mathbb{E}_{\hat{\eta}} T$ if such η exists

Because $\nabla^2 \ell(\eta; X) = -\mathrm{Var}_\eta(T)$ is negative definite unless $\nu' T \overset{a.s.}{=} 0$ (in which case param. redundant)

$\Rightarrow$ at most 1 solution exists

Let $\mu = \psi(\eta) = \nabla A(\eta)$, $\hat{\eta} = \psi^{-1}(T)$

Ex $\quad X_i \overset{iid}{\sim} e^{\eta T(x) - A(\eta)} h(x) \qquad \eta \in \Xi \subseteq \mathbb{R}$

$$\hat{\eta} = \psi^{-1}(\bar{T}), \quad \bar{T} = \frac{1}{n}\sum T(X_i)$$

Assume $\eta \in \Xi^\circ$. $\quad \dot{\psi}(\eta) = \ddot{A}(\eta) > 0 \quad \forall \eta \in \Xi^\circ$

so ψ^{-1} cts, $\quad (\dot{\psi^{-1}})(\mu) = \dfrac{1}{\dot{\psi}(\psi(\mu))} = \dfrac{1}{\ddot{A}(\eta)}$

Consistency: $\quad \bar{T} \xrightarrow{P_\eta} \mu$

Cts mapping: $\quad \psi^{-1}(\bar{T}) \xrightarrow{P_\eta} \psi^{-1}(\mu) = \eta$

Since $\quad \sqrt{n}(\bar{T} - \mu) \Rightarrow N(0, \mathrm{Var}_\eta(T(X_1)))$

$$= N(0, \ddot{A}(\eta))$$

(Recall $J_1(\mu) = \mathrm{Var}(T)^{-1} = \ddot{A}(\eta)^{-1}$)

Delta method:

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(\psi^{-1}(\bar{T}) - \eta)$$

$$\Rightarrow N\left(0, \frac{1}{\ddot{A}(\eta)}^2 \cdot \ddot{A}(\eta)\right)$$

$$= N\left(0, \frac{1}{\ddot{A}(\eta)}\right)$$

Recall $J_1(\eta) = \mathrm{Var}_\eta(T(X_i)) = \ddot{A}(\eta)$

= Fisher info from 1 obs

$$\hat{\eta} \approx N\left(\eta, \frac{1}{nJ_1(\eta)}\right)$$

Asymptotically unbiased, Gaussian, achieves CRLB

($\mathrm{corr}(\bar{T}, \hat{\eta}) \to 1$)

Ex $X_1, \ldots, X_n \overset{iid}{\sim} \text{Pois}(\theta)$, $\eta = \log \theta$

$\hat{\eta} = \log \bar{X}$, $\sqrt{n}(\bar{X} - \theta) \Rightarrow N(0, \theta)$

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(\log \bar{X} - \log \theta)$$

$$\Rightarrow N\left(0, \theta \cdot \frac{1}{\theta^2}\right) \quad \text{(Delta method)}$$

$$= N(0, \theta^{-1})$$

But $\forall$ finite n, $\forall \theta > 0$:

$$\mathbb{P}_\theta(\hat{\eta} = -\infty) = \mathbb{P}_\theta(X_1 = 0)^n$$

$$= e^{-\theta n} > 0$$

$$\Rightarrow \mathbb{E}\hat{\eta} = -\infty \qquad \text{Var}(\hat{\eta}) = \infty$$

[MLE can have embarassing finite-sample performance despite being asy. optimal!]

Prop: If $\mathbb{P}(B_n) \to 0$, $X_n \Rightarrow X$, Z_n arbitrary

then $X_n 1_{B_n^c} + Z_n 1_{B_n} \Rightarrow X$

Proof $\mathbb{P}(\|Z_n 1_{B_n}\| > \varepsilon) \le \mathbb{P}(B_n) \to 0$ so $Z_n 1_{B_n} \xrightarrow{p} 0$

Also $1_{B_n^c} \xrightarrow{p} 1$, apply Slutsky $\boxtimes$

[So zany behavior has no effect on cvg. in dist]

Asymptotic Efficiency

[The nice behavior of MLE we found in the exponential family case generalizes to a much broader class of models]

Setting $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta(x) \qquad \theta \in \Theta \subseteq \mathbb{R}^d$

p_θ "smooth" in θ, e.g. 2 cts integrable derivs
(can be relaxed)

Let $\ell_1(\theta; X_i) = \log p_\theta(X_i)$, $\ell_n(\theta; X) = \sum_{i=1}^n \ell_1(\theta; X_i)$

$$J_1(\theta) = \mathrm{Var}_\theta(\nabla \ell_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta; X_i)]$$

$$J_n(\theta) = \mathrm{Var}_\theta(\nabla \ell_n(\theta; X)) = n J_1(\theta)$$

We say an estimator $\hat{\theta}_n$ is <u>asymptotically efficient</u> if

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{P_\theta}{\Rightarrow} N(0, J_1(\theta)^{-1})$$

Delta method for differentiable estimand $g(\theta)$ $\quad (g : \Theta \to \mathbb{R})$

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \overset{P_\theta}{\Rightarrow} N(0, \nabla g(\theta)' J_1(\theta)^{-1} \nabla g(\theta))$$

also achieves CRLB if $\hat{\theta}_n$ does; g diff.

Asymptotic Dist. of MLE

Under mild conditions, $\hat{\theta}_{MLE}$ is asy. Gaussian, efficient

We will be interested in $\hat{\ell}(\theta; X)$ as a function of θ

Notate "true" value as θ_0 $(X \sim P_{\theta_0})$

Derivatives of ℓ_n at θ_0: $(\theta_0 \in \Theta^\circ)$

$$\nabla \ell_1(\theta_0; X_i) \overset{iid}{\sim} (0, J_1(\theta_0))$$

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) = \sqrt{n} \cdot \frac{1}{n} \sum \nabla \ell_1(\theta_0; X_i) \overset{P_{\theta_0}}{\Rightarrow} N(0, J_1(\theta_0))$$

$$\frac{1}{n} \nabla^2 \ell_n(\theta_0; X) \xrightarrow{P_{\theta_0}} \mathbb{E}_{\theta_0} \nabla^2 \ell_1(\theta_0; X_i) = -J_1(\theta_0)$$

Informal Proof:

between $\theta_0, \hat{\theta}_n$

$$0 = \nabla \ell_n(\hat{\theta}_n; X) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\underbrace{\left(\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1}}_{\text{(Want)} \xrightarrow{p} J(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)}_{\Rightarrow N_d(0, J(\theta_0))}$$

$$\Rightarrow N_d(0, J(\theta_0)^{-1})$$

More rigorous proof later, but note we need consistency of $\hat{\theta}_n$ first to even justify Taylor expansion

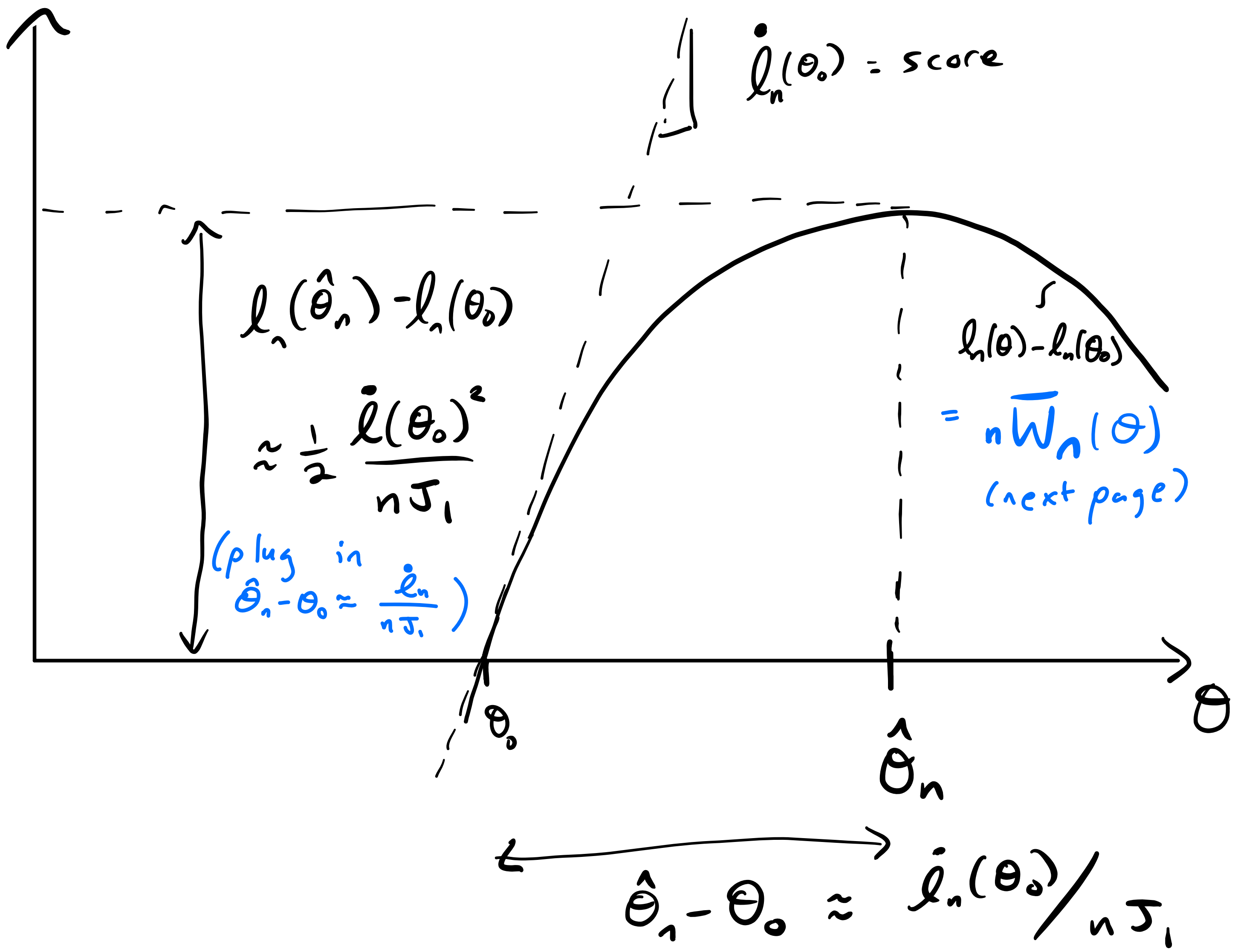
Asymptotic Picture $(d=1)$

Recall $\left(\ell_n(\theta) - \ell_n(\theta_0)\right)_{\theta \in \Theta}$ is minimal suff.

Quadratic approximation near θ_0:

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \underbrace{\dot{\ell}_n(\theta_0)}_{\approx N(0,\, n J_1(\theta_0))} (\theta - \theta_0) + \frac{1}{2} \underbrace{\ddot{\ell}_n(\theta_0)}_{\approx -n J_1(\theta_0)} (\theta - \theta_0)^2$$

Gaussian linear term   Deterministic curvature



Consistency of MLE

$X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0}$, $\quad \hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \ell_n(\theta; X)$

[Will be OK if $\hat{\theta}_n$ comes close to maximizing ℓ_n]

Question: When does $\hat{\theta}_n \xrightarrow{p} \theta_0$?

Assume model identifiable ($P_\theta \neq P_{\theta_0}$ for $\theta \neq \theta_0$)

Recall **KL Divergence**:

$$D_{KL}(\theta_0 \| \theta) = \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$$

$$-D_{KL}(\theta_0 \| \theta) \leq \log \mathbb{E}_{\theta_0} \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \quad \text{(note switch)}$$

$$= \log \int_{x : p_{\theta_0}(x) > 0} \frac{p_\theta(x)}{\cancel{p_{\theta_0}(x)}} \cancel{p_{\theta_0}(x)} \, d\mu(x)$$

$$\leq \log 1 = 0$$

(Jensen)

strict ineq unless $\frac{p_\theta}{p_{\theta_0}}$ const. (i.e., unless $P_\theta = P_{\theta_0}$)

Let $W_i(\theta) = \ell_1(\theta; X_i) - \ell(\theta_0; X_i)$, $\overline{W}_n = \frac{1}{n} \sum W_i$

Note $\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \overline{W}_n(\theta)$ too

$$\overline{W}_n(\theta) \xrightarrow{p} \mathbb{E}_{\theta_0} W_i(\theta)$$
$$= -D_{KL}(\theta_0 \,\|\, \theta)$$
$$\leq 0, \quad \text{equality iff } \theta = \theta_0$$

<u>But</u> not enough:

- MLE $\hat{\theta}_n$ depends on entire function $\overline{W}_n(\cdot)$
- need uniform convergence in θ

<u>Def</u> For compact K let $C(K) = \{ f : K \to \mathbb{R},\ \text{cts} \}$

For $f \in C(K)$ let $\|f\|_\infty = \sup_{t \in K} |f(t)|$

$f_n \underset{(\xrightarrow{p})}{\to} f$ in this norm if $\|f_n - f\|_\infty \underset{(\xrightarrow{p})}{\to} 0$

<u>Thm</u> (LLN for random functions)

Assume K compact, $W_1, W_2, \ldots \in C(K)$ iid.

$\mathbb{E}\|W_i\|_\infty < \infty$, $\quad \mu(t) = \mathbb{E} W_i(t)$

Then $\mu(t) \in C(K)$

and $\mathbb{P}\left(\| \frac{1}{n} \Sigma W_i - \mu \|_\infty > \varepsilon \right) \to 0$

(i.e., $\overline{W}_n \xrightarrow{p} \mu$ in $\|\cdot\|_\infty$, or $\|\overline{W}_n - \mu\|_\infty \xrightarrow{p} 0$)

Theorem (Keener 9.4):

Let $G_1, G_2, \ldots$ random functions in $C(K)$, K cpt.

$\|G_n - g\|_\infty \xrightarrow{P} 0$, some fixed $g \in C(K)$. Then

① If $t_n \xrightarrow{P} t^* \in K$ (t^* fixed) then $G_n(t_n) \xrightarrow{P} g(t^*)$

② If g maximized at unique value t^*,

and $G_n(t_n) = \max G_n(t)$ then $t_n \xrightarrow{P} t^*$

$G_n(t_n) \geq \max G_n - \alpha_n$, $\alpha_n \to 0$ (mod.s of proof in purple)

③ If $K \subseteq \mathbb{R}$, $g(t) = 0$ has unique sol. t^*,

and t_n solve $G_n(t_n) = 0$ then $t_n \xrightarrow{P} t^*$

$|G_n(t_n)| \leq \alpha_n$, $\alpha_n \to 0$

Note we need ① for $\tilde{\theta}_n$ from MVT in Taylor expansion

② for consistency

Proof

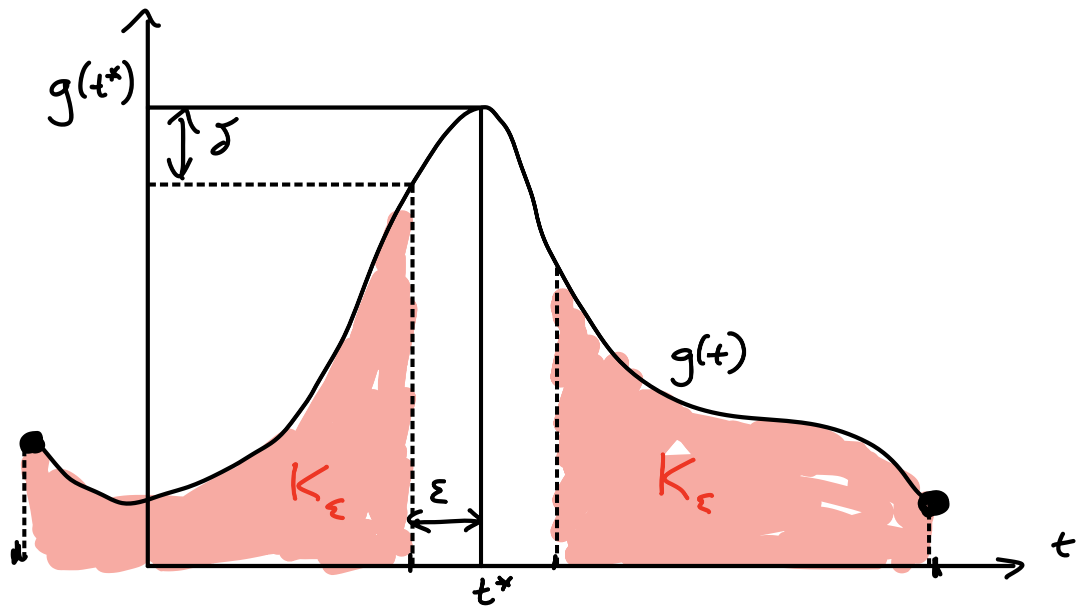
① $|G_n(t_n) - g(t^*)| \leq |G_n(t_n) - g(t_n)| + |g(t_n) - g(t^*)|$

$$\leq \underbrace{\|G_n - g\|_\infty}_{\xrightarrow{P} 0 \text{ (by assumption)}} + \underbrace{|g(t_n) - g(t^*)|}_{\xrightarrow{P} 0 \text{ (by cts mapping)}}$$

② Fix $\varepsilon > 0$, let $B_\varepsilon(t^*) = \{t : \|t - t^*\| < \varepsilon\}$

Let $K_\varepsilon = K \setminus B_\varepsilon(t^*) = K \cap B_\varepsilon^c(t^*)$ (compact)

$\delta = g(t^*) - \max_{t \in K_\varepsilon} g(t) > 0$



If $t_n \in K_\varepsilon$ then $G_n(t_n) \leq \underbrace{g(t^*) - \delta}_{> \max_{K_\varepsilon} g(t)} + \|G_n - g\|_\infty$

and $G_n(t_n) \geq G_n(t^*) - \alpha_n \geq g(t^*) - \|G_n - g\|_\infty - \alpha_n$

then $2\|G_n - g\|_\infty \geq \delta - \alpha_n$

$$\mathbb{P}(\|t_n - t^*\| \geq \varepsilon) \leq \mathbb{P}\left(\|G_n - g\|_\infty \geq \frac{\delta - \alpha_n}{2}\right) \to 0$$

③ Analogous to ②

Theorem (Consistency of MLE for compact Θ)

$X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0}$, $\mathcal{P}$ has densities p_θ, $\theta \in \Theta$

Assume
- p_θ cts in θ
- Θ compact
- $\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \Theta} |W_i(\theta)|\right] < \infty$ $= \mathbb{E}_{\theta_0} \sup_\theta |\ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)|$
- Model identifiable

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$ if $\hat{\theta}_n \in \operatorname{argmax} \ell_n(\theta; X)$

Proof $W_i \in C(\Theta)$ iid, mean $\mu(\theta) = -D_{KL}(\theta_0 \| \theta)$

$\mu(\theta_0) = 0$, $\mu(\theta) < 0$ $\forall \theta \neq \theta_0$ $(\theta_0 = \operatorname{argmin} \mu)$

By definition, $\hat{\theta}_n$ maximizes $\overline{W}_n$,

$\|\overline{W}_n - \mu\|_\infty \xrightarrow{p} 0$, apply 9.4, ②

We usually care about non-compact parameter spaces, need some extra assumption to get us there.

<u>Thm</u> ($\approx$ Keener 9.11, but stronger conditions)

$X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0}$, $\mathcal{P}$ has cts densities p_θ, $\theta \in \Theta = \mathbb{R}^d$

Assume
- Model identifiable
- For all compact $K \subseteq \mathbb{R}^d$, $\mathbb{E}\left[\sup_{\theta \in K} |W_i(\theta)|\right] < \infty$
- $\exists\, r > 0$ s.t. $\mathbb{E}\left[\sup_{\|\theta - \theta_0\| \geq r} W_i(\theta)\right] < 0$

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$ if $\hat{\theta}_n \in \operatorname{argmax} \ell_n(\theta; X)$

<u>Proof</u> Let $K = \{\theta : \|\theta - \theta_0\| \leq r\}$, $\beta = \mathbb{E} \sup_{\theta \notin K} W_i(\theta) < 0$

$$\sup_{\theta \notin K} \overline{W}_n(\theta) \leq \frac{1}{n} \sum_{i=1}^n \sup_{\theta \notin K} W_i(\theta) \xrightarrow{p} \beta < 0$$

Hence, $$\mathbb{P}(\hat{\theta}_n \notin K) \leq \mathbb{P}\Big(\underbrace{\overline{W}_n(\theta_0)}_{\xrightarrow{p} 0} < \underbrace{\sup_{\theta \notin K} \overline{W}_n(\theta)}_{\xrightarrow{p} \beta}\Big) \to 0$$

Let $\hat{\theta}_n^K = \sup_{\theta \in K} \overline{W}_n(\theta) \xrightarrow{p} \theta_0$ by prev. theorem (K compact)

$$\hat{\theta}_n = \hat{\theta}_n^K 1\{\hat{\theta}_n \in K\} + \hat{\theta}_n 1\{\hat{\theta}_n \notin K\}$$

$$\xrightarrow{p} \theta_0 \quad \text{since } \mathbb{P}(\hat{\theta}_n \notin K) \to 0$$

Asymptotic Dist. of MLE

Theorem

$X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0}$ for $\theta_0 \in \Theta^\circ \subseteq \mathbb{R}^d$

Assume
- $\hat{\theta}_n \in \underset{\theta \in \Theta}{\arg\max}\, \ell_n(\theta; X)$, $\hat{\theta}_n \xrightarrow{p} \theta_0$
- In a neighborhood $\bar{B}_\varepsilon(\theta_0) = \{\theta : \|\theta - \theta_0\| \le \varepsilon\} \subseteq \Theta^\circ$:

  (i) $\ell_1(\theta; x)$ has 2 cts deriv.s on $\bar{B}_\varepsilon(\theta_0)$, $\forall x$

  (ii) $\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \bar{B}_\varepsilon} \|\nabla^2 \ell_1(\theta; X_i)\|\right] < \infty$

  any norm on $\mathbb{R}^{d \times d}$, e.g. Frobenius

- Fisher info:

  $$\mathbb{E}_{\theta_0} \nabla \ell_1(\theta_0; X_1) = 0$$

  (positive def.) $v' J_1 v > 0 \; \forall v$

  $$\mathrm{Var}_{\theta_0} \nabla \ell_1(\theta_0; X) = -\mathbb{E}_{\theta_0} \nabla^2 \ell_1(\theta_0; X) \succ 0$$

  (enough to have 3^{rd} deriv. of ℓ_1 bdd in $B_\varepsilon(\theta_0)$)

Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, J_1(\theta_0)^{-1})$

Proof $\sup_{\theta \in \bar{B}_\varepsilon} \| \frac{1}{n} \nabla^2 \ell_n(\theta) - J_1(\theta) \| \xrightarrow{P} 0$ since $\bar{B}_\varepsilon$ compact, $\nabla^2 \ell_1$ cts

Let $A_n = \{ \| \hat{\theta}_n - \theta_0 \| > \varepsilon \}$,

$\mathbb{P}_{\theta_0}(A_n) \to 0$ by assumption

On A_n^c, $\hat{\theta}_n \in \bar{B}_\varepsilon(\theta_0)$ and we have

$$
\begin{aligned}
0 &= \nabla \ell_n(\hat{\theta}_n; X) \\
&= \nabla \ell_n(\theta_0; X) + \nabla^2 \ell_n(\tilde{\theta}_n; X)(\hat{\theta}_n - \theta_0),
\end{aligned}
$$

for some $\tilde{\theta}_n$ between θ_0 and $\hat{\theta}_n$ (MVT)

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) = \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n) \right)^{-1}}_{\xrightarrow{P_{\theta_0}} J_1(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)}_{\Rightarrow N(0, J_1(\theta_0))}
$$

By 9.4 ① + cts mapping

$$
\Rightarrow N_d(0, J_1(\theta_0)^{-1})
$$

Behavior on A_n irrelevant to asymptotic limit $\boxtimes$