

Math 4261: Mathematical Statistics I

Lecture 13: Properties of MLE (References: Rice 8.5; DS 7.5)

Fall 2026

Outline

Recap

Properties of MLE

Examples

Proof ideas

Outline

Recap

Properties of MLE

Examples

Proof ideas

Recap: Asymptotic normality

Consider an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ of θ . Here $X_1, \dots, X_n$ are data from some distribution parametrized by θ .

- ▶ **Asymptotic normality** means

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)).$$

- ▶ In particular, $\hat{\theta} \xrightarrow{p} \theta$: **consistency**.
- ▶ Large-sample approximation: $\hat{\theta} \approx \mathcal{N}(\theta, v(\theta)/n)$.
- ▶ Sampling error has **$1/\sqrt{n}$ scale**; approximate standard error $\sqrt{v(\theta)/n}$.

Recap: Delta method and Pareto

- ▶ If $\sqrt{n}(S - m) \xrightarrow{d} \mathcal{N}(0, v)$, the **delta method** gives for smooth g :

$$\sqrt{n}(g(S) - g(m)) \xrightarrow{d} \mathcal{N}(0, g'(m)^2 v).$$

- ▶ For **Pareto($\theta, 1$)**, we found the MoM estimator

$$\mathbb{E}_\theta[X] = \frac{\theta}{\theta - 1} \quad \Longrightarrow \quad \hat{\theta}_{\text{MoM}} = \frac{\overline{X}}{\overline{X} - 1}.$$

We showed it was consistent for $\theta > 1$. Furthermore, based on the delta method, we showed for $\theta > 2$, **asymptotic normality**: specifically

$$\sqrt{n}(\hat{\theta}_{\text{MoM}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta(\theta - 1)^2}{\theta - 2}\right).$$

Recap: Plug-in estimators and coin log odds

Suppose we want to estimate $g(\theta)$, where θ is some unknown parameter. If we already have an estimator $\hat{\theta}$ of θ , then the **plug-in** estimator of $g(\theta)$ is $g(\hat{\theta})$.

Example:

$X_1, \dots, X_n \sim \text{Bernoulli}(p)$. Want to estimate **log-odds**
 $g(p) = \log(p/(1-p))$.

- ▶ Used plug-in $g(\bar{X})$
- ▶ Delta method gives

$$\sqrt{n}(g(\bar{X}) - g(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).$$

Outline

Recap

Properties of MLE

Examples

Proof ideas

Theoretical properties of MLE

Theorem

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x|\theta)$, where $\theta \in \Omega \subseteq \mathbb{R}$ is a single parameter. Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ denote the MLE. Under regularity conditions^a on $f(x | \theta)$,

- ▶ $\hat{\theta}$ is **consistent**: $\hat{\theta} \rightarrow \theta$ in probability.
- ▶ $\hat{\theta}$ is **asymptotically normal**: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I(\theta)})$.

Here $I(\theta)$ is the single observation **Fisher information** for $X \sim f(x | \theta)$:

$$I(\theta) = \text{Var}_{\theta} \left[\frac{\partial}{\partial \theta} \log f(X | \theta) \right] = -\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(X | \theta) \right],$$

where $\mathbb{E}_{\theta}$, Var_{θ} denote expectation/variance over $X \sim f(x | \theta)$.

^aIn this course we won't discuss these exact conditions. Three of them are (i) θ is not on the boundary of the parameter space, (ii) $\theta \mapsto \log f(x | \theta)$ is twice differentiable, and (iii) $I(\theta)$ is nonzero.

Score

For one observation $X \sim f(x \mid \theta)$, define the **score** to be

$$s(X, \theta) = \frac{\partial}{\partial \theta} \log f(X \mid \theta).$$

For i.i.d. observations, $\ell_n(\theta) = \sum_{i=1}^n \log f(X_i \mid \theta)$ and so

$$\ell'_n(\theta) = \sum_{i=1}^n s(X_i, \theta).$$

Under usual regularity conditions, the score has **mean zero**:

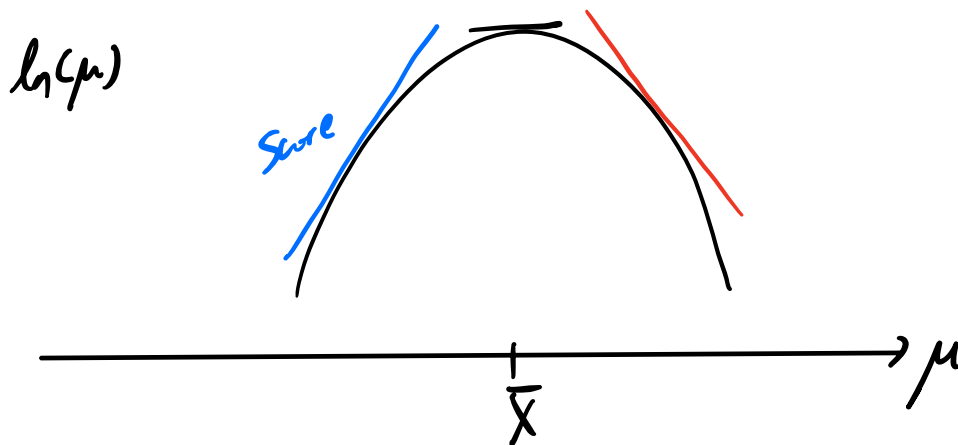
$$\mathbb{E}_\theta[s(X, \theta)] = 0$$

Interpreting the score

The score is the **slope of the log-likelihood**.

Example: Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\sigma^2 > 0$ known. Then

$$s(X, \mu) = \frac{X - \mu}{\sigma^2}, \quad \ell'_n(\mu) = \frac{n(\bar{X} - \mu)}{\sigma^2}.$$



- ▶ If $\mu < \bar{X}$, the score is **positive**: increase μ .
- ▶ If $\mu > \bar{X}$, it is **negative**: decrease μ .
- ▶ At the MLE $\hat{\mu} = \bar{X}$, the score is zero.

Fisher information

The function

$$\begin{aligned} I(\theta) &= \text{Var}_\theta \left[\frac{\partial}{\partial \theta} \log f(X \mid \theta) \right] \\ &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X \mid \theta) \right] \end{aligned}$$

is called the **Fisher information**.

- ▶ Fisher information is the variance of the score.
- ▶ Since the score has mean zero, it is also $I(\theta) = \mathbb{E}_\theta[s(X, \theta)^2]$.
- ▶ Here $I(\theta)$ is for **one observation**; n i.i.d. observations have information $nI(\theta)$.

Using the MLE theorem

Under the conditions of the MLE theorem, for large n ,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right).$$

- ▶ The approximate standard error is $1/\sqrt{nI(\theta)}$.
- ▶ Replacing θ by the consistent MLE gives the **estimated standard error**:

$$\widehat{\text{SE}} = \frac{1}{\sqrt{nI(\hat{\theta})}}.$$

- ▶ More observations give **smaller sampling error**, with $1/\sqrt{n}$ scale.

Outline

Recap

Properties of MLE

Examples

Proof ideas

Example: Poisson

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda)$. The MLE is $\hat{\lambda} = \bar{X}$ (for a positive total count), where $\lambda > 0$.

$$f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \Rightarrow \log f(x|\lambda) = -\lambda + x \log \lambda - \log(x!)$$

Score $\frac{\partial}{\partial \lambda} \log f(x|\lambda) = -1 + \frac{x}{\lambda}$

(check: $\mathbb{E}_\lambda \left[\frac{\partial}{\partial \lambda} \log f(x|\lambda) \right] = \mathbb{E}_\lambda \left[-1 + \frac{x}{\lambda} \right] = -1 + \frac{1}{\lambda} = 0$)

fisher info $\mathcal{I}(\lambda) = \text{Var}_\lambda \left[\frac{\partial}{\partial \lambda} \log f(x|\lambda) \right] = \text{Var}_\lambda \left[-1 + \frac{x}{\lambda} \right]$
 $= \frac{1}{\lambda^2} \text{Var}_\lambda X = \frac{1}{\lambda}$

Alternatively, $\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) = -\frac{x}{\lambda^2} \Rightarrow \mathcal{I}(\lambda) = -\mathbb{E}_\lambda \left[\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) \right] = \frac{1}{\lambda}$

Hence $\sqrt{n} (\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \frac{1}{\mathcal{I}(\lambda)}) = \mathcal{N}(0, \lambda)$

Example: Pareto

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Pareto}(\theta, 1)$, where $\theta > 0$.

Pareto($\theta, 1$) has PDF $f(x | \theta) = \frac{\theta}{x^{\theta+1}}$ for $x > 1$.



(1) Show that the MLE is $\hat{\theta} = \frac{n}{\sum_{i=1}^n \log X_i}$.

$$\log f(x|\theta) = \log \theta - (\theta+1) \log x$$

log-likelihood $\ln(\theta) = \sum_{i=1}^n \log f(X_i|\theta) = n \log \theta - (\theta+1) \sum_{i=1}^n \log X_i$

$$\ln'(\theta) = \frac{n}{\theta} - \sum_{i=1}^n \log X_i \stackrel{\text{set}}{=} 0$$

$$\Rightarrow \text{MLE } \hat{\theta} = \frac{n}{\sum_{i=1}^n \log X_i}$$

$$\left[\begin{array}{l} \text{check: } \ln'(\theta) > 0 \text{ for } \theta < \hat{\theta} \\ \ln'(\theta) < 0 \text{ for } \theta > \hat{\theta} \end{array} \right]$$

Example: Pareto

$$\Rightarrow \log f(x|\theta) = \log \theta - (\theta+1) \log x$$

For $\theta > 0$, $\text{Pareto}(\theta, 1)$ has PDF $f(x | \theta) = \frac{\theta}{x^{\theta+1}}$ for $x > 1$.

(2) Compute the Fisher information and give the asymptotic normality statement.

Score $\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{1}{\theta} - \log x$

$$\left[\begin{array}{l} \text{check } E_{\theta}[\log X] = \int_1^{\infty} \frac{\theta}{x^{\theta+1}} \cdot \log x \, dx = -\frac{\log x}{x^{\theta}} \Big|_1^{\infty} + \int_1^{\infty} \frac{1}{x^{\theta+1}} \, dx \\ \text{So } E_{\theta} \left[\frac{\partial}{\partial \theta} \log f(x|\theta) \right] = \frac{1}{\theta} - \frac{1}{\theta} = 0 \cdot \left[-\frac{1}{\theta x^{\theta}} \Big|_1^{\infty} = \frac{1}{\theta} \right] \end{array} \right]$$

$$\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = -\frac{1}{\theta^2} \Rightarrow I(\theta) = -E_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] = \frac{1}{\theta^2}$$

$$\text{here } T_n(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I(\theta)}) = \mathcal{N}(0, \theta^2)$$

Example: Pareto MoM and MLE

Recall from Lecture 12: $\hat{\theta}_{\text{MoM}} = \frac{\overline{X}}{\overline{X}-1}$.

For $\theta > 2$, compare the two estimators of θ :

$$\sqrt{n}(\hat{\theta}_{\text{MoM}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta(\theta-1)^2}{\theta-2}\right),$$

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

- ▶ Both are consistent and asymptotically normal.
- ▶ The MLE has **smaller asymptotic variance**:

$$\frac{\theta(\theta-1)^2}{\theta-2} - \theta^2 = \frac{\theta}{\theta-2} > 0.$$

Example: Pareto mean from MoM (recap)

- ▶ Last class, we found the MoM estimator for θ :

$$\hat{\theta}_{\text{MoM}} = \frac{\overline{X}}{\overline{X} - 1}.$$

- ▶ We showed consistency of $\hat{\theta}_{\text{MoM}}$ for $\theta > 1$ and asymptotic normality for $\theta > 2$.
- ▶ To estimate the **population mean** $\mu = g(\theta) = \frac{\theta}{\theta-1}$, we used the plug-in estimator

$$g(\hat{\theta}_{\text{MoM}}) = \frac{\hat{\theta}_{\text{MoM}}}{\hat{\theta}_{\text{MoM}} - 1} = \overline{X}.$$

Let's see what happens when we use **MLE** instead for the plug-in.

Example: Pareto mean from MLE

A **plug-in** estimator of the mean $\frac{\theta}{\theta-1}$ based on the MLE $\hat{\theta}$ would be

$$\hat{\mu} = \frac{\hat{\theta}}{\hat{\theta} - 1}, \quad \hat{\theta} > 1.$$

To compute its asymptotic variance:

$$\text{MLE: } \sqrt{n} (\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta^2)$$

$$\text{Plug-in est for mean: } \sqrt{n} \left(\frac{\hat{\theta}}{\hat{\theta} - 1} - \frac{\theta}{\theta - 1} \right) \xrightarrow{d} ?$$

Example: Pareto mean from MLE

Delta method with $g(x) = \frac{x}{x-1} = 1 + \frac{1}{x-1}$, $g'(x) = -\frac{1}{(x-1)^2}$

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \rightarrow g'(\theta) \mathcal{N}(0, \theta^2) \\ = \mathcal{N}(0, \theta^2 (g'(\theta))^2)$$

$$\text{Thus } ? = \boxed{\mathcal{N}\left(0, \frac{\theta^2}{(\theta-1)^4}\right)}$$

Practical consequences

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\theta^2}{n(\theta-1)^4}\right), \\ \text{approximately}$$

We can report approximate
Standard error as

$$\widehat{SE}(\hat{\mu}) = \frac{\hat{\theta}}{\sqrt{n}(\hat{\theta}-1)^2}.$$

Outline

Recap

Properties of MLE

Examples

Proof ideas

Proof sketch: Consistency of MLE

To explain why the MLE $\hat{\theta}$ is consistent, recall that $\hat{\theta}$ is the value of θ which maximizes (a scaled version) of the log-likelihood:

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i \mid \theta)$$

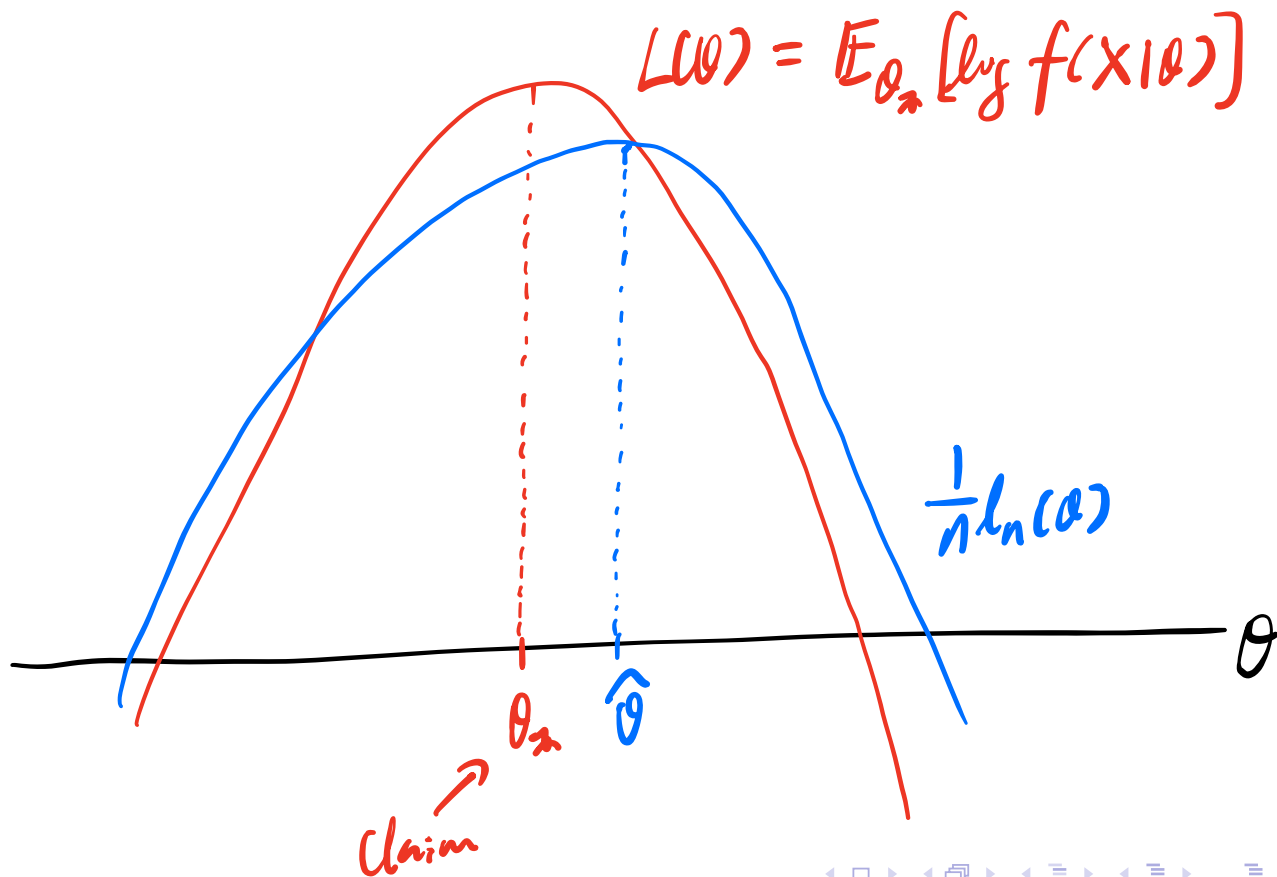
Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x \mid \theta_*)$, where θ_* is the true parameter. Fixing any θ (not necessarily θ_*), the above is the sample average of n i.i.d. random variables. By LLN,

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i \mid \theta) \rightarrow \mathbb{E}_{\theta_*} [\log f(X \mid \theta)] \quad \text{in probability.}$$

Here $\mathbb{E}_{\theta_*} [\log f(X \mid \theta)]$ is the expected value of the function $\log f(X \mid \theta)$ under $X \sim f(x \mid \theta_*)$ (the log-likelihood is evaluated at an arbitrary parameter θ which may be different from θ_*).

Proof sketch: Consistency of MLE

Under suitable conditions, the value of θ maximizing $\frac{1}{n}\ell_n(\theta)$ (which is the MLE $\hat{\theta}$) converges in probability to the value of θ maximizing the limiting function $L(\theta) = \mathbb{E}_{\theta_*}[\log f(X | \theta)]$.



Proof sketch: Consistency of MLE

Claim

For an identifiable model, the θ maximizing $L(\theta) = \mathbb{E}_{\theta_*} [\log f(X | \theta)]$ is exactly the true parameter θ_* .

$$\begin{aligned} L(\theta) - L(\theta_*) &= \mathbb{E}_{\theta_*} [\log f(X|\theta) - \log f(X|\theta_*)] \\ &= \mathbb{E}_{\theta_*} \left[\log \frac{f(X|\theta)}{f(X|\theta_*)} \right] \end{aligned}$$

The function $\log x$ is concave.

$$L(\theta) - L(\theta_*) \stackrel{\text{Jensen}}{\leq} \log \mathbb{E}_{\theta_*} \left[\frac{f(X|\theta)}{f(X|\theta_*)} \right]$$

$$\stackrel{\text{Wtms}}{=} \log \int \frac{f(x|\theta)}{\cancel{f(x|\theta_*)}} \cdot \cancel{f(x|\theta_*)} dx$$

Proof sketch: Consistency of MLE

Cont'd,

$$L(\theta) - L(\theta_n) \leq \log \int \overbrace{f(x|\theta)}^{=1} dx = 0$$

Hence $L(\theta) \leq L(\theta_n) \quad \forall \theta$,

so $L(\theta)$ is maximized at θ_n .

This gives an explanation of the consistency of $\hat{\theta}$.

Proof sketch: Asymptotic normality

Let us give some ideas for why the following is true:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

(Recall) Slutsky's lemma

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables. Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a **constant**. Then

- ▶ $X_n + Y_n \xrightarrow{d} X + c$
- ▶ $X_n Y_n \xrightarrow{d} cX$
- ▶ $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$, provided $c \neq 0$.

Proof sketch: Asymptotic normality

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

The MLE $\hat{\theta}$ minimizes $\ln(\ell) = \sum_{i=1}^n \log f(x_i | \theta)$.

So $\ln'(\theta) = 0$. Since $\hat{\theta} \rightarrow \theta$ in prob., we expect that for n large, $\hat{\theta}$ is close to θ .

Taylor expand around θ :

$$0 = \ln'(\hat{\theta}) \approx \ln'(\theta) + \ln''(\theta)(\hat{\theta} - \theta)$$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta) \approx -\frac{\sqrt{n}}{\ln''(\theta)} \ln'(\theta) = \frac{\ln'(\theta)/\sqrt{n}}{-\ln''(\theta)/n}$$

Proof sketch: Asymptotic normality

$$\sqrt{n}(\hat{\theta} - \theta) \approx \frac{\ell'_n(\theta)/\sqrt{n}}{-\ell''_n(\theta)/n}$$

Denominator

$$-\frac{\ell''_n(\theta)}{n} = \frac{1}{n} \sum_{i=1}^n \left(-\frac{\partial^2}{\partial \theta^2} \log f(X_i | \theta) \right)$$

$$\xrightarrow{\text{prob.}} \mathbb{E}_{\theta} \left[-\frac{\partial^2}{\partial \theta^2} \log f(X | \theta) \right]$$

LLN

hence also $\xrightarrow{d}$

$$= \mathbf{I}(\theta)$$

Proof sketch: Asymptotic normality

Numerator $\frac{\ln'(\theta)}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{\frac{\partial}{\partial \theta} \log f(X_i | \theta)}_{\text{score evaluated at } X_i}$

$$\xrightarrow{d} \mathcal{N}(0, \text{Var}_{\theta} \left[\frac{\partial}{\partial \theta} \log f(X | \theta) \right]) \quad \text{CLT}$$

$$= \mathcal{N}(0, \mathbf{I}(\theta))$$

Slutsky's : $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \frac{1}{\mathbf{I}(\theta)} \mathcal{N}(0, \mathbf{I}(\theta))$

$$= \mathcal{N}(0, \frac{1}{\mathbf{I}(\theta)})$$

