Lecture 5: Maximum Likelihood Estimation

5.1 The maximum likelihood estimator (MLE)

Remark 5.1 Let $X_1, \ldots, X_n$ be a random sample from some population pdf $f(x;\theta)$, $\theta \in \Omega$. If we are interested in the probability of a realized value of $X_1, \ldots, X_n$, namely $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, this can be estimated by

$$f_{X_1,\ldots,X_n}(\boldsymbol{x};\theta) \cdot |\Delta \boldsymbol{x}| = \prod_{i=1}^{n} f(x_i;\theta) \cdot |\Delta x_i|.$$

This value can actually be interpreted as the "likelihood of $\boldsymbol{x}$ given the model parameter θ." This begs the question, what value of θ *maximizes* the likelihood of the realization $\boldsymbol{x}$ being generated? If we are between two values θ_1 and θ_2, we can answer this by comparing

$$\prod_{i=1}^{n} f(x_i;\theta_1) > \prod_{i=1}^{n} f(x_i;\theta_2) \text{ or } \prod_{i=1}^{n} f(x_i;\theta_1) < \prod_{i=1}^{n} f(x_i;\theta_2).$$

In general, we want to estimate θ by finding the value which maximizes $\prod_{i=1}^{n} f(x_i;\theta)$.

Definition 5.1 Let $X_1, \ldots, X_n$ be a random sample from some population pdf $f(x;\theta)$, $\theta \in \Omega$. The **likelihood function** given a fixed realization $\boldsymbol{x} = (x_1, \ldots, x_n)$ is defined as

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Omega.$$

Often it is more convenient to use the log-likelihood function, denoted

$$\ell(\theta;\boldsymbol{x}) = \log L(\theta;\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i;\theta), \quad \theta \in \Omega.$$

We may omit $\boldsymbol{x}$ and denote $L(\theta)$ or $\ell(\theta)$ as the likelihood or log-likelihood function, respectively.

Definition 5.2 For a given $\boldsymbol{x}$, call $\hat{\theta}_{MLE}(\boldsymbol{x})$ the **maximum likelihood estimate** if

$$L(\hat{\theta}_{\text{MLE}}(\boldsymbol{x});\boldsymbol{x}) = \max_{\theta\in\Omega} L(\theta;\boldsymbol{x}), \quad \text{or} \quad \hat{\theta}_{\text{MLE}}(\boldsymbol{x}) = \arg\max_{\theta\in\Omega} L(\theta;\boldsymbol{x}).$$

If $\hat{\theta}_{\text{MLE}}(\boldsymbol{x})$ is defined for any $\boldsymbol{x}$, then call $\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MLE}}(X_1, \ldots, X_n) = \hat{\theta}_{\text{MLE}}(\boldsymbol{X})$ the **maximum likelihood estimator (MLE)**.

Remark 5.2 It is often more convenient to work with the log-likelihood because $\hat{\theta}_{\text{MLE}}$ is often found by differentiating with respect to θ, and sums are easier to differentiate than products.

Example 5.1 Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$ with $\lambda \geq 0$. Find the MLE of λ.

Solution. The joint pmf of the random sample is given by:

$$\text{pdf}_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad \lambda \geq 0.$$

The likelihood function is then given by

$$L(\lambda; x_1, \ldots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!}, \quad \lambda \geq 0.$$

Thus, the log-likelihood for $\lambda > 0$ is

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \log(x_1! \cdots x_n!)$$

The first and second derivatives are

$$\ell'(\lambda) = -n + \frac{n\bar{x}}{\lambda},$$
$$\ell''(\lambda) = -\frac{n\bar{x}}{\lambda^2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\lambda > 0$. Then, by examining the increase and decrease of the log-likelihood function we have the following cases

(a) If $\bar{x} > 0$ we know $\ell''(\lambda) < 0$, so $\ell(\lambda)$ is concave and because $\bar{x}$ is a stationary point ($\ell'(\bar{x}) = 0$), we know

$$\max_{\lambda > 0} \ell(\lambda) = \ell(\bar{x})$$

(b) If $\bar{x} = 0$, then $\ell(\lambda)$ is a decreasing function, namely

$$\ell(\lambda) = -n\lambda - \log(x_1! \cdots x_n!).$$

Since the log-likelihood function is defined and continuous at $\lambda = 0$, we have

$$\max_{\lambda \geq 0} \ell(\lambda) = \ell(0) = \ell(\bar{x}) = \log L(\bar{x}; x_1, \ldots, x_n).$$

Therefore, the maximum likelihood estimator for λ is given by $\hat{\lambda}^{\text{MLE}} = \bar{X}$. ■

Remark 5.3 We use the derivative of the log-likelihood function to examine the increase and decrease of the function, as shown in Example 5.1. Specifically, the equation that sets the first derivative to zero to find the value of the parameter is called the **likelihood equation** $[\ell'(\theta) = 0]$. We can determine the maximum likelihood estimates $\hat{\theta}(x)$ among the roots of the equation.

Theorem 5.1 *Suppose a likelihood function $l(\theta)$ defined on a parameter space Ω on the real line, and suppose it is twice differentiable and its second derivative be continuous. If*

$$l''(\theta) < 0, \quad \forall\theta \in \Omega \text{ and } l'(\hat{\theta}) = 0, \quad \hat{\theta} \in \Omega,$$

then, $\max_{\theta\in\Omega} l(\theta) = l(\hat{\theta})$.

Proof. From the Taylor expansion of $l(\theta)$ around $\hat{\theta}$:

$l(\theta) = l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}l''(\theta^*)(\theta - \hat{\theta})^2, \quad \exists\theta^* \in \Omega$ by Taylor's theorem (ref. mean value theorem).

Therefore, from the given conditions:

$$l(\theta) = l(\hat{\theta}) + \frac{1}{2}l''(\theta^*)(\theta - \hat{\theta})^2 < l(\hat{\theta}), \quad \forall\theta \in \Omega : \theta \neq \hat{\theta}.$$

□

Theorem 5.2 *Let $l(\theta)$ be a function defined on an open interval Ω_0 on a real line, which is twice differentiable and continuous. If*

$$l''(\theta) < 0, \quad \forall\theta \in \Omega_0, \quad \lim_{\theta\to\partial(\Omega_0)} l(\theta) = -\infty,$$

where $\partial(\Omega_0)$ is the boundary set of Ω_0. Then the equation $l'(\theta) = 0$ has one and only one solution $\hat{\theta} = \hat{\theta}^{\text{MLE}}$.

Proof. For simplicity, let the domain be $\Omega_0 = (-\infty, +\infty)$, and assume by $l''(\theta) < 0$, $\forall\theta \in \Omega_0$, and $\lim_{\theta\to\partial(\Omega_0)} l(\theta) = -\infty$ that

$$\lim_{\theta\to-\infty} l(\theta) = -\infty, \quad \lim_{\theta\to\infty} l(\theta) = -\infty.$$

Since $l(\theta)$ is continuous, it attains a maximum value. That is,

$$l(\hat{\theta}) = \max_{\theta\in\Omega_0} l(\theta), \quad \hat{\theta} \in \Omega_0,$$

which proves the existence. Now, we show the uniqueness. Assume by way of contradiction that $l(\theta)$ attains its maximum at two different points, $\hat{\theta}_1$ and $\hat{\theta}_2$, then

$$l(\hat{\theta}_1) = l(\hat{\theta}_2) = \max_{\theta\in\Omega_0} l(\theta), \quad \hat{\theta}_1 < \hat{\theta}_2.$$

Since $l(\theta)$ is a concave function, we have

$$l\left(\frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_2\right) > \frac{1}{2}l(\hat{\theta}_1) + \frac{1}{2}l(\hat{\theta}_2) = \max_{\theta\in\Omega_0} l(\theta),$$

which is a contradiction. Therefore, $l(\theta)$ attains its maximum at a single point $\hat{\theta} \in \Omega_0$, and $l'(\hat{\theta}) = 0$ holds. Thus, $\hat{\theta}$ is the unique critical point and the maximizer of $l(\theta)$. □

Remark 5.4 Recall Example 5.1 where our boundary set is $\partial(\Omega_0) = \{0\}$. In particular, when $\lambda \to 0$, $l(\lambda) \to -\infty$. So in this case, we could apply Theorem 5.2.

Theorem 5.3 *Suppose $\eta = g(\theta)$ is a one-to-one transformation of θ for $\theta \in \Omega$. If the maximum likelihood estimator of the parameter θ exists, the MLE of the transformed parameter η is given by $\hat{\eta}^{MLE} = g(\hat{\theta}^{MLE})$.*

Proof. That is, for the reparameterization $\eta = g(\theta)$, where $\theta \in \Omega$ and $\eta \in g(\Omega)$, we have:

$$\text{pdf}(x_1, \ldots, x_n; \theta) = \text{pdf}(x_1, \ldots, x_n; g^{-1}(\eta)), \quad \eta \in g(\Omega)$$

Therefore, when $\eta = g(\theta)$, the likelihood function becomes:

$$L(g^{-1}(\eta); x_1, \ldots, x_n) = \text{pdf}(x_1, \ldots, x_n; g^{-1}(\eta))$$

When we maximize the likelihood,

$$\max_{\eta \in g(\Omega)} L(g^{-1}(\eta); x_1, \ldots, x_n) = \max_{\theta \in \Omega} L(\theta; x_1, \ldots, x_n) = L(\hat{\theta}^{MLE}) = L(g^{-1}(g(\hat{\theta}^{MLE}))).$$

Thus, the likelihood is maximized at $g(\hat{\theta}^{MLE})$, and we conclude that $\hat{\eta}^{MLE} = g(\hat{\theta}^{MLE})$. □

Remark 5.5 The theorem is especially useful when the transformation makes the likelihood function concave.

Example 5.2 Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$, with $\theta > 0$. Find the MLE of the parameter θ.

Solution. The pdf of the exponential distribution $\text{Exp}(\theta)$ is:

$$\text{pdf}(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0.$$

Thus, given the random sample, the log-likelihood function and its derivatives are given by

$$l(\theta) = -n \log \theta - n\bar{x}/\theta, \quad l'(\theta) = -n/\theta + n\bar{x}/\theta^2, \quad l''(\theta) = n/\theta^2 - 2n\bar{x}/\theta^3$$

Since the log-likelihood function is not concave for all $\theta > 0$ (inflection point at $\theta = 2\bar{x}$), we cannot apply Theorem 5.1. In this case, we need to investigate the increase and decrease of the likelihood function, which might not be easy. Now, using $\lambda = 1/\theta$ as the new parameter, the log-likelihood

function and its derivatives become

$$l(\lambda) = n\log\lambda - n\lambda\bar{x}, \quad l'(\lambda) = n/\lambda - n\bar{x}, \quad l''(\lambda) = -n/\lambda^2$$

It follows that:

$$l''(\lambda) = -n/\lambda^2 < 0, \quad \forall\lambda : 0 < \lambda < \infty$$

Next, we check the limits:

$$\lim_{\lambda\to 0} l(\lambda) = -\infty, \quad \lim_{\lambda\to\infty} l(\lambda) = -\infty.$$

Therefore, when using $\lambda = 1/\theta$, the log-likelihood function is concave, and we can apply Theorem 5.1. Solving the likelihood equation $l'(\lambda) = 0$, the maximum likelihood estimator for λ is $\hat{\lambda}^{MLE} = 1/\bar{X}$. Thus, from Theorem 5.3, the MLE of θ is

$$\hat{\theta}^{MLE} = \frac{1}{\hat{\lambda}^{MLE}} = \bar{X}.$$

■

Remark 5.6 In all previous examples, we have dealt with cases where the likelihood function is *differentiable*. However, there are cases where the likelihood function is non-differentiable or discontinuous. In such cases, we need to investigate the increase and decrease of the likelihood or log-likelihood function to find the MLE.

Example 5.3 Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{DE}(\theta, 1)$, the double exponential distribution (aka the "Laplace distribution") with $-\infty < \theta < \infty$ and suppose $n = 2m + 1$. Find the MLE of θ.

Solution. The pdf of $\text{DE}(\theta, 1)$ is

$$\text{pdf}(x;\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty.$$

Thus, the log-likelihood function is

$$l(\theta) = -\sum_{i=1}^{n} |x_i - \theta| - n\log 2.$$

From this, we observe that the function is not differentiable at $\theta = x_1, \ldots, \theta = x_n$, so we need to examine which value maximizes the log-likelihood function. Denote the order statistics $x_{(1)} < \cdots < x_{(n)}$, and define $x_{(0)} = -\infty$, $x_{(n+1)} = +\infty$. For $x_{(r)} \le \theta < x_{(r+1)}$ (where $r = 0, \ldots, n$), we have

$$l(\theta) = -\sum_{i=1}^{r}(\theta - x_{(i)}) - \sum_{i=r+1}^{n}(x_{(i)} - \theta) - n\log 2,$$

which can be simplified to

$$l(\theta) = (n-2r)\theta + \sum_{i=1}^{r} x_{(i)} - \sum_{i=r+1}^{n} x_{(i)} - n\log 2.$$

From this, when $n = 2m+1$, we can conclude the following:

(i) If $n - 2r > 0$, i.e., $r = 0, \ldots, m$, then $l(\theta)$ increases in the interval $[x_{(r)}, x_{(r+1)})$.

(ii) If $n - 2r < 0$, i.e., $r = m+1, \ldots, n$, then $l(\theta)$ decreases in the interval $[x_{(r)}, x_{(r+1)})$.

Therefore, the log-likelihood function $l(\theta)$ reaches its maximum at $\theta = x_{(m+1)}$. Hence, the maximum likelihood estimator is $\hat{\theta}^{MLE} = X_{(m+1)}$. ■

Remark 5.7 Similar to there being exceptions to the likelihood being differentiable, we also have to consider the cases where the maximum likelihood estimator is *not uniquely determined.*

Example 5.4 Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[\theta - 1, \theta + 1]$, with $(-\infty < \theta < \infty)$. Find the maximum likelihood estimator (MLE) of θ.

Solution. The joint probability density function (pdf) of the random sample is

$$\text{pdf}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \frac{1}{2} \mathbf{1}\{x_i \in [\theta - 1, \theta + 1]\} = 2^{-n} \mathbf{1}\{\theta - 1 \le x_{(1)}, x_{(n)} \le \theta + 1\},$$

where $x_{(1)} = \min_{1 \le i \le n} x_i$ and $x_{(n)} = \max_{1 \le i \le n} x_i$. Thus, the likelihood function is given by

$$L(\theta) = 2^{-n} \cdot \mathbf{1}\{x_{(n)} - 1 \le \theta \le x_{(1)} + 1\},$$

which we see that the likelihood function attains its maximum at all points in the interval $[x_{(n)} - 1, x_{(1)} + 1]$. Therefore, in this case, the MLE of θ is not uniquely determined, and any value of θ within this interval satisfies the maximum likelihood condition. For example, any value of the form

$$\hat{\theta}^{MLE} = \alpha(x_{(n)} - 1) + (1 - \alpha)(x_{(1)} + 1), \quad 0 \le \alpha \le 1,$$

is a valid MLE of θ. ■

Example 5.5 Consider the normal distribution $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$, $\sigma^2 > 0$. Find the maximum likelihood estimator of $\boldsymbol{\theta} = (\mu, \sigma^2)^{\mathrm{T}}$.

Solution. The probability density function (pdf) of the normal distribution $N(\mu, \sigma^2)$ is:

$$\text{pdf}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < \infty,$$

thus, the log-likelihood function is

$$l(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi.$$

Taking the partials with respect to μ and $\zeta = \sigma^2$ we get

$$\frac{\partial}{\partial\mu}l(\mu,\zeta) = \frac{1}{\zeta}\left[\sum_{i=1}^{n}(x_i - \mu)\right], \qquad \frac{\partial^2}{\partial\mu^2}l(\mu,\zeta) = -\frac{n}{\zeta} < 0$$
$$\frac{\partial}{\partial\zeta}l(\mu,\zeta) = \frac{1}{2\zeta^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\zeta}, \qquad \frac{\partial^2}{\partial\zeta^2}l(\mu,\zeta) = -\frac{1}{\zeta^3}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{n}{2\zeta^2}$$

By setting the first partial w.r.t. μ equal to zero, we get $\hat{\mu} = \bar{x}$ which obtains the maximum because the second partial is always negative. Plugging in our estimate for μ and following the same procedure with $\zeta = \sigma^2$, we see that $\hat{\zeta} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. To check if this is a maximum, we find that

$$\frac{\partial}{\partial\zeta}l(\hat{\mu},\zeta) > 0 \quad \text{if and only if} \quad \zeta < \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

With this we know $l(\hat{\mu}, \zeta)$ achieves its maximum at $\hat{\zeta} = \hat{\sigma}^2$. Thus,

$$\max_{\sigma^2>0}\max_{-\infty<\mu<+\infty} l(\mu, \sigma^2) = l(\bar{x}, \hat{\sigma}^2)$$

with the maximum likelihood estimator

$$\hat{\boldsymbol{\theta}}^{MLE} = (\hat{\mu}, \hat{\sigma}^2)^{\mathrm{T}} = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{\mathrm{T}}.$$

■

5.2 Consistency of the MLE

Remark 5.8 Consistency is the most fundamental property of an estimator which states that the estimated value should approach the true value of the parameter as the sample size increases. In special cases, the method of moment estimator (MME) and the maximum likelihood estimator (MLE) coincide, such as when a random sample $X_1, \ldots, X_n$ follows Bernoulli, Poisson, Exponential, or Gaussian distribution. In these cases, the MLE is consistent (by WLLN) and also asymptotically normal (by CLT). But what about the general case? What conditions do we need for consistency?

Remark 5.9 The following assumptions are often referred to as **regularity conditions**:

(R0) The cdfs are distinct; i.e., $\theta_1 \neq \theta_2 \implies F(x_i;\theta_1) \neq F(x_i;\theta_2)$.

(R1) The pdfs have common support for all $\theta \in \Omega$ (i.e., the support of X_i does not depend on θ).

(R2) The point θ_0 (the *true value* of θ) is an interior point in Ω.

(R3) The pdf of $f(x;\theta)$ is twice differentiable as a function of θ (note: this is given if R5 is true).

(R4) The integral $\int f(x;\theta)dx$ can be differentiated twice under the integral sign as a function of θ.

(R5) The pdf $f(x;\theta)$ is three times differentiable as a function of θ. Further, for all $\theta \in \Omega$, there exists a constant c and a function $M(x)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)\right| \leq M(x)$$

with $\mathbb{E}_{\theta_0}[M(x)] < \infty$ for all $\theta \in (\theta_0 - c, \theta_0 + c)$ and all x in the support of X.

Theorem 5.4 *Assume $X_1, \ldots, X_n$ satisfy regularity conditions R0–R2, where θ_0 is the true parameter, and further that $f(x;\theta)$ is differentiable with respect to $\theta \in \Omega$. Then the likelihood equation*

$$\frac{\partial}{\partial\theta}L(\theta) = 0 \quad \textit{or equivalently} \quad \frac{\partial}{\partial\theta}l(\theta) = 0$$

has a solution $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Corollary 5.1 *Assuming the same conditions as Theorem 5.4, if $\hat{\theta}_n$ is a **unique** solution, then $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Example 5.6 The probability density function (pdf) of the logistic distribution $L(\theta, 1)$ is:

$$f(x;\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2} = \frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2}, \quad -\infty < x < +\infty$$

Then, we have

$$\begin{aligned}
l_n(\theta) &= -n\bar{x} + n\theta - 2\sum_{i=1}^{n}\log\Big(1+e^{-x_i+\theta}\Big),\\
l_n'(\theta) &= n - 2\sum_{i=1}^{n}\frac{e^{-x_i+\theta}}{1+e^{-x_i+\theta}},\\
l_n''(\theta) &= -2\sum_{i=1}^{n}\frac{e^{-x_i+\theta}}{(1+e^{-x_i+\theta})^2},
\end{aligned}$$

which gives

$$l_n''(\theta) < 0, \quad \forall\theta : -\infty < \theta < +\infty, \quad \lim_{\theta\to-\infty} l(\theta) = -\infty, \quad \lim_{\theta\to+\infty} l(\theta) = -\infty.$$

Since the conditions for Theorem 5.2 are satisfied, we know that $l'(\theta) = 0$ has only one solution. The MLE is the root of the likelihood equation

$$l_n'(\theta) = n - 2\sum_{i=1}^{n} \frac{e^{-x_i+\theta}}{(1+e^{-x_i+\theta})} = 0,$$

and the MLE $\hat{\theta}_n^{MLE}$ is *consistent* by Theorem 5.4 and Corollary 5.1 (need numerical methods to actually solve for $\hat{\theta}_n^{MLE}$.

5.3 Asymptotic normality of the MLE

Theorem 5.5 ***(Asymptotic normality of the MLE)*** *Assume $X_1, \ldots, X_n$ are iid with pdf $f(x;\theta_0)$ for $\theta_0 \in \Omega$ such that the regularity conditions R0–R5 are satisfied. Suppose further that the Fisher Information satisfies $0 < I(\theta_0) < \infty$. Then if $\hat{\theta}$ satisfies $S(\hat{\theta}) = l'(\hat{\theta}) = 0$ and $\hat{\theta} \xrightarrow{P} \theta_0$ as $n \to \infty$, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Proof. Omitted (see Hogg & Craig Theorem 6.2.2 for proof). □

Remark 5.10 **(Intuition and guidance for Theorem 5.5)**

(i) Regularity conditions R1–R4 allow us to derive the following

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x;\theta)dx \\
0 &= \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial\theta}dx \quad \text{(take partial w.r.t. } \theta \text{ on both sides)} \\
0 &= \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)/\partial\theta}{f(x;\theta)} \cdot f(x;\theta)dx \quad \text{(multiple and divide by } f(x;\theta)) \\
0 &= \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial\theta} \cdot f(x;\theta)dx \quad \left(\text{notice } \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{1}{f(x;\theta)} \cdot \partial f(x;\theta)/\partial\theta\right) \\
0 &= \mathbb{E}_\theta\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right] \qquad (\star)
\end{aligned}$$

In other words, the mean of the random quantity $\frac{\partial}{\partial\theta}\log f(X;\theta)$ is 0. Now, let us differentiate this quantity twice (R4) to obtain

$$\begin{aligned}
0 &= \int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} \cdot f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial\theta} \cdot \frac{\partial f(x;\theta)}{\partial\theta}dx \\
&= \int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} \cdot f(x;\theta)dx + \int_{-\infty}^{\infty} \left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2 \cdot f(x;\theta)dx \quad \text{(reference lines 2–4 above)}
\end{aligned}$$

Putting everything in terms of expectations, we see that

$$0 = \mathbb{E}_\theta\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right] + \mathbb{E}_\theta\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right] \qquad (\star\star)$$

(ii) *Score function*, $S(\theta)$:

$$S(\theta) := \frac{\partial}{\partial\theta}\log f(x;\theta)$$

The importance of the score function is it determines the estimating equation for $\hat{\theta}_{MLE}$,

$$\frac{\partial}{\partial\theta}\sum_{i=1}^n \log f(x_i;\theta) = 0 \iff \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i;\theta) = 0$$

(iii) *Fisher information*, $I(\theta)$:

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

Remark 5.11 **(Using Remark 5.10 to sketch the proof of Theorem 5.5)**

Proof. Let $\hat{\theta}$ denote our MLE and $\theta_0 \in \Omega$ is the true parameter. We know $S(\hat{\theta}) = 0$, where in this proof, $S(\theta) = \sum_{i=1}^n \log f(x_i;\theta)$. Now, recall Taylor's Theorem which states that if a function $f(x)$ is differentiable at $x = a$, then

$$f(x) = f(a) + f'(a)\cdot(x-a) + \frac{f''(a)}{2}\cdot(x-a)^2 + \ldots$$

Applying this to our scenario, we know

$$0 = S(\hat{\theta}) = S(\theta_0) + S'(\theta_0)\cdot(\hat{\theta}-\theta_0) + \frac{S''(\theta_0)}{2}\cdot(\hat{\theta}-\theta_0)^2 + R_n$$

Without going into too much detail (again see proof in Hogg & Craig Theorem 6.2.2), regularity condition (R5) lets $\frac{S''(\theta_0)}{2}\cdot(\hat{\theta}-\theta_0)^2 + R_n$ to go to 0. Then, we are left with

$$0 = S(\theta_0) + S'(\theta_0)\cdot(\hat{\theta}-\theta_0).$$

Recall that the goal is to show the asymptotic normality of $\hat{\theta}$, so if we want to apply CLT, we need something of the from $\sqrt{n}(\hat{\theta} - \theta_0)$. Using the Taylor approximation, we know

$$\begin{aligned} S'(\theta_0) \cdot (\hat{\theta} - \theta_0) &= -S(\theta_0) \\ \hat{\theta} - \theta_0 &= -\frac{S(\theta_0)}{S'(\theta_0)} \\ \sqrt{n}(\hat{\theta} - \theta_0) &= -\sqrt{n}\frac{S(\theta_0)}{S'(\theta_0)} \\ &= \frac{\sqrt{n}\left(\frac{1}{n}S(\theta_0)\right)}{\frac{1}{n} \cdot -S'(\theta_0)} \end{aligned}$$

In the numerator, we can apply the CLT because we showed in Remark 5.10(i) that $\mathbb{E}_\theta(S(\theta)) = 0$, therefore,

$$\sqrt{n}\left(\frac{1}{n}S(\theta_0) - 0\right) \xrightarrow{D} Y$$

where $Y \sim \mathcal{N}(0, \text{Var}(S(\theta_0)) \overset{D}{\equiv} \mathcal{N}(0, I(\theta_0))$. Then, if we write out the denominator, we see that we can apply WLLN

$$\frac{1}{n} \cdot -S'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta)\Big|_{\theta=\theta_0} \xrightarrow{P} \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)\Big|_{\theta=\theta_0}\right] = I(\theta_0)$$

Then, by application of Slutsky's Theorem, we see that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \frac{1}{I(\theta_0)} \cdot Y \sim \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$

□