

# ML Estimation

Aleksandar Dogandžić

2017-02-16

## Contents

Introduction and Examples	1
ML Decoding	2
Invariance to invertible parameterizations	3
Properties of ML Estimators for I.I.D. Measurements	4
Consistency	5
Asymptotic distribution	6
ML Estimation of Vector Parameters	9
Conditions for asymptotic Gaussian distribution	9
Sensor array processing	10
Delta Method	11
Computing ML Estimates	12
Newton-Raphson iteration	13
Concentrated Likelihood	16
Summary of Classical Estimation	17

READING: §7 and 8 in the textbook and (Hero 2015, §4.4 and 4.5), (Rao 1973, pp. 364–366, Ch. 5f.2) and (Ferguson 1996) for properties.

## Introduction and Examples

SOLVE the following optimization problem:

$$\hat{\theta}(x) = \arg \max_{\theta} f_{X|\Theta}(x | \theta).$$

\* COMMENTS:

- $f_{X|\Theta}(x | \theta)$ , viewed as function of  $\theta$ , is the *likelihood function* of  $\theta$ .
- For a given  $\theta$  and discrete case,  $p_{X|\Theta}(x | \theta)$  is the probability of observing the point  $x$ . In the continuous case,  $f_{X|\Theta}(x | \theta)$  is approximately proportional to probability of observing a point in a small rectangle around  $x$ . However, when we think of  $p_{X|\Theta}(x | \theta)$

or  $f_{X|\Theta}(x|\theta)$  as functions of  $\theta$ , they give, for a given observed  $x$ , the *likelihood or plausibility* of various  $\theta$ .

- The maximum-likelihood (ML) estimate of  $\theta$  is value of the parameter  $\theta$  that *makes the probability of the data as great as it can be under the assumed model*.
- When finding the ML estimator, examine the boundary of the parameter space<sup>1</sup> to check if the global maximum of the likelihood function lies on this boundary.

<sup>1</sup> if such a boundary exists

\* EXAMPLE.

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}(x - H\theta)^T \Sigma^{-1}(x - H\theta)\right]$$

where  $\Sigma$  is a known positive definite covariance matrix. The ML estimate of  $\theta$  is

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} (x - H\theta)^T \Sigma^{-1}(x - H\theta) \\ &= (H^T \Sigma^{-1} H)^{-1} H^T \Sigma^{-1} x.\end{aligned}\tag{1}$$

In this example, the ML estimator of  $\theta$  coincides with its minimum-variance unbiased (MVU) and best linear unbiased estimator (BLUE) estimators.

- \* EXAMPLE. Consider  $X = (X[n])_{n=0}^{N-1}$  conditionally independent, identically distributed (i.i.d.)  $\mathcal{N}(\theta, \sigma^2)$  given  $\theta$ , where  $\theta$  is the unknown parameter and  $\sigma^2 > 0$  is a known constant. Maximize the log-likelihood function  $\ln f_{X|\Theta}(x|\theta)$  with respect to  $\theta$ , where

$$\ln f_{X|\Theta}(x|\theta) = \text{const} - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \theta)^2.$$

const denotes terms that are not functions of  $\theta$ .

The ML estimate is the sample mean

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$$

which is a special case of (1).

## ML Decoding

For a symmetric channel, the ML decoder is the minimum Hamming distance decoder.

*Proof.*  $x$  and  $\theta$  are the received and transmitted vectors from a binary symmetric channel. The elements of  $x$  and  $\theta$  are zeros and ones. Note

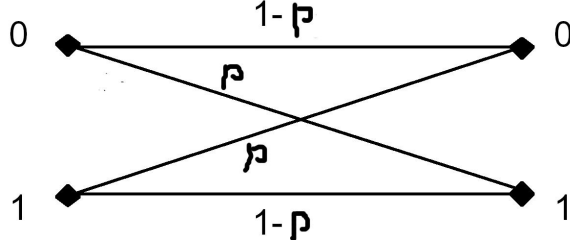


Figure 1: Symmetric channel.

that  $\theta$  belongs to a finite set of codewords, denoted by  $\text{sp}_{\Theta}$ . We wish to find which  $\theta$  was transmitted based on the received  $x$ . We have

$$X = \theta + W \pmod{2} \triangleq \theta \oplus W$$

where  $W = (W)_{n=0}^{N-1}$  and  $W[n]$  are i.i.d. Bernoulli random variables taking value 1 with probability  $p$ :

$$\Pr_W(W[n] = 1) = p.$$

The likelihood function of  $\theta$  for data  $x$  is given by

$$\begin{aligned} p_{X|\Theta}(x|\theta) &= \Pr_{X|\Theta}(X = x|\theta) \\ &= \Pr_W(\theta \oplus W = x) \\ &= \Pr_W(W = x \oplus \theta) \\ &= p^{\sum_{n=0}^{N-1} x[n] \oplus \theta[n]} (1-p)^{N - \sum_{n=0}^{N-1} x[n] \oplus \theta[n]} \\ &= \left( \frac{p}{1-p} \right)^{d_H(x, \theta)} (1-p)^N \end{aligned}$$

where

$$d_H(x, \theta) = \sum_{n=0}^{N-1} x[n] \oplus \theta[n]$$

is the *Hamming distance* between  $x$  and  $\theta$ , i.e., the number of bits that are different between the two vectors. Hence, if  $p < 0.5$ , then  $\max_{\theta \in \text{sp}_{\Theta}} p_{X|\Theta}(x|\theta)$  is equivalent to

$$\min_{\theta \in \text{sp}_{\Theta}} d_H(x, \theta).$$

□

Invariance to invertible parameterizations

THE ML estimate of  $\alpha = g(\theta)$  for an invertible function  $g(\cdot)$  is

$$\hat{\varphi} = g(\hat{\theta})$$

where  $\hat{\theta}$  is the ML estimate of  $\theta$ , obtained by maximizing the likelihood function  $f_{X|\Theta}(x|\theta)$  or  $p_{X|\Theta}(x|\theta)$  with respect to  $\theta$ .

$\oplus$  stands for the exclusive or (XOR) operator

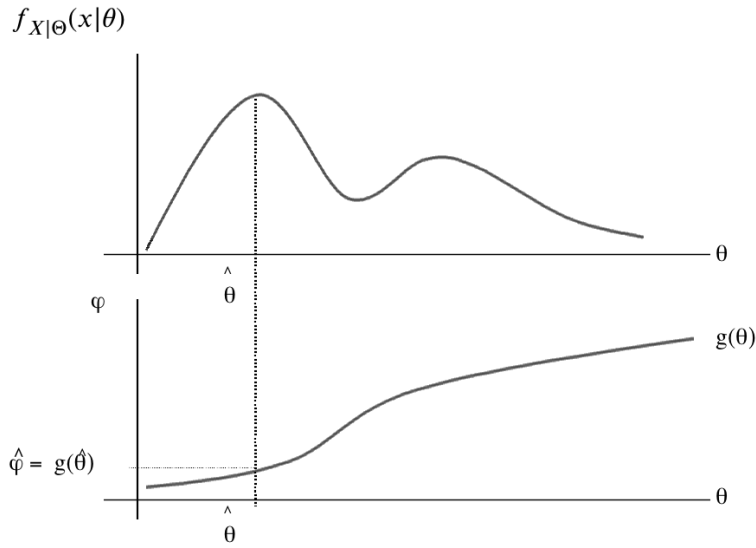


Figure 2: Invariance of ML estimators to invertible functional transformation  $g(\cdot)$ .

- ✱ **EXAMPLE.** i.i.d. measurements  $(X[n])_{n=0}^{N-1}$  follow the  $\text{Poisson}(\lambda)$  probability mass function (pmf). Find the ML estimate of the probability that a measurement  $X \sim \text{Poisson}(\lambda)$  is greater than  $\lambda$ .

$$\begin{aligned}\varphi &= g(\lambda) = \Pr_{X|\Lambda}(X > \lambda \mid \lambda) \\ &= \sum_{k=\lfloor \lambda + 1 \rfloor}^{+\infty} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= 1 - \sum_{k=0}^{\lfloor \lambda \rfloor} e^{-\lambda} \frac{\lambda^k}{k!}.\end{aligned}$$

The ML estimate of  $\varphi$  is

$$\hat{\varphi} = 1 - \sum_{k=0}^{\lfloor \hat{\lambda} \rfloor} e^{-\hat{\lambda}} \frac{\hat{\lambda}^k}{k!}$$

where  $\hat{\lambda}$  is the ML estimate of  $\lambda$ :

$$\hat{\lambda} = \bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n].$$

### Properties of ML Estimators for I.I.D. Measurements

WE outline proofs of properties of ML estimators under the i.i.d. measurement model.

## Consistency

We have collected  $N$  i.i.d. observations  $\mathbf{x} = (x[n])_{n=0}^{N-1}$  from a distribution  $f_{X|\Theta}(x|\theta^*)$ , where  $\theta^*$  denotes the true value of the parameter  $\theta$ .

Consider the following log likelihood ratio:

$$\text{LR}_N(\theta, X) = \frac{1}{N} \sum_{n=0}^{N-1} \ln \frac{f_{X|\Theta}(X[n]|\theta^*)}{f_{X|\Theta}(X[n]|\theta)}, \quad \theta \in \text{sp}_{\Theta}$$

and define its expectation

$$\begin{aligned} \overline{\text{LR}}(\theta) &= \mathbb{E}_{X|\Theta}[\text{LR}_N(\theta, X) | \theta^*] \\ &= \mathbb{E}_{X|\Theta} \left[ \ln \frac{f_{X|\Theta}(X[0]|\theta^*)}{f_{X|\Theta}(X[0]|\theta)} \mid \theta^* \right] \\ &= \int \ln \frac{f_{X|\Theta}(x|\theta^*)}{f_{X|\Theta}(x|\theta)} f_{X|\Theta}(x|\theta^*) dx \\ &= D(f_{X|\Theta}(x|\theta^*) \parallel f_{X|\Theta}(x|\theta)). \end{aligned}$$

Kullback-Leibler (KL) divergence from  $f_{X|\Theta}(x|\theta^*)$  to  $f_{X|\Theta}(x|\theta)$

Suppose that the following assumptions hold:

**Assumption 1.** The log likelihood ratio converges uniformly (with respect to  $\theta$ ) to the KL divergence as  $N \nearrow +\infty$ :

$$\sup_{\theta \in \text{sp}_{\Theta}} |\text{LR}_N(\theta, X) - \overline{\text{LR}}(\theta)| \xrightarrow{\text{P}} 0.$$

**Assumption 2.** Locally, the true value  $\theta^*$  of the parameter is strictly better (in KL divergence) than  $\theta$ :

$$\inf_{\theta: \|\theta - \theta^*\| \geq \epsilon} \overline{\text{LR}}(\theta) > \overline{\text{LR}}(\theta^*), \quad \forall \epsilon > 0.$$

Then, the ML estimator of  $\theta$

$$\begin{aligned} \hat{\theta}_N(X) &= \hat{\theta}_N(X) \\ &= \arg \max_{\theta} \prod_{n=0}^{N-1} f_{X|\Theta}(X[n]|\theta) \\ &= \arg \min_{\theta} \text{LR}_N(\theta, X) \end{aligned}$$

converges in probability to the true parameter  $\theta^*$ :

$$\hat{\theta}_N(X) \xrightarrow{\text{P}} \theta^*.$$

estimator consistency

*Proof:* Note that

$$\text{LR}_N(\hat{\theta}_N, X) \leq \text{LR}_N(\theta^*, X) \quad (2)$$

because  $\hat{\theta}_N$  minimizes  $\text{LR}_N(\theta, X)$ . Therefore,

by (2)

$$\begin{aligned}
\overline{\text{LR}}(\hat{\theta}_N) - \overline{\text{LR}}(\theta^*) &= \overline{\text{LR}}(\hat{\theta}_N) - \text{LR}_N(\theta^*, X) + \text{LR}_N(\theta^*, X) - \overline{\text{LR}}(\theta^*) \\
&\leq \overline{\text{LR}}(\hat{\theta}_N) - \text{LR}_N(\hat{\theta}_N, X) + \text{LR}_N(\theta^*, X) - \overline{\text{LR}}(\theta^*) \\
&\leq \sup_{\theta \in \text{sp}_{\Theta}} |\overline{\text{LR}}(\theta) - \text{LR}_N(\theta, X)| + \text{LR}_N(\theta^*, X) - \overline{\text{LR}}(\theta^*) \\
&\xrightarrow{p} 0
\end{aligned}$$

by Assumption 1 and the Law of Large Numbers (LLN). Consequently, for any  $\delta > 0$ ,

$$\Pr\{\overline{\text{LR}}(\hat{\theta}_N) > \overline{\text{LR}}(\theta^*) + \delta\} \rightarrow 0 \quad \text{as } N \nearrow +\infty.$$

Pick any  $\epsilon > 0$ . Now, by Assumption 2, there exists a  $\delta > 0$  such that

$$\overline{\text{LR}}(\theta) > \overline{\text{LR}}(\theta^*) + \delta \quad \text{if } \|\theta - \theta^*\| \geq \epsilon.$$

Hence,

$$\Pr\{\|\hat{\theta}_N - \theta^*\| \geq \epsilon\} \leq \Pr\{\overline{\text{LR}}(\hat{\theta}_N) > \overline{\text{LR}}(\theta^*) + \delta\} \rightarrow 0. \quad \square$$

Almost sure convergence can be shown as well:

$$\hat{\theta}_N(X) \xrightarrow{\text{a.s.}} \theta^*.$$

strong estimator consistency

Asymptotic distribution

For simplicity, we focus on the case of scalar parameter  $\theta$ .

We have collected  $N$  i.i.d. observations  $\mathbf{x} = (x[n])_{n=0}^{N-1}$  from a distribution  $f_{X|\Theta}(x | \theta^*)$ , where  $\theta^*$  denotes the true value of the parameter  $\theta$ . The ML estimate of  $\theta$  is

$$\begin{aligned}
\hat{\theta}_N(\mathbf{x}) &= \hat{\theta}_N \\
&= \arg \max_{\theta} \prod_{n=0}^{N-1} f_{X|\Theta}(x[n] | \theta) \\
&= \arg \max_{\theta} \sum_{n=0}^{N-1} \ln f_{X|\Theta}(x[n] | \theta).
\end{aligned}$$

Suppose that Assumptions 1–3 from handout crb hold, as well as additional conditions (see the following section). Then,

asymptotic efficiency

$$\begin{aligned}
\sqrt{N}[\hat{\theta}_N(\mathbf{x}) - \theta^*] &\xrightarrow{d} \mathcal{N}(0, N\mathcal{I}^{-1}(\theta^*)) \\
&= \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta^*))
\end{aligned} \tag{3}$$

where  $\mathcal{I}(\theta)$  is the Fisher information for  $\theta$  and  $N$  measurements and  $\mathcal{I}_1(\theta)$  is the Fisher information for  $\theta$  and a single measurement.

*Rough proof.* Define the log-likelihood function

See also Appendix 7b in the textbook.

$$L(\theta) = \sum_{n=0}^{N-1} \ln f_{X|\Theta}(x[n] | \theta).$$

By the definition of  $\hat{\theta}_N$ ,

$$\frac{dL(\hat{\theta}_N)}{d\theta} = 0$$

and, by the mean-value theorem,

$$\frac{dL(\hat{\theta}_N)}{d\theta} = \frac{dL(\theta^*)}{d\theta} + \frac{d^2L(\tilde{\theta})}{d\theta^2}(\hat{\theta}_N - \theta^*)$$

mean-value theorem

where  $\tilde{\theta}$  is some value between  $\theta^*$  and  $\hat{\theta}_N$ . Hence,

$$\hat{\theta}_N - \theta^* = -\frac{dL(\theta^*)/d\theta}{d^2L(\tilde{\theta})/d\theta^2}.$$

Consider  $\sqrt{N}(\hat{\theta}_N - \theta^*)$ , where the scaling by  $\sqrt{N}$  stabilizes the limiting distribution; then

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = -\frac{\frac{1}{\sqrt{N}}dL(\theta^*)/d\theta}{\frac{1}{N}d^2L(\tilde{\theta})/d\theta^2}. \quad (4)$$

By the Central Limit Theorem (CLT), we have

$$\begin{aligned} \frac{1}{\sqrt{N}} \frac{dL(\theta^*)}{d\theta} &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \left. \frac{\partial \ln f_{X|\Theta}(x[n]|\theta)}{\partial \theta} \right|_{\theta=\theta^*} \\ &\xrightarrow{d} \mathcal{N}\left(\mathbb{E}_{X|\Theta}\left[\frac{\partial \ln f_{X|\Theta}(X|\theta)}{\partial \theta} \middle| \theta^*\right], \text{var}_{X|\Theta}\left[\frac{\partial \ln f_{X|\Theta}(X|\theta)}{\partial \theta} \middle| \theta^*\right]\right). \end{aligned}$$

For example, if  $(X[n])_{n=0}^{N-1}$  are i.i.d.  $\mathcal{N}(\theta^*, 1)$ , then  $\sqrt{N}(\hat{\theta}_N - \theta^*) \sim \mathcal{N}(0, 1)$ . Here,  $\hat{\theta}_N = \bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$ , the sample mean.

CLT: If  $(Z_n)_{n=1}^N$  are i.i.d. random variables with  $\mathbb{E}(Z_1) = \mu$  and  $\mathbb{E}[(Z_1 - \mu)^2] = \sigma^2$ , then  $\frac{1}{\sqrt{N}} \sum_{n=1}^N Z_n \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ .

By Lemma 1 from handout crb,

$$\begin{aligned} \mathbb{E}_{X|\Theta}\left[\frac{\partial \ln f_{X|\Theta}(X|\theta)}{\partial \theta} \middle| \theta^*\right] &= 0 \\ \text{var}_{X|\Theta}\left[\frac{\partial \ln f_{X|\Theta}(X|\theta)}{\partial \theta} \middle| \theta^*\right] &= \mathcal{I}_1(\theta^*) \end{aligned}$$

which implies

$$\frac{1}{\sqrt{N}} \frac{dL(\theta^*)}{d\theta} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\theta^*)). \quad (5a)$$

By the Strong Law of Large Numbers (SLLN),

$$\begin{aligned} \frac{1}{N} \frac{d^2L(\tilde{\theta})}{d\theta^2} &= \frac{1}{N} \sum_{n=0}^{N-1} \left. \frac{\partial^2 \ln f_{X|\Theta}(X[n]|\theta)}{\partial \theta^2} \right|_{\theta=\tilde{\theta}} \\ &\xrightarrow{\text{a.s.}} \mathbb{E}_{X|\Theta}\left[\frac{\partial^2 \ln f_{X|\Theta}(X[n]|\theta)}{\partial \theta^2} \middle| \theta^*\right] \\ &= -\mathcal{I}_1(\theta^*). \end{aligned} \quad (5b)$$

SLLN: Consider i.i.d. random variables  $(X_n)_{n=1}^N$  with  $\mathbb{E}X_i$  ( $|X_i|$ )  $< +\infty$ . Then,  $\sum_{n=1}^N X_n/N$  converges almost surely to  $\mathbb{E}X_n$ .

Hence, for large  $N$ , the numerator of (4) 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:

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \xrightarrow{d} \frac{1}{\mathcal{I}_1(\theta^*)} \mathcal{N}(0, \mathcal{I}_1(\theta^*)) = \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta^*)).$$

This type of convergence is argued using Slutsky's Theorem.  $\square$

See also (Rao 1973, §5f) for the case of independent observations.

☞ NOTE: At lower signal-to-noise ratios (SNRs), a threshold effect occurs: Outliers cause increased mean-square error (MSE) (than predicted by the Cramér-Rao bound (CRB)). This behavior is characteristic of practically all (good) parametric nonlinear estimators.

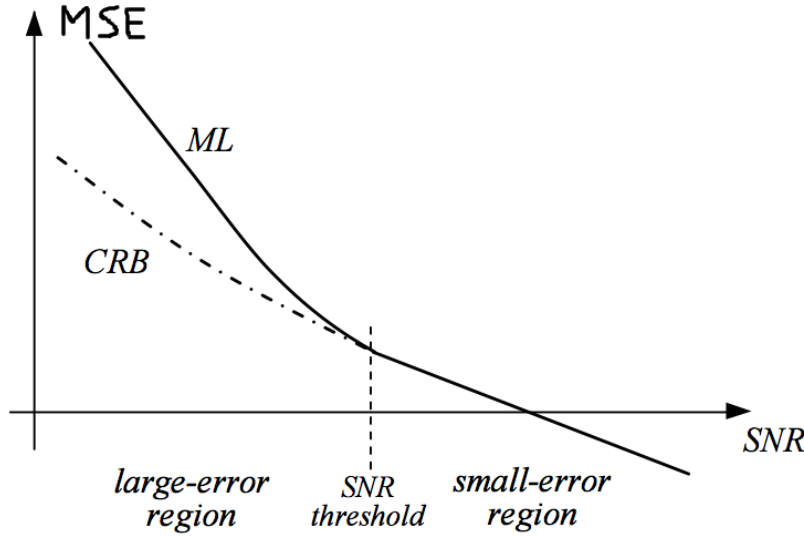


Figure 3: The threshold effect.

☞ NOTE: When estimation error cannot be made small as  $N \nearrow +\infty$ , the asymptotic pdf in (3) is invalid. For asymptotics to work, there *has to be* an averaging effect.

\* EXAMPLE. Estimation of the DC level in *fully dependent non-Gaussian noise*:

$$X[n] = a + W[n].$$

We measure  $(X[n])_{n=0}^{N-1}$ , but all noise samples are the same:

$$W[0] = W[1] = \dots = W[N-1].$$

Hence, we can discard  $(X[n])_{n=0}^{N-1}$ ; then,

$$\hat{a} = X[0].$$

Clearly, the probability density function (pdf) of this  $\hat{a}$  remains non-Gaussian as  $N \nearrow +\infty$  and, therefore, (3) cannot hold. Furthermore,  $\hat{a}$  is not consistent because

$$\text{var}_{X|A}(\hat{a} | a) = \text{var}_{X|A}(X[0] | a) \not\rightarrow 0 \quad \text{as } N \nearrow +\infty.$$

Example 7.7 in the textbook



## ML Estimation of Vector Parameters


THE ML estimate of  $\theta$  is a solution to the following optimization problem:

$$\hat{\theta}(x) = \arg \max_{\theta} f_{X|\Theta}(x | \theta).$$


Under appropriate regularity conditions (e.g., in the previous section), this estimator is consistent and

$$\sqrt{N}[\hat{\theta}(X) - \theta^*] \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\mathcal{I}^{-1}(\theta^*))$$

where  $\mathcal{I}(\theta)$  is now the *Fisher information matrix (FIM)* for  $N$  measurements and  $\theta^*$  denotes the true value of the parameter  $\theta$ .

 THE FIM is the expected value of the negative Hessian matrix of the log-likelihood function at the true parameter vector  $\theta^2$ . The Hessian is the curvature of the log-likelihood surface. For example, for scalar  $\theta$ , the FIM is simply the second derivative of the log-likelihood function. Since we are maximizing the log likelihood, the curvature should be negative. The more negative the curvature, the more sharply defined the location of the maximum. Therefore, more negative curvatures lead to less variable estimates, as revealed by the limiting distribution above.

<sup>2</sup> if Assumptions 1–3 from handout `multipar_gauss_crb` hold

 THE FIM and CRB matrix are *local* performance measures. Ambiguity functions have been used in, e.g., radar, to quantify global performance measures, see (Rendas and Moura 1998) and references therein. We can think of FIM as the curvature of the ambiguity function.

### Conditions for asymptotic Gaussian distribution

REGULARITY conditions vary for different measurement models.

Here are typical regularity conditions for i.i.d. measurements. Consider  $(X[n])_{n=0}^{N-1}$  i.i.d. given  $\theta$  that follow

$$f_{X|\Theta}(x[n] | \theta)$$

where  $\theta = [\theta_1, \theta_2, \dots, \theta_d]^T$ :

i)  $\theta$  is identifiable under the model

$$f_{X|\Theta}(x | \theta)$$

and the support of  $f_{X|\Theta}(x | \theta)$  is not a function of  $\theta$ ;

ii) The true value of the parameter  $\theta$  lies in an open subset of the parameter space  $\text{sp}_{\Theta}$ ;

- iii) For almost all  $x$ , the pdf  $f_{X|\Theta}(x|\theta)$  has continuous derivatives to order three with respect to all elements of  $\theta$  and all values in the open subset of  $\text{sp}_{\Theta}$ ;
- iv) The following are satisfied:

$$\mathbb{E}_{X|\Theta} \left[ \frac{\partial}{\partial \theta} \ln f_{X|\Theta}(X|\theta) \mid \theta \right] = \mathbf{0}$$

and

$$\begin{aligned} \mathcal{I}(\theta) &= \mathbb{E}_{X|\Theta} \left[ \frac{\partial}{\partial \theta} \ln f_{X|\Theta}(X|\theta) \frac{\partial}{\partial \theta^\top} \ln f_{X|\Theta}(X|\theta) \mid \theta \right] \\ &= -\mathbb{E}_{X|\Theta} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \ln f_{X|\Theta}(X|\theta) \mid \theta \right]. \end{aligned}$$

- v) the FIM  $\mathcal{I}(\theta)$  is positive definite;
- vi) Bounding functions  $m_{i,k,l}(\cdot)$  exist such that

$$\left| \frac{\partial^3}{\partial \theta_i \partial \theta_k \partial \theta_l} \ln f_{X|\Theta}(x|\theta) \right| \leq m_{i,k,l}(x)$$

for all  $\theta$  in the open subset of  $\text{sp}_{\Theta}$ , and

$$\mathbb{E}_{X|\Theta} [m_{i,k,l}(X) \mid \theta] < \infty.$$

### Sensor array processing

CONSIDER

$$(\mathbf{x}[n])_{n=0}^{N-1} = A(\phi)s[n] + \mathbf{W}[n]$$

with the set of unknown parameters

$$\theta_N = (\phi, \sigma^2, (s[n])_{n=0}^{N-1})$$

where  $\mathbf{W}[n]$  is zero-mean additive white Gaussian noise (AWGN) with unknown variance  $\sigma^2$ .

\* NOTE:

- $X[n]$  are not i.i.d. given  $\theta$  and, therefore, the usual conditions do not apply;
- the size of  $\theta$  grows with  $N$ .

trouble

It is well known that CRB cannot be attained asymptotically in this case (Stoica and Nehorai 1989).

☞ WHAT if

$$(X[n])_{n=0}^{N-1} = A(\phi)\mathbf{S}[n] + \mathbf{W}[n]$$

where  $\mathbf{S}[n]$  are random  $\mathcal{N}(\mathbf{0}, \Gamma)$  conditional on  $\Gamma$ ? Here, the parameters  $\theta$  are

$$\theta = (\phi, \Gamma, \sigma^2)$$

and  $(X[n])_{n=0}^{N-1}$  are i.i.d. given  $\theta$ , following

$$\{X[n] \mid \theta\} \sim \mathcal{N}(\mathbf{0}, A(\theta)\Gamma A^\top(\theta) + \sigma^2 I).$$

Here, the number of parameters *does not grow with the number of measurements*  $N$ . If the regularity conditions that we stated for the i.i.d. case hold, CRB will be attained asymptotically. Furthermore, CRB for  $\phi$  will be different (smaller) than the CRB for  $\phi$  when  $s[n]$  is deterministic.

## Delta Method

ASSUME that the invertible transform  $\alpha = g(\theta)$  has bounded derivatives up to the second order. Then, if  $\hat{\theta}$  is consistent, so is

$$\hat{\alpha} = g(\hat{\theta}).$$

Moreover, the asymptotic covariance matrices  $\text{cov}_{X|\Theta}(\hat{\theta} \mid \theta)$  and  $\text{cov}_{X|\alpha}(\hat{\alpha} \mid \alpha)$  are asymptotically equal to the corresponding MSE matrices (due to consistency) and are related as follows:

$$\text{cov}_{X|\alpha}(\hat{\alpha} \mid \alpha) = \frac{\partial g}{\partial \theta^\top} \text{cov}_{X|\Theta}(\hat{\theta} \mid \theta) \frac{\partial g^\top}{\partial \theta}.$$

*Proof.* Follows from the Taylor expansion around the true value  $\alpha = g(\theta)$ :

$$\hat{\alpha} = g(\theta) + \frac{\partial g(\theta)}{\partial \theta^\top}(\hat{\theta} - \theta) + o(\|\hat{\theta} - \theta\|).$$

□

\* EXAMPLE: Amplitude and phase estimation of a sinusoid. Assume

$$(X[n])_{n=0}^{N-1} = a \cos(\omega_0 n + \phi) + W[n] \quad (6)$$

where  $\omega_0$  is a known frequency and  $W[n]$  is zero-mean AWGN with variance  $\sigma^2$ . We wish to estimate the amplitude  $a$  and phase  $\phi$ .

Rewrite (6) as a linear model:

$$X = \begin{bmatrix} X[0] \\ \vdots \\ X[N-1] \end{bmatrix} = H\theta + W$$

where  $\theta = [\theta_1, \theta_2]^\top$  and

$$\begin{aligned} (H_{i,1})_{i=1}^N &= \cos(\omega_0(i-1)) \\ (H_{i,2})_{i=1}^N &= \sin(\omega_0(i-1)) \\ (a \cos \phi, -a \sin \phi) &\leftrightarrow (\theta_1, \theta_2) \end{aligned}$$

$\sigma^2$  can be known or unknown, because of decoupling between  $\theta$  and  $\sigma^2$

We have

$$\{\hat{\boldsymbol{\theta}}(X) \mid \boldsymbol{\theta}\} = (H^\top H)^{-1} H^\top X \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (H^\top H)^{-1}).$$

By the ML invariance principle,  $\hat{a}$  and  $\hat{\phi}$  can be found from  $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2]^\top$  via rectangular-to-polar coordinate conversion:

$$(\hat{\theta}_1, \hat{\theta}_2) \leftrightarrow (\hat{A} \cos \hat{\phi}, -\hat{A} \sin \hat{\phi}).$$

Define  $\boldsymbol{\alpha} = [a, \phi]^\top = \mathbf{g}(\boldsymbol{\theta})$ . Then, the delta method yields

$$\text{cov}_{X|\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}) = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}^\top} \text{cov}_{X|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) \frac{\partial \mathbf{g}^\top}{\partial \boldsymbol{\theta}}.$$

Here,  $\text{cov}_{X|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$  and  $\text{cov}_{X|\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha})$  are the asymptotic covariance (and MSE) matrices of  $\hat{\boldsymbol{\theta}}(X)$  and  $\hat{\boldsymbol{\alpha}}(X)$ .

(Ferguson 1996) is a good reference for large-sample theory.

## Computing ML Estimates

FINDING the ML estimate of a  $d$ -dimensional parameter vector  $\boldsymbol{\theta}$  often requires a nonlinear  $d$ -dimensional optimization:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$$

where

$$V(\boldsymbol{\theta}) = -\ln f_{X|\boldsymbol{\theta}}(\mathbf{x} \mid \boldsymbol{\theta})$$

is the negative log-likelihood (NLL) function.

\* EXAMPLE. Consider  $(X[n])_{n=0}^{N-1}$  i.i.d. given  $\boldsymbol{\theta}$  following

$$f_{X|\boldsymbol{\theta}}(x \mid \boldsymbol{\theta}) = p\mathcal{N}(x \mid \mu_0, \sigma_0^2) + (1-p)\mathcal{N}(x \mid \mu_1, \sigma_1^2)$$

where  $\boldsymbol{\theta} = (p, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ .

The likelihood is a complicated nonlinear and nonconvex function of  $\boldsymbol{\theta}$ .

$$f_{X|\boldsymbol{\theta}}(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{n=0}^{N-1} f_{X|\boldsymbol{\theta}}(x[n] \mid \boldsymbol{\theta})$$

a product of sums of exponentials

where  $\mathbf{x} = (x[n])_{n=0}^{N-1}$ . Taking the logarithm does not simplify the problem:

$$\ln f_{X|\boldsymbol{\theta}}(\mathbf{x} \mid \boldsymbol{\theta}) = \text{a sum of logs of sums of exponentials.}$$

Furthermore, the sufficient statistic in this case is the whole measurement vector  $\mathbf{x}$ ; i.e., there is no small sufficient statistic that summarizes the measurements.

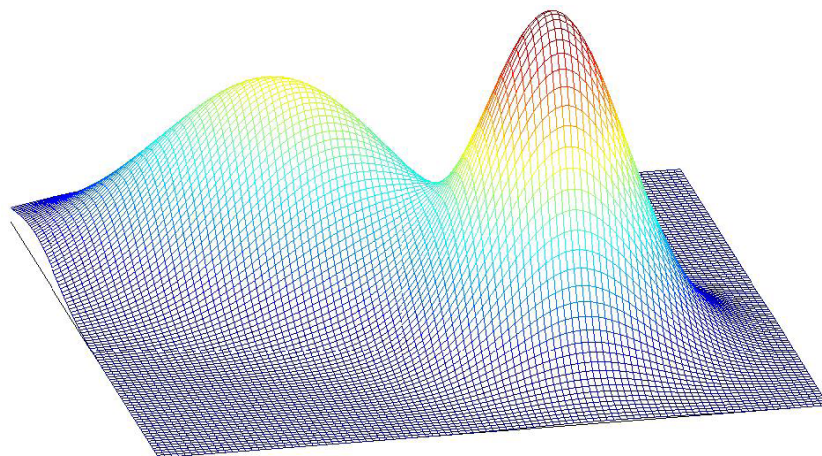


Figure 4: Two-dimensional Gaussian mixture density.

What can we do in such situations? We need a computational method to maximize the likelihood function. There are two common approaches:

☞ GRADIENT/Newton methods:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \underbrace{\Delta_t}_{\text{step size}} \mathbf{g}^{(t)} \quad \Delta_t > 0$$

called gradient methods if  $\Delta_t$  scalar, Newton/quasi-Newton if  $\Delta_t$  a matrix.

where

$$\begin{aligned} \mathbf{g}^{(t)} &= \left. \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= - \left. \frac{\partial \ln f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \end{aligned}$$

☞ Expectation-maximization (EM) methods.

Gradient descent methods for minimization should be familiar to most. The EM algorithm is a specialized approach designed for ML estimation problems, see handout em.

Newton-Raphson iteration

ASSUME that a guess  $\boldsymbol{\theta}^{(t)}$  of the parameter vector  $\boldsymbol{\theta}$  is available. We wish to improve  $\boldsymbol{\theta}^{(t)}$  by moving to  $\boldsymbol{\theta}^{(t+1)}$ . Apply quadratic Taylor expansion:

$$V(\boldsymbol{\theta}) \approx V(\boldsymbol{\theta}^{(t)}) + (\mathbf{g}^{(t)})^\top \Delta \boldsymbol{\theta}_t + \frac{1}{2} (\Delta \boldsymbol{\theta}_t)^\top H^{(t)} \Delta \boldsymbol{\theta}_t$$

Reading: §7.7 in the textbook.

where the gradient  $\mathbf{g}^{(t)}$  has been defined in (7) and

$$\begin{aligned}\Delta\boldsymbol{\theta}^{(t)} &= \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \\ H^{(t)} &= \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= - \frac{\partial^2 \ln f_{X|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.\end{aligned}$$

Complete the squares:

$$V(\boldsymbol{\theta}) \approx [\Delta\boldsymbol{\theta}^{(t)} + (H^{(t)})^{-1} \mathbf{g}^{(t)}]^\top \frac{1}{2} H^{(t)} [\Delta\boldsymbol{\theta}^{(t)} + (H^{(t)})^{-1} \mathbf{g}^{(t)}] + \text{const.}$$

const denotes terms that are not functions of  $\boldsymbol{\theta}$ .

We assume that  $H^{(t)} > 0$  (positive definite)<sup>3</sup> and choose

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - (H^{(t)})^{-1} \mathbf{g}^{(t)}.$$

<sup>3</sup>  $H^{(t)} \not> 0$  means problems for our iteration and we should not run the Newton-Raphson step in this case

If the underlying Taylor approximation is good, the Newton-Raphson iteration achieves *quadratic convergence* near the optimum, i.e.,

$$\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}\|_2 \leq c \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}\|_2^2$$

where  $c$  is a positive constant. Therefore, we gain approximately one significant digit per iteration.

However, the algorithm can diverge if we start too far from the optimum. To facilitate convergence (to a *local optimum*, in general), we can apply a damped Newton-Raphson algorithm. Here is one such damped algorithm:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mu_t (H^{(t)})^{-1} \mathbf{g}^{(t)} \quad (7)$$

where the step length  $\mu_t$  is  $\mu_t = 1, \frac{1}{2}, \frac{1}{4}, \dots$ . In particular, in the  $i$ th iteration, start with the step length  $\mu_t = 1$ , compute  $\boldsymbol{\theta}^{(t+1)}$  using (7), and check if

$$V(\boldsymbol{\theta}^{(t+1)}) < V(\boldsymbol{\theta}^{(t)}) \quad (8)$$

holds; if yes, go to the  $(t+1)$ st iteration. If no, *keep halving  $\mu_t$* <sup>4</sup> and recomputing  $\boldsymbol{\theta}^{(t+1)}$  using (7) until (8) is first satisfied; then, go to the  $(t+1)$ st iteration. Once in the  $(t+1)$ st iteration, reset  $\mu^{(t+1)}$  to 1 and continue in the same manner.

<sup>4</sup> or we may apply a different update of  $\mu_t$  such as line search

✱ **FISHER scoring.** Use an approximate form of the Hessian matrix of  $V(\boldsymbol{\theta})$ . In the case of ML estimation, the algorithm (7) with Hessian replaced by the expected Hessian<sup>5</sup>, i.e.,

$$H^{(t)} = \mathcal{I}(\boldsymbol{\theta}^{(t)})$$

<sup>5</sup> which is the FIM!

is particularly popular. The resulting algorithm is called *Fisher scoring*. This choice of  $H^{(t)}$  guarantees positive semidefiniteness of  $H^{(t)}$  because  $\mathcal{I}(\boldsymbol{\theta}^{(t)}) \geq 0$ <sup>6</sup>. It can be written as

<sup>6</sup> recall that the FIM is expected score squared

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \mu^{(t)} \mathcal{I}^{-1}(\boldsymbol{\theta}^{(t)}) \mathbf{g}^{(t)} \\
&= \boldsymbol{\theta}^{(t)} + \mu^{(t)} \mathcal{I}^{-1}(\boldsymbol{\theta}) \left. \frac{\partial \ln f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \quad (9)
\end{aligned}$$

✱ COMMENTS on Newton-Raphson schemes.

- The convergence point is a *local* minimum of the NLL  $V(\boldsymbol{\theta})$ . It is the global minimum if  $V(\boldsymbol{\theta})$  is a unimodal function of  $\boldsymbol{\theta}$  or if the initial estimate is sufficiently good.
- If (we suspect that) there are multiple local minima of  $V(\boldsymbol{\theta})$  (i.e., multiple local maxima of the likelihood function), we should try multiple (wide-spread/different) starting values and select as our (best guess of the) ML estimate the convergence point that yields the smallest  $V(\boldsymbol{\theta})$ .

✱ EXAMPLE: Fisher scoring. Consider the following model:

$$(X[n])_{n=0}^{N-1} = s[n; \boldsymbol{\theta}] + W[n]$$

where  $W[n]$  is zero-mean AWGN with known variance  $\sigma^2$ ,

$$s[n; \boldsymbol{\theta}] = \sin(\omega_1 n) + \sin(\omega_2 n)$$

and the unknown parameters are  $\boldsymbol{\theta} = (\omega_1, \omega_2)$ . Now,

$$\ln f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}])^2 + \text{const}$$

and the score function and FIM of  $\boldsymbol{\theta}$  are

$$\begin{aligned}
\frac{\partial \ln f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}]) \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
\mathcal{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}^\top}.
\end{aligned}$$

The damped Fisher-scoring iteration becomes [see (9)]

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \mu_t \left( \sum_{n=0}^{N-1} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}^\top} \right)^{-1} \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}]) \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
&= \boldsymbol{\theta}^{(t)} + \mu_t \left\{ \sum_{n=0}^{N-1} n^2 \begin{bmatrix} \cos^2(\omega_1^{(t)} n) & \cos(\omega_1^{(t)} n) \cos(\omega_2^{(t)} n) \\ \cos(\omega_1^{(t)} n) \cos(\omega_2^{(t)} n) & \cos^2(\omega_2^{(t)} n) \end{bmatrix} \right\}^{-1} \\
&\quad \cdot \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}^{(t)}]) \begin{bmatrix} n \cos(\omega_1^{(t)} n) \\ n \cos(\omega_2^{(t)} n) \end{bmatrix}.
\end{aligned}$$

## Concentrated Likelihood

WE introduce the concept of concentrated likelihood via an example.

Consider a situation in which a small-scale disease epidemic has been observed, with individuals exposed to the disease (e.g., a virus) at a common place and time (Harter and Moore 1966; Hill 1963).

Or, consider computers infected by a virus. We assume that a time interval is known for exposure, but not the exact time.

We collect times at which infection was detected at various computers ('incubation times'), say, with time 0 corresponding to the start of a known interval in which exposure occurred. We model the collected infection times after the exposure  $(X[n])_{n=0}^{N-1}$  as i.i.d. given  $\theta$ , following

$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{(x-\alpha)\sigma\sqrt{2\pi}} \exp\{-[\ln(x-\alpha) - \mu]^2/(2\sigma^2)\}, & x \geq \alpha \\ 0, & \text{otherwise} \end{cases}$$

with parameters  $\theta = (\alpha, \mu, \sigma)$ , where  $\alpha > 0$  represents the time at which the exposure took place. Since the support of the above distribution *depends on the parameter  $\alpha$* , regularity condition **i)** does not hold.

\* NOTE: This measurement model is equivalent to

$$\{\ln(X[n] - \alpha) | \theta\} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \quad (10)$$

useful for simulating the data from the above model, as well as for finding the ML estimates of  $\mu$  and  $\sigma^2$  (see the discussion below).

The log-likelihood function of  $\theta$  for the data  $\mathbf{x} = (x[n])_{n=0}^{N-1}$  is

$$\begin{aligned} l(\theta) &= \ln f_{X|\Theta}(\mathbf{x} | \theta) = \sum_{n=0}^{N-1} \ln f_{X|\Theta}(x[n] | \theta) \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \sum_{n=0}^{N-1} \ln(x[n] - \alpha) \\ &\quad - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \{\ln(x[n] - \alpha) - \mu\}^2 \end{aligned} \quad (11)$$

where  $\alpha < \min_{n=0, \dots, N-1} x[n]$ . For a fixed  $\alpha$ , we can find easily  $\mu$  and  $\sigma^2$  that maximize (11):

$$\hat{\mu}(\alpha) = \frac{1}{N} \sum_{n=0}^{N-1} \ln(x[n] - \alpha) \quad (12a)$$

$$\hat{\sigma}^2(\alpha) = \frac{1}{N} \sum_{n=0}^{N-1} \{\ln(x[n] - \alpha) - \hat{\mu}(\alpha)\}^2 \quad (12b)$$



which can be obtained directly by maximizing (11) or from (10), using our knowledge of the ML estimates of the mean and variance of i.i.d. Gaussian measurements. Now, we can *concentrate* the log-likelihood function of  $\theta$  with respect to  $\mu$  and  $\sigma^2$  by substituting (12a) and (12b) into (11):

$$l\left(\alpha, \hat{\mu}(\alpha), \hat{\sigma}^2(\alpha)\right) = -\frac{N}{2} \ln[2\pi\hat{\sigma}^2(\alpha)] - \sum_{n=0}^{N-1} \ln(x[n] - \alpha) - \frac{N}{2}. \quad (13)$$

concentrated log-likelihood function of  $\alpha$

In statistics, concentrated likelihood function is sometimes referred to as *profile likelihood*.

## Summary of Classical Estimation

THE CRB is a lower bound on the covariance of all unbiased estimators of an unknown parameter vector (information inequality). The information inequality holds only if the regularity conditions from handouts `crb` or `multipar_gauss_crb` hold.<sup>7</sup>

The information-inequality theorem also gives us estimator that attains the bound, if it exists. An unbiased estimator that attains the CRB is termed *efficient*. An efficient estimator is an MVU estimator. If an efficient estimator exists, it coincides with the ML estimator.

When computing ML estimators, make sure you examine the boundary of the parameter space.

Typically, an MVU estimator does not exist. We have also seen that unbiasedness is often too restrictive and not needed.

The CRB is generally used as a

- measure of the potential performance attainable from the system,
- benchmark for assessing algorithm performance,
- measure for system design,
- tool for constructing confidence regions for the unknown parameters,
- tool for constructing noninformative priors in Bayesian applications.

Under certain regularity conditions, the CRB is attained *asymptotically* (for large numbers of measurements) by the ML estimator. Hence, under certain regularity conditions, the ML method is *asymptotically efficient*.

<sup>7</sup> Assumption 1 in `crb` or its multivariate extension in `multipar_gauss_crb` are easy to check.

## Acronyms

*AWGN* additive white Gaussian noise. 10, 11, 15  
*BLUE* best linear unbiased estimator. 2  
*CLT* Central Limit Theorem. 7  
*CRB* Cramér-Rao bound. 8–11, 17  
*EM* expectation-maximization. 13  
*FIM* Fisher information matrix. 9, 10, 14, 15  
*i.i.d.* independent, identically distributed. 2–7, 9–12, 16, 17  
*KL* Kullback-Leibler. 5  
*LLN* Law of Large Numbers. 6  
*ML* maximum-likelihood. 2–6, 9, 12–17  
*MSE* mean-square error. 8, 11, 12  
*MVU* minimum-variance unbiased. 2, 17  
*NLL* negative log-likelihood. 12, 15  
*pdf* probability density function. 8, 10  
*pmf* probability mass function. 4  
*LLN* Strong Law of Large Numbers. 7  
*SNR* signal-to-noise ratio. 8  
*XOR* exclusive or. 3

## References

- Ferguson, Thomas Shelburne (1996). *A Course in Large Sample Theory*. New York: Chapman & Hall (cit. on pp. 1, 12).  
 Harter, H. L. and A. H. Moore (1966). “Local-maximum-likelihood estimation of parameters of 3-parameter lognormal populations from complete and censored samples”. In: *J. Am. Stat. Assoc.* 61.315, 842–851 (cit. on p. 16).  
 Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI (cit. on p. 1).  
 Hill, B. M. (1963). “3-parameter lognormal-distribution and Bayesian-analysis of a point-source epidemic”. In: *J. Am. Stat. Assoc.* 58.301, 72–84 (cit. on p. 16).

- Rao, C. R. (1973). *Linear Statistical Inference and Its Application*. 2nd ed. New York: Wiley (cit. on pp. 1, 8).
- Rendas, M. J. D. and J. M. F. Moura (1998). "Ambiguity in radar and sonar". In: *IEEE Trans. Signal Process.* 46.2, pp. 294–305 (cit. on p. 9).
- Stoica, P. and A. Nehorai (1989). "MUSIC, maximum likelihood, and Cramer-Rao bound". In: *IEEE Trans. Signal Process.* 37.5, pp. 720–741 (cit. on p. 10).