14. Maximum Likelihood Estimation

ECE 830, Spring 2014

Maximum Likelihood Estimation

MLEs are a very important type of estimator for the following reasons:

- MLE occurs naturally in composite hypothesis testing and signal detection (i.e., GLRT)
- The MLE is often simple and easy to compute
- MLEs often have asymptotic optimal properties (consistency (MSE $\to 0$ as $N \to \infty$) and efficiency (achieves CRLB))
- MLEs are invariant under reparameterization

Estimation Using the Likelihood

Definition: Likelihood function

$p(x|\theta)$ as a function of θ with x fixed is called the "likelihood function".

If the likelihood function carries the information about θ brought by the observation x, how do we use it to obtain an estimator?

Definition: Maximum Likelihood Estimation

$$\widehat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta \in \Theta} p(x|\theta)$$

is the value of θ that maximizes the density at x. Intuitively, we are choosing θ to maximize the probability of occurrence for x.

ML Estimation and Density Estimation

ML Estimation is equivalent to Density Estimation.

Assume

$$X_i \overset{\text{iid}}{\sim} p, \quad i = 1, \cdots, n, \quad p \in \{p_\theta\}_{\theta \in \Theta}$$

The ML Estimation is equivalent to finding the density in $\{p_\theta\}_{\theta \in \Theta}$ that best fits the data. i.e., the generative model with the highest density/probability value at the point x.

Computing the MLE

If the likelihood function is differentiable, then $\widehat{\theta}$ is found from

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = 0$$

If multiple solutions exist, then the MLE is the solution that maximizes $\log p(x|\theta)$. That is, take the global maximizer.

Note: It is possible to have multiple global maximizers that are all MLEs!

Example: DC Level in Gaussian white noise

$$x_n = A + w_n, \qquad w_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \qquad n = 1, \cdots, N$$

$$\theta = [A, \sigma^2]^T$$

$$\frac{\partial \log p(x|\theta)}{\partial A} =$$

$$\frac{\partial \log p(x|\theta)}{\partial \sigma^2} =$$

$$\Rightarrow \widehat{A} =$$

$$\Rightarrow \widehat{\sigma^2} =$$

Note: $\widehat{\sigma^2}$ is biased!

MLE and Linear Models

Example:

$$\begin{aligned} X &\sim \mathcal{N}(H\theta, \Sigma), \quad \theta \in \mathbb{R}^k, \quad \Sigma, H \text{ known} \\ p(x|\theta) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x - H\theta)^T \Sigma^{-1}(x - H\theta)\} \end{aligned}$$

The value of $\widehat{\theta}$ is given by,

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\theta} -\log p(x|\theta) \\ &= \arg\min_{\theta} (x - H\theta)^T \Sigma^{-1}(x - H\theta) \\ &= \end{aligned}$$

Note: $\widehat{\theta}$ is also the MVUE estimator in this special case!

Invariance of MLE

Suppose we wish to estimate the function $g = G(\theta)$ and not θ itself. Intuitively we might try

$$\widehat{g} = G(\widehat{\theta})$$

where $\widehat{\theta}$ is the MLE of θ.

Remarkably, it turns out that $\widehat{g}$ is the MLE of g.

This very special invariance principle is summarized in the following theorem.

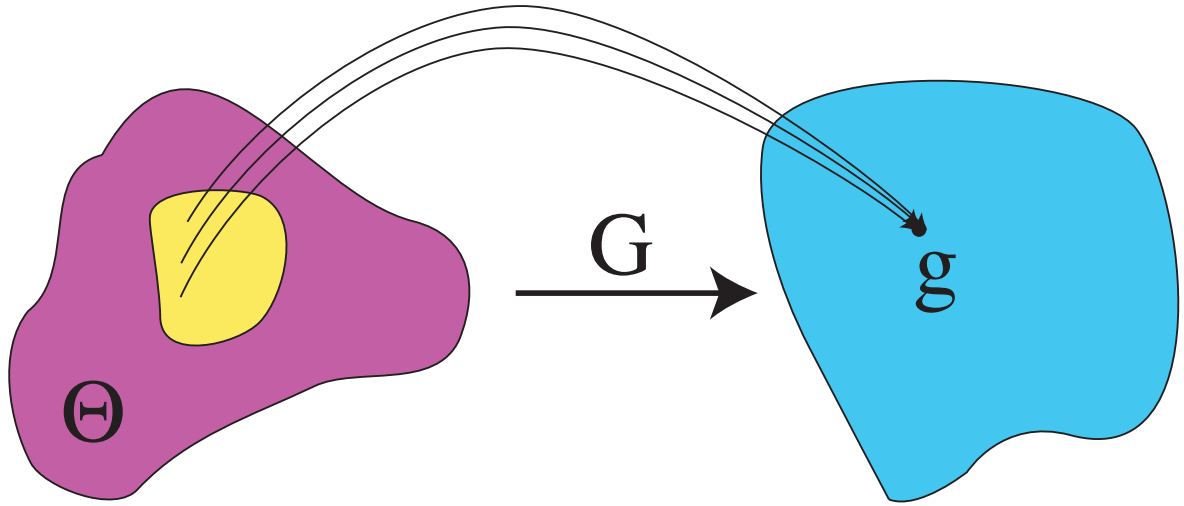
Theorem: Invariance of the MLE

Let $\widehat{\theta}$ denote the MLE of θ. Then $\widehat{g} = G(\widehat{\theta})$ is the MLE of $g = G(\theta)$.

Proof:

Define the "induced" log likelihood function:

$$L(x|g) \equiv \max_{\theta:G(\theta)=g} \log p(x|\theta)$$



The MLE of g is

$$\begin{aligned}
\widehat{g} &= \arg\max_{g} L(x|g) \\
&= \arg\max_{g} \max_{\theta:G(\theta)=g} \log p(x|\theta) \\
&= G(\widehat{\theta}) \quad , \quad \text{where } \widehat{\theta} = \text{MLE of } \theta
\end{aligned}$$

Example:

Let $x = [x_1, \cdots, x_N]^T$ where $x_i \sim \text{Poisson}(\lambda)$. Given x, find the MLE of the probability that $x \sim \text{Poisson}(\lambda)$ exceeds the mean λ.

$$\begin{aligned} G(\lambda) &= \mathbb{P}(x > \lambda) \\ &= \end{aligned}$$

The MLE of g is

where $\widehat{\lambda}$ is the MLE of λ:

ML Estimation as Loss Minimization

$$\widehat{\theta}_{\mathsf{MLE}} \quad = \quad \arg\min_{\theta} \frac{1}{p(x|\theta)} = \arg\min_{\theta} -\log p(x|\theta)$$

MLE and loss minimization

We can view the MLE as minimizing the loss

$$\ell_{\mathsf{MLE}}(\theta^*, \theta) := -\log p(x(\theta^*)|\theta)$$

where dependence on θ^* is embodied in $x \sim p_{\theta^*}$. The expected loss is

$$\begin{aligned} R_{\mathsf{MLE}}(\theta^*, \theta) &= \mathbb{E}[\ell_{\mathsf{MLE}}(\theta^*, \theta)] \\ &= \mathbb{E}\left[-\log p(x|\theta)\right] \\ &= \int p(x|\theta^*)\left(-\log p(x|\theta)\right) dx \end{aligned}$$

Excess Risk ("Regret")

Let θ be any value of the parameter and θ^* be the true value that generates x. Then we can compare

$$R_{\mathsf{MLE}}(\theta^*, \theta) - R_{\mathsf{MLE}}(\theta^*, \theta^*)$$

which quantifies how much larger the expected loss is when we use θ instead of θ^*. Note that

$$\begin{aligned}
R_{\mathsf{MLE}}(\theta^*, \theta) - R_{\mathsf{MLE}}(\theta^*, \theta^*) &= \mathbb{E}\left[\log p(x|\theta^*) - \log p(x|\theta)\right] \\
&= \mathbb{E}\left[\log \frac{p(x|\theta^*)}{p(x|\theta)}\right] \\
&= \int p(x|\theta^*) \left(\log \frac{p(x|\theta^*)}{p(x|\theta)}\right) dx \\
&= \\
&\geq 0
\end{aligned}$$

with equality if $\theta = \theta^*$

Multiple iid observations

In general

$$X_i \overset{\text{iid}}{\sim} p(x|\theta^*), \quad \theta^* \in \Theta, \quad i = 1, \cdots, n$$

Then for any $\theta \in \Theta$

Loss: $$\ell(\theta^*, \theta) = -\log\left(\prod_{i=1}^{n} p(x_i|\theta)\right) = -\sum_{i=1}^{n} \log p(x_i|\theta)$$

MLE: $$\widehat{\theta} = \arg\min_{\theta} -\sum_{i=1}^{n} \log p(x_i|\theta)$$

Excess Risk: $$R_{\mathsf{MLE}}(\theta^*, \theta) - R_{\mathsf{MLE}}(\theta^*, \theta^*) = nD(p(x|\theta^*)\|p(x|\theta))$$

Convergence of log likelihood to KL

Suppose $X_i \overset{\text{iid}}{\sim} p(x|\theta^*)$, then by the strong law of large numbers for any fixed $\theta \in \Theta$

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)} \xrightarrow{\text{a.s.}} D\left(p(x|\theta^*)\|p(x|\theta)\right)$$

We would like to show that the MLE

$$\widehat{\theta}_n = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta)$$

converges to θ^* in the following sense:

$$D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) \longrightarrow 0$$

Note that since $\widehat{\theta}_n$ maximizes $\sum_{i=1}^n \log p(x_i|\theta)$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} \leq 0$$

Thus we have

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} - D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) + D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) \leq 0$$

$$\implies D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) \leq D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) - \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)}$$

So $D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right) \longrightarrow 0$ if

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} \longrightarrow D\left(p(x|\theta^*)\|p(x|\widehat{\theta}_n)\right)$$

The subtle issue here is that $\widehat{\theta}_n$ is a random variable, not a fixed $\theta \in \Theta$, so we cannot just appeal to the SLLN.

Convergence of log likelihood to KL

Assume $X_i \overset{\text{iid}}{\sim} p(x|\theta^*), i = 1, \cdots, n$. Define

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)}, \quad \forall \theta \in \Theta$$

$$L(\theta) := \mathbb{E}\left[L_n(\theta)\right]$$

Suppose the following assumptions hold:

$$\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\text{P}} 0 \tag{A1}$$

$$\inf_{\theta : \|\theta - \theta^*\| \geq \epsilon} L(\theta) > L(\theta^*), \quad \forall \epsilon > 0 \tag{A2}$$

then

$$\|\widehat{\theta}_n - \theta^*\| \xrightarrow{\text{P}} 0.$$

A1 says that the LR converges uniformly (wrt θ) to the KL divergence. A2 says that locally θ^* is strictly better (in KL) than θ.

Proof: Since $\widehat{\theta}_n$ minimizes $L_n(\theta)$ we have

$$L_n(\widehat{\theta}_n) \leq L_n(\theta^*)$$

Hence,

$$\begin{aligned} L(\widehat{\theta}_n) - L(\theta^*) &= L(\widehat{\theta}_n) - L_n(\theta^*) + L_n(\theta^*) - L(\theta^*) \\ &\leq L(\widehat{\theta}_n) - L_n(\widehat{\theta}_n) + L_n(\theta^*) - L(\theta^*) \\ &\leq \sup_\theta |L(\theta) - L_n(\theta)| + L_n(\theta^*) - L(\theta^*) \\ &\xrightarrow{\mathrm{P}} 0, \quad \text{by A1} \end{aligned}$$

It follows that for any $\delta > 0$

$$\mathbb{P}\left(L(\widehat{\theta}_n) > L(\theta^*) + \delta\right) \longrightarrow 0, \quad \text{as} \quad n \longrightarrow \infty$$

Now pick any $\epsilon > 0$. By A2 $\exists \delta > 0$ such that

$$\|\theta - \theta^*\| \geq \epsilon \quad \Rightarrow \quad L(\theta) > L(\theta^*) + \delta$$

Hence

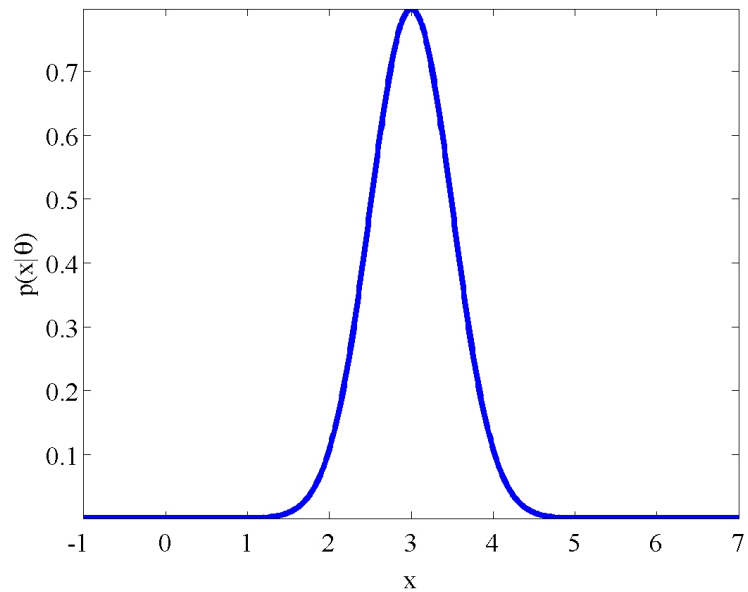
$$\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| \geq \epsilon) \leq \mathbb{P}(L(\widehat{\theta}_n) > L(\theta^*) + \delta) \longrightarrow 0$$

□

We just saw that under reasonable assumptions $\widehat{\theta}_n$ converges in probability to the value θ^* that generated the observations. Next we will study the asymptotic distribution of $\widehat{\theta}_{\text{MLE}}$.

Estimator Accuracy

Consider the likelihood function $p(x|\theta)$ where θ is a scalar unknown (parameter).



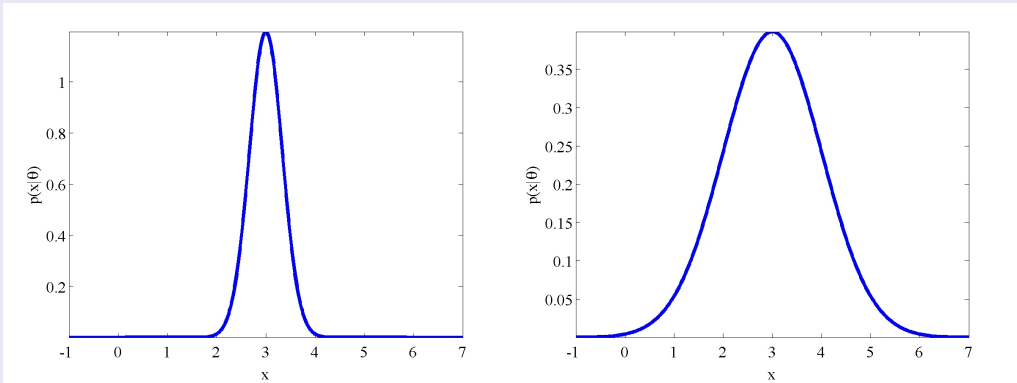
We can plot the likelihood as a function of the unknown. The more "peaky" or "spiky" the likelihood function, the easier it is to determine the unknown parameter.

Example:

Suppose we observe

$$x = A + w$$

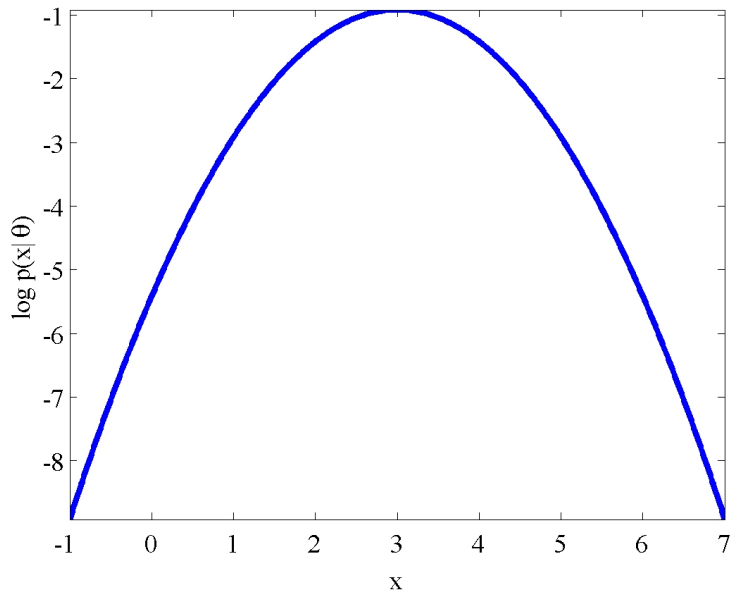
where $w \sim \mathcal{N}(0, \sigma^2)$ and A is an unknown parameter. The "smaller" the noise w is, the easier it will be to estimate A from the observation x.



$A = 3$ and $\sigma = 1/3$ $\quad$ $A = 3$ and $\sigma = 1$

Given the left density function we can easily rule out estimates of A greater than 4 or less than 2, since it is very unlikely that such A could give rise to our observation. On the other hand, when $\sigma = 1$ the noise power is larger and it is very difficult to estimate A accurately.

The key thing to notice is that the estimation accuracy of A depends on σ, which in effect determines the peakiness of the likelihood. The more peaky, the better localized the data is about the true parameter.



The peakiness is effectively measured by the negative of the second derivative of the log-likelihood at its peak.

Example:

$$x = A + w, \qquad w \sim \mathcal{N}(0, \sigma^2)$$

$$\log p(x|A) = -\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}(x - A)^2$$

$$\frac{\partial \log p(x|A)}{\partial A} =$$

$$-\frac{\partial^2 \log p}{\partial A^2} =$$

Curvature increases as σ^2 decreases (curvature = peakiness)

Fisher Information

In general, the curvature will depend on the observation data:

$$-\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \quad \text{is a function of } x$$

Thus an average measure of curvature is more appropriate.

$$-\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$$

The expectation averages out randomness due to the data and is a function of θ alone.

Definition: Fisher Information

$$I(\theta) := \mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$$

is the *Fisher Information*. Here the derivative is evaluated at the true value of θ and the expectation is with respect to $p(x|\theta)$.

Asymptotic Distribution of MLE

Let $x_1, \ldots, x_n$ be iid observations from $p(x|\theta^*)$, where $\theta^* \in \mathbb{R}^d$,

$$L_n(\theta) := \sum_{i=1}^{n} \log p(x_i|\theta) \qquad \text{and} \qquad \widehat{\theta}_n = \arg\max_{\theta} \; L_n(\theta),$$

assume $\frac{\partial L_n(\theta)}{\partial \theta_j}$ and $\frac{\partial^2 L_n(\theta)}{\partial \theta_j \partial \theta_k}$ exist for all j, k, and the "regularity contition" $\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right] = 0$ for all θ holds. Then

$$\widehat{\theta}_n \overset{\text{asymp.}}{\sim} \mathcal{N}(\theta^*, n^{-1} I^{-1}(\theta^*))$$

where $I(\theta^*)$ is the Fisher-Information Matrix (FIM), whose elements are given by

$$[I(\theta^*)]_{j,k} = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta_j \partial \theta_k}\Big|_{\theta=\theta^*}\right]$$

Regularity condition

The regularity condition amounts to assuming that we can interchange order of differentiation and integration to compute

$$\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right] = \int \frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(x|\theta^*)dx$$

$$= \int \frac{1}{p(x|\theta^*)} \frac{\partial p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(x|\theta^*)dx$$

$$=$$

since $\int p(x|\theta)dx = 1$ for all θ and the derivative of a constant is 0. The last line, where integration and differentiation are interchanged, is only possible for "regular" likelihood functions. This is simply the Fundamental Theorem of Calculus applied to $p(x|\theta)$. As long as $p(x|\theta)$ is absolutely continuous w.r.t. Lebesgue measure (i.e., when the derivative is well-defined), this is possible.

This is true for many distributions, but not true when the support of X depends on θ (e.g. $X \sim \text{Unif}(0, \theta)$).

Note:

$$\mathbb{E}[\widehat{\theta}] \to \theta^*$$

$$\text{Cov}(\widehat{\theta}) \to \frac{1}{n} I^{-1}(\theta^*)$$

$\Rightarrow \widehat{\theta}$ is consistent and efficient asymptotically (i.e., asymptotically achieves CRLB)

Example:

$$x \sim \mathcal{N}(A \cdot \mathbf{1}_{N\times 1}, \sigma^2 I_{N\times N})$$

$$\theta = [A, \sigma^2]^T$$

$$\widehat{A} = \frac{1}{N}\sum_{n=1}^{N} x_n \sim$$

$$s := \sum_{n=1}^{N} \frac{(x_n - \widehat{A})^2}{\sigma^2} \sim \chi^2_{N-1}$$

$$\widehat{\sigma^2} = \left(\frac{\sigma^2}{N}\right) s$$

Example: (cont.)

For large N, the CLT tells us that

$$\chi^2_N \approx \mathcal{N}(N, 2N).$$

Therefore,

$$s \approx \mathcal{N}(N-1, 2(N-1)) \leftarrow \text{approximately distributed}$$

Hence,

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{A})^2 = \frac{\sigma^2}{N}s$$

$$\approx$$

Example: (cont.)

Moreover, for large N

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}\right] &= \begin{bmatrix} A \\ \frac{N-1}{N}\sigma^2 \end{bmatrix} \to \begin{bmatrix} A \\ \sigma^2 \end{bmatrix} = \theta^* \\
C_{\widehat{\theta}} &= \begin{bmatrix} \sigma^2/N & 0 \\ 0 & \frac{2(N-1)\sigma^4}{N^2} \end{bmatrix} \to \begin{bmatrix} \sigma^2/N & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} \\
&= I^{-1}(\theta^*) \leftarrow \text{ inverse Fisher Info. Matrix}
\end{aligned}$$

Hence,

$$\widehat{\theta} \sim \mathcal{N}(\theta^*, I^{-1}(\theta^*)) \text{ for large } N$$

Proof:

We will prove the theorem for the special case when θ is scalar. The proof for multidimensional vectors follows the same steps using multivariable calculus. By the mean value theorem,

$$\frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}_n} = \frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} + \frac{\partial^2 L_n(\theta)}{\partial \theta^2}\Big|_{\theta=\tilde{\theta}}(\widehat{\theta}_n - \theta^*),$$

where $\tilde{\theta}$ is some value between θ^* and $\widehat{\theta}_n$. By definition, $\frac{\partial L_n(\theta)}{\partial \theta}|_{\theta=\widehat{\theta}_n} = 0$, so

$$0 = \frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} + \frac{\partial^2 L_n(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}\ (\widehat{\theta}_n - \theta^*).$$

From the equation above we have

$$\widehat{\theta}_n - \theta^* \quad = \quad -\frac{\frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*}}{\frac{\partial^2 L_n(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}}$$

Next consider $\sqrt{n}(\widehat{\theta}_n - \theta^*)$[1]. From above we have

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) = -\frac{\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*}}{\frac{1}{n}\frac{\partial^2 L_n(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}}.$$

[1] The reason scaling the difference by $\sqrt{n}$ is that this is the normalization needed to stabilize the limiting distribution. For example, if $x_1, \ldots, x_n$ were iid observations from the distribution $N(\theta^*, 1)$, then it is easy to see that $\sqrt{n}(\widehat{\theta}_n - \theta^*) \sim N(0, 1)$.

First let's study the numerator.

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log p(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}$$

Central Limit Theorem

If $z_1, \ldots, z_n$ are iid random variables with $\mathbb{E}[z_1] = \mu$ and $\mathbb{E}[(z_1 - \mu)^2] = \sigma^2$, then $\frac{1}{\sqrt{n}} \sum_i z_i \xrightarrow{D} \mathcal{N}(\mu/\sqrt{n}, \sigma^2)$, meaning the the random variable defined by summation has a distribution that tends to the Gaussian as $n \to \infty$.

Therefore, by the CLT we have

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} \xrightarrow{D} \mathcal{N}\left(\mathbb{E}\left[\frac{\frac{1}{\sqrt{n}} \partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} \right], \text{var}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right.\right.$$

By assumption,

$$\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right] = 0.$$

Since the mean is zero, the variance is

$$\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right)^2\right].$$

The variance can be related to the curvature of the log-likelihood function as follows. First observe that

$$\begin{aligned}\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial \theta}\right) \\ &= -\frac{1}{p^2(x|\theta)}\left(\frac{\partial p(x|\theta)}{\partial \theta}\right)^2 + \frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial \theta^2}\end{aligned}$$

Now let's take the expectation

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^*}\right] \\
&= -\int \left(\frac{1}{p(x|\theta^*)}\frac{\partial p(x|\theta)}{\partial\theta}\Big|_{\theta=\theta^*}\right)^2 p(x|\theta^*)\,dx \\
&\quad + \int \frac{1}{p(x|\theta^*)}\frac{\partial^2 p(x|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^*} p(x|\theta^*)\,dx \\
&= -\int \left(\frac{\partial \log p(x|\theta)}{\partial\theta}\Big|_{\theta=\theta^*}\right)^2 p(x|\theta^*)\,dx + \int \frac{\partial^2 p(x|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^*} dx \\
&= -\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial\theta}\Big|_{\theta=\theta^*}\right)^2\right] + \frac{\partial^2}{\partial\theta^2}\left(\int p(x|\theta)\,dx\right)\Big|_{\theta=\theta^*}.
\end{aligned}
$$

Since $\int p(x|\theta)\,dx = 1$ the second term is 0.

Therefore, the variance is equal to the negative expected curvature:

$$\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\Big|_{\theta=\theta^*}\right] = I(\theta^*).$$

Now consider the denominator. By the Strong Law of Large Numbers (SLLN), this average converges to its mean value, the negative Fisher Information, almost surely:

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 L_n(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log p(x_i|\theta)}{\partial \theta^2}\Big|_{\theta=\theta^*} \\ &\xrightarrow{\text{a.s.}} \mathbb{E}\left[\frac{\partial^2 \log p(x_i|\theta)}{\partial \theta^2}\Big|_{\theta=\theta^*}\right] \\ &= -I(\theta^*).\end{aligned}$$

To summarize, the numerator converges in distribution to a Gaussian

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} \xrightarrow{D} \mathcal{N}(0, I(\theta^*)),$$

and the denominator $\frac{1}{n}\frac{\partial^2 L_n(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}} \xrightarrow{a.s.} -I(\theta^*)$. So for large n, the numerator behaves like a Gaussian random variable and the denominator is almost constant. The ratio therefore converges in distribution to a Gaussian rescaled by the limiting constant of the denominator

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \quad \xrightarrow{D} \quad \frac{1}{I(\theta^*)}\mathcal{N}(0, I(\theta^*)) \equiv \mathcal{N}(0, I^{-1}(\theta^*)).$$

This type of convergence is rigorously proved by *Slutsky's Theorem* (for more information see `http://en.wikipedia.org/wiki/Slutsky's_theorem`). □

Finding the MLE

In certain cases (e.g., exponential family of parameterized distributions) the MLE can be easily solved for. That is,

$$\frac{\partial}{\partial \theta} \log p(x|\theta) = 0$$

can be solved using calculus and standard linear algebra.

In general, however, we may have to resort to more advanced numerical maximization techniques:

- Newton-Raphson Iteration
- Iteration by the Scoring Method

$$\theta_{k+1} = \theta_k + \left[I^{-1}(\theta) \, \frac{\partial \log p(x|\theta)}{\partial \theta} \right]_{\theta=\theta_k}$$

- Expectation-Maximization Algorithm