

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

$$\begin{aligned}\hat{\theta}_{MLE}(x) &= \operatorname{argmax}_{\theta \in \Theta} p_\theta(x) \\ &= \operatorname{argmax}_{\theta \in \Theta} l(\theta; x)\end{aligned}$$

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 $g(\hat{\theta}_{MLE})$

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

$$l(\gamma; x) = \gamma' T(x) - A(\gamma) + \log h(x)$$

$$\nabla l(\gamma; x) = T(x) - \mathbb{E}_\gamma T(x)$$

$$\Rightarrow \hat{\gamma}_{MLE} \text{ solves } T = \mathbb{E}_{\hat{\gamma}} T \quad \text{if such } \gamma \text{ exists}$$

Because $\nabla^2 l(\gamma; x) = -\text{Var}_\gamma(T)$ is negative definite unless $\gamma' T \stackrel{a.s.}{=} 0$ (in which case param. redundant)

\Rightarrow at most 1 solution exists

$$\text{Let } M = \varphi(\gamma) = \nabla A(\gamma), \hat{\gamma} = \varphi^{-1}(T)$$

$$\underline{\text{Ex}} \quad X_i \stackrel{\text{iid}}{\sim} e^{\gamma T(x) - A(\gamma)} h(x) \quad \gamma \in \Xi \subseteq \mathbb{R}$$

$$\hat{\gamma} = \tilde{\psi}(\bar{T}), \quad \bar{T} = \frac{1}{n} \sum T(x_i)$$

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

$$\text{so } \psi^{-1} \text{ cts, } (\dot{\psi}^{-1})(\mu) = \frac{1}{\dot{\psi}(\psi(\mu))} = \frac{1}{\ddot{A}(\gamma)}$$

Consistency: $\bar{T} \xrightarrow{P} \mu$

Cts mapping: $\tilde{\psi}(\bar{T}) \xrightarrow{P} \tilde{\psi}(\mu) = \gamma$

$$\text{Since } \sqrt{n}(\bar{T} - \mu) \Rightarrow N(0, \text{Var}_n(T(x_i))) \\ = N(0, \ddot{A}(\mu))$$

Delta method: (Recall \quad J_1(\mu) = \text{Var}(T)^{-1} = \ddot{A}(\mu)^{-1})

$$\sqrt{n}(\hat{\gamma} - \gamma) = \sqrt{n}(\tilde{\psi}(\bar{T}) - \gamma) \\ \Rightarrow N(0, \frac{1}{\ddot{A}(\mu)^2} \cdot \ddot{A}(\gamma)) \\ = N(0, \frac{1}{\ddot{A}(\mu)})$$

Recall $J_1(\gamma) = \text{Var}_{\gamma}(T(x_i)) = \ddot{A}(\gamma)$
 $= \text{Fisher info from 1 obs}$

$$\hat{\gamma} \approx N(\gamma, \frac{1}{n J_1(\gamma)})$$

Asymptotically unbiased, Gaussian, achieves CRLB
(\text{corr}(\bar{T}, \hat{\gamma}) \rightarrow 1)

Ex $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\theta)$, $\gamma = \log \theta$

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

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

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

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

$$\begin{aligned} P_\theta(\hat{\gamma} = -\infty) &= P_\theta(X_1 = 0)^n \\ &= e^{-\theta n} \rightarrow 0 \end{aligned}$$

$$\Rightarrow E \hat{\gamma} = -\infty \quad \text{Var}(\hat{\gamma}) = \infty$$

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

Prop: If $P(B_n) \rightarrow 0$, $X_n \Rightarrow X$, Z_n arbitrary
then $X_n 1_{B_n^c} + Z_n 1_{B_n} \Rightarrow X$

Proof $P(\|Z_n 1_{B_n}\| > \varepsilon) \leq P(B_n) \rightarrow 0$ so $Z_n 1_{B_n} \xrightarrow{P} 0$

Also $1_{B_n^c} \xrightarrow{P} 1$, apply Slutsky \square

[So zany behavior has no effect on cug. 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 $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta(x) \quad \theta \in \Theta \subseteq \mathbb{R}^d$

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

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

$$J_1(\theta) = \text{Var}_\theta(\nabla l_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 l_1(\theta; X_i)]$$

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

We say an estimator $\hat{\theta}_n$ is asymptotically efficient

$$\text{if } \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{P_\theta} N(0, J_1(\theta)^{-1}) \quad (g: \Theta \rightarrow \mathbb{R})$$

Delta method for differentiable estimand $g(\theta)$

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

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 $\ell(\theta; X)$ as a function of θ

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

Derivatives of ℓ_n at θ_0 : ($\theta_0 \in \Theta^0$)

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

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

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

Informal Proof:

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

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

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

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

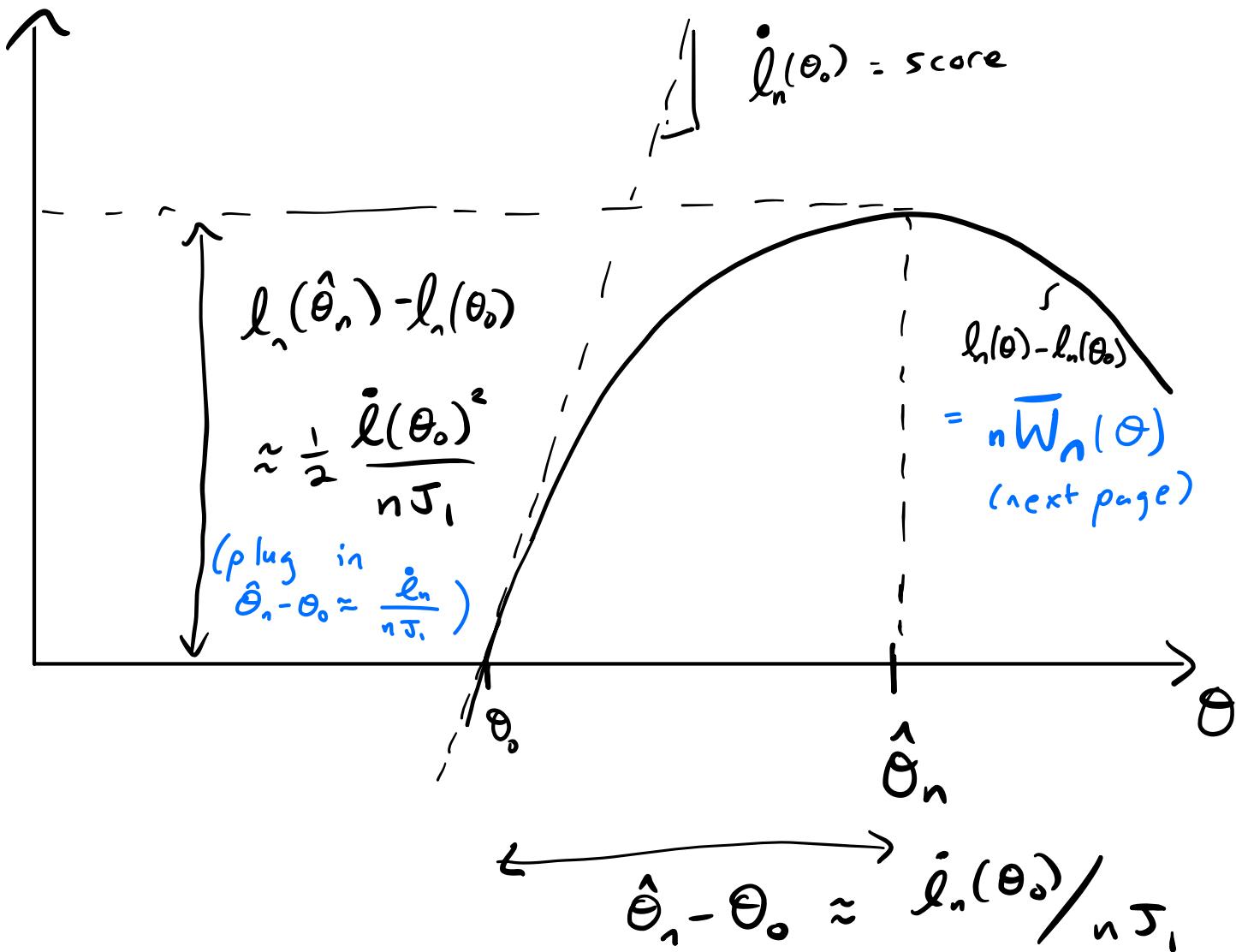
Asymptotic Picture ($d=1$)

Recall $(l_n(\theta) - l_n(\theta_0))_{\theta \in \mathbb{U}}$ is minimal suff.

Quadratic approximation near θ_0 :

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

Gaussian linear term Deterministic curvature



Consistency of MLE

$$X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta_0}, \quad \hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} l_n(\theta; X)$$

[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:

$$\begin{aligned} D_{KL}(\theta_0 \parallel \theta) &= E_{\theta_0} \log \frac{P_{\theta_0}(X_i)}{P_\theta(X_i)} \\ -D_{KL}(\theta_0 \parallel \theta) &\leq \log E_{\theta_0} \frac{P_\theta(X_i)}{P_{\theta_0}(X_i)} \quad \leftarrow \text{(note switch)} \\ &= \log \int_{x: P_{\theta_0}(x) > 0} \frac{P_\theta(x)}{P_{\theta_0}(x)} P_{\theta_0}(x) d\mu(x) \\ &\leq \log 1 = 0 \end{aligned}$$

(Jensen)

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

Let $W_i(\theta) = l_i(\theta; X_i) - l(\theta_0; X_i)$, $\bar{W}_n = \frac{1}{n} \sum W_i$

Note $\hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} \bar{W}_n(\theta)$ too

$$\begin{aligned}
 \bar{W}_n(\theta) &\xrightarrow{P} \mathbb{E}_{\theta_0} W_i(\theta) \\
 &= -D_{KL}(\theta_0 \parallel \theta) \\
 &\leq 0, \text{ equality iff } \theta = \theta_0
 \end{aligned}$$

But not enough:

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

Def For compact K let $C(K) = \{f: K \rightarrow \mathbb{R}, \text{cts}\}$

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

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

Thm (LLN for random functions)

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

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

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

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

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

Theorem (Keener 9.4):

Let G_1, G_2, \dots 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 $t_n \xrightarrow{P} t^*$
 $G_n(t_n) \geq \max G_n - \alpha_n, \alpha_n \rightarrow 0$ (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 $t_n \xrightarrow{P} t^*$

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

Note we need ① for $\hat{\Theta}_n$ from MVT in Taylor expansion
 ② for consistency

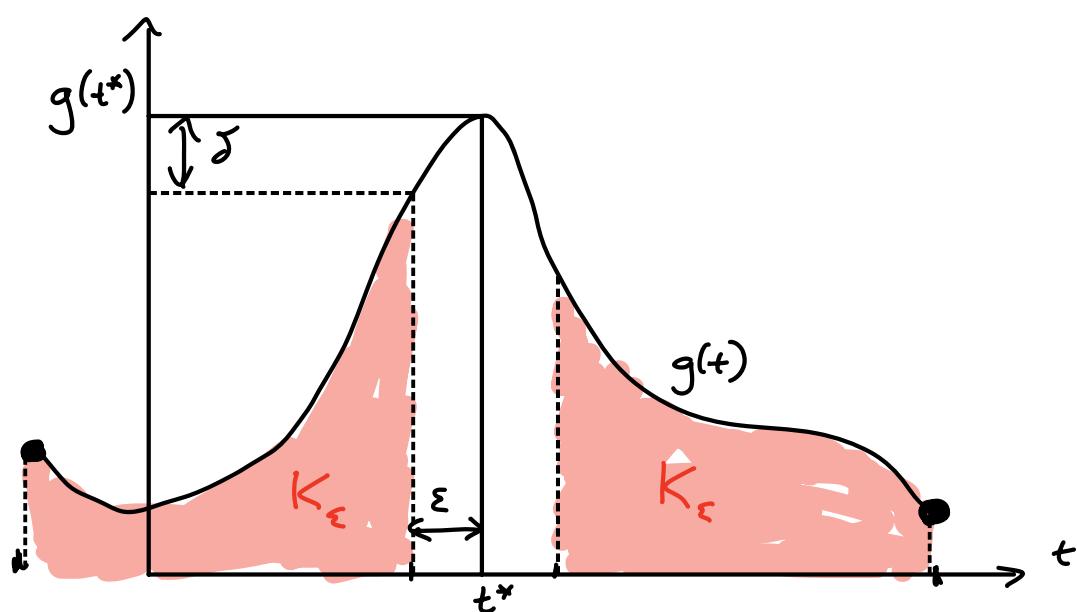
Proof

$$\begin{aligned} ① |G_n(t_n) - g(t^*)| &\leq |G_n(t_n) - g(t_n)| + |g(t_n) - g(t^*)| \\ &\leq \|G_n - g\|_\infty + |g(t_n) - g(t^*)| \\ &\xrightarrow{P} 0 \quad \text{(by assumption)} \quad \xrightarrow{P} 0 \quad \text{(by cts mapping)} \end{aligned}$$

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

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

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

③ Analogous to ②

Theorem (Consistency of MLE for compact Θ)

$X_1, \dots, X_n \stackrel{iid}{\sim} \rho_{\theta_0}$, \mathcal{P} has densities ρ_θ , $\theta \in \Theta$

Assume ρ_θ cts in θ

- Θ compact $\Rightarrow \mathbb{E}_{\theta_0} \sup_{\theta \in \Theta} |\ell(\theta; x_i) - \ell(\theta_0; x_i)| < \infty$
- $\mathbb{E}_{\theta_0} \left[\sup_{\theta \in \Theta} |W_i(\theta)| \right] < \infty$
- 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 $\mu(\theta) = -D_{KL}(\theta_0 \parallel \theta)$

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

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

$$\|\bar{W}_n - \mu\|_\infty \xrightarrow{P} 0, \text{ apply 9.4, ②}$$

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

Thm (\approx Keener 9.11, but stronger conditions)

$X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta_0}$, P has cts densities p_{θ} , $\Theta \in \mathbb{R}^d$

Assume . Model identifiable

- For all compact $K \subseteq \mathbb{R}^d$, $E \left[\sup_{\theta \in K} |W_i(\theta)| \right] < \infty$

- $\exists r > 0$ s.t. $E \left[\sup_{\|\theta - \theta_0\| \geq r} W_i(\theta) \right] < 0$

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

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

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

Hence, $P(\hat{\theta}_n \notin K) \leq P(\underbrace{\bar{W}_n(\theta_0)}_{\xrightarrow{P} \beta_0} < \underbrace{\sup_{\theta \notin K} \bar{W}_n(\theta)}_{\xrightarrow{P} \beta}) \rightarrow 0$

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

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

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

Asymptotic Dist. of MLE

Theorem

$X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta_0}$ for $\theta_0 \in \Theta^0 \subseteq \mathbb{R}^d$

Assume • $\hat{\theta}_n \in \arg\max_{\theta \in \Theta} l_n(\theta; x)$, $\hat{\theta}_n \xrightarrow{P} \theta_0$

• In a neighborhood $\bar{B}_\varepsilon(\theta_0) = \{\theta : \|\theta - \theta_0\| \leq \varepsilon\} \subseteq \Theta$:

(i) $l_1(\theta; x)$ has 2 cts deriv.s on $\bar{B}_\varepsilon(\theta_0)$, $\forall x$

(ii) $\mathbb{E}_{\theta_0} \left[\sup_{\theta \in \bar{B}_\varepsilon} \|\nabla^2 l_1(\theta; X_i)\| \right] < \infty$

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

• Fisher info :

$$\mathbb{E}_{\theta_0} \nabla l_1(\theta_0; X_i) = 0$$

$$\text{Var}_{\theta_0} \nabla l_1(\theta_0; X_i) = -\mathbb{E}_{\theta_0} \nabla^2 l_1(\theta_0; X_i) \succ 0$$

(enough to have 3rd 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 $\sup_{\theta \in \bar{B}_\varepsilon} \left\| \frac{1}{n} \nabla^2 \ell(\theta) - J_1(\theta) \right\| \xrightarrow{P} 0$ since \bar{B}_ε compact,
 $\nabla^2 \ell_1$ cts

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

$P_{\theta_0}(A_n) \rightarrow 0$ by assumption

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

$$0 = \nabla \ell_n(\hat{\theta}_n; x)$$

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

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

$$\nabla \ell_n(\hat{\theta}_n - \theta_0) = \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n) \right)^{-1}}_{P_{\theta_0} \rightarrow J_1(\theta_0)^{-1}} \underbrace{\frac{1}{n} \nabla \ell_n(\theta_0)}_{\text{By 9.4 ① + cts mapping}} \Rightarrow N(0, J_1(\theta_0))$$

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

Behavior on A_n irrelevant to asymptotic limit \square