Lecture 6:

Recall that our setting was:

Data: $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta_0)$

Here, θ_0 denotes the true parameter that generates the data.

Input and Goal: We only have access to a random sample $(X_1, \ldots, X_n)$, and we know that the data is generated according to a pdf $f(x|\theta)$. However, we don't know the true value of θ (i.e. we don't know θ_0) and our goal is to **estimate** θ_0 from the data $(X_1, \ldots, X_n)$. Formally speaking

we'd like to design an estimator $\hat{\theta}(x_1, \ldots, x_n)$ that gives us a good approximation for θ_0.

In the previous lecture, we introduced the Maximum Likelihood Estimator (mle): Given the sample data $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ we define the likelihood function

$$lik(\theta) = f(x_1, x_2, \ldots, x_n | \theta)$$

and the mle estimator, $\hat{\theta}_{mle}$, is given as:

$$\hat{\theta}_{mle} = \underset{\theta}{\arg\max}\ lik(\theta),$$

that is, $\hat{\theta}_{mle}$ is the value of θ that maximizes the likelihood, or equivalently, makes the observed data sample "most probable" or "most likely".

Note that when X_i's are ciid, we have

$$f(X_1, \dots, X_n | \theta) = f(X_1|\theta) \times f(X_2|\theta) \times \cdots \times f(X_n|\theta)$$
$$= \prod_{i=1}^{n} f(X_i|\theta)$$

So $$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(X_i|\theta)$$

Note also that since $\log(\cdot)$ is an increasing function, we can equivalently write:

$$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}} \log\left(\prod_{i=1}^{n} f(X_i|\theta) \right)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(X_i|\theta)$$

In summary, if we define

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$\ell(\theta)$ is called the log-likelihood function.

we have

$$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}} \ell(\theta)$$

Example: Assume $X_1, \ldots, X_n$ are generated iid according to exponential distribution with parameter λ_0. Derive the mle estimator for λ_0?

So in this example, we have:

$$X_1, \ldots, X_n \overset{iid}{\sim} f(x|\lambda_0)$$

$$= \lambda_0 \exp(-\lambda_0 x) \mathbb{1}\{x \geq 0\}$$

We don't know the true value λ_0, and we'd like to estimate it from data. We know, however, that data is generated according to $f(x|\lambda) = \lambda \exp(-\lambda x) \mathbb{1}\{x \geq 0\}$ for some (unknown) λ.

The log-likelihood function is:

$$\ell(\lambda) = \sum_{i=1}^{n} \log f(x_i|\lambda)$$

$$= \sum_{i=1}^{n} \log(\lambda e^{-\lambda x_i}) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

In order to find the maximizer of $\ell(\lambda)$ we need to solve the equation:

$$\ell'(\lambda) = 0$$

and the solution will be λ_{mle}.

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

$$\Rightarrow \quad \ell'(\lambda) = 0 \longrightarrow \lambda_{mle} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Note here that λ_{mle} is a **biased** estimate of λ_0 because

$$E[\lambda_{mle}] = E\left[\frac{1}{\bar{x}}\right] \neq \frac{1}{\lambda_0}.$$

From the previous example we learned at least two facts:

1) To find the mle estimate

$$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}}\ \ell(\theta)$$

we need to solve the equation

$$\ell'(\theta) = 0,$$

and θ_{mle} is a solution to this equation.

2) In general $\hat{\theta}_{mle}$ is not an unbiased estimator for the true parameter θ_0.

The next natural question that we should ask is how good is the mle estimator? Why should we trust that the mle estimator would give us a good estimate? We'll find an answer to these questions in the next section.

Large Sample Theory For mle:

In this section we'd like find out how good the mle estimator becomes as we increase the number of data samples n.

The first thing we sould expect is that $\hat{\theta}_{mle}$ will become close to the true value θ_0 as the number of samples n increases. Let's first verify this.

Lemma: $\lim_{n \to \infty} \hat{\theta}_{mle} = \theta_0$

The full proof of this lemma is beyond the scope of our class. Here, I'll provide the intuition. Recall that $\hat{\theta}_{mle}$ is the solution of the following:

$$\ell'(\hat{\theta}_{mle}) = \frac{d}{d\theta} \ell(\theta)\Big|_{\theta=\hat{\theta}_{mle}} = 0.$$

So if $\hat{\theta}_{mle}$ approaches θ_0 as n increases, we should expect that $\ell'(\theta_0)$ becomes very small (converges to zero)

as n increases. Instead of $\ell(\theta)$, let's look at a normalized version of it, $\frac{1}{n}\ell(\theta)$, and see how $\frac{1}{n}\ell'(\theta_0)$ behaves.

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta).$$

$$\Rightarrow \frac{1}{n}\ell'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\theta}\log f(x_i|\theta)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)}$$

Let $Y_i = \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)}$

then by the Law of Large Numbers (LLN) we have:

$\frac{1}{n}\sum_{i=1}^{n} Y_i \to E[Y]$

$$\xrightarrow{n\to\infty} E\left[\frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)}\right]$$

Now, recall that data is generated according to $x_i \sim f(x_i|\theta_0)$.

we have:

$$\frac{1}{n}\ell'(\theta_0) \xrightarrow{n\to\infty} E\left[\frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta_0)}\right]$$

$$= \int \frac{\frac{d}{d\theta} f(x|\theta_0)}{f(x|\theta_0)} f(x|\theta_0)\, dx$$

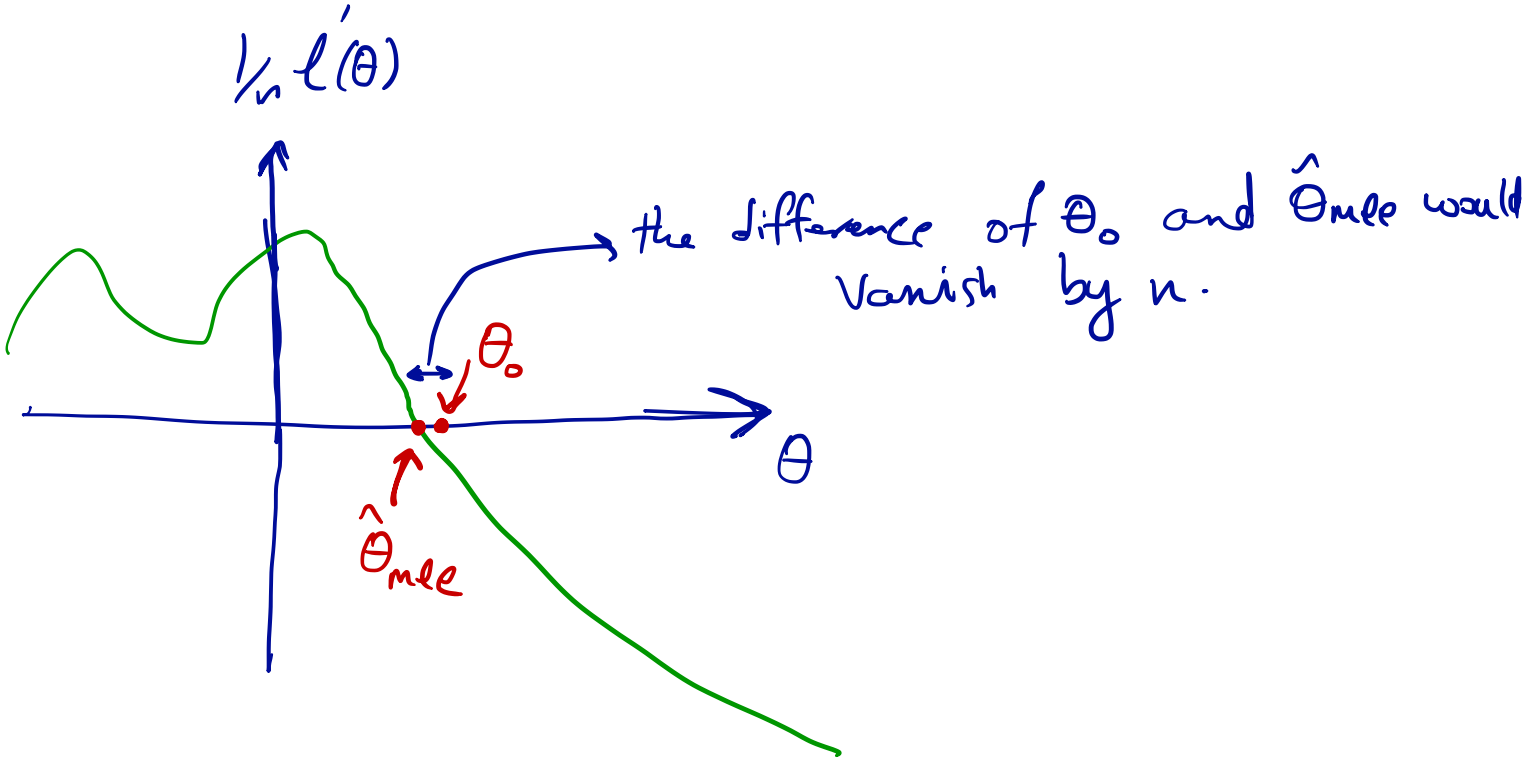
$$= \int \frac{d}{d\theta} f(x|\theta_0)\, dx$$

$$= \frac{d}{d\theta}\left(\int f(x|\theta_0)\, dx\right)$$

$$= \frac{d}{d\theta}(1) = 0$$

So, we have

$$\frac{1}{n} \ell'(\theta_0) \xrightarrow{n \to \infty} 0$$



So we now that θ_0 and $\hat{\theta}_{mle}$ will become closer and closer as n increases. We now want to refine this picture a litte further and see how small is the difference $\hat{\theta}_{mle} - \theta_0$ in terms of n.

Theorem (Central-limit type relation for $\hat{\theta}_{mle} - \theta_0$):

For n sufficiently large, we have:

normal distribution; zero-mean; variance is of order $1/n$

$$\hat{\theta}_{mle} - \theta_0 \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n\, I(\theta_0)}}\right),$$

where $I(\theta_0)$, is the so-called Fisher Information, given as follows: Assuming $X \sim f(x|\theta_0)$:

$$I(\theta_0) = \mathbb{E}\left[\left(\frac{d}{d\theta}\left(\log f(x|\theta_0)\right)\right)^2\right]$$

$$= \int \left(\frac{\frac{d}{d\theta} f(x|\theta_0)}{f(x|\theta_0)}\right)^2 f(x|\theta_0)\, dx.$$

proof: Again, the full proof is beyond the scope of this class and needs some extra smoothness assumptions on f. In the following, I will provide an intuitive sketch.

Recall that $\ell'(\hat{\theta}_{mle}) = 0$. Using Taylor Series, we can write:

$$0 = \ell'(\hat{\theta}_{mle}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta}_{mle} - \theta_0)$$

As a result, we have:

$$\hat{\theta}_{mle} - \theta_0 = -\frac{\ell'(\theta_0)}{\ell''(\theta_0)}$$

$$= -\frac{\frac{1}{n}\ell'(\theta_0)}{\frac{1}{n}\ell''(\theta_0)} \quad (*)$$

Let's try to compute both terms $\frac{1}{n}\ell'(\theta_0)$ and $\frac{1}{n}\ell''(\theta_0)$. We have:

$$\frac{1}{n}\ell'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\theta} f'(x_i|\theta_0)$$

$$\xrightarrow{CLT} \text{Normal}\left(\text{mean}, \sqrt{\frac{\text{Var}}{n}}\right)$$

where:

$$\text{mean} = E\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right]$$

$$= \int \frac{\frac{\partial}{\partial\theta} f(x|\theta_0)}{f(x|\theta_0)} f(x|\theta_0)\,dx = \int \frac{\partial}{\partial\theta} f(x|\theta_0)\,dx$$

$$= \frac{\partial}{\partial\theta}\left(\int f(x|\theta_0)\,dx\right)$$

$$= \frac{\partial}{\partial\theta}(1) = 0.$$

$$\Rightarrow \text{mean} = 0$$

$$\Rightarrow \text{Var} = E\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta_0)\right)^2\right] = I(\theta_0)$$

Hence,

$$\frac{1}{n}\ell'(\theta_0) \sim N\left(0, \sqrt{\frac{I(\theta_0)}{n}}\right). \quad (**)$$

Let's compute $\frac{1}{n}\ell''(\theta_0)$. We have:

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0)$$

$$\xrightarrow{\text{law of large numbers}} E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right].$$

We will prove later on that:

$$E\left[\frac{d^2}{d\theta^2}\log f(X|\theta_0)\right] = -I(\theta_0)$$

(to-be-proven)

Hence:

$$\frac{1}{n}\,\ell''(\theta_0) \longrightarrow -I(\theta_0) \qquad (***)$$

Let us now put together the three relations (*), (**), and (***), to obtain:

$$\hat{\theta}_{mle} - \theta_0 \sim N\left(0, \frac{1}{I(\theta_0)}\sqrt{\frac{I(\theta_0)}{n}}\right)$$

$$\sim N\left(0, \frac{1}{\sqrt{n\,I(\theta_0)}}\right).$$

And we get the result of the theorem.

Let us now show the relation (to-be-proven):

Recall that for every θ we have

$$\int f(x|\theta)\,dx = 1,$$

as $f(x|\theta)$ are probability density functions.

Let us define the function

$$g(\theta) = \int \frac{d}{d\theta}\left(\log f(x|\theta)\right) \cdot f(x|\theta)\, dx$$

note that

$$g(\theta) = \int \frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)} \cdot f(x|\theta)\, dx$$

$$= \int \frac{d}{d\theta} f(x|\theta) = \frac{d}{d\theta} \int f(x|\theta)\, dx$$

$$= \frac{d}{d\theta}(1) = 0$$

Hence, $g(\theta) = 0$ for every θ. As a result, we can also say that:

$$\frac{d}{d\theta} g(\theta) = 0, \text{ which gives:}$$

$$0 = \frac{d}{d\theta}\left[\int \left(\frac{d}{d\theta} \log f(x|\theta)\right) \cdot f(x|\theta)\, dx\right.$$

$$= \int \frac{d}{d\theta}\left[\left(\frac{d}{d\theta} \log f(x|\theta)\right) \cdot f(x|\theta)\right] dx$$

and hence, using the relation $(h_1 h_2)' = h_1' h_2 + h_2' h_1$, we have:

$$0 = \int \left(\frac{d^2}{d\theta^2} \log f(x|\theta) \right) f(x|\theta)\, dx + \int \left[\frac{d}{d\theta} \log f(x|\theta) \right]^2 f(x|\theta)\, dx$$

and hence:

$$\int \left(\frac{d^2}{d\theta^2} \log f(x|\theta) \right) f(x|\theta) = - \int \left(\frac{d}{d\theta} \log f(x|\theta) \right)^2 f(x|\theta)\, dx$$

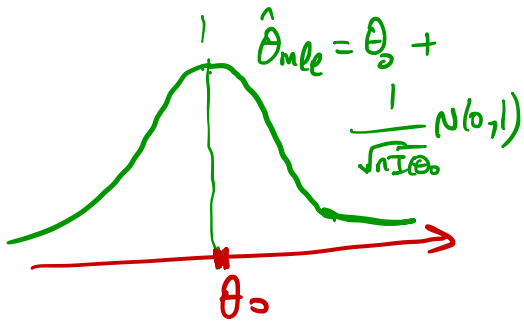
letting $\theta = \theta_0$ we get:

$$E\left[\frac{d^2}{d\theta^2} \log f(x|\theta_0) \right] = - I(\theta_0) .$$

Confidence intervals for mle:

We found from the previous section that

$$\hat{\theta}_{mle} - \theta_0 \sim \frac{1}{\sqrt{nI(\theta_0)}} N(0,1) \rightarrow$$



Let's now find β such that

$$Pr\left\{\hat{\theta}_{mle} - \beta \leq \theta_0 \leq \hat{\theta}_{mle} + \beta\right\} \approx 1 - \alpha$$

Let $\beta \triangleq \frac{\gamma}{\sqrt{nI(\theta_0)}}$, we have:

$$Pr\left\{\hat{\theta}_{mle} - \beta \leq \theta_0 \leq \hat{\theta}_{mle} + \beta\right\}$$

$$= Pr\left\{-\beta \leq \hat{\theta}_{mle} - \theta_0 \leq \beta\right\}$$

$$= Pr\left\{-\frac{\gamma}{\sqrt{nI(\theta_0)}} \leq \hat{\theta}_{mle} - \theta_0 \leq \frac{\gamma}{\sqrt{nI(\theta_0)}}\right\}$$

$$= Pr\left\{-\frac{\gamma}{\sqrt{nI(\theta_0)}} \leq \frac{1}{\sqrt{nI(\theta_0)}} N(0,1) \leq \frac{\gamma}{\sqrt{nI(\theta_0)}}\right\}$$

$$= \quad Pr\left\{ -\gamma \leq N(0,1) \leq \gamma \right\}$$

we know that by letting $\gamma = z_{\alpha/2}$

we have that

$$Pr\left\{ -z_{\alpha/2} \leq N(0,1) \leq +z_{\alpha/2} \right\} = 1-\alpha$$

$$\Rightarrow \quad \beta = z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\theta_0)}}$$

or

$$Pr\left\{ \hat{\theta}_{mle} - \frac{z_{\alpha/2}}{\sqrt{nI(\theta_0)}} \leq \theta_0 \leq \hat{\theta}_{mle} + \frac{z_{\alpha/2}}{\sqrt{nI(\theta_0)}} \right\} \approx 1-\alpha$$

Now, in reality, we don't know what θ_0 is (in fact we are estimating θ_0!). For finding confidence intervals when we only have access to data $(X_1, \ldots, X_n)$ we can use the estimate

$\hat{\theta}_{mle}$ instead of θ_0 (in the formula for the confidence interval). That is,

$$Pr\left\{\hat{\theta}_{mle} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{mle})}} \leq \theta_0 \leq \hat{\theta}_{mle} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{mle})}}\right\} \approx 1-\alpha$$

Cramer-Rao Bound:

Recall that for our estimators, we typically prefer the one which have smaller variance. In this regard, a fundamental question is what is the minimum possible variance in terms of the number of samples n. Given our setting $X_1, X_2, \ldots, X_n \sim f(x|\theta_0)$, the Cramer-Rao bound tells us that for any unbiased estimator of θ_0, call it $T(X_1, \ldots, X_n)$, the variance is bound by

$$Var(T) \geq \frac{1}{nI(\theta_0)}.$$

Recall that the mle estimator has the distribution (for n large):

$$\hat{\theta}_{mle} - \theta_0 \sim \frac{1}{\sqrt{nI(\theta_0)}} N(0,1)$$

So, when n is large, we have $\mathrm{Var}(\hat{\theta}_{mle}) = \frac{1}{nI(\theta_0)}$ which matches the Cramer-Rao lower bound. In this sense, $\hat{\theta}_{mle}$ has the minimum possible variance when n is large.

(An intuitive sketch of) the proof of Cramer-Rao:

Define the random variable

$$L = \sum_{i=1}^{n} \frac{d}{d\theta} \left(\log f(X_i | \theta_0) \right)$$

We saw before that:

$$E[L] = nE\left[\frac{d}{d\theta} \log f(X|\theta_0) \right]$$

$$= nE\left[\frac{\frac{d}{d\theta} f(X|\theta_0)}{f(X|\theta_0)} \right] = 0$$

Also,

$$\mathrm{Var}(L) = \mathrm{Var}\left(\frac{d}{d\theta}(\log f(X_1|\theta_0))\right) + \cdots + \mathrm{Var}\left(\frac{d}{d\theta}(\log f(X_n|\theta_0))\right)$$

$$= n\,\mathrm{Var}\left(\frac{d}{d\theta}(\log f(X|\theta_0))\right)$$

$$= n\,E\left[\left(\frac{\frac{d}{d\theta}f(X|\theta_0)}{f(X|\theta_0)}\right)^2\right]$$

$$= n\,I(\theta_0)$$

(note that if $Y_1, \ldots, Y_n$ are iid, then $\mathrm{Var}(Y_1+Y_2+\cdots+Y_n) = \mathrm{Var}(Y_1)+\cdots+\mathrm{Var}(Y_n) = n\,\mathrm{Var}(Y_1)$)

So, to summarize:

$$\begin{cases} E[L] = 0 \\ \mathrm{Var}(L) = n I(\theta_0). \end{cases}$$

Now, take any unbiased estimator of θ_0; i.e.

$E[T] = \theta_0$, when $X_i \overset{iid}{\sim} f(X|\theta_0)$ for any value of θ_0.

We will first show that $E[L \cdot T] = 1$.

We have:

$$E[L \cdot T]$$

$$= \int_{(x_1,\dots,x_n)} T(x_1,\dots,x_n) \cdot L(x_1,\dots,x_n)\, f(x_1,\dots,x_n \mid \theta_0)\, dx_1 dx_2 \cdots dx_n$$

$$= \int_{(x_1,\dots,x_n)} T(x_1,\dots,x_n)\, L(x_1,\dots,x_n) \prod_{i=1}^{n} f(x_i \mid \theta_0)\, dx_i$$

$$= \int_{(x_1,\dots,x_n)} T(x_1,\dots,x_n) \left(\sum_{i=1}^{n} \frac{\frac{d}{d\theta} f(x_i \mid \theta_0)}{f(x_i \mid \theta_0)} \right) \prod_{i=1}^{n} f(x_i \mid \theta_0)\, dx_i$$

$$= \int_{(x_1,\dots,x_n)} T(x_1,\dots,x_n)\, \frac{d}{d\theta} \left(\prod_{i=1}^{n} f(x_i \mid \theta_0) \right) dx_1 \cdots dx_n$$

$$= \frac{d}{d\theta} \left(\int_{(x_1,\dots,x_n)} T(x_1,\dots,x_n) \cdot \prod_{i=1}^{n} f(x_i \mid \theta_0)\, dx_1 \cdots x_n \right)$$

$$= \frac{d}{d\theta} \left(\theta \right) \Big|_{\theta=\theta_0} = 1$$

note that $\frac{d}{d\theta} \prod_{i=1}^{n} f(x_i \mid \theta) = \sum_{i=1}^{n} \frac{\frac{d}{d\theta} f(x_i \mid \theta)}{f(x_i \mid \theta)} \prod_{i=1}^{n} f(x_i \mid \theta)$

Now that we have $E[L \cdot T] = 1$, we can use the Cauchy–Schwartz inequality to deduce that:

$$\left(E[L \cdot T]\right)^2 \leq E[L^2] \cdot E[T^2]$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

$$1 \leq n I(\theta_0) \cdot \mathrm{Var}(T)$$

$$\Rightarrow \mathrm{Var}(T) \geq \frac{1}{n I(\theta_0)}.$$