Max-Likelihood estimator Lecture Notes

Gleb Pantileev

March 2025

Contents

1 Preface

The origin of this article was quite usual. During my regular studies, I have faced a problem with one of the properties of MLE. I asked several people, but no one could give me an answer, so together with my teacher, Yaroslav Alexandrovich, we decided to make this project.

2 Introduction

In this work I want to understand why Max-Likelihood estimators are widely used and prove their properties. In order to do so I will use (and firstly prove) Cramer-Rao lower bound, Kullback-Leibler distance and Cauchy-Swartz inequality (without proof).

3 Basic of estimators

3.1 Definitions

First of all, it is important to understand what is meant by population parameters, estimators, etc.

Population parameters- parameters which define the distribution of the random variable: $X \sim f(x, \overrightarrow{\theta})$. For Normal distribution, these parameters are μ, σ^2, for Binomial: n and p.

Statistic- any function of the sample $(X_1, \ldots, X_n)$. For example, $\bar{X}, s^2, X_i, 0, g(X_1, \ldots, X_n)$

Estimator- statistic used as an approximation for the population parameter θ, based on $(X_1, \ldots, X_n)$-random sample. Whereas estimate- estimator based on sample obtained $(x_1, \ldots, x_n)$.

3.2 Basic Properties

3.2.1 Bias

Estimator is called unbiased if $E(\hat{\theta}) = \theta$. If not, then bias of the statistics can be found as $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

3.2.2 Efficiency

Efficiency of the estimator is connected with its Mean Squared Error (MSE).

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + Bias^2(\hat{\theta})$$

Consider two estimators: $\hat{\theta}$ and $\tilde{\theta}$. Then, if $MSE(\hat{\theta}) < MSE(\tilde{\theta})$, $\hat{\theta}$ is more efficient estimator of θ. According to Cramer-Rao inequality theorem:

$$Var(\hat{\theta}_n) \geqslant \frac{(1 + bias(\hat{\theta}))^2}{I_n(\theta)}, \text{ where } I_n(\theta) = Var(\frac{dL(\overrightarrow{X}, \theta)}{d\theta})$$

But about this later... (proof in the appendix (3))

3.2.3 Consistency

Consistency is an asymptotic property which means that estimator tends to the value of the true parameter in probability as n (size of the sample) tends to infinity.
$\hat{\theta}$ is consistent estimator if θ if:

$$P(|\hat{\theta} - \theta| > \epsilon) \to 0 \ \forall \epsilon > 0 \iff \hat{\theta} \xrightarrow{P} \theta$$

3.2.4 Sufficient conditions fo consistency

Theorem states that if MSE($\hat{\theta}$) $\to 0$ as n $\to \infty$ then $\hat{\theta}$ is a consistent estimator of θ.

$$MSE(\hat{\theta}) \to 0 \text{ as n} \to \infty \iff \begin{cases} E(\hat{\theta}) \to \theta \\ Var(\hat{\theta}) \to 0 \end{cases}$$

In order to prove this property, Chebychev's inequality should be used : $P(|\hat{\theta} - \theta| \geqslant \epsilon) \leqslant \frac{MSE(\hat{\theta})}{\epsilon^2}$ (proof is in the appendix(1)). So as $MSE(\hat{\theta}) \to 0$, $P(|\hat{\theta} - \theta| \geqslant \epsilon) \to 0$ too.

4 Maximum-Likelihood Estimator

4.1 Way to obtain

The method of finding the Max-Likelihood estimator is next:

Knowing the distribution of X_i: $X \sim f(x, \overrightarrow{\theta})$, with unknown $\overrightarrow{\theta}$ and getting the sample of n observations $(X_1, \ldots, X_n)$ find such $\hat{\theta}_{ml}$ which maximises the probability of such sample.

Consider Likelihood function $L(\theta)$:

$$L(\overrightarrow{\theta}) = \log f(\overrightarrow{x}, \overrightarrow{\theta})$$

However, life is not such simple and there could be two possible cases:

- $\hat{\theta}_{ml}$ can be obtained by equating the derivative of Likelihood Function with respect to parameter to 0.
- But if it impossible to do, $\hat{\theta}_{ml}$ should be obtained analytically.

Consider two examples:

First:

$$X_i \sim N(\mu, \sigma^2) \ \sigma \text{ is known} \Rightarrow f(\overrightarrow{X}, \mu) = \prod_{i=1}^{n} f(X_i, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(X_i-\mu)^2}{2\sigma^2}}$$

$$L(\mu) = \log(f(\overrightarrow{X}, \mu)) = \frac{-n}{2} \cdot \log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}$$

$$\frac{dL}{d\mu} = \frac{\sum_{i=1}^{n} X_i - n \cdot \hat{\mu}}{\sigma^2} = 0 \Rightarrow \hat{\mu}_{ml} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

Second:

$$Y_i \sim U[0, \theta] \text{ need to estimate } \theta \ f(y, \theta) = \begin{cases} \frac{y}{\theta}, \ \text{if } 0 \leqslant \text{y} \leqslant \theta \\ 0, \text{ otherwise} \end{cases}$$

$$f(\overrightarrow{Y}, \theta) = \prod_{i=1}^{n} f(Y_i, \theta) = \prod_{i=1}^{n} \frac{Y_i}{\theta}$$

$$L(\theta) = \sum_{i=1}^{n} \log(Y_i) - n \log \theta$$

$$\frac{dL}{d\theta} = -\frac{n}{\theta}$$

This case is more inconvenient since θ is non-negative and thus derivative of likelihood function is negative($L(\theta)$ constantly decreases with respect to θ), thus, in order to be maximized, θ should be minimized.

$$\hat{\theta}_{ml} \rightarrow min_{s.t. \theta \geqslant X_i} \Rightarrow \hat{\theta}_{ml} = X_{max}$$

4.2 Properties of MLE

4.2.1 Consistency

When it is said that estimator is consistent, it means that as size of the sample is large enough, it converges to the value of true parameter in distribution.

$$\hat{\theta}_n \xrightarrow{P} \theta$$

In order to prove this property I would use Kullback-Leibler distance between f and g, where f,g- pdfs:

$$D(f,g) = \int f(x) \cdot \log(\frac{f(x)}{g(x)})dx$$
$$D(f,g) \geqslant 0 \text{ (proof in the appendix(2))}$$
$$D(\theta,\phi) \equiv D(f(x,\theta), f(x,\phi)) \ \forall \theta, \phi \in \Theta$$

Let $\theta^\star$- true value of θ

Consider new function $M(\theta)$:

$$M_n(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta^\star)} = \frac{1}{n}(L_n(\theta) - L_n(\theta^\star))$$
$$\text{thus if } M_n(\theta) \to max \Leftrightarrow L_n(\theta) \to max$$

By Law of Large Numbers:

$$M_n(\theta) \to M(\theta) = E_{\theta^\star}(\log \frac{f(X_i,\theta)}{f(X_i,\theta^\star)}) =$$
$$\int \log \frac{f(x,\theta)}{f(x,\theta^\star)} f(x,\theta^\star)dx = -\int \log \frac{f(x,\theta^\star)}{f(x,\theta)} f(x,\theta^\star)dx = -D(\theta^\star,\theta)$$
$$M(\theta) = -D(\theta^\star,\theta) \leqslant 0 \Rightarrow M(\theta) \ max \ at \ \theta = \theta^\star$$
$$M_n(\theta) \ max \ at \ \theta = \hat{\theta}_{ml}, \ M_n(\theta) \to M(\theta), \ thus \ \hat{\theta}_{ml} \to \theta^\star$$

But this proof is dirty, now consider more formal proof! Stay hard) Uniform convergence should be proved (for all values of $\theta \in \Theta$) estimator converges to the value of true parameter.

$M(\theta) = -D(\theta^\star, \theta) \leqslant 0$
$M(\theta^\star) = -D(\theta^\star, \theta^\star) = 0$

$M(\theta) = E_{\theta^\star}(M_n) = E_{\theta^\star}(M_i)$

$M_n(\theta) = \frac{\sum_{i=1}^n M_i}{n} = \frac{\sum_{i=1}^n \log \frac{f(X_i,\theta)}{f(X_i,\theta^\star)}}{n} = \frac{\log f(\vec{x},\theta) - \log f(\vec{x},\theta^\star)}{n}$
$M_n(\theta^\star) = 0$
Assumptions made:

1) $M_n(\theta)$ converges to $M(\theta)$ $\forall \theta \in \Theta$:
$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$

2) For all $\theta \neq \theta^\star$ $M(\theta)$ is less than the value of $M(\theta^\star)$:
$\sup_{\theta:|\theta-\theta^\star|\geqslant\epsilon} M(\theta) < M(\theta^\star)$ $\forall \epsilon > 0$

3) For the finite samples $\hat{\theta}_{ml}$ determine the distribution better (due to a "bad sample" with outliers or due to a small size).
$M_n(\hat{\theta}_{ml}) \geqslant M_n(\theta^\star)$ (for finite samples only)

Proof:

$$M(\theta^\star) - M(\hat{\theta}_{ml}) = M_n(\theta^\star) - M(\hat{\theta}_{ml}) + M(\theta^\star) - M_n(\theta^\star) \leqslant M_n(\hat{\theta}_{ml}) - M(\hat{\theta}_{ml}) + M(\theta^\star) - M_n(\theta^\star) =$$
$$= M_n(\hat{\theta}_{ml}) - M(\hat{\theta}_{ml}) \leqslant \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$$
$$\text{Thus, } M(\theta^\star) - M(\hat{\theta}_{ml}) \xrightarrow{P} 0 \Leftrightarrow P(|M(\theta^\star) - M(\hat{\theta}_{ml}| > \delta) \to 0 \ \forall \delta > 0$$

Now consider two events:

$$A = \left\{|\hat{\theta}_{ml} - \theta^\star| > \epsilon\right\} \ \forall \epsilon > 0$$
$$B = \left\{M(\theta^\star) - M(\hat{\theta}_{ml}) > 0\right\}$$

If event A happens, then $\hat{\theta}_{ml} \neq \theta^\star$ so under the second assumption it is true that $M(\hat{\theta}_{ml}) < M(\theta^\star)$ (B happens too)$\Rightarrow$ A- subset of B $\Rightarrow$ $P(A) \leqslant P(B)$

$$P(|\hat{\theta}_{ml} - \theta^\star| > \epsilon) \leqslant P(M(\theta^\star) - M(\hat{\theta}_{ml}) > 0) \to 0$$

Thus $P(|\hat{\theta}_{ml} - \theta^\star| > \epsilon) \to 0$ - proved.

4.2.2 Invariance

The invariance property means that if $\hat{\theta}_{ml}$- max-likelihood estimator of θ, then $g(\hat{\theta}_{ml}) = \hat{g}_{ml}$- max-likelihood estimator of $g(\theta)$.

Proof:

• Let $L(\hat{\theta})$- is maximum of $L(\theta)$: $\frac{dL}{d\theta}(\hat{\theta}) = 0$

• Let $\theta = g(\eta)$, so in order to get the exact value of η, function g should be invertible: $\eta = g^{-1}(\theta)$.

$$Let\ \frac{dL}{d\eta} = \frac{\partial L}{\partial g(\eta)} \cdot \frac{dg(\eta)}{d\eta} = 0 \Rightarrow$$

$$\left[\begin{array}{l} \dfrac{dL}{dg(\eta)} = 0 \\ \dfrac{dg(\eta)}{d\eta} = 0 \end{array}\right.$$

Consider $g(x)$ such that its derivative does not equal to 0 at any point (so there would be no lie points). Then $\frac{dL}{dg(\eta)} = 0$ at some certain point $\hat{\eta}$. So $\hat{\eta}$ is a maximiser of likelihood function and there is only one such point, then $\hat{\theta}$=$g(\hat{\eta})$. Taking the inverse function $(g^{-1}(x))$ of both sides, we obtain $\hat{\eta} = g^{-1}(\hat{\theta})$. Overall, limitations of the function: g(ϵ) should be invertible and strictly monotonic (its derivative should not be equal to 0 at any point, otherwise it would be necessary to check the obtained point analytically).

It is a very useful property because usually it is necessary to estimate probability of event, which is a function of θ also. For example, Poisson distribution:
$X \sim Pois(\lambda), \hat{\lambda}_{ml} = \bar{X}$, and $P(X = 5) = \frac{\lambda^5}{5!}e^{-\lambda}$, so $\hat{\lambda}_{ml} = \bar{X}$ can be used to estimate this probability:

$$\hat{p} = \frac{\hat{\lambda}_{ml}^5}{5!}e^{-\hat{\lambda}_{ml}} = \frac{\bar{X}^5}{5!}e^{-\bar{X}}$$

4.2.3 Asymptotically Normal

Refresh previous results:
$S_n(\theta) \equiv S(\overrightarrow{X}, \theta)$
$E(S_n) = 0$
$I_n(\theta) \equiv Var(S_n) = E(-S'_n)$(see appendix(2))

Now let's approximate the $S_n(\theta)$ function by Taylor series:

$$S_n(\theta) = S_n(\theta_0) + S'_n(\theta_0)(\theta - \theta_0) + o(\theta - \theta_0)$$

Let $\theta = \hat{\theta}_n$ (approximate the value of $S_n(\theta)$ at point $\hat{\theta}_n$)
Let $\theta_0 = \theta$- true value of parameter.

$$S_n(\hat{\theta}_n) = S_n(\theta) + S'_n(\theta)(\hat{\theta}_n - \theta) + o(\hat{\theta}_n - \theta 1)$$

Now consider the case in which $\hat{\theta}_{ml}$ is obtained by nullifying the $S_n(\theta)$. In this case $S_n(\hat{\theta}_{ml})$=0

$$0 = S_n(\theta) + S'_n(\theta)(\hat{\theta}_n - \theta) + o(\hat{\theta}_n - \theta)$$

Due to a consistency of $\hat{\theta}_n$ the first order precision is enough since $|\hat{\theta}_n - \theta| \xrightarrow{P} 0$

$$\hat{\theta}_n - \theta \approx \frac{S_n(\theta)}{-S'_n(\theta)}$$

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}} S_n(\theta)}{-\frac{1}{n} S'_n(\theta)}$$

$\frac{1}{\sqrt{n}} S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\theta) = \sqrt{n} \cdot \bar{S}$, then by CLT:

$$\frac{\bar{S} - E(\bar{S})}{Var(\bar{S})} \sim N(0, 1)$$

$E(\bar{S}) = E(S_i) = 0$
$Var(\bar{S}) = \frac{1}{n^2} Var(S_n) = \frac{1}{n^2} \cdot I_n(\theta) = \frac{1}{n} I(\theta)$,
Thus,

$$\bar{S} \sim N(0, \frac{1}{n} I(\theta)) \Rightarrow \sqrt{n} \cdot \bar{S} \sim N(0, I(\theta))$$

$-\frac{1}{n}S_n'(\theta) = -\frac{1}{n}\sum_{i=1}^n S_i'(\theta) = -\overline{S'}$ by CLT is distributed

$$-\overline{S'} \sim N(E(-\frac{1}{n}S_n'(\theta)), Var\ (-\frac{1}{n}S_n'(\theta))) = N(E(-\overline{S'}), Var(-\overline{S'}))$$

$E(-\overline{S'}) = E(-S_i') = I(\theta)$
$Var(-\overline{S'}) = \frac{1}{n} \cdot Var(S_i'(\theta)) \to 0$ as $n \to \infty$.

$$\text{Finally,} \frac{1}{\sqrt{n}}S_n(\theta) \sim N(0, I(\theta))$$
$$-\frac{1}{n}S_n'(\theta) \xrightarrow{P} I(\theta)$$

Then, by theorem from probability theory and statistics if $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow const$, then $X_n \cdot Y_n \rightsquigarrow cX$

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}}S_n(\theta)}{-\frac{1}{n}S_n'(\theta)} \rightsquigarrow \frac{N(0, I(\theta))}{I(\theta)} = N(0, \frac{1}{I(\theta)})$$

Meaning that $\hat{\theta}_n \sim N(\theta, I_n^{-1}(\theta))$

So, the unicorn has been proved, but this is not useful because in the real world, true value of θ cannot be obtained and used. So the question arises, having only estimator for theta, how can we use it in calculating its variance?

Def se- standard error of the estimator: $se = \sqrt{\frac{1}{I_n(\theta)}}$ and $\hat{se}$- Max-likelihood estimator of se: $\hat{se}$=$\sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$(That is my guess, maybe will be proved later)). Then, let's find the distribution of the estimator using $\hat{se}$:

$$\hat{\theta}_n \sim N(\theta, se^2)$$

$$\frac{\hat{\theta}_n - \theta}{\hat{se}} = (\hat{\theta} - \theta) \cdot \sqrt{n} \cdot \sqrt{I(\hat{\theta}_n)} = \sqrt{n} \cdot (\hat{\theta} - \theta) \cdot \sqrt{I(\theta)} \cdot \frac{\sqrt{I(\hat{\theta}_n)}}{\sqrt{I(\theta)}} =$$

$= |$since $\hat{\theta}_n$ is a consistent estimator of θ, then $\hat{\theta}_n \xrightarrow{P} \theta$; assume $I(\theta)$ is a continuous function, so $\lim\limits_{\hat{\theta}_n \to \theta} I(\hat{\theta}_n) \xrightarrow{P} I(\theta)| =$

$$\Rightarrow \frac{\sqrt{I(\hat{\theta}_n)}}{\sqrt{I(\theta)}} \xrightarrow{P} 1 \text{ and } \sqrt{n} \cdot (\hat{\theta} - \theta) \cdot \sqrt{I(\theta)} \rightsquigarrow N(\theta, I_n^{-1}(\theta))$$

It means that $\sqrt{n} \cdot (\hat{\theta} - \theta) \cdot \sqrt{I(\theta)} \cdot \frac{\sqrt{I(\hat{\theta}_n)}}{\sqrt{I(\theta)}} \sim N(0, 1)$

$$\hat{\theta}_n \sim N(\theta, I_n^{-1}(\hat{\theta}_n)) = N(\theta, \hat{se}^2)$$

Example: in order to prove what we have just found, I decided to use MLE for the population mean ($\bar{X}$), which distribution parameters each of us knows:

$$X_i \sim N(\mu, \sigma^2), \sigma \text{ is known} \Leftrightarrow f(x_i, (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\text{Need } I(\overrightarrow{\theta}) : S_i(\mu) = \frac{\partial}{\partial\mu}(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{(X_i-\mu)^2}{2\sigma^2}) = \frac{X_i-\mu}{\sigma^2}$$

$$-S_i'(\mu) = -\frac{\partial}{\partial\mu}(\frac{X_i-\mu}{\sigma^2}) = \frac{1}{\sigma^2}$$

$$\text{Recall the previous result: } \hat{\theta}_n \sim N(\theta, \frac{1}{n \cdot I(\theta)}) \Rightarrow \overline{X} \sim N(\mu, \frac{1}{n \cdot \frac{1}{\sigma^2}}) = N(\mu, \frac{\sigma^2}{n})$$

$$\text{Also, since } \hat{\theta}_n \sim N(\theta, I_n^{-1}(\hat{\theta}_n)), \text{then } \overline{X} \sim N(\mu, \frac{\hat{\sigma}^2_{ml}}{n})$$

this is written in blue, because I'm not sure about it.=)

4.2.4 Asymptotically efficient

If an estimator has this property, it means that its variance is the smallest among all other estimators (as $n \to \infty$)
Consider two estimators of parameter θ: Max-Likelihood estimator ($\hat{\theta}_{ml}$) and another estimator $\tilde{\theta}$. Asymptotical efficiency means that:

$$Var(\hat{\theta}_{ml}) \leqslant Var(\tilde{\theta}) \ \forall \tilde{\theta} \in \Theta$$

In order to prove this property, it is enough to recall the Cramer-Rao lower bound inequality:

$$Var(\hat{\theta}_n) \geqslant \frac{(1 + bias'(\theta))^2}{I_n(\theta)}$$

Then, as it was found before, if $S_n(\hat{\theta}) = 0$, then $\hat{\theta}_n \sim N(\theta, \frac{1}{I_n(\theta)})$, meaning that this estimator is unbiased and its variance is equal to the Craner-Rao lower bound. However, there is a piece of uncertainty, which I want to explain in my future work: consider a harder example of the $\hat{\theta}_{ml}$ for the bound of uniform distribution(it was calculated in "Way to obtain" part of the notes): $X_i \sim U[0, \theta]$, then, $\hat{\theta}_{ml}$=X_{max}, but at this point $S_n(\theta) \neq 0$. So it changes the proof of the asymptomatically Normal property, but I have not understood, how this should be done... YET!!!

5 Appendix

(1) Proof of Chebychev's inequality

Consider g(x)- non-decreasing, non-negative function

$$E(g(x)) = \int_{-\infty}^{+\infty} g(x)f(x)dx = \int_{g(x)\geqslant\epsilon} g(x)f(x)dx + \int_{g(x)<\epsilon} g(x)f(x)dx$$

$$\int_{g(x)\geqslant\epsilon} g(x)f(x)dx + \int_{g(x)<\epsilon} g(x)f(x)dx \geqslant \epsilon \cdot \int_{g(x)\geqslant\epsilon} f(x)dx + \int_{g(x)<\epsilon} g(x)f(x)dx =$$

$$= |\text{since g(x)-non-negative, second integral is a positive constant}| =$$

$$E(g(x)) \geqslant \epsilon \cdot \int_{g(x)\geqslant\epsilon} f(x)dx = \epsilon \cdot P(g(x) \geqslant \epsilon) \Rightarrow P(g(x) \geqslant \epsilon) \leqslant \frac{E(g(x))}{\epsilon}$$

$$Let\ g(x) = |\hat{\theta} - \theta|,\ then\ P(|\hat{\theta} - \theta| \geqslant \epsilon) = P((\hat{\theta} - \theta)^2 \geqslant \epsilon^2) \leqslant \frac{E(\hat{\theta} - \theta)^2}{\epsilon} = \frac{MSE(\hat{\theta})}{\epsilon}$$

(2) Proof of $D(f,g) \geqslant 0$

$$D(f,g) = \int \log\frac{f(x)}{g(x)} f(x)dx = \int -\log\frac{g(x)}{f(x)} f(x)dx = E(-\log\frac{g(x)}{f(x)})$$

Consider y=-ln(x)- convex function, then, by Jensen's inequality:

$$E(y(x)) \geqslant y(E(x))$$

$$E(-\log\frac{g(x)}{f(x)}) \geqslant -\log E(\frac{g(x)}{f(x)})$$

$$D(f,g) \geqslant -log(\int \frac{g(x)}{f(x)} \cdot f(x)dx$$

$$D(f,g) \geqslant -log(1) = 0$$

(3) Proof of Cramer-Rao lower bound inequality:

- Consider the score function: $S_n(\theta) = S(\overrightarrow{X},\theta) \equiv \frac{\partial}{\partial\theta}\log f(\overrightarrow{X},\theta) = n \cdot S_i(\theta)$
- $I_n(\theta) = Var(S_n(\theta)) = n \cdot Var(S_i(\theta)) = nI(\theta)$, where $I(\theta) \equiv I_1(\theta)$

First of all, need to find $E(S_n(\theta))$:

$$1 = \int f(\vec{x},\theta)d\vec{x} \Leftrightarrow 0 = \frac{\partial}{\partial\theta}\int f(\vec{x},\theta)d\vec{x} = |\text{under regularity conditions}| = \int \frac{\partial}{\partial\theta} f(\vec{x},\theta)d\vec{x} =$$

$$= |\frac{\partial}{\partial\theta}\log f(\vec{x},\theta) = \frac{\frac{\partial f}{\partial\theta}}{f} \Leftrightarrow f\frac{\partial}{\partial\theta}\log f(\vec{x},\theta) = \frac{\partial f}{\partial\theta}| = \int \frac{\partial}{\partial\theta}\log f(\vec{x},\theta) f(\vec{x},\theta)d\vec{x} = E(S_n(\theta))$$

Proof for the most common case: biased estimator of n observations $\hat{\theta}_n(\vec{x}) \equiv \hat{\theta}_n$

$$\text{According to Cauchy-Swartz inequality: } Cov^2(a,b) \leqslant Var(a) \cdot Var(b)$$
$$Cov^2(\hat{\theta}_n, S_n(\theta)) = (\int (\hat{\theta}_n - E(\hat{\theta}_n))(S_n(\theta) - E(S_n(\theta)))d\vec{x})^2 =$$
$$= (\int (\hat{\theta}_n - E(\hat{\theta}_n))(S_n(\theta) - E(S_n(\theta)))d\vec{x})^2 \leqslant \int (\hat{\theta}_n - E(\hat{\theta}_n))^2 d\vec{x} \cdot \int (S_n(\theta) - E(S_n(\theta)))^2 d\vec{x} = Var(\hat{\theta}_n) \cdot I_n(\theta)$$

$$Cov(\hat{\theta}, S_n(\theta)) = E(\hat{\theta}_n \cdot S_n) - E(\hat{\theta}_n) \cdot E(S_n) = |E(S_n) = 0| = E(\hat{\theta}_n \cdot S_n)$$
$$E(\hat{\theta}_n \cdot S_n) = \int \hat{\theta}_n \cdot S_n f(\vec{x})d\vec{x} = \int \hat{\theta}_n \cdot \frac{\partial}{\partial\theta} \log f(\vec{X}) \cdot f(\vec{x})d\vec{x} = \int \hat{\theta}_n \cdot \frac{\partial f(\vec{x})}{\partial\theta}\, d\vec{x}$$
$$\text{Under regularity conditions: } \int \hat{\theta}_n \cdot \frac{\partial f(\vec{x})}{\partial\theta}\, d\vec{x} = \frac{\partial}{\partial\theta} \int \hat{\theta}_n \cdot f(\vec{x})d\vec{x} = \frac{\partial}{\partial\theta} E(\hat{\theta}_n)$$

Since $\hat{\theta}_n$ could be a biased estimator of θ, then $E(\hat{\theta}_n) = \theta + bias(\hat{\theta}_n)$,

where $bias(\hat{\theta}_n)$ is a function of θ, so:

$$Cov(\hat{\theta}, S_n(\theta)) = \frac{\partial}{\partial\theta} E(\hat{\theta}_n) = \frac{\partial}{\partial\theta}(\theta + bias(\hat{\theta}_n)) = 1 + bias'(\hat{\theta}_n)$$
$$\text{Finally, } Cov^2(\hat{\theta}, S_n(\theta)) = (1 + bias'(\hat{\theta}_n))^2 \leqslant Var(\hat{\theta}_n) \cdot I_n(\theta)$$
$$Var(\hat{\theta}_n) \geqslant \frac{(1 + bias'(\hat{\theta}_n))^2}{I_n(\theta)}$$

Another interesting fact about the score function: Let's calculate the Var(S_n):

$$Var(S_n) = E(S_n)^2 - E^2(S_n) = E(S_n)^2 = E(\frac{\partial}{\partial\theta} \log f(\vec{x}))^2$$

In order to obtain an interesting result, we have to make some rearrangements:

$$\frac{\partial^2}{\partial\theta^2} \log f = \frac{\partial}{\partial\theta}(\frac{\frac{\partial f}{\partial\theta}}{f}) = \frac{\frac{\partial^2 f}{\partial\theta^2} \cdot f - (\frac{\partial f}{\partial\theta})^2}{f} = \frac{\partial^2 f}{\partial\theta^2} \cdot \frac{1}{f} - (\frac{\frac{\partial f}{\partial\theta}}{f})^2 = \frac{\partial^2 f}{\partial\theta^2} \cdot \frac{1}{f} - (\frac{\partial}{\partial\theta} \log f)^2 \Rightarrow$$
$$E(\frac{\partial}{\partial\theta} \log f)^2 = \int (\frac{\partial}{\partial\theta} \log f)^2 d\vec{x} = \int \frac{\partial^2 f}{\partial\theta^2} \cdot \frac{1}{f} \cdot f\, d\vec{x} - \int \frac{\partial^2}{\partial\theta^2} \log f \cdot f\, d\vec{x} =$$
$$= |\text{Under regularity conditions}| = \frac{\partial^2}{\partial\theta^2} \int f\, d\vec{x} + E(-\frac{\partial^2}{\partial\theta^2} \log f) = E(-\frac{\partial^2}{\partial\theta^2} \log f)$$

And we obtained the unexpected result: on average the second derivative of the likelihood function is negative and thus it is a concave function so if it has a critical point, it is a point of maximum. Also, if $S_n(\vec{x}, \theta)$- continuous function, then the point of maximum will be the only one.