Outline

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

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

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

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

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

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

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

Proof sketch:

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

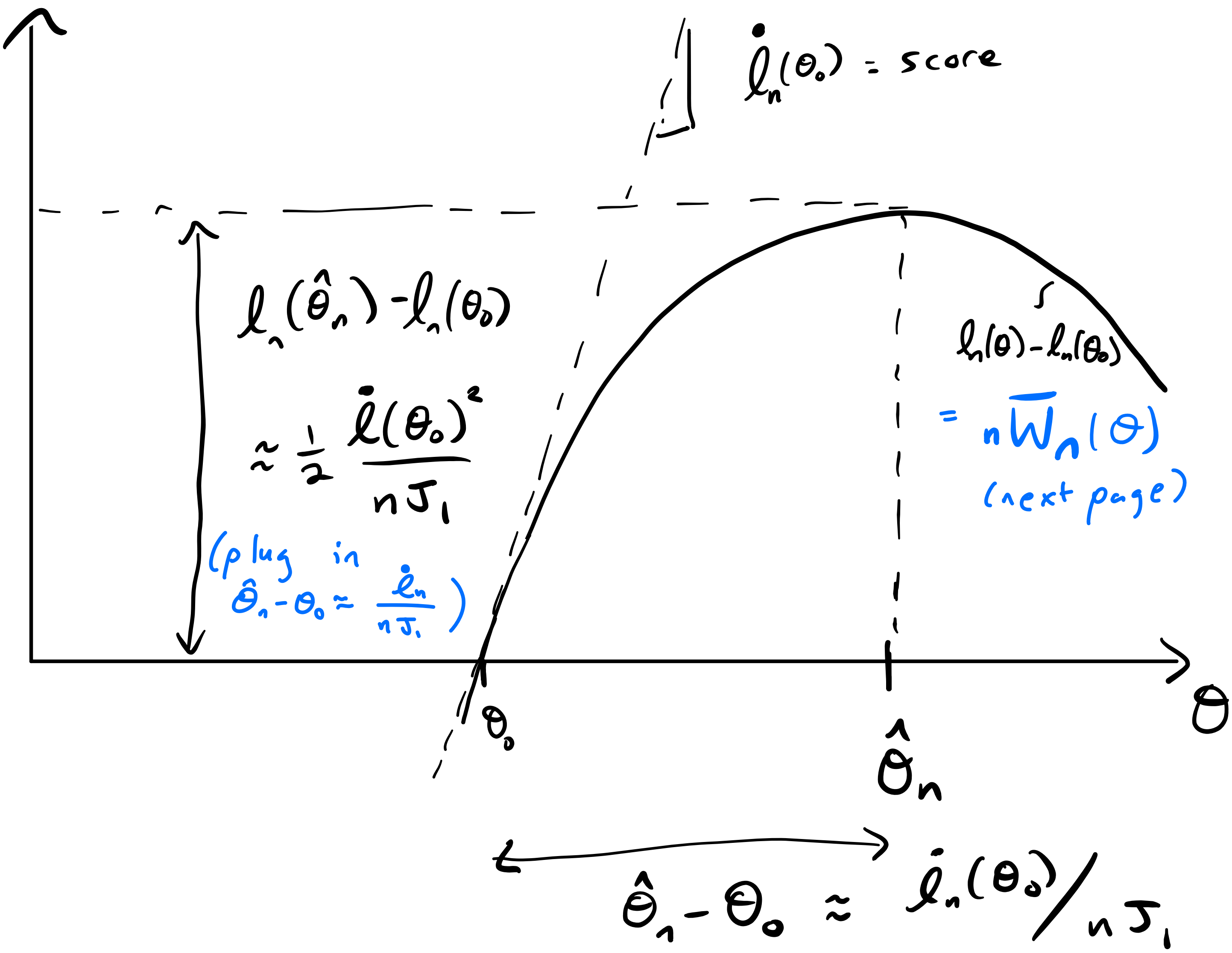
Asymptotic Picture $(d=1)$

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

Quadratic approximation near θ_0:

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \underbrace{\dot{\ell}_n(\theta_0)}_{\approx N(0,\, nJ_1(\theta_0))} (\theta - \theta_0) + \frac{1}{2} \underbrace{\ddot{\ell}_n(\theta_0)}_{\approx -nJ_1(\theta_0)} (\theta - \theta_0)^2$$

Gaussian linear term — Deterministic curvature



Consistency of MLE

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

[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**:

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

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

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

$$\leq \log 1 = 0$$

(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 $\hat{\theta}_n \in \underset{\theta \in \Theta}{\arg\max}\ \overline{W}_n(\theta)$ 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 $\|f\|_\infty = \sup_{t \in K} |f(t)|$

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

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

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

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

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 (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
- $\log p_\theta(x)$ cts in θ, all $x \in \mathcal{X}$
- 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 $\overline{W}_n$,

$$\delta_n = \| \overline{W}_n - \mu \|_\infty \xrightarrow{p} 0.$$

Fix $\varepsilon > 0$, want to show $\mathbb{P}_{\theta_0}(\|\theta - \theta_0\| \geq \varepsilon) \to 0$

Let $\tilde{\Theta}_\varepsilon = \Theta \setminus B_\varepsilon(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}$ (compact)

Let $\mu^*_\varepsilon = \max_{\theta \in \tilde{\Theta}_\varepsilon} \mu(\theta) < 0 = \mu(\theta_0)$

$W^*_\varepsilon = \max_{\theta \in \tilde{\Theta}_\varepsilon} \overline{W}_n(\theta)$

$$\mathbb{P}_{\theta_0}(\|\theta-\theta_0\| \geq \varepsilon) \leq \mathbb{P}_{\theta_0}\Big(\underbrace{W_\varepsilon^*}_{\leq \mu_\varepsilon^* + \delta_n} \geq \underbrace{\overline{W}_n(\theta_0)}_{\geq -\delta_n}\Big)$$

$$\leq \mathbb{P}_{\theta_0}\Big(2\delta_n \geq \underbrace{-\mu_\varepsilon^*}_{>0}\Big) \to 0$$

$\boxtimes$

<u>Note</u> We usually care about non-compact parameter spaces, need some extra assumption to get us there.

<u>Corollary</u> Same assumptions except now $\Theta = \mathbb{R}^d$, (non-compact)

but there's some $R < \infty$ large enough so

$$\mathbb{P}_{\theta_0}(\|\hat{\theta}_n - \theta_0\| > R) \to 0$$

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$

<u>Proof</u> Let $\tilde{\Theta} = \{\theta : \|\theta - \theta_0\| \leq R\}$, $\tilde{\theta}_n = \underset{\theta \in \tilde{\Theta}}{\arg\max}\, p_\theta(x)$

Then $\tilde{\theta}_n \to \theta_0$ by assumption

$$\mathbb{P}_{\theta_0}(\hat{\theta}_n \neq \tilde{\theta}_n) = \mathbb{P}_{\theta_0}(\hat{\theta}_n \notin \tilde{\Theta}) \to 0$$

so $\hat{\theta}_n - \tilde{\theta}_n \xrightarrow{p} 0 \Rightarrow \hat{\theta}_n = \tilde{\theta}_n + (\hat{\theta}_n - \tilde{\theta}_n) \xrightarrow{p} \theta_0$ $\boxtimes$

So the only thing we actually need to worry about is if $\hat{\theta}_n$ is extremely far away from θ_0 with non-negligible Prob.

Theorem (Asymptotic distribution of MLE)

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

Assume
- $\mathcal{P}$ identifiable
- Θ compact
- $\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \Theta} |W_i(\theta)|\right] < \infty$
- $\ell(\theta; x) = \log p_\theta(x)$ has two cts derivatives in θ
- $\mathbb{E}_{\theta_0} \sup_{\theta \in \Theta} \|\nabla^2 \ell_1(\theta; x_i)\| < \infty$
- $J_1(\theta_0) = \mathbb{E}_{\theta_0} \nabla^2 \ell_1(\theta_0; x_i) \succ 0$ positive definite

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

Proof From before, we had

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left(-\tfrac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1} \nabla \ell_n(\theta_0)$$

for $\tilde{\theta}_n$ between θ_0 and $\hat{\theta}_n$

Previous result shows $\hat{\theta}_n \xrightarrow{p} \theta_0$, so $\tilde{\theta}_n \xrightarrow{p} \theta_0$ also

Define $V_i(\theta) = -\nabla^2 \ell_1(\theta; X_i) \in C(\Theta)$, $\quad \mathbb{E}_{\theta_0} \|V_i\|_\infty < \infty$ by assumption

Then $v(\theta) = \mathbb{E}_{\theta_0} V_i(\theta) \in C(\Theta)$, $\quad v(\theta_0) = J_1(\theta_0)$

$\bar{V}_n(\theta) = \frac{1}{n} \sum V_i(\theta)$, $\quad \|\bar{V}_n - v\|_\infty \xrightarrow{p} 0$

$$\left\|-\tfrac{1}{n}\nabla^2 \ell_n(\tilde{\theta}_n) - J_1(\theta_0)\right\| \le \left\|\bar{V}_n(\tilde{\theta}_n) - v(\tilde{\theta}_n)\right\| + \left\|v(\tilde{\theta}_n) - v(\theta_0)\right\|$$
$$\le \underbrace{\left\|\bar{V}_n - v\right\|_\infty}_{\xrightarrow{P} 0} + \underbrace{\left\|v(\tilde{\theta}_n) - v(\theta_0)\right\|}_{\xrightarrow{P} 0 \text{ (cts mapping)}}$$

Hence $\left(-\tfrac{1}{n}\nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1} \xrightarrow{P} J_1(\theta_0)^{-1}$ (cts mapping)

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

<u>Note</u> In this proof we played a bit fast and loose with our LLN for random functions, which we only stated for real-valued W_n

So technically we have only justified for $d=1$

But proof works for $d>1$