Chapter 2

Elements on Estimation Theory

The estimation theory deals with the basic problem of infering some relevant features of a random experiment based on the observation of the experiment outcomes. In some cases, the experiment mechanism is totally unknown to the observer and the use of *nonparametric* estimation methods is necessary. The term "nonparametric" means that the observed experiment cannot be modelled mathematically. Let us consider, for instance, the classical problem of spectral analysis that consists in computing the power spectral density of the observed signal from a finite sample.

The performance of nonparametric methods is usually unsatisfactory when the observed time is limited. This situation is actually very usual because the experiment output is only temporally available; the experiment is not stationary; or the observer is due to supply the estimate in a short time. To design more *efficient* estimation techniques, it is recommended to find previously a convenient mathematical model for the studied experiment. The result of the experiment is thus a function of a *finite* number of unknow parameters, say $\boldsymbol{\theta}$, and other random terms forming the vector $\mathbf{w}$. The vector $\mathbf{w}$ collects all the *nuisance* terms in the model that vary randomly during the observation time as, for example, the measurement noise.

The objective is therefore finding the minimal parameterization in order to concentrate the most the uncertainty about the experiment. In those fields dealing with natural phenomena, the parametrization of the problem is definitely the most difficult point and, actually, the ultimate goal of scientists working in physics, sociology, economics, among others. Fortunately, the parameterization of human-made systems is normally accesible. In particular, in communication engineering, the received signal is known except for a finite set of parameters that must be estimated before recovering the transmitted information. Likewise, in radar applications, the received signal is know except for the time of arrival and, possibly, some other nuisance parameters. In the following, we will focus exclusively on *parametric estimation* methods assuming that we are provided with a convenient parameterization or signal model.

In some of the examples above, it is possible to act on the experiment by introducing an excitation signal. In that case, the random experiment can be seen as an unknown system that is identified by observing how the system reacts to the applied excitation. This alternative perspective is normally adopted in system engineering and, specifically, in the field of automatic control. Unfortunately, in some scenarios, the observer is unaware of the existing input signal and *blind* system identification is required. For example, in digital communications, the transmitted symbols are usually unknown at the receiver side. This thesis is mainly concerned with blind estimation problems in which the problem parameterization includes the unknown input.

Thus far, the formulation is rather general; the observation $\mathbf{y} \in \mathbb{C}^M$ is a given function of the input $\mathbf{x} \in \mathbb{C}^K$, the vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ and the random vector $\mathbf{w} \in \mathbb{C}^M$ of arbitrary known distribution. Formally, the general problem representation is considered in the following equation:

$$\mathbf{y} = \mathbf{a}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{w}) \tag{2.1}$$

where the function $\mathbf{a}(\cdot)$ should be univoque with respect to $\boldsymbol{\theta}$ and $\mathbf{x}$, that is, it should be possible to recover $\boldsymbol{\theta}$ and $\mathbf{x}$ from $\mathbf{y}$ if the value of $\mathbf{w}$ were known. In that case, the estimation problem is not ambiguous. The basic problem is that multiple values of $\boldsymbol{\theta}$, $\mathbf{x}$ and $\mathbf{w}$ yield the same observation $\mathbf{y}$. Otherwise, it would not be an estimation problem but an inversion problem consisting in finding the inverse of $\mathbf{a}(\cdot)$.

Then, the objective is to estimate the value of $\boldsymbol{\theta}$ based on the observation of $\mathbf{y}$ without knowing the input $\mathbf{x}$ and the random vector $\mathbf{w}$. Thus, the entries of $\mathbf{x}$ appear as *nuisance parameters* increasing the uncertainty on the vector of parameters $\boldsymbol{\theta}$. In general, the vector of nuisance parameters would include all the existing parameters which are not of the designer's interest, including the unknown inputs. For example, the signal amplitude is a nuisance parameter when estimating the time-of-arrival in radar applications. This thesis is mainly concerned with the treatment of these nuisance parameters in the context of digital communications.

An estimator of $\boldsymbol{\theta}$ is a given function $\mathbf{z}(\cdot)$ of the random observation $\mathbf{y}$,

$$\widehat{\boldsymbol{\theta}} = \mathbf{z}(\mathbf{y}),$$

yielding a random error $\mathbf{e} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ with $\boldsymbol{\theta}_o$ the true value of $\boldsymbol{\theta}$. Evidently, the aim is to minimize the magnitude of $\mathbf{e}$. Several criteria are listed in the literature minimizing a different *cost function* $C(\mathbf{e})$ as, for example, the mean square error $\|\mathbf{e}\|^2$, or the maximum error $\max\{\mathbf{e}\}$. On the other hand, a vast number of estimators have been formulated by proposing *ad hoc* functions $\mathbf{z}(\cdot)$ whose performance is evaluated next. Some of them are briefly presented in the following sections. For more details, the reader is referred to the excellent textbooks on parametric estimation in the bibliography [Tre68][Sch91a][Kay93b].

2.1 Classical vs. Bayesian Approach

There are two important questions that must be addressed before designing a convenient estimator. The first one is why some terms of the signal model are classified as random variables ($\mathbf{w}$) whereas others are deterministic parameters ($\boldsymbol{\theta}$). The second question is whether the nuisance parameters in $\mathbf{x}$ should be modelled as random or deterministic variables.

In the *classical estimation theory,* the wanted parameters $\boldsymbol{\theta}$ are deterministic unknowns that are *constant* along the observation interval. On the other hand, those unwanted terms varying "chaotically" along the observation interval are usually modelled as random variables (e.g., the measurement noise, the signal amplitude in fast fading scenarios, the received symbols in a digital receiver, etc.).

Regarding the vector $\mathbf{x}$, the nuisance parameters can be classified as deterministic constant unknowns, say $\mathbf{x}_c$, or random variable unknowns, say $\mathbf{x}_u$. In the random case, we will assume hereafter that the probability density function of $\mathbf{x}_u$ is known. However, if this information were not available, the entries of $\mathbf{x}_u$ could be considered deterministic unknowns and estimated together with $\boldsymbol{\theta}$.[1]

In the *classical estimation theory,* the likelihood function $f_{\mathbf{y}}(\mathbf{y}; \mathbf{x}_c, \boldsymbol{\theta})$ supplies all the statistical information for the joint estimation of $\mathbf{x}_c$ and $\boldsymbol{\theta}$. If some nuisance parameters are random, say $\mathbf{x}_u$, the *conditional* likelihood function $f_{\mathbf{y}/\mathbf{x}_u}(\mathbf{y}/\mathbf{x}_u; \mathbf{x}_c, \boldsymbol{\theta})$ must be averaged with respect to the prior distribution of $\mathbf{x}_u$, as indicated next

$$f_{\mathbf{y}}(\mathbf{y}; \mathbf{x}_c, \boldsymbol{\theta}) = E_{\mathbf{x}_u}\left\{f_{\mathbf{y}/\mathbf{x}_u}(\mathbf{y}/\mathbf{x}_u; \mathbf{x}_c, \boldsymbol{\theta})\right\} = \int f_{\mathbf{y}/\mathbf{x}_u}(\mathbf{y}/\mathbf{x}_u; \mathbf{x}_c, \boldsymbol{\theta}) f_{\mathbf{x}_u}(\mathbf{x}_u) d\mathbf{x}_u. \tag{2.2}$$

On the other hand, modeling the *constant* nuisance parameters as random variables is rather controversial. For example, the received carrier phase is almost constant when estimating the signal timing in static communication systems. Even if these parameters come from a random experiment and their p.d.f. is perfectly known, we are only observing a particular realization of $\mathbf{x}_c$, which is most probably different from their mean value. Therefore, modeling these nuisance parameters as random variables might yield biased estimates of $\boldsymbol{\theta}$. Evidently, this bias will be cancelled out if several realizations of $\mathbf{y}$ were averaged, but only one realization is available!

This controversy is inherent to the *Bayesian* philosophy [Kay93b, Ch. 10]. In the Bayesian or stochastic approach, *all* the parameters –including the vector of wanted parameters $\boldsymbol{\theta}$– are modelled as random variables of known a priori distribution or *prior.* Then, the resulting estimators are designed to be optimal *"on the average"*, that is, averaging $\widehat{\boldsymbol{\theta}}$ with respect to the prior distributions of $\boldsymbol{\theta}$ and $\mathbf{x}$. Actually, all the classical concepts such as bias, variance, MSE, consistency and efficiency must be reinterpreted in the Bayesian sense.

[1] Notice that this is not the only solution. For example, we can assume a non-informative prior for $\mathbf{x}_u$ or, alternatively, we can apply Monte Carlo methods to evaluate numerically the unknow distribution of $\mathbf{x}_u$ [Mer00][Mer01].

Bayesian estimators are able to outperform classical estimators when they are evaluated "on the average", mainly when the observation $\mathbf{y}$ is severely degraded in noisy scenarios. This is possible because Bayesian estimators are able to exploit the a priori information on the unknown parameters. Anyway, as S. M. Kay states in his book, *'It is clear that comparing classical and Bayesian estimators is like comparing apples and oranges'* [Kay93b, p. 312].

Bearing in mind the above explaination, let us consider that $\mathbf{y}$, $\mathbf{x}$, and $\boldsymbol{\theta}$ are jointly distributed random vectors. In that case, the whole statistical information about the parameters is given by the joint p.d.f.

$$f_{\mathbf{y},\mathbf{x},\boldsymbol{\theta}}\left(\mathbf{y},\mathbf{x},\boldsymbol{\theta}\right) = f_{\mathbf{y}/\mathbf{x},\boldsymbol{\theta}}\left(\mathbf{y}/\mathbf{x},\boldsymbol{\theta}\right) f_{\mathbf{x}}\left(\mathbf{x}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right), \tag{2.3}$$

assuming that $\mathbf{x}$ and $\boldsymbol{\theta}$ are statistically independent. The first conditional p.d.f. in (2.3) is numerically identical to the conditional likelihood function $f_{\mathbf{y}/\mathbf{x}_u}\left(\mathbf{y}/\mathbf{x}_u;\mathbf{x}_c,\boldsymbol{\theta}\right)$ in (2.2) but it highlights the randomness of $\mathbf{x}_c$ and $\boldsymbol{\theta}$ in the adopted Bayesian model. The other terms $f_{\mathbf{x}}\left(\mathbf{x}\right) = f_{\mathbf{x}_c}\left(\mathbf{x}_c\right) f_{\mathbf{x}_u}\left(\mathbf{x}_u\right)$ and $f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)$ are the a priori distributions of $\mathbf{x}$ and $\boldsymbol{\theta}$, respectively.

Notice that the classical and Bayesian theories coincide in case of non-informative priors, i.e., when $f_{\mathbf{y}}\left(\mathbf{y};\mathbf{x}_c,\boldsymbol{\theta}\right)$ is significantly narrower than $f_{\mathbf{x}_c}\left(\mathbf{x}_c\right)$ and $f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)$ [Kay93b, Sec. 10.8].

In the sequel and for the sake of simplicity, *all* the nuisance parameters will be modelled as random variables or, in other words, $\mathbf{x} = \mathbf{x}_u$ and $\mathbf{x}_c = \emptyset$. Thus,

$$f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) = E_{\mathbf{x}}\left\{f_{\mathbf{y}/\mathbf{x},\boldsymbol{\theta}}\left(\mathbf{y}/\mathbf{x},\boldsymbol{\theta}\right)\right\}$$

will be referred to as the *unconditional* or stochastic likelihood function in opposition to the joint or *conditional* likelihood function

$$f_{\mathbf{y}}\left(\mathbf{y};\mathbf{x},\boldsymbol{\theta}\right) = f_{\mathbf{y}/\mathbf{x},\boldsymbol{\theta}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right),$$

which is also referred to as the deterministic likelihood function in the literature.

2.2 MMSE and MVU Estimation

The ultimate goal in the classical estimation theory is the minimization of the estimator mean square error (MSE), that is given by

$$MSE(\boldsymbol{\theta}) \triangleq E_{\mathbf{y}}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 = E_{\mathbf{y}}\left\|\mathbf{z}(\mathbf{y}) - \boldsymbol{\theta}\right\|^2$$

where $E_{\mathbf{y}}\left\{\cdot\right\}$ involves, implicitly, the expectation over the random vectors $\mathbf{w}$ and $\mathbf{x}$. The MSE can be decomposed as

$$MSE(\boldsymbol{\theta}) = BIAS^2(\boldsymbol{\theta}) + VAR(\boldsymbol{\theta})$$

where the estimator bias and variance are given by

$$\begin{aligned} BIAS^2(\boldsymbol{\theta}) &= \left\| E_{\mathbf{y}}\left\{\widehat{\boldsymbol{\theta}}\right\} - \boldsymbol{\theta} \right\|^2 \\ VAR(\boldsymbol{\theta}) &= E_{\mathbf{y}}\left\|\widehat{\boldsymbol{\theta}} - E_{\mathbf{y}}\left\{\widehat{\boldsymbol{\theta}}\right\}\right\|^2 \end{aligned}$$

The minimum MSE (MMSE) estimator finds a trade-off between the bias and the variance for every value of $\boldsymbol{\theta}$. Unfortunately, the bias term is usually a function of $\boldsymbol{\theta}$ and, consequently, the MMSE estimator is generally not realizable because it depends on $\boldsymbol{\theta}_o$ [Kay93b, Sec. 2.4.]. In general, any estimator depending on the bias term will be unrealizable in the classical framework. This limitation suggests to focus uniquely on unbiased estimators holding that $BIAS^2(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta}$. Thus, the estimator MSE coincides with its variance and the resulting estimator is usually referred to as the minimum variance unbiased (MVU) estimator [Kay93b, Ch. 2]. The MVU estimator minimizes the variance subject to the unbiased constraint *for every* $\boldsymbol{\theta}$.

The Rao-Blackwell-Lehmann-Scheffe theorem facilitates a procedure for finding the MVU estimator [Kay93b, Ch.5]. Unfortunately, this method is usually tedious and sometimes fails to produce the MVU estimator. Notice that the existence of the MVU estimator is not guaranteed either. Despite these difficulties, the MVU formulation is widely adopted because the maximum likelihood principle is known to provide approximatelly the MVU estimator under mild regularity conditions [Kay93b, Ch. 7].

If the classical framework is abandonned in favour of the Bayesian approach, the dependence of $MSE(\boldsymbol{\theta})$ on the true parameter $\boldsymbol{\theta}$ can be solved by averaging with respect to the prior $f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)$. Therefore, the *Bayesian* MMSE estimator can be formulated as the minimizer of

$$E_{\boldsymbol{\theta}}\left\{MSE(\boldsymbol{\theta})\right\} = E_{\boldsymbol{\theta}}\left\{E_{\mathbf{y}}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2\right\} = \int E_{\mathbf{y}}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \tag{2.4}$$

that is known to be the mean of the posterior p.d.f. $f_{\boldsymbol{\theta}/\mathbf{y}}\left(\boldsymbol{\theta}/\mathbf{y}\right)$ [Kay93b, Eq. 10.5], i.e.,

$$\widehat{\boldsymbol{\theta}}_{MMSE} = E_{\boldsymbol{\theta}/\mathbf{y}}\left\{\boldsymbol{\theta}/\mathbf{y}\right\} = f_{\mathbf{y}}^{-1}\left(\mathbf{y}\right)\int \boldsymbol{\theta} f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \tag{2.5}$$

where the Bayes' rule is applied to write $f_{\boldsymbol{\theta}/\mathbf{y}}\left(\boldsymbol{\theta}/\mathbf{y}\right)$ in terms of the likelihood function and the prior:

$$f_{\boldsymbol{\theta}/\mathbf{y}}\left(\boldsymbol{\theta}/\mathbf{y}\right) = \frac{f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)}{f_{\mathbf{y}}\left(\mathbf{y}\right)} = \frac{f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)}{\int f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}}.$$

The Bayesian MMSE estimator is known to minimize the MSE "on the average" (2.4). This means that the actual MSE will be high if the actual parameter $\boldsymbol{\theta}_o$ is unlikely, and small if $f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)$ is distributed around the true parameter $\boldsymbol{\theta}_o$.

2.3 Maximum Likelihood Estimation

Although there are other relevant criteria, the maximum likelihood (ML) principle has become the most popular parametric method for deducing statistically optimal estimators of $\boldsymbol{\theta}$. In the studied signal model (2.1), the observation is clearly a random variable due to the presence of the random vectors $\mathbf{w}$ and $\mathbf{x}$. Actually, we have a single observation $\mathbf{y}_o$ of this random variable from which the value of $\boldsymbol{\theta}$ must be inferred. The ML estimator is the one chosing the value of $\boldsymbol{\theta}$ –and implicitly the value of $\mathbf{w}$ and $\mathbf{x}$– that makes $\mathbf{y}_o$ the most likely observation. Formally, if $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is the probability density function of the random vector $\mathbf{y}$ parameterized by $\boldsymbol{\theta}$, the ML estimator is given by

$$\widehat{\boldsymbol{\theta}}_{UML} = \arg\max_{\boldsymbol{\theta}} \left\{ f_{\mathbf{y}}(\mathbf{y}_o;\boldsymbol{\theta}) \right\}, \tag{2.6}$$

where $\mathbf{y}_o$ is the vector of observed data[2] and

$$f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) = E_{\mathbf{x}}\left\{ f_{\mathbf{y}/\mathbf{x},\boldsymbol{\theta}}(\mathbf{y}/\mathbf{x},\boldsymbol{\theta}) \right\} = \int f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta})\, f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} \tag{2.7}$$

is known as the *unconditional likelihood function.* Likewise, the estimator in (2.6) is known as the unconditional or stochastic maximum likelihood (UML) estimator because the nuisance parameters are modelled as random unknowns (Section 2.1). If the nuisance parameters are really random variables, the UML estimator is actually the *true* ML estimator of $\boldsymbol{\theta}$.

Alternatively, the nuisance parameters can be modelled as deterministic unknowns –as done for $\boldsymbol{\theta}$. In the context of the ML theory, the deterministic or conditional model is unavoidable when $\mathbf{x}$ is a constant unknown or there is no prior information about $\mathbf{x}$ (Section 2.1). Moreover, even if the nuisance parameters are actually random, the CML approach is often adopted if the expectation in (2.7) cannot be solved analytically. In that case, however, the CML solution is generally suboptimal because it ignores the prior information about $\mathbf{x}$. Thus, the deterministic or conditional maximum likelihood (CML) estimator is formulated as follows

$$\widehat{\boldsymbol{\theta}}_{CML} = \arg\max_{\boldsymbol{\theta}} \left\{ \max_{\mathbf{x}} f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta}) \right\} = \arg\max_{\boldsymbol{\theta}} \left\{ f_{\mathbf{y}}(\mathbf{y};\widehat{\mathbf{x}}_{ML},\boldsymbol{\theta}) \right\} \tag{2.8}$$

where $f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta})$ is the joint or *conditional likelihood function* and

$$\widehat{\mathbf{x}}_{ML} = \arg\max_{\mathbf{x}} \left\{ \max_{\boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta}) \right\} \tag{2.9}$$

is the ML estimator of $\mathbf{x}$.

Comparing the UML and CML solutions in (2.6) and (2.8), we observe that in the unconditional model the nuisance parameters are averaged out using the prior $f_{\mathbf{x}}(\mathbf{x})$ whereas in the

[2] In the sequel, the random variable $\mathbf{y}$ and the observation $\mathbf{y}_o$ will be indistinctly named $\mathbf{y}$ for the sake of simplicity.

conditional model $f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta})$ is compressed by means of the ML estimate of $\mathbf{x}$, namely $\widehat{\mathbf{x}}_{ML}$. Also, it is worth noting that, if the nuisance parameters belong to a discrete alphabet, we are dealing with a detection problem and $\widehat{\mathbf{x}}_{ML}$ is actually the *ML detector*. It is found that the estimation of $\boldsymbol{\theta}$ is significantly improved by exploiting the discrete[3] nature of $\mathbf{x}$. This aspect is crucial when designing estimation techniques for digital communications in which $\mathbf{x}$ is the vector of transmitted symbols.

Finally, the following alternative estimator is proposed now:

$$\widehat{\boldsymbol{\theta}}_{CML2} = \arg\max_{\boldsymbol{\theta}} \left\{ \max_{\mathbf{x}} f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta}) f_{\mathbf{x}}(\mathbf{x}) \right\} = \arg\max_{\boldsymbol{\theta}} \{ f_{\mathbf{y}}(\mathbf{y};\widehat{\mathbf{x}}_{MAP},\boldsymbol{\theta}) \} \tag{2.10}$$

where

$$\widehat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} \left\{ \max_{\boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta}) f_{\mathbf{x}}(\mathbf{x}) \right\} \tag{2.11}$$

is the *Maximum a Posteriori* (MAP) detector exploiting the prior distribution of $\mathbf{x}$. Notice that $\widehat{\boldsymbol{\theta}}_{CML} = \widehat{\boldsymbol{\theta}}_{CML2}$ in case of equally likely nuisance parameters.

2.3.1 Decision Directed ML Estimation

Focusing on those estimation problems dealing with *discrete* nuisance parameters, the conditional ML estimators in equations (2.8) and (2.10) exploit the *hard* decisions provided by the ML or MAP detectors of $\mathbf{x}$, respectively. In the context of digital communications, these estimation techniques are referred to as *decision directed* (DD). Decision-directed estimators are usually implemented iterating equations (2.8) and (2.9) for the ML detector, or (2.10) and (2.11) for the MAP detector. The main drawback of iterative algorithms is the uncertain convergence to the global maximum of $f_{\mathbf{y}}(\mathbf{y};\mathbf{x},\boldsymbol{\theta})$.

In some kind of problems, decision directed methods are efficient at high SNR. For example, in digital communications, DD synchronizers are known to attain the Cramér-Rao bound at high SNR [And94][Moe98]. However, when the noise variance is high, hard decisions are unreliable and it is better to compute *soft* decisions on the nuisance parameters. In digital communications, the estimation techniques based on soft decisions about the transmitted symbols are usually known as *non-data-aided* (NDA) [Men97]. Indeed, this interpretation is adopted in [Vaz00][Rib01b] to describe some ML-based NDA synchronizers.

In [Noe03], the Expectation-Maximization (EM) algorithm [Dem77][Fed88] is invoked to prove that UML estimation requires soft decisions from the MAP detector. More specifically, the nuisance parameters soft information is introduced by means of the *a posteriori probabilities*

[3] In order to unify the study of continuous and discrete nuisance parameters, the prior $f_{\mathbf{x}}(\mathbf{x})$ will be used indistinctly in both cases. To do so, if $\mathbf{x} \in \{\mathbf{a}_1, \ldots, \mathbf{a}_I\}$ with I the alphabet size, $f_{\mathbf{x}}(\mathbf{x})$ will be a finite number of Dirac's deltas, i.e, $f_{\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^{I} p(\mathbf{a}_i)\delta(\mathbf{x} - \mathbf{a}_i)$ with $p(\mathbf{a}_i)$ the probability of $\mathbf{a}_i$.

$f_{\mathbf{x}}(\mathbf{a}_1/\mathbf{y},\boldsymbol{\theta}),...,f_{\mathbf{x}}(\mathbf{a}_I/\mathbf{y},\boldsymbol{\theta})$ where $\{\mathbf{a}_i\}_{i=1,\dots,I}$ are all the possible values of $\mathbf{x}$. The EM algorithm is then applied to obtain an *iterative* implementation of the UML estimator following the so-called Turbo principle [Mie00][Ber93]. In these schemes, the estimator (2.6) is assisted with the *decoder soft decisions* and vice versa. The EM foundation ensures the convergence to the UML solution under fairly general conditions. The required soft decisions are provided by the optimal MAP decoder proposed in [Bah74], that supplies at each iteration the a posteriori probability $f_{\mathbf{x}}(\mathbf{x}/\mathbf{y},\boldsymbol{\theta})$ for every possible value of $\mathbf{x}$.

It is worth noting that the UML estimator is able to exploit the *statistical dependence* introduced by the encoder whereas the conditional approach in (2.8)-(2.10) does not. In the conditional model, the estimator is only informed that the codeword $\mathbf{x}$ is a redundant vector and, thus, it belongs to a *reduced* subset or codebook. In addition, the UML estimator is able to exploit the statistical dependence of the nuisance parameters in order to reduce their uncertainty at low SNR.

Another suitable implementation of the conditional estimators (2.8)-(2.10) is to assign a different estimator $\widehat{\boldsymbol{\theta}}$ to each survivor path in the Viterbi decoder, corresponding to a tentative sequence of symbols $\mathbf{x}$. The estimator output is then used to recompute the metric of the associated path. These kind of methods are usually referred to in the literature as *Per Survivor Processing* (PSP) [Pol95]. It can be shown that this approach attains the performance of the CML estimator in (2.8).

2.3.2 Asymptotic properties

The importance of the ML theory is that it supplies the minimum variance unbiased (MVU) estimator if the observed vector is sufficiently large under mild conditions. This result is a consequence of the asymptotic *efficiency* of the ML criterion, which is known to attain the Cramér-Rao lower bound as the number of observations increases (Section 2.6.1). Therefore, the ML theory facilitates a systemmatic procedure to formulate the MVU estimator in most estimation problems of interest.

In this section, the most relevant properties of the ML estimator are enunciated [Kay93b, Sec. 7B]. If the observation size goes to infinity ($M \to \infty$), it can be shown that

Property 1. The ML estimator is asymptotically Gaussian distributed with mean $\boldsymbol{\theta}_o$ and covariance $\mathbf{B}_{CRB}(\boldsymbol{\theta}_o)$ where $\boldsymbol{\theta}_o$ is the true parameter and $\mathbf{B}_{CRB}(\boldsymbol{\theta}_o)$ is the Crámer-Rao lower bound evaluated at $\boldsymbol{\theta}_o$ (Section 2.6.1). This means that the ML estimator is asymptotically *unbiased* and *efficient* or, in other words, the ML estimator leads asymptotically ($M \to \infty$) to

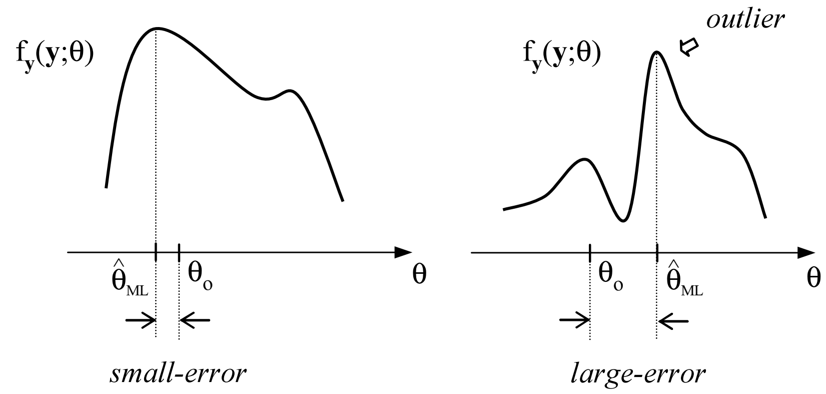


Figure 2.1: This picture illustrates the significance of the term *outlier* in the context of ML estimation.

the minimum variance unbiased (MVU) estimator with

$$E_{\mathbf{y}}\left\{\widehat{\boldsymbol{\theta}}_{ML}\right\} \longrightarrow \boldsymbol{\theta}_o$$

$$E_{\mathbf{y}}\left\{\left(\widehat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_o\right)\left(\widehat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_o\right)^H\right\} \longrightarrow \mathbf{B}_{CRB}\left(\boldsymbol{\theta}_o\right).$$

Property 2. The ML estimator is asymptotically *consistent* meaning that $\widehat{\boldsymbol{\theta}}_{ML} \rightarrow \boldsymbol{\theta}_o$ as the observation size goes to infinity. This property implies that the CRB tends to zero as M is increased, i.e., $\mathbf{B}_{CRB}\left(\boldsymbol{\theta}_o\right) \rightarrow \mathbf{0}$.

These properties are verified if the regularity condition

$$E_{\mathbf{y}}\left\{\left.\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right\}=\mathbf{0} \qquad (2.12)$$

is guaranteed for every $\boldsymbol{\theta}_o$. Fortunately, most problems of interest verify the above regularity condition. The implicit requirement is that the function support on $\mathbf{y}$ of $f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right)$ does not depend on the parameter $\boldsymbol{\theta}$ so that the integral limits of $E_{\mathbf{y}}\{\cdot\}$ are independent of $\boldsymbol{\theta}$. This condition is needed to have *unbiased* estimates since (2.12) guarantees that $E_{\mathbf{y}}\left\{\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right)\right\}$ has a maximum at the true parameter $\boldsymbol{\theta}_o$ whatever the value of $\boldsymbol{\theta}_o$.

As proved in [Kay93b, Theorem 7.5], the first property on the optimality of the ML estimator is satisfied even for finite observations provided that the signal model is *linear* in $\boldsymbol{\theta}$ and $\mathbf{x}$. However, a large number of estimation problems are *nonlinear* in the parameter vector $\boldsymbol{\theta}$. In that case, it is very important to determine how many samples (M) are required to guarantee the ML asymptotic efficiency (property 1). Fortunately, in most problems of interest this value is not excessive. It is found that the minimum M depends on the signal model at hand as well as the variance of the noise term $\mathbf{w}$, say σ_w^2. If the value of σ_w^2 is low and/or M is large, the

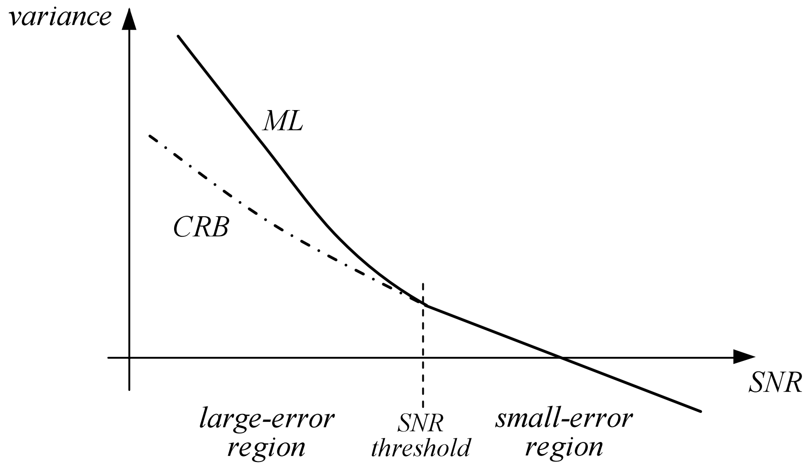


Figure 2.2: This picture illustrates the existence of a SNR threshold in nonlinear estimation problems. This threshold divides the SNR axis into the small-error regime, in which the CRB is attained, and the large-error regime, in which efficient estimators does not exist. Notice that the threshold position can be moved to the left by increasing the vector of observations.

log-likelihood function $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ exhibits a parabolic shape –quadratic form– with a unique maximum near the true parameter $\boldsymbol{\theta}_o$. Only in this *small-error regime,* the ML estimator is statistically efficient holding property 1.

On the other hand, if the value of σ_w^2 is high and/or M is not sufficiently large, the likelihood function $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ becomes multimodal and large errors are committed when the level of a distant maximum or *outlier* exceeds the true parameter maximum (Fig. 2.1). In this *large-error regime,* the variance of the ML estimator departs *abruptly* from the CRB. It is found that the estimator enters in the large-error regime if the noise variance σ_w^2 exceeds a given *threshold.* This threshold can be augmented (i.e., σ_w^2 greater) if the observation size is increased and, therefore, the large-error region disappears as long as the observation size goes to infinity. This is actually the sense of the ML asymptotic efficiency (property 1).

The existence of a low-SNR threshold in nonlinear estimation problems suggests to distinguish between the small-error and large-error scenario (Fig. 2.2). In the first case, ML estimators are efficient and, hence, they attain the CRB (Section 2.6.1). Thus, the ML principle becomes the systematic way of deducing the MVU estimator in the small-error regime. Moreover, in the small-error case, the ML estimator is also optimal in terms of mean square error (Section 2.2). This conclusion is important because MMSE estimators are generally not realizable since they depend on the unknown parameter $\boldsymbol{\theta}_o$.

On the other hand, efficient estimators do not exist in the large-error case and, other lower bounds are needed to take into account the existence of large errors –or outliers– and predict the threshold effect (Section 2.6). In this large-error regime, unbiased estimators are generally not optimal from the MSE point of view and the MMSE solution establishes a trade-off between the variance and bias contribution (Section 2.2). In this context, the Bayesian theory allows deducing *realizable* estimators minimizing the so-called *Bayesian MSE*, which is the MSE averaged over all the possible values of the parameter $\boldsymbol{\theta}$ [Kay93b, Sec. 10].

In this thesis, Chapter 3 is devoted to design optimal second-order large-error estimators whereas these results are particularized in Chapter 4 to formulate the optimal second-order small-error estimator.

To conclude this brief introduction to the maximum likelihood theory, two additional properties are presented next. These properties are satisfied even if the observation interval is finite.

Property 3. Whenever an efficient estimator exists, it corresponds to the ML estimator. In other words, if the MVU estimator attains the CRB, the ML estimator is also the MVU estimator. Otherwise, if the MVU variance is higher than the CRB, nothing can be stated about the optimality of the ML estimator for finite observations.

Property 4. The ML estimator is invariant in the sense that, if $\widehat{\boldsymbol{\theta}}_{ML}$ stands for the ML estimator of $\boldsymbol{\theta}$, the ML estimator of $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ is simply $\widehat{\boldsymbol{\alpha}}_{ML} = \mathbf{g}\left(\widehat{\boldsymbol{\theta}}_{ML}\right)$ for any one-to-one function $\mathbf{g}(\cdot)$. Otherwise, if $\mathbf{g}(\cdot)$ is not univoque, $\widehat{\boldsymbol{\alpha}}_{ML}$ maximizes $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\alpha})$, that is obtained as $\max_{\boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ subject to $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\alpha}$ [Kay93b, Th. 7.2].

2.4 Linear Signal Model

The formulation of parameter estimation techniques from the general model introduced in (2.1) is mostly fruitless. Accordingly, in the following, the focus will be on those *linear systems* corrupted by an additive Gaussian noise, holding that

$$\mathbf{y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x} + \mathbf{w} \tag{2.13}$$

where $\mathbf{x} \in \mathbb{C}^K$ is the system input forming the vector of nuisance parameters, $\mathbf{w} \in \mathbb{C}^M$ is the Gaussian noise vector and, $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{C}^{M\times K}$ is the system response parameterized by the vector $\boldsymbol{\theta} \in \mathbb{R}^P$. Despite its simplicity, the adopted linear model is really important because it appears in a vast number of engineering applications. In the context of digital communications, this model applies for any linear modulation as well as for continuous phase modulations (CPM) thanks to the Laurent's expansion [Lau86][Men97, Sec. 4.2] (Section 6.1.2).

We will assume that the noise vector in (2.13) is zero-mean and its covariance matrix is a priori known, that is,

$$\begin{aligned} E\left\{\mathbf{w}\right\} &= \mathbf{0} \\ E\left\{\mathbf{w}\mathbf{w}^H\right\} &= \mathbf{R}_w, \end{aligned}$$

with $\mathbf{R}_w$ a given full-rank matrix. Furthermore, we will assume that $\mathbf{w}$ is a proper or circular random vector holding that $E\left\{\mathbf{w}\mathbf{w}^T\right\} = \mathbf{0}$ [Sch03][Pic96]. The statistical distribution of the noise samples is normal (Gaussian), although the results in the following chapters could be easily extended to admit any other noise distribution. Finally, the noise variance is defined as

$$\sigma_w^2 \triangleq \frac{\mathrm{Tr}\left(\mathbf{R}_w\right)}{M},$$

which is the variance of the noise samples $\left[\mathbf{w}\right]_m$, if they are identically distributed. Additionally, we introduce the matrix $\mathbf{N}$, which is defined as

$$\mathbf{N} \triangleq \sigma_w^{-2}\mathbf{R}_w.$$

In the unconditional model, the nuisance parameters are modelled as random variables of known probability density function $f_{\mathbf{x}}\left(\mathbf{x}\right)$ with zero-mean and uncorrelated entries[4], meaning that

$$\begin{aligned} E\left\{\mathbf{x}\right\} &= \int \mathbf{x} f_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x} = \mathbf{0} \\ E\left\{\mathbf{x}\mathbf{x}^H\right\} &= \int \mathbf{x}\mathbf{x}^H f_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x} = \mathbf{I}_K \end{aligned}$$

where $f_{\mathbf{x}}\left(\mathbf{x}\right)$ would be composed of a finite number of Dirac's deltas in case of discrete nuisance parameters. On the other hand, the nuisance parameters are possibly improper random variables with $E\left\{\mathbf{x}\mathbf{x}^T\right\} \neq \mathbf{0}$ [Sch03][Pic96]. This consideration is specially important in digital communications because some relevant modulations (e.g., BPSK and CPM) are actually improper or noncircular, i.e., $E\left\{\mathbf{x}\mathbf{x}^T\right\} \neq \mathbf{0}$.

In the linear signal model, the conditional or joint likelihood function is given by

$$\begin{aligned} f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta},\mathbf{x}\right) &= \frac{1}{\pi^M \det\left(\mathbf{R}_w\right)} \exp\left(-\left\|\mathbf{y} - \mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{x}\right\|^2_{\mathbf{R}_w^{-1}}\right) \\ &= C_1 \exp\left(2\,\mathrm{Re}\left(\mathbf{x}^H\mathbf{A}^H\left(\boldsymbol{\theta}\right)\mathbf{R}_w^{-1}\mathbf{y}\right) - \mathbf{x}^H\mathbf{A}^H\left(\boldsymbol{\theta}\right)\mathbf{R}_w^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{x}\right) \end{aligned} \qquad (2.14)$$

with

$$C_1 \triangleq \frac{\exp(-\mathbf{y}^H\mathbf{R}_w^{-1}\mathbf{y})}{\pi^M \det\left(\mathbf{R}_w\right)}$$

an irrelevant factor independent of $\boldsymbol{\theta}$.

[4] Notice that there is no loss of generality because the correlation of $\mathbf{x}$ can always be included into the matrix $\mathbf{A}(\boldsymbol{\theta})$ in (2.13).

On the other hand, the unconditional likelihood function in (2.7) does not admit a *general* analytical solution, even for the linear model presented in this section. By replacing (2.14) into (2.7), it is found that the unconditional likelihood function is given by[5]

$$f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) = C_1 E_{\mathbf{x}}\left\{\exp\left(2\operatorname{Re}\left(\mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{y}\right) - \mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\mathbf{x}\right)\right\}. \tag{2.15}$$

Moreover, in case of i.i.d. nuisance parameters, the expectation with respect to $\mathbf{x}$ results in the following expressions:

$$\begin{aligned}
E_{\mathbf{x}}\left\{\exp\left(2\operatorname{Re}\left(\mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{y}\right)\right)\right\} &= \prod_{k=1}^{K} E_x\left\{\exp\left(2\operatorname{Re}\left(x_k^*\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{y}\right)\right)\right\}\\
E_{\mathbf{x}}\left\{\exp\left(\mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\mathbf{x}\right)\right\} &= \prod_{k=1}^{K}\prod_{l=1}^{K} E_x\left\{\exp\left(x_k^* x_l\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_l\right)\right\}\\
&= \prod_{k=1}^{K}\left(\prod_{l>k}^{K} E_x\left\{\exp\left(2\operatorname{Re}\left(x_k^* x_l\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_l\right)\right)\right\} + \right.\\
&\quad \left. E_x\left\{\exp\left(\left|x_k\right|^2\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_k\right)\right\}\right)
\end{aligned}$$

where $x_k \triangleq [\mathbf{x}]_k$ and $\mathbf{a}_k \triangleq [\mathbf{A}]_k$ are the k-th element and column of $\mathbf{x}$ and $\mathbf{A}$, respectively. The above expectations over the nuisance parameters have only been solved analytically in case of Gaussian nuisance parameters (Section 2.4.3) and polyphase discrete alphabets as shown in Appendix 2.A. However, a general closed-form solution is not available. In the next subsections, some alternative criteria are proposed to circumvent the computation of the exact unconditional likelihood function (2.15).

2.4.1 Low-SNR Unconditional Maximum Likelihood

The usual way of finding the UML estimator is the evaluation of (2.15) assuming a very low SNR [Vaz00][Men97]. The low-SNR constitutes a worst-case situation leading to robust estimators of $\boldsymbol{\theta}$. When the noise variance increases, the exponent of (2.15) is very small and, therefore, the exponential can be expanded into the following Taylor series:

$$f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) \simeq C_2 E_{\mathbf{x}}\left\{1 + \chi(\mathbf{y};\boldsymbol{\theta},\mathbf{x}) + \chi^2(\mathbf{y};\boldsymbol{\theta},\mathbf{x})\right\} \tag{2.16}$$

where $\chi(\mathbf{y};\boldsymbol{\theta},\mathbf{x}) \triangleq 2\operatorname{Re}\left(\mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{y}\right) - \mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\mathbf{x}$ is the exponent of (2.15) [Vaz00]. Assuming that the nuisance parameters are *circular*, zero-mean, unit-power and uncorrelated, the expectation in (2.16) is evaluated obtaining that

$$\begin{aligned}
E_{\mathbf{x}}\left\{\chi(\mathbf{y};\boldsymbol{\theta},\mathbf{x})\right\} &= -\operatorname{Tr}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right) = -\sigma_w^{-2}\operatorname{Tr}\left(\mathbf{A}^H\mathbf{N}^{-1}\mathbf{A}\right)\\
E_{\mathbf{x}}\left\{\chi^2(\mathbf{y};\boldsymbol{\theta},\mathbf{x})\right\} &= 2\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right) + \zeta(\boldsymbol{\theta})\\
&= 2\sigma_w^{-4}\operatorname{Tr}\left(\mathbf{N}^{-1}\mathbf{A}\mathbf{A}^H\mathbf{N}^{-1}\widehat{\mathbf{R}}\right) + \zeta(\boldsymbol{\theta})
\end{aligned}$$

[5] For the sake of clarity, the dependence on $\boldsymbol{\theta}$ is omitted from $\mathbf{A}(\boldsymbol{\theta})$ in the following expressions.

where $\widehat{\mathbf{R}} \triangleq \mathbf{y}\mathbf{y}^H$ is the sample covariance matrix and, $\zeta(\boldsymbol{\theta}) \triangleq \sigma_w^{-4} E_{\mathbf{x}}\left\{\left(\mathbf{x}^H\mathbf{A}^H\mathbf{N}^{-1}\mathbf{A}\mathbf{x}\right)^2\right\}$ has not been expanded because it is negligible compared to $E_{\mathbf{x}}\left\{\chi(\mathbf{y};\boldsymbol{\theta},\mathbf{x})\right\}$ for $\sigma_w^2 \to \infty$.

Finally, having in mind that $\ln(1+x) \simeq x$ for $x \simeq 0$ and omitting constant terms, the *low-SNR* log-likelihood function becomes

$$\begin{aligned}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) &\propto -\operatorname{Tr}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right) + \operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right) \\ &= \operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\left(\widehat{\mathbf{R}} - \mathbf{R}_w\right)\right), \end{aligned} \tag{2.17}$$

proving that the sample covariance matrix $\widehat{\mathbf{R}} \triangleq \mathbf{y}\mathbf{y}^H$ is a *sufficient statistic* for the estimation of $\boldsymbol{\theta}$ in the studied linear model, if the SNR goes to zero. More precisely, the log-likelihood function in (2.17) is an affine transformation of the sample covariance matrix with

$$\begin{aligned} b(\boldsymbol{\theta}) &= -\operatorname{Tr}\left(\mathbf{A}^H(\boldsymbol{\theta})\,\mathbf{R}_w^{-1}\mathbf{A}(\boldsymbol{\theta})\right) \\ \mathcal{M}(\boldsymbol{\theta}) &= \mathbf{R}_w^{-1}\mathbf{A}(\boldsymbol{\theta})\,\mathbf{A}^H(\boldsymbol{\theta})\,\mathbf{R}_w^{-1} \end{aligned}$$

the independent term and the kernel of $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$, respectively. Notice that this result is independent of the actual distribution of the nuisance parameters $f_{\mathbf{x}}(\mathbf{x})$. Actually, the result is valid for any circular distribution having zero mean and unitary variance.

Finally, the explicit formula for the UML estimator at low SNR is given by

$$\widehat{\boldsymbol{\theta}}_{lowSNR} = \arg\max_{\boldsymbol{\theta}} \operatorname{Tr}\left(\mathbf{N}^{-1}\mathbf{A}\mathbf{A}^H\mathbf{N}^{-1}\left(\widehat{\mathbf{R}} - \mathbf{R}_w\right)\right). \tag{2.18}$$

This result is relevant because it states that in low SNR scenarios, *second-order* techniques are asymptotically efficient for *any* estimation problem following the linear model in (2.13). Actually, this conclusion was the starting point of this thesis.

Unfortunately, the low-SNR solution has some important inconveniences. In Appendix 2.B, it is shown that the low SNR approximation usually yields biased estimates for any positive SNR. Moreover, the low-SNR UML estimator might yield a significant variance floor when applied in high SNR scenarios due to the variance induced by the random nuisance parameters (Appendix 2.B). This variability is usually referred to as *self-noise* or pattern-noise in digital synchronization [Men97].

Despite these potential problems, the low SNR approximation is extensively used in the context of digital communications and *ad hoc* methods are introduced to mitigate or cancel the self-noise contribution at high SNR. On the other hand, the *ML-based* estimators proposed in the following sections are suitable candidates to cancel out the bias and self-noise terms at high SNR. However, our main contribution in Chapter 4 is proving that all of them are suboptimal in terms of self-noise cancelation when applied to polyphase alphabets such as MPSK.

To conclude this section, we notice that the term depending on $\widetilde{\mathbf{R}} \triangleq \mathbf{y}\mathbf{y}^T$ also appears in $E_{\mathbf{x}}\left\{\chi^2\left(\mathbf{y};\boldsymbol{\theta},\mathbf{x}\right)\right\}$ when dealing with noncircular nuisance parameters. Therefore, the low-SNR log-likelihood function should be modified in the following way:

$$\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) \simeq -\operatorname{Tr}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right) + \operatorname{Tr}\operatorname{Re}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}} + \mathbf{R}_w^{-T}\mathbf{A}^*\Gamma^H\mathbf{A}^H\mathbf{R}_w^{-1}\widetilde{\mathbf{R}}\right),$$

with

$$\Gamma \triangleq E\left\{\mathbf{x}\mathbf{x}^T\right\}$$

the improper covariance matrix of $\mathbf{x}$. Furthermore, if $\mathbf{x}$ is real-valued (e.g., in baseband communications or for the BPSK modulation), it follows that $\Gamma = E\left\{\mathbf{x}\mathbf{x}^H\right\} = \mathbf{I}_K$. Notice that this second term is the one exploited in Section 6.2 to estimate the carrier phase because the term on $\widehat{\mathbf{R}}$ does not provide information about the signal phase.

2.4.2 Conditional Maximum Likelihood (CML)

In this section, the CML criterion in (2.8) is formulated for the linear signal model in (2.13). In that case, the conditional likelihood function in (2.14) can be compressed with respect to $\mathbf{x}$ if the nuisance parameters are *continuous variables*, i.e., $\mathbf{x} \in \mathbb{C}^K$. If the nuisance parameters are discrete (e.g., in digital communications), this compression strategy yields a *suboptimal* version of the CML estimator formulated in (2.8). This suboptimal CML estimator has been successfully applied to different estimation problems in digital communications such as timing synchronization [Rib01b]. Some degradation is incurred because the estimator does not exploit the fact that $\mathbf{x}$ belongs to a finite alphabet. As it is shown in Section 2.4.1, this information is irrelevant at low SNR but it is crucial when the noise term vanishes at high SNR. Nonetheless, in the following, we will refer to this estimator as the CML estimator regardless of having discrete or continuous nuisance parameters.

Therefore, if there is absolutely no information about $\mathbf{x}$, the nuisance parameters must be assumed *deterministic, continuous* unknowns. Then, the ML estimator of $\mathbf{x}$ in (2.9) is obtained in the linear case by solving a classical weighted least squares (WLS) problem leading to

$$\widehat{\mathbf{x}}_{ML}\left(\boldsymbol{\theta}\right) = \left(\mathbf{A}^H\left(\boldsymbol{\theta}\right)\mathbf{R}_w^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right)\right)^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right)^H\mathbf{R}_w^{-1}\mathbf{y},$$

assuming that $\mathbf{A}\left(\boldsymbol{\theta}\right)$ is a tall matrix, i.e., $M > K$ [Sch91a, Sec. 9.12]. After some algebra, the corresponding log-likelihood function is given by

$$\begin{aligned}\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}\left(\boldsymbol{\theta}\right)\right) &\propto -\left\|\mathbf{y}-\mathbf{A}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{x}}_{ML}\left(\boldsymbol{\theta}\right)\right\|^2_{\mathbf{R}_w} \\ &\propto \operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right),\end{aligned} \tag{2.19}$$

becoming a linear transformation of the sample covariance matrix and, thus, a *quadratic* function of the observation $\mathbf{y}$. Finally, the CML estimator of $\boldsymbol{\theta}$ is computed as follows:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{CML} &= \arg\max_{\boldsymbol{\theta}} \ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}\left(\boldsymbol{\theta}\right)\right) \\ &= \arg\max_{\boldsymbol{\theta}} \operatorname{Tr}\left(\mathbf{N}^{-1}\mathbf{A}\left(\mathbf{A}^{H}\mathbf{N}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^{H}\mathbf{N}^{-1}\widehat{\mathbf{R}}\right) \\ &= \arg\max_{\boldsymbol{\theta}} \operatorname{Tr}\left(\mathcal{M}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{R}}\right), \end{aligned} \tag{2.20}$$

with

$$\mathcal{M}\left(\boldsymbol{\theta}\right) = \mathbf{N}^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right)\left(\mathbf{A}^{H}\left(\boldsymbol{\theta}\right)\mathbf{N}^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right)\right)^{-1}\mathbf{A}^{H}\left(\boldsymbol{\theta}\right)\mathbf{N}^{-1}$$

the associated kernel.

The resulting estimator is actually projecting the *whitened* observation $\mathbf{N}^{-1/2}\mathbf{y}$ onto the orthogonal subspace generated by the columns of $\mathbf{N}^{-1/2}\mathbf{A}\left(\boldsymbol{\theta}\right)$. Clearly, the above solution is related to subspace methods like MUSIC [Sch79][Bie80][Sto89][Sto97]. In fact, the CML estimator in (2.20) is equivalent to a variant of the MUSIC algorithm proposed in [Sto89].

It can be seen that the CML estimator in (2.20) corresponds to the low-SNR UML estimator in (2.18) if $\mathbf{R}_w^{-1/2}\mathbf{A}\left(\boldsymbol{\theta}\right)$ is unitary or, in other words,

$$\mathbf{A}^{H}(\boldsymbol{\theta})\,\mathbf{N}^{-1}\mathbf{A}\left(\boldsymbol{\theta}\right) \propto \mathbf{I}_K. \tag{2.21}$$

If the above equation is not fulfilled, the CML estimator might suffer from *noise-enhancement* at low SNR when the observation length is limited. In that case, the low-SNR UML estimator deduced in Section 2.4.1 outperforms the CML estimator in the low SNR regime because the former exploits the a priori statistical knowledge about $\mathbf{x}$.

The CML solution is shown in Appendix 2.C to hold the following regularity condition:

$$E_{\mathbf{y}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}\left(\boldsymbol{\theta}\right)\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right\} = \mathbf{0}$$

and, therefore, the CML estimator is always *unbiased* and *self-noise free* even for finite observations. Another significative feature of the CML solution is that it is not necessary to know the variance of the noise samples σ_w^2.

2.4.3 Gaussian Maximum Likelihood (GML)

The Gaussian assumption on the nuisance parameters is generally adopted when the actual distribution is unknown or becomes an obstacle to compute the expectation in (2.15). The Gaussian assumption is known to yield *almost* optimal second-oder estimators on account of the Central Limit Theorem. This subject is addressed throughout this dissertation and the asymptotic efficiency of the Gaussian assumption is studied in Chapter 7.

If the nuisance parameters are Gaussian, the observed vector $\mathbf{y}$ is also Gaussian in the studied linear signal model. Thus, we have that

$$f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) = \frac{\exp\left(-\mathbf{y}^H\mathbf{R}^{-1}\left(\boldsymbol{\theta}\right)\mathbf{y}\right)}{\pi^M \det\left(\mathbf{R}\left(\boldsymbol{\theta}\right)\right)}, \tag{2.22}$$

where $\mathbf{y}$ is zero-mean and

$$\mathbf{R}\left(\boldsymbol{\theta}\right) \triangleq E\left\{\widehat{\mathbf{R}}\right\} = E\left\{\mathbf{y}\mathbf{y}^H\right\} = \mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{A}^H\left(\boldsymbol{\theta}\right) + \mathbf{R}_w \tag{2.23}$$

is the covariance matrix of $\mathbf{y}$. Once again, the log-likelihood solution is an affine transformation of the sample covariance matrix that, omitting constant additive terms, is given by

$$\ln f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) = \ln E_{\mathbf{x}}\left\{f_{\mathbf{y}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)\right\} = -\ln\det\left(\mathbf{R}\left(\boldsymbol{\theta}\right)\right) - \operatorname{Tr}\left(\mathbf{R}^{-1}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{R}}\right) \tag{2.24}$$

Therefore, having in mind that $\ln\det\left(\mathbf{M}\right) = \operatorname{Tr}\ln\left(\mathbf{M}\right)$, it is found that

$$\begin{aligned} b\left(\boldsymbol{\theta}\right) &= -\ln\det\left(\mathbf{R}\left(\boldsymbol{\theta}\right)\right) = -\operatorname{Tr}\left(\ln\mathbf{R}\left(\boldsymbol{\theta}\right)\right) \\ \mathcal{M}\left(\boldsymbol{\theta}\right) &= -\mathbf{R}^{-1}\left(\boldsymbol{\theta}\right) = -\left(\mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{A}^H\left(\boldsymbol{\theta}\right) + \mathbf{R}_w\right)^{-1} \end{aligned}$$

are the independent term and the kernel of the GML likelihood function, respectively. Consequently, the GML estimator is computed as follows:

$$\widehat{\boldsymbol{\theta}}_{GML} = \arg\min_{\boldsymbol{\theta}} \operatorname{Tr}\left(\ln\mathbf{R}\left(\boldsymbol{\theta}\right) + \mathbf{R}^{-1}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{R}}\right) \tag{2.25}$$

In Appendix 2.D, we prove that the GML estimator converges to the low-SNR UML solution (2.18) for $\sigma_w^2 \to \infty$ and to the CML solution (2.20) for $\sigma_w^2 \to 0$. Therefore, the GML estimator is asymptotically efficient at low SNR and, evidently, for any SNR if the nuisance parameters are Gaussian. Indeed, any statistical assumption about the nuisance parameters leads to the UML solution (2.18) at low SNR. Consequently, the GML estimator can only be outperformed using quadratic techniques in the medium-to-high SNR interval if the nuisance parameters are non-Gaussian random variables. This subject is addressed thoroughly in subsequent chapters.

2.5 Maximum Likelihood Implementation

Generally, the ML-based estimators presented in the last section does not admit an analytical solution[6] and the maximization of the associated log-likelihood function must be carried out using numerical techniques. In that case the log-likelihood function should be sampled. If

[6] An exception is the estimation of the carrier phase in digital communications (see Section 6.2).

the samples separation is decided according to the sampling theorem, the ML estimate can be determined by means of ideal interpolation. Otherwise, if the sampling rate violates the Nyquist criterion, a gradient-based algorithm can be applied to find the maximum of $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$. Moreover, if the gradient of $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ has a single root in the parameter space, a gradient-based algorithm is able to look for the maximum of $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ without any assistance. Nonetheless, in a multimodal problem, the same gradient-based method might converge to a local maximum unless a preliminary search of the global maximum is performed.

The utilization of a gradient-based or iterative algorithm is generally preferred because it has a lower complexity than the grid search implementation[7]. The convergence of gradient-based methods is guaranteed if and only if the Hessian matrix is negative definite –and lower bounded– in the closed subset $\Theta = \left\{\boldsymbol{\theta} \mid f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) \geq f_{\mathbf{y}}\left(\mathbf{y};\widehat{\boldsymbol{\theta}}_0\right)\right\}$ with $\widehat{\boldsymbol{\theta}}_0$ the initial guess [Boy04, Sec. 8.3.]. Among the existing gradient-based methods, the Newton-Raphson algorithm is extensively adopted because its convergence is quadratic –instead of linear– when the recursion approaches the log-likelihood maximum ($\widehat{\boldsymbol{\theta}}_{ML}$) [Boy04, Sec. 8.5.]. Other methods are the steepest descent method, conjugate gradient, quasi-Newton method, among many others (see [Boy04][Lue84] and references therein).

The Newton-Raphson iteration is given by

$$\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_k - \mathbf{H}^{-1}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k) \tag{2.26}$$

where k is the iterate index and

$$\begin{aligned}\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}) &\triangleq \frac{\partial \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \mathbf{H}(\mathbf{y};\boldsymbol{\theta}) &\triangleq \frac{\partial^2 \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\end{aligned}$$

are the gradient and the Hessian of the log-likelihood function, respectively. Notice that, in a low-SNR scenario (2.17) and/or if the nuisance parameters are Gaussian (2.24), $\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta})$ is linear in the sample covariance matrix $\widehat{\mathbf{R}} \triangleq \mathbf{y}\mathbf{y}^H$. In that case, the Newton-Raphson recursion in (2.26) is quadratic in the observation $\mathbf{y}$.

The quadratic convergence of the Newton-Raphson algorithm is accelerated when approaching $\widehat{\boldsymbol{\theta}}_{ML}$ because $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ becomes approximatelly parabolic around the current estimate $\widehat{\boldsymbol{\theta}}_k$, that is,

$$\begin{aligned}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) \simeq \ln f_{\mathbf{y}}\left(\mathbf{y};\widehat{\boldsymbol{\theta}}_k\right) + \boldsymbol{\nabla}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_k\right) + \\ \frac{1}{2}\operatorname{Tr}\left\{\mathbf{H}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_k\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_k\right)^T\right\}\end{aligned}$$

[7] Recall that the parameter $\boldsymbol{\theta} \in \mathbb{R}^P$ is a continuous variable and we are assuming that $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$.

and, therefore, (2.26) yields approximatelly the ML solution for $\left\|\widehat{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_{ML}\right\|$ sufficiently small. Notice that $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is strictly quadratic in case of linear estimation problems having additive Gaussian noise [Kay93b, Theorem 3.5]. In that case, the ML estimate is obtained after a single iteration of the Raphson-Newton algorithm. Otherwise, the convergence rate is slow if the log-likelihood curvature is large near the maximum $\widehat{\boldsymbol{\theta}}_{ML}$. In that case, however, the estimation accuracy is found to be superior.

The Newton-Raphson method in (2.26) can be generalized to estimate a given transformation of the parameter [Sto01][Kay93b, Sec. 3.8] as follows

$$\widehat{\boldsymbol{\alpha}}_{k+1} = \widehat{\boldsymbol{\alpha}}_k - \mathbf{D}_g(\widehat{\boldsymbol{\theta}}_k)\mathbf{H}^{-1}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k) \tag{2.27}$$

where $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ is the referred transformation, $\mathbf{D}_g(\boldsymbol{\theta}) \triangleq \partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ is the Jacobian of $\mathbf{g}(\boldsymbol{\theta})$ and, $\widehat{\boldsymbol{\alpha}}_{ML} = \mathbf{g}(\widehat{\boldsymbol{\theta}}_{ML})$ holds from the invariance property of the ML estimator (Section 2.3.2).

According to the asymptotic properties of the ML estimator (Section 2.3.2), it follows that *any* iterative method converging to the ML solution is asymptotically ($M \to \infty$) consistent and efficient, if the ML regularity condition is satisfied (2.12). In the asymptotic case, the *small-error* condition is verified and the ML estimator attains the Cramer-Rao bound (Section 2.6.1), which is given by

$$\mathbf{B}_{CRB}(\boldsymbol{\theta}_o) \triangleq \mathbf{D}_g(\boldsymbol{\theta}_o)\mathbf{J}^{-1}(\boldsymbol{\theta}_o)\mathbf{D}_g^H(\boldsymbol{\theta}_o),$$

where

$$\mathbf{J}(\boldsymbol{\theta}) \triangleq -E_{\mathbf{y}}\{\mathbf{H}(\mathbf{y};\boldsymbol{\theta})\} = E_{\mathbf{y}}\left\{\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta})\boldsymbol{\nabla}^H(\mathbf{y};\boldsymbol{\theta})\right\} \tag{2.28}$$

is the Fisher's information matrix (FIM) and the expectation is computed with respect to the random observation $\mathbf{y}$. The last equality is a consequence of the regularity condition (2.12) [Kay93b, Appendix 3A].

The asymptotic efficiency is also guaranteed if the Newton-Raphson method (2.27) is substituted by the following *scoring* method:

$$\widehat{\boldsymbol{\alpha}}_{k+1} = \widehat{\boldsymbol{\alpha}}_k + \mathbf{D}_g(\widehat{\boldsymbol{\theta}}_k)\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}(\mathbf{y};\widehat{\boldsymbol{\theta}}_k), \tag{2.29}$$

in which the Hessian matrix is replaced by the negative of its expected value (2.28). The method of scoring is preferred because it improves the convergence to the ML solution for short data records, mainly in multiparametric problems. However, both methods are equivalent if the observation size goes to infinity.

2.5.1 ML-Based Closed-Loop Estimation

Conventionally, ML estimators are developed in *batch* mode, that is, the M samples of $\mathbf{y}$ are recorded first and, afterwards, $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is iteratively maximized in order to find the ML

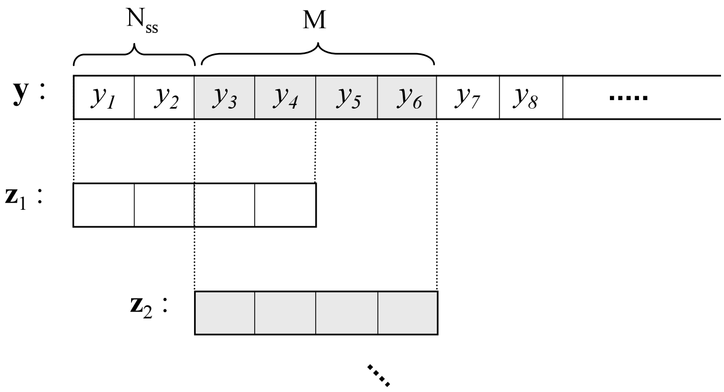


Figure 2.3: Sequential processing of the received vector $\mathbf{y}$ in the context of digital communications. The observed blocks $\{\mathbf{z}_n\}$ last $M = 4$ samples and are taken every $N_{ss} = 2$ samples where N_{ss} is the number of samples per symbol.

estimate $\widehat{\boldsymbol{\theta}}_{ML}$. Unfortunately, the complexity and latency of this *batch-mode* implementation is excessive when long observations are required to comply with the specifications. To ameliorate this problem, the long observation $\mathbf{y}$ is fragmented into smaller blocks $\{\mathbf{z}_n\}_{n=1,\ldots,N}$ that are ergodic realizations of the same distribution $f_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})$. The minimum block size is one sample, in which case the estimator would work in a sample-by-sample basis. Identically distributed blocks are feasible if the observation is (cyclo-)stationary.

In Appendix 2.E, it is shown that the following *closed-loop* estimator,

$$\widehat{\boldsymbol{\alpha}}_{n+1} = \widehat{\boldsymbol{\alpha}}_n + \mu \mathbf{D}_g(\widehat{\boldsymbol{\theta}}_n)\mathbf{J}_{\mathbf{z}}^{-1}(\widehat{\boldsymbol{\theta}}_n)\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\widehat{\boldsymbol{\theta}}_n), \tag{2.30}$$

is efficient in the small-error regime if the N partial observations $\mathbf{z}_n$ are *statistically independent*, where

$$\begin{aligned}\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta}) &\triangleq \frac{\partial \ln f_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta}) &\triangleq -E_{\mathbf{z}}\left\{\frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\} = E_{\mathbf{z}}\left\{\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})\,\boldsymbol{\nabla}^H(\mathbf{z};\boldsymbol{\theta})\right\} = \frac{1}{N}\mathbf{J}(\boldsymbol{\theta})\end{aligned}$$

is the gradient and the FIM for the block-size observations $\{\mathbf{z}_n\}_{n=1,\ldots,N}$, respectively. The *step-size* or *forgetting factor* μ is selected to achieve the same performance than the off-line recursions in (2.27) and (2.29). If N is sufficiently large, the parameter μ must be set to approximatelly $2/N$ (Appendix 2.E).

Although closed-loop estimators have the same aspect as their off-line versions in (2.27) and (2.29), the closed-loop scheme in (2.30) aims at maximizing the *stochastic* likelihood function $f_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})$, which has a time-varying shape. Therefore, the gradient $\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})$ is also a random vector pointing into the direction of the maximum of $f_{\mathbf{z}}(\mathbf{z};\boldsymbol{\theta})$. Thus, the ML-based closed-loop estimator proposed in (2.30) belongs to the family of *stochastic gradient* algorithms. Indeed, equation (2.30) is referred to as the *natural gradient* in the context of neural learning [Ama98].

Despite the closed-loop estimator in (2.29) has been deduced assuming N independent blocks, the necessary and sufficient condition for efficiency is more general and is formulated next.

Proposition 2.1 *The closed loop estimator proposed in (2.30) is efficient in the small-error regime if and only if there is at least one block* $\mathbf{z}_n$ *in which each sample* $[\mathbf{y}]_m$ *(*$m = 1, ..., M$*) is jointly processed with all the samples that are statistically dependent on it.*

The above proposition implies in most cases the partial overlapping of the observed blocks. This means that the same sample is processed more than once. For example, in digital communications the received signal is cyclostationary if we have $N_{ss} > 1$ samples per symbol. The data symbols are usually i.i.d. random variables that modulate a known pulse $p(t)$ lasting LN_{ss} samples. In that case, the optimal performance is achieved if the block size is equal to LN_{ss} samples and the block separation is one sample. However, in order to have identically distributed blocks, the block separation is usually set to N_{ss}, taking into account the signal cyclostationarity (see Fig. 2.3).

As it has been previously stated, closed-loop estimators yield efficient estimates if the *small-error* regime is attained in the steady-state. However, the initial guess $\widehat{\boldsymbol{\theta}}_0$ is usually far away from the true parameter $\boldsymbol{\theta}_o$ and the algorithm has to converge towards $\boldsymbol{\theta}_o$. The initial convergence constitutes the estimator *acquisition* and has been studied for a long time [Mey90]. Unfortunately, only approximated results are available on the acquisition mean time, lock-in and lock-out probability, etc. [Mey90]. The step-size μ in equation (2.30) can be adjusted to trade acquisition speed –large μ– and steady-state performance –small μ.

Closed Loop Architecture

The ML-based closed loop proposed in (2.30) has two components (Fig. 2.4): a nonlinear *discriminator (or detector)* of the estimation error, and a first-order loop filter. The discriminator input-output response is given by

$$\widehat{\mathrm{e}}(\mathbf{z}_n;\boldsymbol{\theta}) = \mathbf{D}_g(\boldsymbol{\theta})\mathbf{J}_{\mathbf{z}}^{-1}(\boldsymbol{\theta})\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$ is the current estimate serving as a reference to infere the estimation error $\mathbf{g}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{g}(\boldsymbol{\theta}_o)$ at time n.

The mean value of the discriminator output is given by

$$E_{\mathbf{z}}\left\{\widehat{\mathrm{e}}(\mathbf{z}_n;\boldsymbol{\theta})\right\} = \mathbf{D}_g(\boldsymbol{\theta})\mathbf{J}_{\mathbf{z}}^{-1}(\boldsymbol{\theta})E_{\mathbf{z}}\left\{\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})\right\}. \tag{2.31}$$

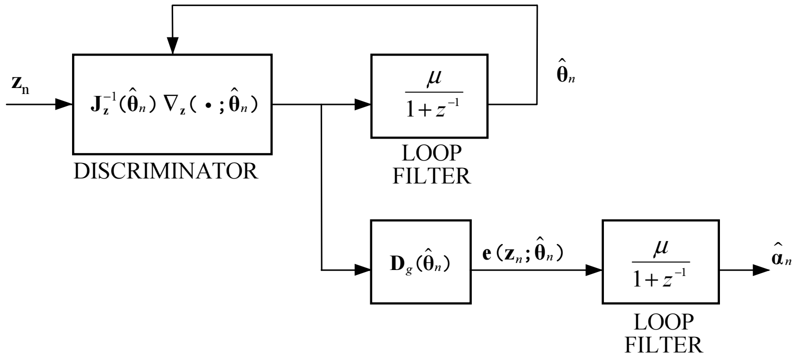

$\mathbf{z}_n$
$\mathbf{J}_z^{-1}(\hat{\boldsymbol{\theta}}_n)\,\nabla_z(\bullet\,;\hat{\boldsymbol{\theta}}_n)$
DISCRIMINATOR
$\frac{\mu}{1+z^{-1}}$
LOOP FILTER
$\hat{\boldsymbol{\theta}}_n$
$\mathbf{D}_g(\hat{\boldsymbol{\theta}}_n)$
$\mathbf{e}(\mathbf{z}_n;\hat{\boldsymbol{\theta}}_n)$
$\frac{\mu}{1+z^{-1}}$
LOOP FILTER
$\hat{\boldsymbol{\alpha}}_n$

Figure 2.4: Block diagram for the ML closed-loop estimator in equation (2.30). The same scheme is applicable to any other closed-loop estimator or tracker if the discriminator and/or the loop filters are conveniently modified.

It can be shown that the discriminator output is unbiased in the neighbourhood of the equilibrium point $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ because

$$\begin{aligned} E_{\mathbf{z}}\left\{\widehat{\mathbf{e}}\left(\mathbf{z}_n;\boldsymbol{\theta}_o\right)\right\} &= \mathbf{0} \\ \frac{\partial}{\partial\boldsymbol{\theta}_o^T} E_{\mathbf{z}}\left\{\widehat{\mathbf{e}}\left(\mathbf{z}_n;\boldsymbol{\theta}\right)\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= \mathbf{D}_g\left(\boldsymbol{\theta}_o\right), \end{aligned}$$

taking into account that

$$\begin{aligned} E_{\mathbf{z}}\left\{\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta}_o)\right\} &= \mathbf{0} \\ \left(\frac{\partial}{\partial\boldsymbol{\theta}_o^T} E_{\mathbf{z}}\left\{\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})\right\}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= -\left(\frac{\partial}{\partial\boldsymbol{\theta}^T} E_{\mathbf{z}}\left\{\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})\right\}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ &= -E_{\mathbf{z}}\left\{\frac{\partial \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right\} = \mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta}_o) \end{aligned}$$

is always verified in the studied linear signal model (Section 2.4). The first equation is the classical regularity condition introduced in (2.12) and the second equation is the Fisher's information matrix $\mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta})$. Precisely, $\mathbf{J}_{\mathbf{z}}^{-1}(\boldsymbol{\theta})$ normalizes the discriminator slope in (2.31) to have unbiased estimates of $\boldsymbol{\theta} - \boldsymbol{\theta}_o$. The Jacobian matrix $\mathbf{D}_g\left(\boldsymbol{\theta}\right)$ is then used to obtain unbiased estimates of $\mathbf{g}\left(\boldsymbol{\theta}\right) - \mathbf{g}\left(\boldsymbol{\theta}_o\right)$ taking into account that $\mathbf{g}\left(\boldsymbol{\theta}\right)$ can be linearized around $\boldsymbol{\theta} \simeq \boldsymbol{\theta}_o$ using the first-order Taylor expansion $\mathbf{g}\left(\boldsymbol{\theta}\right) \simeq \mathbf{g}\left(\boldsymbol{\theta}_o\right) + \mathbf{D}_g\left(\boldsymbol{\theta}_o\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}_o\right)$.

In some problems of digital communications, the discriminator mean value (2.31) only depends on the estimation error $\boldsymbol{\theta} - \boldsymbol{\theta}_o$ and is named the discriminator *S-curve* because it looks like an "S" rotated by 90^o [Men97][Mey90].

2.5.2 ML-based Tracking

An important feature of the stochastic gradient methods previously presented is the ability of *tracking* the evolution of slowly time-varying parameters. Thus, let us consider that $\boldsymbol{\theta}_n$ is a time-varying parameter and $\boldsymbol{\alpha}_n = \mathbf{g}(\boldsymbol{\theta}_n)$ a given transformation. The closed loop in (2.30) must be modified to track the parameter evolution and supply unbiased estimates of $\boldsymbol{\theta}_n$ in the steady-state.

A first-order loop filter was used in the last section because the parameter was constant, i.e., $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ (Fig. 2.4). However, if $\boldsymbol{\theta}_n$ has a polynomial evolution in time, i.e.,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_o + \sum_{r=1}^{R-1} \boldsymbol{\delta}_r n^r$$

a Rth-order loop filter is required to track $\boldsymbol{\theta}_n$ without systemmatic or pursuit errors [Men97][Mey90]. For example, if θ_o is the carrier phase and we are designing a phase-lock loop (PLL), δ_1 corresponds to the Doppler frequency and δ_2 to the Doppler rate.

Another alternative to take into account the parameter dynamics is the one adopted in the Kalman filter theory [Kay93b, Ch.13]. In this framework, a dynamical model (or state-equation) is assumed for the parameters of interest

$$\boldsymbol{\theta}_{n+1} = \mathbf{f}(\boldsymbol{\theta}_n),$$

where $\boldsymbol{\theta}_n$ stacks all the parameters involved in the dynamical model, i.e., $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_o^T, \boldsymbol{\delta}_1^T, ..., \boldsymbol{\delta}_{R-1}^T\right]^T$ for the polynomial model above. Although the parameter dynamics are generally nonlinear, they can be linearized around the actual estimate $\widehat{\boldsymbol{\theta}}_n$, leading to the following approximation

$$\mathbf{f}(\boldsymbol{\theta}_n) \simeq \mathbf{f}\left(\widehat{\boldsymbol{\theta}}_n\right) + \mathbf{D}_f\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}_n\right)$$

where $\mathbf{D}_f(\boldsymbol{\theta}) \triangleq \partial\mathbf{f}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ is the Jacobian of $\mathbf{f}(\boldsymbol{\theta})$.

If the parameter dynamics are incorporated into the original closed loop (2.30), we obtain the following higher-order tracker

$$\widehat{\boldsymbol{\alpha}}_{n+1} = \mathbf{h}(\widehat{\boldsymbol{\theta}}_n) + \operatorname{diag}(\boldsymbol{\mu})\,\mathbf{D}_h\left(\widehat{\boldsymbol{\theta}}_n\right)\mathbf{J}_{\mathbf{z}}^{-1}(\widehat{\boldsymbol{\theta}}_n)\boldsymbol{\nabla}_n(\mathbf{z}_n;\widehat{\boldsymbol{\theta}}_n) \tag{2.32}$$

where $\mathbf{h}(\boldsymbol{\theta}) \triangleq \mathbf{g}(\mathbf{f}(\boldsymbol{\theta}))$ and

$$\mathbf{D}_h(\boldsymbol{\theta}) \triangleq \partial\mathbf{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T = \mathbf{D}_g(\boldsymbol{\theta})\,\mathbf{D}_f(\boldsymbol{\theta})$$

is the Jacobian of the composite function $\mathbf{h}(\boldsymbol{\theta})$ [Gra81, Sec. 4.3.][8]. The vector of forgetting factors $\boldsymbol{\mu}$ sets the (noise equivalent) loop bandwidth of each parameter in $\boldsymbol{\alpha}_n$. In Appendix 2.E

[8] If the dynamical model is specified for $\boldsymbol{\alpha}_n$, i.e., $\widehat{\boldsymbol{\alpha}}_{n+1} = \mathbf{f}(\widehat{\boldsymbol{\alpha}}_n)$, the composition must be reversed having that $\mathbf{h}(\boldsymbol{\theta}) \triangleq \mathbf{f}(\mathbf{g}(\boldsymbol{\theta}))$.

it is shown that $B_n \simeq \mu/4$. The loop bandwidth determines the maximum variability of the parameters that the closed loop is able to track as well as the closed loop effective observation time that, approximately, is equal to $N \simeq 0.5/B_n$ samples [Men97, Sec. 3.5.6] (see also Appendix 2.E).

A vast number of tracking techniques have been proposed in the field of automatic control [Kai00][Söd89], signal processing [Kay93b] and communications [Men97][Mey90], e.g., least mean squares (LMS) and recursive least squares (RLS) [Hay91][Kai00], Kalman-Bucy filtering [And79][Hay91], machine learning [Mit97], etc. In fact, filtering, smoothing, prediction, deconvolution, source separation and other applications can be seen as particular cases of parameter estimation or tracking in which the aim is to determine the input data at time n, say $\boldsymbol{\theta}_n$, from a vector of noisy observations.

2.6 Lower Bounds in Parameter Estimation

The calculation of an *attainable* benchmark for the adopted performance criterion is necessary to identify whether a given estimation technique is efficient or not. For example, the ML estimator is known to be optimal in the small-error regime because it attains the Cramer-Rao lower bound. Once the optimal performance is known, suboptimal techniques can be devised trading-off performance and complexity. Moreover, lower bounds usually give insight into the contribution of the different parameters onto the estimator performance (e.g., SNR, observation size and others). In the following sections, some important lower bounds are briefly described.

Focusing on the mean squared error (MSE), lower bounds can be classified as Bayesian or deterministic depending on whether the prior statistics of the parameters are exploited or not. On the other hand, lower bounds are also classified into *small-error* (or local) bounds and *large-error* (global) bounds. Furthermore, the lower bounds in the literature are derived from either the Cauchy-Schwarz or Kotelnikov inequalies.

From the above classification criteria, the most important lower bounds in the literature are described and interconnected in the following subsections. Finally, all these bounds are organized and presented in a concluding table at the end of the section (Fig. 2.5).

NOTE: *the material in the following section is not essential to understand the central chapters of the dissertation. Only those lower bounds derived from the CRB in the presence of nuisance parameters will be extensively used throughout the thesis. Thus, we recommend the reader to skip Section 2.6 in the first reading.*

2.6.1 Deterministic Bounds based on the Cauchy-Schwarz Inequality

A large number of deterministic lower bounds on the mean square error (MSE) have been derived from the Cauchy-Schwarz inequality, e.g., [Gor90, Eq. 10][Abe93, Eq. 5][McW93, Eq. 2][Rif75, Eq. 13]. The Cauchy-Schwarz inequality states that

$$E\left\{\mathbf{e}\mathbf{e}^H\right\} \geq E\left\{\mathbf{e}\mathbf{s}^H\right\}\left(E\left\{\mathbf{s}\mathbf{s}^H\right\}\right)^{\#} E\left\{\mathbf{s}\mathbf{e}^H\right\} \tag{2.33}$$

for two arbitrary random vectors $\mathbf{e}$ and $\mathbf{s}$.[9] The Moore-Penrose pseude-inverse operator was introduced in [Gor90, Eq. 10][Sto01] to cover those cases in which $E\left\{\mathbf{s}\mathbf{s}^H\right\}$ is singular. Notice that the expectation is computed with respect to the random components of $\mathbf{e}$ and $\mathbf{s}$. Furthermore, equation (2.33) holds with equality if and only if the vector $\mathbf{e}$ and $\mathbf{s}$ are connected as

[9] For the scalar case, we have the conventional Cauchy-Schwarz inequality, $E\left|e\right|^2 \geq \left|E\left\{es\right\}\right|^2 / E\left|s\right|^2$, as it appears in [Wei88b, Eq. 7]

follows

$$\mathbf{e} = E\left\{\mathbf{e}\mathbf{s}^H\right\}\left(E\left\{\mathbf{s}\mathbf{s}^H\right\}\right)^{\#}\mathbf{s}. \tag{2.34}$$

Indeed, the Cauchy-Schwarz inequality is a consequence of the more general relation

$$\begin{bmatrix} \mathbf{A}\,\mathbf{B}^H \\ \mathbf{B}\,\mathbf{C} \end{bmatrix} \geq \mathbf{0} \Leftrightarrow \mathbf{A} \geq \mathbf{B}^H\mathbf{C}^{\#}\mathbf{B} \tag{2.35}$$

which is valid if $\mathbf{C}$ is non-negative definite [Mag98, Ex. 3, p. 25]. This property is used in [Gor90, Lemma 1] to prove the vectorial Cauchy-Schwarz inequality (2.33). The proof is straightforward if (2.35) is applied to the matrix $E\left\{\mathbf{z}\mathbf{z}^H\right\}$ with $\mathbf{z} \triangleq \left[\mathbf{e}^T, \mathbf{s}^T\right]^T$. Also, this matrix inequality is adopted in [McW93, Eq. 2] to analyze the geometry of several "quadratic covariance bounds".

Based on the Cauchy-Schwarz inequality (2.33), lower bounds on the estimation mean square error can be formulated considering that $\mathbf{e} = \widehat{\boldsymbol{\alpha}} - \mathbf{g}(\boldsymbol{\theta})$ is the estimation error and $\mathbf{s}$ an arbitrary *score function*. In the deterministic case, both $\mathbf{e}$ and $\mathbf{s}$ are functions of the random observation $\mathbf{y}$, which is distributed as $f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})$. Various deterministic lower bounds on the MSE have been deduced by selecting different score functions $\mathbf{s}$ as, for instance, the following well-known bounds; Cramér-Rao [Kay93b, Chapter 3], Battacharyya [Bat46], Barankin [Bar49], Hammersley-Chapman-Robbins [Cha51][Ham50], Abel [Abe93] and Kiefer [Kie52], among others.

Because (2.33) is valid for any score function, the aim is to find the score function leading to the *highest* lower bound on the estimator MSE and, if possible, the estimator attaining the resulting bound. Conversely, if an estimator satisfies (2.33) with equality for a given score function, the resulting bound is the tightest, attainable lower bound on the MSE. Furthermore, this estimator is the one holding (2.34).

In [McW93], it is shown that tight lower bounds are obtained provided that

P1: the score function is zero-mean, i.e.,

$$E\left\{\mathbf{s}(\mathbf{y}, \boldsymbol{\theta})\right\} = \int \mathbf{s}(\mathbf{y}, \boldsymbol{\theta})\, f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})\, d\mathbf{y} = \mathbf{0}$$

for every value of $\boldsymbol{\theta}$. Thus, we are only concerned with unbiased estimators since the estimation error is proportional to $\mathbf{s}(\mathbf{y}, \boldsymbol{\theta})$ (2.34);

P2: the score function is a function of the sufficient statistics of the estimation problem at hand. Recall that $\mathbf{t}(\mathbf{y})$ is a sufficient statistic if and only if $f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})$ depends on the parameter vector $\boldsymbol{\theta}$ *uniquely* throught a function of the sufficient statistic $\mathbf{t}(\mathbf{y})$. Consequently, $\mathbf{s}(\mathbf{y}, \boldsymbol{\theta})$ can be any biunivoque function of the likelihood function $f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})$, as for instance, its gradient $\boldsymbol{\nabla}(\mathbf{y}; \boldsymbol{\theta})$. See the Neyman-Fisher factorization theorem in [Kay93b, Th. 5.3];

P3: the score function must hold (2.34). This mean that $\mathbf{s}(\mathbf{y},\boldsymbol{\theta})$ must span the estimation error subspace.

The first property is really important because it states that we only have to consider unbiased estimators of the parameter. In fact, it can be shown that the bias term always increases the overall MSE and it is not possible to trade bias for variance in the deterministic case. To show this result, we have to decompose the estimation error as

$$\mathbf{e}(\mathbf{y};\boldsymbol{\theta}) \triangleq \mathbf{b}(\boldsymbol{\theta}) + \mathbf{v}(\mathbf{y};\boldsymbol{\theta})$$

with $\mathbf{b}(\boldsymbol{\theta}) = E\{\widehat{\boldsymbol{\alpha}}(\mathbf{y})\} - \mathbf{g}(\boldsymbol{\theta})$ the estimator bias and $\mathbf{v}(\mathbf{y};\boldsymbol{\theta}) = \widehat{\boldsymbol{\alpha}}(\mathbf{y}) - E\{\widehat{\boldsymbol{\alpha}}(\mathbf{y})\}$ the deviation with respect to the estimator mean. Consequently, the estimator MSE can be written as

$$\Sigma_{\mathbf{ee}}(\boldsymbol{\theta}) = \mathbf{b}(\boldsymbol{\theta})\,\mathbf{b}^H(\boldsymbol{\theta}) + \Sigma_{\mathbf{vv}}(\boldsymbol{\theta})$$

where

$$\Sigma_{\mathbf{xy}} \triangleq E\left\{\mathbf{x}\mathbf{y}^H\right\}$$

stands for the cross correlation matrix[10]. Then, the Cauchy-Schwarz inequality (2.33) can be applied to the covariance matrix $\Sigma_{\mathbf{vv}}(\boldsymbol{\theta})$ in order to obtain the following lower bound on the MSE:

$$\Sigma_{\mathbf{ee}}(\boldsymbol{\theta}) \geq \mathbf{b}(\boldsymbol{\theta})\,\mathbf{b}^H(\boldsymbol{\theta}) + \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) \tag{2.36}$$

in which the bias function has been set to $\mathbf{b}(\boldsymbol{\theta})$ [Abe93, Eq. 6]. Equation (2.36) is usually referred to as the “covariance inequality” [Gor90][McW93][Abe93, Eq. 6]. Therefore, if the covariance inequality in (2.36) is compared with the original bound,

$$\Sigma_{\mathbf{ee}}(\boldsymbol{\theta}) \geq \Sigma_{\mathbf{es}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{se}}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}),$$

it follows that the bias term $\mathbf{b}(\boldsymbol{\theta})\,\mathbf{b}^H(\boldsymbol{\theta})$ can never reduce the MSE matrix $\Sigma_{\mathbf{ee}}(\boldsymbol{\theta})$. In the last expression, we take into account that $\Sigma_{\mathbf{es}} = \Sigma_{\mathbf{vs}}$ because the score function is zero-mean.

The Cauchy-Schwarz inequality can be used then to extend the concept of *efficiency* to other lower bounds besides the usual Cramér-Rao bound. Thus, $\widehat{\boldsymbol{\alpha}}(\mathbf{y})$ is an efficient estimator of $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ if and only if it holds that

$$E\{\mathbf{e}(\mathbf{y},\boldsymbol{\theta})\} = \mathbf{0} \tag{2.37}$$

$$\Sigma_{\mathbf{ee}}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vv}}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) \tag{2.38}$$

for (at least) a score function $\mathbf{s}(\mathbf{y},\boldsymbol{\theta})$.

[10] Notice that the transpose conjugate will be considered in the sequel for both real and complex vectors.

Additionally, we know from (2.34) that the estimator $\widehat{\boldsymbol{\alpha}}(\mathbf{y})$ is efficient if and only if it verifies that

$$\widehat{\boldsymbol{\alpha}}(\mathbf{y}) = \mathbf{g}(\boldsymbol{\theta}) + \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\mathbf{s}(\mathbf{y},\boldsymbol{\theta}) \tag{2.39}$$

for any value of $\boldsymbol{\theta}$.

An important question is whether a *realizable*[11], unbiased estimator can attain the covariance inequality or not for a given score function. If this estimator was found, the resulting covariance would constitute the highest lower bound. Therefore, any other score function $\mathbf{s}(\mathbf{y},\boldsymbol{\theta})$ would yield a weaker bound on the MSE, which will not be attainable. Next, a sufficient condition on $\mathbf{s}(\mathbf{y},\boldsymbol{\theta})$ leading to realizable estimators is shown.

Proposition 2.2 *If the zero-mean score function can be factorized as*

$$\mathbf{s}(\mathbf{y},\boldsymbol{\theta}) \triangleq \mathbf{H}(\boldsymbol{\theta})\mathbf{z}(\mathbf{y}) - \mathbf{u}(\boldsymbol{\theta})$$

with $\mathbf{z}(\mathbf{y})$ *a function of the sufficient statistics* $\mathbf{t}(\mathbf{y})$ *and*

$$\begin{aligned}\Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\mathbf{H}(\boldsymbol{\theta}) &\triangleq \mathbf{M}^{H}\\ \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\mathbf{u}(\boldsymbol{\theta}) &= \mathbf{g}(\boldsymbol{\theta}),\end{aligned}$$

the estimator $\widehat{\boldsymbol{\alpha}}(\mathbf{y}) = \mathbf{M}^{H}\mathbf{z}(\mathbf{y})$ *is efficient and its covariance matrix is given by*

$$\Sigma_{\mathbf{ee}}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vv}}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) = \mathbf{M}^{H}\Sigma_{\mathbf{zz}}\mathbf{M} - \mathbf{g}(\boldsymbol{\theta})\mathbf{g}^{H}(\boldsymbol{\theta})$$

that becomes therefore the highest lower bound on the estimation error covariance.

Unfortunately, most score functions of interest cannot be factorized as in the last proposition for all the values of $\boldsymbol{\theta}$. Consequently, efficient estimators are usually *unrealizable* in the deterministic framework. In that case, efficient deterministic estimators are only feasible in the *small-error regime* once the value of $\boldsymbol{\theta}$ has been iteratively learnt using a suitable gradient-based method. Notice that this was the adopted approach in the case of the ML estimator and the associated Cramér-Rao bound. Thus, the following scoring method

$$\widehat{\boldsymbol{\alpha}}_{k+1} = \mathbf{g}\left(\widehat{\boldsymbol{\theta}}_k\right) + \Sigma_{\mathbf{vs}}\left(\widehat{\boldsymbol{\theta}}_k\right)\Sigma_{\mathbf{ss}}^{\#}\left(\widehat{\boldsymbol{\theta}}_k\right)\mathbf{s}\left(\mathbf{y},\widehat{\boldsymbol{\theta}}_k\right)$$

is efficient in the small-error regime (i.e., $\widehat{\boldsymbol{\theta}}_k \simeq \boldsymbol{\theta}$) *for any score function.*

Consequently, all the deterministic bounds will converge to the Cramér-Rao bound in the small-error regime. However, the Cramér-Rao bound is not attained when the estimator operates in the *large-error regime.* In that case, tighter bounds can be formulated by using a better score function. Next, the score functions associated to the most important large-error and small-error deterministic bounds are presented.

[11] The adjective "realizable" means that $\widehat{\boldsymbol{\alpha}}(\mathbf{y})$ does not depend on the vector of unknown parameters $\boldsymbol{\theta}$.

Barankin Bound (BB)

The Barankin bound was originally formulated in [Bar49] for scalar, real-valued estimation problems. The Barankin bound is constructed looking for the estimator minimizing its s^{th}-order absolute central moment subject to the unbiased constraint over all the parameter space Θ, i.e.,

$$\widehat{\alpha}_{BB} = \arg\min_{\widehat{\alpha}} E\left|\widehat{\alpha} - g\left(\theta_o\right)\right|^s \text{ subject to } E\left\{\widehat{\alpha}\right\} = g\left(\theta\right) \tag{2.40}$$

for every $\theta \in \Theta$. Focusing on the estimator variance ($s = 2$), it can be only stated that $\widehat{\alpha}_{BB}$ is the minimum variance unbiased estimator in the neighbourhood of θ_o. Furthermore, if the obtained local solution is independent of θ_o, $\widehat{\alpha}_{BB}$ turns out to be the *global* minimum variance unbiased estimator.

The Barankin bound has been extended in [Mar97] to multivariate estimation problems adopting a simpler formulation than the original one. In [Mar97, Eq. 9] the Barankin bound was shown to be a covariance inequality bound (2.38) with

$$\mathbf{s}\left(\mathbf{y};\boldsymbol{\theta}\right) = \int_{\widetilde{\boldsymbol{\theta}}\in\Theta} \frac{f_{\mathbf{y}}(\mathbf{y};\widetilde{\boldsymbol{\theta}})}{f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right)} \mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}} \tag{2.41}$$

the adopted score function, and $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}) \in \mathbb{R}^P$ an arbitrary function that must be selected to supply the tightest covariance lower bound. Notice that tighter lower bounds will be obtained if the mean of the score function is null (property 1), i.e.,

$$E\left\{\mathbf{s}\left(\mathbf{y};\boldsymbol{\theta}\right)\right\} = \int_{\widetilde{\boldsymbol{\theta}}\in\Theta} f_{\mathbf{y}}(\mathbf{y};\widetilde{\boldsymbol{\theta}})\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}} = \mathbf{0}.$$

Therefore, the functions $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ leading to the tightest lower bound must be proportional to the difference of two vectors of probability density functions $\mathbf{f}_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ and $\mathbf{f}_2(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$, i.e.,

$$\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}) = \kappa\left[\mathbf{f}_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}) - \mathbf{f}_2(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})\right]$$

with κ an arbitrary constant (e.g., $\kappa = 1$) and $\int \mathbf{f}_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}} = \int \mathbf{f}_2(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}} = \mathbf{1}$. This relevant property of $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ was taken into account in [Tre68, Pr. 2.4.18] to derive the Barankin bound in a different way. Also, a multidimensional version of the Kiefer bound [Kie52] can be obtained replacing $\mathbf{f}_2(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ by a multivariate delta measure $\boldsymbol{\delta}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})$.

Using now the covariance inequality, we have that the Barankin bound for the estimation of $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ is given by

$$\mathbf{B}_{BB}\left(\boldsymbol{\theta}\right) = \sup_{\mathbf{f}(\widetilde{\boldsymbol{\theta}})} \Sigma_{\mathbf{vs}}\left(\boldsymbol{\theta}\right)\Sigma_{\mathbf{ss}}^{\#}\left(\boldsymbol{\theta}\right)\Sigma_{\mathbf{sv}}\left(\boldsymbol{\theta}\right) \leq \Sigma_{\mathbf{vv}}\left(\boldsymbol{\theta}\right) \tag{2.42}$$

with

$$\Sigma_{\mathbf{vs}}\left(\boldsymbol{\theta}\right) = \int_{\widetilde{\boldsymbol{\theta}}\in\Theta} \left[\mathbf{g}(\widetilde{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta})\right]\mathbf{f}^H(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}}$$

$$\Sigma_{\mathbf{ss}}\left(\boldsymbol{\theta}\right) = \int_{\widetilde{\boldsymbol{\theta}}_1,\widetilde{\boldsymbol{\theta}}_2\in\Theta} E\left\{\frac{f_{\mathbf{y}}\left(\mathbf{y};\widetilde{\boldsymbol{\theta}}_1\right) f_{\mathbf{y}}\left(\mathbf{y};\widetilde{\boldsymbol{\theta}}_2\right)}{f_{\mathbf{y}}^2\left(\mathbf{y};\boldsymbol{\theta}\right)}\right\}\mathbf{f}(\widetilde{\boldsymbol{\theta}}_1,\boldsymbol{\theta})\mathbf{f}^H(\widetilde{\boldsymbol{\theta}}_2,\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}}_1 d\widetilde{\boldsymbol{\theta}}_2.$$

Notice that the original bound [Bar49] is somewhat more involved because the integral on $\widetilde{\boldsymbol{\theta}}$ is formulated as a Riemann integration, that is,

$$\int_{\widetilde{\boldsymbol{\theta}}\in\Theta} \xi\left(\widetilde{\boldsymbol{\theta}}\right) \mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})d\widetilde{\boldsymbol{\theta}} = \lim_{Q\rightarrow\infty}\sum_{q=1}^{Q}\xi\left(\widetilde{\boldsymbol{\theta}}_q\right)\mathbf{f}(\widetilde{\boldsymbol{\theta}}_q,\boldsymbol{\theta})$$

where the so-called *test points* $\widetilde{\boldsymbol{\theta}}_1,...,\widetilde{\boldsymbol{\theta}}_Q$ are selected to expand the whole parameter domain Θ and take into account the existence of large errors. In fact, we can understand the original approach as the bound obtained when the continuous function $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ is sampled at the test points $\widetilde{\boldsymbol{\theta}}_1,...,\widetilde{\boldsymbol{\theta}}_Q$. From the sampling theory, the separation of the test points should be adjusted according to the variability of the *selected* function $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$. Specifically, a dense sampling –closer test points– should be applied to those regions where $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ is more abrupt and vice versa. An important consequence of the sampling theorem is that infinite test points are needed *if the parameter range is finite* whatever the selected function $\mathbf{f}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$. This comment is related with the fact that unbiased estimators do not exist for all $\boldsymbol{\theta}\in\Theta$ when Θ is a finite set.[12]

If the number of test points is finite, the Barankin bound is only constrained to be unbiased at the test points $\widetilde{\boldsymbol{\theta}}_1,...,\widetilde{\boldsymbol{\theta}}_Q$ [For02]. Consequently, the resulting lower bound is not the highest Barankin bound ($Q\rightarrow\infty$) but it is generally realizable even if the parameter range is finite. The resulting bound can be improved by considering also the bias derivatives at the test points. This idea has been applied to derive other hybrid lower bounds in [Abe93] or [For02]. Also, the same reasoning was applied in [Vil01a] to design second-order almost-unbiased estimators.

The Barankin bound theory has been applied to determine the SNR threshold in a lot of nonlinear estimation problems as, for example, time delay estimation [Zei93][Zei94] or frequency estimation [Kno99]. A geometric interpretation of the Barankin bound is provided in [Alb73] and references therein.

[12] If an estimator were unbiased in the boundary of Θ, this would imply that the estimation error must be zero for these values of $\boldsymbol{\theta}$. Unfortunately, this situation is unreal and biased estimators are unavoidable along the boundary of Θ.

Hammersley-Chapman-Robbins Bound (HCRB)

The simplest Barankin bound was formulated by Chapman and Robbins [Cha51] and Hammersley [Ham50] simultaneously by considering a single test point per parameter, i.e., $Q = P$. This simplified version is by far the most usual variant of the Barankin bound. The original scalar bound was extended to deal with multidimensional problems by Gorman et al. [Gor90]. In that paper, every test point determines a single component of $\mathbf{f}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \in \mathbb{R}^P$ in the following manner:

$$\left[\mathbf{f}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\right]_p = \frac{\delta\left(\widetilde{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_p\right) - \delta\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)}{\left\|\widetilde{\boldsymbol{\theta}}_p - \boldsymbol{\theta}\right\|},$$

where the P vectors $\boldsymbol{\delta}_p \triangleq \widetilde{\boldsymbol{\theta}}_p - \boldsymbol{\theta}$ are *linearly independent* and span the entire parameter space Θ. It can be shown that this is the optimal choice of $\mathbf{f}\left(\widetilde{\boldsymbol{\theta}}\right)$ in case of having P test points [Wei88b, Eq. 33]. Therefore, the p-th element of the score function (2.41) becomes

$$[\mathrm{s}(\mathbf{y}; \boldsymbol{\theta})]_p = \frac{f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta} + \boldsymbol{\delta}_p) - f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})}{\|\boldsymbol{\delta}_p\| \, f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})} \quad \text{for} \quad p = 1, ..., P$$

and the multiparametric Hammersley-Chapman-Robbins bound is given by

$$\mathbf{B}_{HCRB}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_P} \Sigma_{\mathbf{vs}}(\boldsymbol{\theta}) \Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta}) \Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) \leq \Sigma_{\mathbf{vv}}(\boldsymbol{\theta})$$

with

$$\begin{aligned} [\Sigma_{\mathbf{vs}}(\boldsymbol{\theta})]_p &= \frac{\mathbf{g}(\boldsymbol{\theta} + \boldsymbol{\delta}_p) - \mathbf{g}(\boldsymbol{\theta})}{\|\boldsymbol{\delta}_p\|} \\ \Sigma_{\mathbf{ss}}(\boldsymbol{\theta}) &= E\left\{\mathrm{s}(\mathbf{y}; \boldsymbol{\theta}) \, \mathrm{s}^H(\mathbf{y}; \boldsymbol{\theta})\right\}. \end{aligned}$$

Cramér-Rao Bound (CRB)

The Cramér-Rao bound can be obtained from the Hammersley-Chapman-Robbins bound when the P test points converge to the true parameter $\boldsymbol{\theta}$ [Gor90][For02]. This means that the CRB is only able to test the *small-error* region whereas the Barankin-type bounds were able to test the large-error region, as well. The CRB score function is shown to correspond to the projection of the log-likelihood gradient $\boldsymbol{\nabla}_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})$ onto the directions determined by $\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_P$, i.e.,

$$[\mathrm{s}(\mathbf{y}; \boldsymbol{\theta})]_p = \lim_{\|\boldsymbol{\delta}_p\| \to 0} \frac{f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta} + \boldsymbol{\delta}_p) - f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})}{\|\boldsymbol{\delta}_p\| \, f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})} = \boldsymbol{\delta}_p^H \frac{\partial f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}}{f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})} = \boldsymbol{\delta}_p^H \boldsymbol{\nabla}_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) \tag{2.43}$$

and, thus,

$$\mathrm{s}(\mathbf{y}; \boldsymbol{\theta}) = \mathbf{W}^H \nabla_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) \tag{2.44}$$

with $\mathbf{W} \triangleq [\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_P]$ the *non-singular square* matrix stacking the P linearly-independent directions. Therefore, the CRB bound is given by

$$\begin{aligned}\mathbf{B}_{CRB} &= \limsup_{\|\boldsymbol{\delta}_1\|,...,\|\boldsymbol{\delta}_P\| \to 0} \mathbf{B}_{HCRB} = \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\, \Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\, \Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) \\ &= \mathbf{D}_g(\boldsymbol{\theta})\, \mathbf{J}^{\#}(\boldsymbol{\theta})\, \mathbf{D}_g^H(\boldsymbol{\theta}) \leq \Sigma_{\mathbf{vv}}(\boldsymbol{\theta}) \end{aligned} \tag{2.45}$$

where $\Sigma_{\mathbf{vs}}(\boldsymbol{\theta})$ and $\Sigma_{\mathbf{ss}}(\boldsymbol{\theta})$ are given by

$$\begin{aligned}\mathbf{D}_g(\boldsymbol{\theta}) &\triangleq \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ \mathbf{J}(\boldsymbol{\theta}) &\triangleq E\left\{\nabla_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\, \nabla_{\mathbf{y}}^H(\mathbf{y};\boldsymbol{\theta})\right\} = -E\left\{\mathbf{H}_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\right\},\end{aligned}$$

respectively (Appendix 2.F). The matrix $\mathbf{W}$ becomes irrelevant provided that $\mathbf{W}^{-1}$ exists and, thus, we can choose the canonical basis $\mathbf{W} = \mathbf{I}_P$.

Notice that the CRB bound only makes sense in estimation problems, i.e., when the parameter is continuous and the first- and second-order derivatives exist for $\boldsymbol{\theta} \in \Theta$. On the other hand, the above large-error bounds could be also applied to detection problems in which the parameters are discrete variables.

In [Fen59, Th. 1], it is shown that the necessary and sufficient condition for a statistic $\mathbf{z}(\mathbf{y})$ to attain the CRB is that $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ belongs to the exponential family below

$$f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) = \exp\left(\mathbf{h}^T(\boldsymbol{\theta})\mathbf{z}(\mathbf{y}) + u(\boldsymbol{\theta}) + v(\mathbf{y})\right) \tag{2.46}$$

whatever the content of $\mathbf{h}(\boldsymbol{\theta})$, $u(\boldsymbol{\theta})$ or $v(\mathbf{y})$. From the fourth property of the ML estimator in Section 2.3.2, it follows that $\mathbf{z}(\mathbf{y})$ must be the maximum likelihood estimator. This result can also be obtained by introducing the CRB score function (2.43) into Proposition 2.2. The existence of efficient estimates for the exponential family is relevant since the normal, Rayleigh and exponential distributions are members of this family [Kay93b, Pr. 5.14].

Another interpretation of the Cramér-Rao bound is possible [For02] if equation (2.40) is evaluated *locally* for every value of the true parameter $\boldsymbol{\theta}_o$. Thus, the Crámer-Rao bound can be obtained solving the following optimization problem:

$$\min_{\widehat{\alpha}} E\left\|\widehat{\boldsymbol{\alpha}} - \mathbf{g}(\boldsymbol{\theta}_o)\right\|^2 \text{ subject to } \mathbf{b}(\boldsymbol{\theta}_o) = \mathbf{0} \text{ and, } \left.\frac{\partial \mathbf{b}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \mathbf{0}$$

where $\mathbf{b}(\boldsymbol{\theta}) = E\{\widehat{\boldsymbol{\alpha}}\} - \mathbf{g}(\boldsymbol{\theta})$ stands for the estimator bias.

Finally, the Cramér-Rao bound can also be derived by expanding the log-likelihood function in a *quadratic* Taylor series around the true parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ (small-error condition), obtaining that

$$\begin{aligned}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) \simeq{}& \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}_o) + \boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o) + \\ & \frac{1}{2}\operatorname{Tr}\left\{\mathbf{H}(\mathbf{y};\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o)^H\right\}\end{aligned} \tag{2.47}$$

where $\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta})$ and $\mathbf{H}(\mathbf{y};\boldsymbol{\theta})$ are the gradient and Hessian of the log-likelihood function. Thus, the gradient of the log-likelihood is *linear* in the parameter of interest,

$$\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}) \simeq \boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}_o) + \mathbf{H}(\mathbf{y};\boldsymbol{\theta}_o)(\boldsymbol{\theta}-\boldsymbol{\theta}_o),$$

and becomes zero for

$$\widehat{\boldsymbol{\theta}}_{ML} \simeq \boldsymbol{\theta}_o - \mathbf{H}^{-1}(\mathbf{y};\boldsymbol{\theta}_o)\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}_o).$$

Taking now into account the invariance property of the ML estimator, we obtain the following *clairvoyant* estimator of $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$,

$$\widehat{\boldsymbol{\alpha}}_{ML} = \mathbf{g}(\widehat{\boldsymbol{\theta}}_{ML}) \simeq \mathbf{g}(\boldsymbol{\theta}_o) + \mathbf{D}_g(\boldsymbol{\theta}_o)\mathbf{H}^{-1}(\mathbf{y};\boldsymbol{\theta}_o)\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}_o),$$

whose covariance matrix coincides with the CRB (2.45). Although the above estimator does not admit a closed form unless $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ belong to the exponential family (2.43), efficient estimates are approximatelly supplied by the Newton-Raphson and scoring algorithms in the small-error regime, i.e., $\lim_{k\to\infty}\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\theta}}_{ML} \simeq \boldsymbol{\theta}_o$ (Section 2.5).

Bhattacharyya Bound (BHB)

The Bhattacharyya bound constitutes an extension of the CRB when considering the higher-order derivatives in the Taylor expansion of $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ (2.47). Therefore, it is also a *small-error* bound with higher-order derivative constraints on the bias. Indeed, it can be seen as the result of the following optimization problem [For02]:

$$\min_{\widehat{\boldsymbol{\alpha}}} E\left\|\widehat{\boldsymbol{\alpha}} - \mathbf{g}(\boldsymbol{\theta}_o)\right\|^2 \text{ subject to } \mathbf{b}(\boldsymbol{\theta}_o) = \mathbf{0} \text{ and, } \left.\frac{\partial^n \mathbf{b}^H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^n}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \mathbf{0} \quad (i = 1, ..., N)$$

where $\mathbf{b}(\boldsymbol{\theta}) = E\{\widehat{\boldsymbol{\alpha}}\} - \mathbf{g}(\boldsymbol{\theta})$ stands for the estimator bias and, $\partial\boldsymbol{\theta}^n \in \mathbb{R}^{P^n}$ stands for the *vectorized* n-th power of the differential $\partial\boldsymbol{\theta}$, which can be computed recursively as $\partial\boldsymbol{\theta}^n = \operatorname{vec}\left(\partial\boldsymbol{\theta}^{n-1}\partial\boldsymbol{\theta}^T\right)$ with $\partial\boldsymbol{\theta}^1 \triangleq \partial\boldsymbol{\theta}$. Notice that the CRB corresponds to $N = 1$.

To motivate the interest of the Bhattacharyya bound, let us consider that $\widehat{\boldsymbol{\theta}} = \mathbf{z}(\mathbf{y})$ is an efficient, unbiased estimator of $\boldsymbol{\theta}$ and, therefore, the likelihood function is given by (2.46). Let us consider the estimation of the following polynomial in $\boldsymbol{\theta}$ of order I,

$$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) = \sum_{i=0}^{I} \mathbf{G}_i \boldsymbol{\theta}^i,$$

with $\boldsymbol{\theta}^i \in \mathbb{R}^{P^i}$ the vectorized i-th power of $\boldsymbol{\theta}$. It can be shown that the estimator

$$\widehat{\boldsymbol{\alpha}}(\mathbf{y}) = \mathbf{g}(\boldsymbol{\theta}) + \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\mathbf{s}(\mathbf{y};\boldsymbol{\theta})$$

attains the N-th order Bhattacharyya bound for $N \geq I$ [Gor91, Prop. 3][Fen59, Th. 2] with

$$\mathbf{s}_n(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}\frac{\partial^n f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n} = \frac{\partial^n \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n} \quad (n = 1, ..., N),$$

the n-th component of the Bhattacharyya score funtion $\mathbf{s}(\mathbf{y};\boldsymbol{\theta}) = [\mathbf{s}_1^T(\mathbf{y};\boldsymbol{\theta}), ..., \mathbf{s}_N^T(\mathbf{y};\boldsymbol{\theta})]^T$. Accordingly, the Bhattacharyya bound becomes

$$\mathbf{B}_{BHB}(\boldsymbol{\theta}) = \Sigma_{\mathbf{vs}}(\boldsymbol{\theta})\,\Sigma_{\mathbf{ss}}^{\#}(\boldsymbol{\theta})\,\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) \geq \Sigma_{\mathbf{vv}}(\boldsymbol{\theta})$$

where

$$\Sigma_{\mathbf{vs}}(\boldsymbol{\theta}) = \left[\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{(\partial\boldsymbol{\theta})^T}, \frac{\partial^2 \mathbf{g}^H(\boldsymbol{\theta})}{(\partial\boldsymbol{\theta}^2)^T}, \ldots, \frac{\partial^N \mathbf{g}^H(\boldsymbol{\theta})}{(\partial\boldsymbol{\theta}^N)^T}\right]$$
$$\Sigma_{\mathbf{ss}}(\boldsymbol{\theta}) = E\left\{\mathbf{s}(\mathbf{y};\boldsymbol{\theta})\,\mathbf{s}^H(\mathbf{y};\boldsymbol{\theta})\right\}$$

bearing in mind the results in Appendix 2.F [Abe93].

It can be proved that $\widehat{\boldsymbol{\alpha}}(\mathbf{y})$ is unable to attain the N-th Battacharyya bound for any $N < I$ and hence the Cramér-Rao bound $(N = 1)$. Moreover, the ML estimator is not efficient even in the asymptotic case [Fen59].

Finally, the Bhattacharyya bound can also be obtained from the Barankin bound when we have at least $Q = N \times P$ test points that converge to the true parameter $\boldsymbol{\theta}$ following N *linearly-dependent trajectories* per parameter [Gor91, Sec. 4][For02]. In [For02], the N colinear trajectories corresponding to the p-th parameter are $\boldsymbol{\theta}+n\boldsymbol{\delta}_p$ with $\|\boldsymbol{\delta}_p\| \to 0$ and $n = 1, ..., N$. Therefore, we have that

$$\mathbf{B}_{BHB} = \lim_{n\|\boldsymbol{\delta}_p\|\to 0} \mathbf{B}_{HCRB}$$

for $p = 1, ..., P$ and $n = 1, ..., N$.

Deterministic Cramér-Rao Bounds in the presence of Nuisance Parameters

All the above lower bounds are formulated from the likelihood function $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$. If we deal with a blind estimation problem in which there is a vector of unknown *stochastic* nuisance parameters $\mathbf{x}$, we have to calculate $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ from the conditional p.d.f. $f_{\mathbf{y}/\mathbf{x}}(\mathbf{y}/\mathbf{x};\boldsymbol{\theta})$ as explained in Section 2.3 and indicated next

$$f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) = E_{\mathbf{x}}\left\{f_{\mathbf{y}/\mathbf{x}}(\mathbf{y}/\mathbf{x};\boldsymbol{\theta})\right\} = \int f_{\mathbf{y}/\mathbf{x}}(\mathbf{y}/\mathbf{x};\boldsymbol{\theta})\, f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}.$$

Therefore, the same assumptions about the nuisance parameters leading to the conditional and Gaussian ML estimators in Section 2.4.2 and 2.4.3 can be applied now to obtain their asymptotic performance in the small-error regime. In the first case, we obtain the so-called conditional CRB (CCRB) and, in the second case, the Gaussian unconditional CRB (UCRB). The CCRB and UCRB were deduced in [Sto90a][Sto89] [Ott93] in the context of array signal processing and adapted to the field of digital synchronization in [Vaz00][Rib01b][Vaz01] and references therein.

To obtain the (Gaussian) UCRB, the observed vector $\mathbf{y}$ is supposed to be normally distributed (2.24). Likewise, the CCRB is obtained assuming that $\mathbf{y}$ is distributed according to the conditional p.d.f. $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta}))$ (2.19). Therefore, the CCRB and UCRB are not "universal" lower bounds and, in general, they are only meaningful in the ambit of the conditional or the unconditional assumptions.

Thus, the CCRB and UCRB can be derived from the CRB formula (2.45) under the assumption adopted on the nuisance parameters. In the multidimensional case, it is obtained in Appendix 2.G that

$$\mathbf{B}_{CCRB}(\boldsymbol{\theta}) = \mathbf{D}_g(\boldsymbol{\theta})\,\mathbf{J}_c^{\#}(\boldsymbol{\theta})\,\mathbf{D}_g^T(\boldsymbol{\theta}) \tag{2.48}$$

$$\mathbf{B}_{UCRB}(\boldsymbol{\theta}) = \mathbf{D}_g(\boldsymbol{\theta})\,\mathbf{J}_u^{\#}(\boldsymbol{\theta})\,\mathbf{D}_g^T(\boldsymbol{\theta}) \tag{2.49}$$

where

$$\mathbf{J}_c(\boldsymbol{\theta}) \triangleq 2\,\mathrm{Re}\left(\mathbf{D}_a^H(\boldsymbol{\theta})\left(\mathbf{I}_K \otimes \mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}\right)\mathbf{D}_a(\boldsymbol{\theta})\right) \tag{2.50}$$

$$\mathbf{J}_u(\boldsymbol{\theta}) \triangleq \mathbf{D}_r^H(\boldsymbol{\theta})\left(\mathbf{R}^*(\boldsymbol{\theta}) \otimes \mathbf{R}(\boldsymbol{\theta})\right)^{-1}\mathbf{D}_r(\boldsymbol{\theta}) \tag{2.51}$$

are the Fisher's information matrix for the conditional and unconditional model, respectively, and $\mathbf{D}_a(\boldsymbol{\theta})$, $\mathbf{D}_r(\boldsymbol{\theta})$ are defined as

$$[\mathbf{D}_a(\boldsymbol{\theta})]_p \triangleq \mathrm{vec}\left(\frac{\partial \mathbf{A}(\boldsymbol{\theta})}{\partial \theta_p}\right)$$
$$[\mathbf{D}_r(\boldsymbol{\theta})]_p \triangleq \mathrm{vec}\left(\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \theta_p}\right).$$

The CCRB predicts the asymptotic performance of the CML and GML quadratic estimators when the SNR goes to infinity. On the other hand, the UCRB supplies the performance of the GML estimator for Gaussian nuisance parameters or, in general, for infinitely large samples. These two bounds are generally applied to bound the (small-error) variance of second-order estimation methods. However, in this dissertation it is shown that, if the nuisance parameters belong to a polyphase alphabet of constant modulus, this information can be exploited –using exclusively quadratic processing– to improve the CML and GML estimates. The covariance of the resulting estimator is shown in Chapter 4 to be the highest lower bound on the performance of any *second-order* technique. The resulting bound is deduced in Section 4.2 and has the following form

$$\mathbf{B}_{BQUE}(\boldsymbol{\theta}) = \mathbf{D}_g(\boldsymbol{\theta})\,\mathbf{J}_2^{\#}(\boldsymbol{\theta})\,\mathbf{D}_g^H(\boldsymbol{\theta}),$$

where BQUE is the acronym of “Best Quadratic Unbiased Estimator” [Vil01a][Vil05] and

$$\mathbf{J}_2\left(\boldsymbol{\theta}\right) \triangleq \mathbf{D}_r^H\left(\boldsymbol{\theta}\right)\mathbf{Q}^{-1}\left(\boldsymbol{\theta}\right)\mathbf{D}_r\left(\boldsymbol{\theta}\right) \tag{2.52}$$

becomes the Fisher’s information matrix in second-order estimation problems (4.14) with $\mathbf{Q}\left(\boldsymbol{\theta}\right)$ the matrix containing the central fourth-order moments of $\mathbf{y}$ (3.10).

Another useful lower bound is the so-called modified CRB (MCRB). This bound was deduced in the context of digital synchronization by D’Andrea et al. [And94] under the assumption that all the nuisance parameters are known (see also [Men97][Moe98][Vaz00]). This assumption corresponds to data-aided estimation problems in which the input signal is known. Thus, the MCRB allows assessing the performance loss due to the lack of knowledge about the nuisance parameters in blind estimation problems.

In the multidimensional case, the MCRB is given by

$$\mathbf{B}_{MCRB}\left(\boldsymbol{\theta}\right) = \mathbf{D}_g\left(\boldsymbol{\theta}\right)\mathbf{J}_m^{\#}\left(\boldsymbol{\theta}\right)\mathbf{D}_g^H\left(\boldsymbol{\theta}\right) \leq \Sigma_{\mathbf{vv}}\left(\boldsymbol{\theta}\right) \tag{2.53}$$

where

$$\mathbf{J}_m\left(\boldsymbol{\theta}\right) \triangleq -E_{\mathbf{x}}E_{\mathbf{y}/\mathbf{x}}\left\{\frac{\partial^2 \ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\} = 2\,\mathrm{Re}\left(\mathbf{D}_a^H\left(\boldsymbol{\theta}\right)\left(\mathbf{I}_K \otimes \mathbf{R}_w^{-1}\right)\mathbf{D}_a\left(\boldsymbol{\theta}\right)\right). \tag{2.54}$$

is deduced in Appendix 2.G.

To conclude this section, let us explain how the lower bounds above are connected in the studied linear model. It can be shown that

$$\mathbf{B}_{UCRB}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{BQUE}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{CRB}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{MCRB}\left(\boldsymbol{\theta}\right).$$

Additionally, if $\widehat{\boldsymbol{\alpha}}\left(\mathbf{y}\right)$ is a second-order unbiased estimator of $\mathbf{g}\left(\boldsymbol{\theta}\right)$, the associated error covariance matrix holds that

$$\Sigma_{\mathbf{vv}}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{BQUE}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{CRB}\left(\boldsymbol{\theta}\right) \geq \mathbf{B}_{MCRB}\left(\boldsymbol{\theta}\right),$$

and the following statements are verified:

1. $\mathbf{B}_{CRB}\left(\boldsymbol{\theta}\right) = \mathbf{B}_{MCRB}\left(\boldsymbol{\theta}\right)$ if the nuisance parameters are known [And94]. Alternatively, the MCRB could be attained in high-SNR scenarios if the mean of the nuisance parameters were not zero (i.e., semiblind estimation problems).

2. $\mathbf{B}_{BQUE}\left(\boldsymbol{\theta}\right) = \mathbf{B}_{CRB}\left(\boldsymbol{\theta}\right)$ if and only if $\widehat{\mathbf{R}}$ is a sufficient statistic for the estimation problem at hand. This occurs in case of Gaussian nuisance parameters (Section 2.4.3), or in low-SNR scenarios (Section 2.4.1) whatever the distribution of the nuisance parameters $\mathbf{x}$.

3. $\mathbf{B}_{UCRB}(\boldsymbol{\theta}) = \mathbf{B}_{BQUE}(\boldsymbol{\theta})$ if the nuisance parameters are Gaussian or the SNR is sufficiently low. Moreover, if the amplitude of $\mathbf{x}$ is not constant, it is shown in this thesis that the Gaussian assumption supplies asymptotically ($M \to \infty$) second-order efficient estimates, i.e., $\mathbf{B}_{UCRB}(\boldsymbol{\theta}) \to \mathbf{B}_{BQUE}(\boldsymbol{\theta})$. This point is intensively studied in Chapter 7.

4. $\mathbf{B}_{CCRB}(\boldsymbol{\theta}) \leq \mathbf{B}_{UCRB}(\boldsymbol{\theta})$ and $\mathbf{B}_{CCRB}(\boldsymbol{\theta}) = \mathbf{B}_{UCRB}(\boldsymbol{\theta})$ if the SNR tends to infinity (Appendix 2.D).

2.6.2 Bayesian Bounds based on the Cauchy-Schwarz Inequality

In the Bayesian case, lower bounds on the estimator MSE can also be derived from the Cauchy-Schwarz inequality

$$E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{e}\mathbf{e}^H\right\} \geq E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{e}\mathbf{s}^H\right\}\left(E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{s}\mathbf{s}^H\right\}\right)^{\#} E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{s}\mathbf{e}^H\right\}$$

in which the expectation involves also the random parameters and the score function $\mathbf{s}(\mathbf{y},\boldsymbol{\theta})$ is zero-mean for any value of $\mathbf{y}$ [Wei88b, Eq. 1], i.e.,

$$E_{\boldsymbol{\theta}/\mathbf{y}}\left\{\mathbf{s}(\mathbf{y},\boldsymbol{\theta})\right\} = \int \mathbf{s}(\mathbf{y},\boldsymbol{\theta}) f_{\boldsymbol{\theta}/\mathbf{y}}(\boldsymbol{\theta}/\mathbf{y})\, d\boldsymbol{\theta} = \mathbf{0} \tag{2.55}$$

and, therefore, $E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{s}(\mathbf{y},\boldsymbol{\theta})\right\} = E_{\mathbf{y}}E_{\boldsymbol{\theta}/\mathbf{y}}\left\{\mathbf{s}(\mathbf{y},\boldsymbol{\theta})\right\} = \mathbf{0}.$ Once again the bound is attained if and only if the estimation error is proportional to the selected score function, i.e.,

$$\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) = E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{e}\mathbf{s}^H\right\}\left(E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{s}\mathbf{s}^H\right\}\right)^{\#}\mathbf{s}(\mathbf{y},\boldsymbol{\theta}).$$

It is known that the conditional mean estimator yields the highest lower bound on the (Bayesian) MSE [Wei88b, Eq. 9][Kay93b, Sec. 11.4] with

$$\mathbf{s}(\mathbf{y};\boldsymbol{\theta}) = \mathbf{e}(\mathbf{y};\boldsymbol{\theta}) = E_{\boldsymbol{\theta}/\mathbf{y}}\left\{\mathbf{g}(\boldsymbol{\theta})/\mathbf{y}\right\} - \mathbf{g}(\boldsymbol{\theta})$$

the associated score function. However, the conditional mean estimator is often not practical because it usually requires numerical integration. For this reason, some simpler but weaker lower bounds have been proposed in [Wei88b] by adopting a different set of score functions. Accordingly, none of these bounds will be attained unless they coincide with the MMSE bound. Among these bounds, we can find the Bayesian Cramér-Rao [Tre68] [Wei88b], Bayesian Bhattacharyya [Tre68][Wei88b], Bobrovsky-Zakai [Bob76] and Weiss-Weinstein [Wei85][Wei88b]. These bounds are the Bayesian counterparts of the CRB, Bhattacharyya, Hammersley-Chapman-Robbins and Barankin-type deterministic bounds, respectively, in which the likelihood function $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is substituted by the joint p.d.f. $f_{\mathbf{y},\boldsymbol{\theta}}(\mathbf{y},\boldsymbol{\theta})$. Notice that Bayesian bounds are implicitly large-error bounds because the whole range of $\boldsymbol{\theta}$ is considered by means of the parameter prior $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. The Weiss-Weinstein bound is briefly described in the following section since it is the most general one.

Weiss-Weinstein Bound (WWB)

The Weiss-Weinstein bound can be understood as the Bayesian version of a Barankin-type bound in which multiple test points are considered. The score function of the WWB is given by

$$\mathbf{s}\left(\mathbf{y},\boldsymbol{\theta}\right)=\int_{\boldsymbol{\theta}\pm\boldsymbol{\delta}\in\Theta}Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}\right)\mathbf{f}\left(\boldsymbol{\delta}\right)d\boldsymbol{\delta}$$

with $Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}\right)$ defined as

$$Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}\right)\triangleq\left(\frac{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}+\boldsymbol{\delta}\right)}{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)}\right)^{s(\boldsymbol{\delta})}-\left(\frac{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}-\boldsymbol{\delta}\right)}{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)}\right)^{1-s(\boldsymbol{\delta})}$$

and the terms $0<s\left(\boldsymbol{\delta}\right)<1$ and $\mathbf{f}\left(\boldsymbol{\delta}\right)$ selected to produce the tightest lower bound. If we choose $s\left(\boldsymbol{\delta}\right)=1$, we have exactly the Bayesian replica of the Barankin bound. However, the authors showed that tighter lower bounds can be derived with $s\left(\boldsymbol{\delta}\right)<1$.

The above score function verifies the regularity condition $E_{\boldsymbol{\theta}/\mathbf{y}}\left\{\mathbf{s}\left(\mathbf{y},\boldsymbol{\theta}\right)\right\}=\mathbf{0}$ in (2.55) so that the WWB can be computed as

$$\mathbf{B}_{WWB}=\sup_{\mathbf{f}(\boldsymbol{\delta}),s(\boldsymbol{\delta})}\Sigma_{\mathbf{es}}\Sigma_{\mathbf{ss}}^{\#}\Sigma_{\mathbf{se}}\leq\Sigma_{\mathbf{ee}}$$

where

$$\begin{aligned}\Sigma_{\mathbf{es}}&=E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{e}\left(\mathbf{y},\boldsymbol{\theta}\right)\mathbf{s}^H\left(\mathbf{y},\boldsymbol{\theta}\right)\right\}=-E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{g}\left(\boldsymbol{\theta}\right)\mathbf{s}^H\left(\mathbf{y},\boldsymbol{\theta}\right)\right\}\\&=E_{\mathbf{y},\boldsymbol{\theta}}\left\{\int_{\boldsymbol{\theta}\pm\boldsymbol{\delta}\in\Theta}\left[\mathbf{g}\left(\boldsymbol{\theta}+\boldsymbol{\delta}\right)-\mathbf{g}\left(\boldsymbol{\theta}\right)\right]\left(\frac{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}+\boldsymbol{\delta}\right)}{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)}\right)^{s(\boldsymbol{\delta})}\mathbf{f}^H\left(\boldsymbol{\delta}\right)d\boldsymbol{\delta}\right\}\\\Sigma_{\mathbf{ss}}&=E_{\mathbf{y},\boldsymbol{\theta}}\left\{\mathbf{s}\left(\mathbf{y},\boldsymbol{\theta}\right)\mathbf{s}^H\left(\mathbf{y},\boldsymbol{\theta}\right)\right\}\end{aligned}$$

Thus far, infinite test points have been considered as done in the initial approach to the Barankin bound in (2.41). If a finite number of Q test points shall be considered, we can always use a set of delta measures, $\mathbf{f}\left(\boldsymbol{\delta}\right)=\sum_{q=1}^{Q}\mathbf{f}\left(\boldsymbol{\delta}_q\right)\delta\left(\boldsymbol{\delta}-\boldsymbol{\delta}_q\right)$, to obtain the following score function

$$\mathbf{s}\left(\mathbf{y},\boldsymbol{\theta}\right)=\sum_{q=1}^{Q}Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_q\right)\mathbf{f}\left(\boldsymbol{\delta}_q\right),$$

that must be optimized for $\left\{\boldsymbol{\delta}_q\right\}_{q=1,\ldots,Q}$, $\left\{\mathbf{f}\left(\boldsymbol{\delta}_q\right)\right\}_{q=1,\ldots,Q}$ and $\left\{s\left(\boldsymbol{\delta}_q\right)\right\}_{q=1,\ldots,Q}$. In that case, the Qth-order WWB can be obtained as indicated next

$$\begin{aligned}\mathbf{B}_{WWB}&=\sup_{\{\boldsymbol{\delta}_q\},\{\mathbf{f}(\boldsymbol{\delta}_q)\},\{s(\boldsymbol{\delta}_q)\}}\Sigma_{\mathbf{es}}\Sigma_{\mathbf{ss}}^{\#}\Sigma_{\mathbf{se}}=\sup_{\{\boldsymbol{\delta}_q\},\{\mathbf{f}(\boldsymbol{\delta}_q)\},\{s(\boldsymbol{\delta}_q)\}}\mathbf{G}\left(\mathbf{F}\mathbf{F}^{\#}\right)^H\mathbf{Q}^{\#}\left(\mathbf{F}\mathbf{F}^{\#}\right)\mathbf{G}^H\\&=\sup_{\{\boldsymbol{\delta}_q\},\{s(\boldsymbol{\delta}_q)\}}\mathbf{G}\mathbf{Q}^{\#}\mathbf{G}^H\end{aligned}$$

where $\Sigma_{\mathbf{es}} = \mathbf{G}\mathbf{F}^H$ and $\Sigma_{\mathbf{ss}} = \mathbf{F}\mathbf{Q}\mathbf{F}^H$ are given by

$$\begin{aligned}
\left[\mathbf{G}\right]_q &\triangleq E_{\mathbf{y},\boldsymbol{\theta}}\left\{\left[\mathbf{g}\left(\boldsymbol{\theta}+\boldsymbol{\delta}_q\right)-\mathbf{g}\left(\boldsymbol{\theta}\right)\right]\left(\frac{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}+\boldsymbol{\delta}_q\right)}{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)}\right)^{s(\boldsymbol{\delta}_q)}\right\} \\
\mathbf{F} &\triangleq \left[\mathbf{f}\left(\boldsymbol{\delta}_1\right),\ldots,\mathbf{f}\left(\boldsymbol{\delta}_Q\right)\right] \\
\left[\mathbf{Q}\right]_{p,q} &\triangleq E_{\mathbf{y},\boldsymbol{\theta}}\left\{Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_p\right)Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_q\right)\right\}.
\end{aligned}$$

A simpler expression is obtained if $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. In that case, we get the original WWB bound [Wei85], that is given by

$$\mathbf{B}_{WWB} = \sup_{\Delta} \Delta\widetilde{\mathbf{Q}}^{\#}\Delta^H \leq \Sigma_{\mathbf{ee}} \tag{2.56}$$

with $\Delta \triangleq [\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_Q]$ and

$$\left[\widetilde{\mathbf{Q}}\right]_{p,q} \triangleq \frac{E_{\mathbf{y},\boldsymbol{\theta}}\left\{Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_p\right)Q_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_q\right)\right\}}{E_{\mathbf{y},\boldsymbol{\theta}}\left\{L_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_p\right)\right\}E_{\mathbf{y},\boldsymbol{\theta}}\left\{L_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}_q\right)\right\}},$$

using the following definition

$$L_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}\right) \triangleq \left(\frac{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}+\boldsymbol{\delta}\right)}{f_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)}\right)^{s(\boldsymbol{\delta})}.$$

The optimization of $\mathbf{B}_{WWB}$ is normally prohibitive and the authors suggest in [Wei88b, Eq. 39] to work with $s\left(\boldsymbol{\delta}_q\right) = 1/2$ because it is usually the optimal choice in the unidimensional case. In that case, it is possible to write the WWB bound in terms of the distance

$$\mu\left(s,\boldsymbol{\theta},\boldsymbol{\delta}\right) \triangleq \ln E_{\mathbf{y},\boldsymbol{\theta}}\left\{L_s\left(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\delta}\right)\right\} = \ln\int f^{s}_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}+\boldsymbol{\delta}\right)f^{1-s}_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{y},\boldsymbol{\theta}\right)d\mathbf{y}d\boldsymbol{\theta}$$

used to derive the Chernoff bound on the probability of detection error [Tre68, p. 119]. Thus, the matrix $\widetilde{\mathbf{Q}}$ can be represented in terms of the Bhattacharyya distance $\mu\left(1/2,\boldsymbol{\delta}\right)$ as follows

$$\left[\widetilde{\mathbf{Q}}\right]_{p,q} \triangleq 2\frac{e^{\mu(1/2,\boldsymbol{\delta}_p-\boldsymbol{\delta}_q)}-e^{\mu(1/2,\boldsymbol{\delta}_p+\boldsymbol{\delta}_q)}}{e^{\mu(1/2,\boldsymbol{\delta}_p)+\mu(1/2,\boldsymbol{\delta}_q)}}.$$

As it happened in the deterministic case, the Bobrovsky-Zakai, Bayesian Cramér-Rao and Bhattacharyya bounds can be deduced from the more general Weiss-Weinstein bound in (2.56). Specifically, the Bobrovsky-Zakai bound is obtained by setting $s = 1$ and $Q = P$ (i.e., a test point per parameter). The Bayesian CRB is obtained from the Bobrovsky-Zakai bound if the $Q = P$ test points converge to the true parameter along linearly-independent lines. In addition, the *Nth*-order Bhattacharyya bound is obtained when there are $N \times P$ test points converging to the true parameter through P linealy-independent trajectories.

2.6.3 Bayesian Bounds based on the Kotelnikov's Inequality

Other important class of Bayesian lower bounds are obtained from the Kotelnikov's inequality proposed for the first time in [Kot59, p. 91], and used afterwards in [Bel74, Eq. 2] and [Cha75] to bound the MSE in case of a single uniformly distributed parameter. The Kotelnikov's result is extended in [Bel97, Eq. 11] to admit any distribution of the parameter of interest, resulting in the following inequality

$$\Pr\left(|e| \geq \delta\right) \geq \int_{-\infty}^{\infty} P_e\left(\theta-\delta,\theta+\delta\right)\left[f_\theta\left(\theta-\delta\right)+f_\theta\left(\theta+\delta\right)\right]d\theta \triangleq D\left(\delta\right) \tag{2.57}$$

where $e = \widehat{\theta} - \theta_o$ is the estimation error for the scalar case, $f_\theta\left(\theta\right)$ is the parameter prior and $P_e\left(\theta-\delta,\theta+\delta\right)$ is the *minimum* error probability associated to the following binary detection problem:

Definition 1 *Let us assume that the parameter θ_o could take only two possible values, $\theta^- \triangleq \theta - \delta$ and $\theta^+ \triangleq \theta + \delta$ with probabilities*

$$\Pr\left(\theta^-\right) \triangleq \frac{f_\theta\left(\theta^-\right)}{f_\theta\left(\theta^-\right)+f_\theta\left(\theta^+\right)} \quad \textit{and} \quad \Pr\left(\theta^+\right) \triangleq \frac{f_\theta\left(\theta^+\right)}{f_\theta\left(\theta^-\right)+f_\theta\left(\theta^+\right)},$$

respectively. In that case, the estimation problem becomes a binary detection problem consisting in deciding the most likely hypothesis θ^- or θ^+ in view of the observation $\mathbf{y}_o$ and the prior probabilities $\Pr\left(\theta^-\right)$ and $\Pr\left(\theta^+\right)$.

The solution to this classical problem is supplied by the MAP detector or, equivalently, by the likelihood ratio test [Kay93a]. Then, the parameter is decided as follows

$$\widehat{\theta} = \begin{cases} \theta^- & f_{\mathbf{y}}\left(\mathbf{y};\theta^-\right)\Pr\left(\theta^-\right) \geq f_{\mathbf{y}}\left(\mathbf{y};\theta^+\right)\Pr\left(\theta^+\right) \\ \theta^+ & f_{\mathbf{y}}\left(\mathbf{y};\theta^-\right)\Pr\left(\theta^-\right) < f_{\mathbf{y}}\left(\mathbf{y};\theta^+\right)\Pr\left(\theta^+\right) \end{cases}$$

and, thus,

$$P_e\left(\theta^-,\theta^+\right) = \Pr\left(\theta^-\right)\int_{\theta}^{\infty} f_{\mathbf{y}}\left(\mathbf{y};\theta^-\right)d\mathbf{y} + \Pr\left(\theta^+\right)\int_{-\infty}^{\theta} f_{\mathbf{y}}\left(\mathbf{y};\theta^+\right)d\mathbf{y}.$$

If the proposed estimator solves optimally the related detection problem for all the possible values of θ, equation (2.57) is hold with equality. Moreover, if the hypotesis are very close ($\delta \rightarrow 0$), the MAP estimator,

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} f_{\mathbf{y}}\left(\mathbf{y};\boldsymbol{\theta}\right) f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right),$$

attains the Kotelnikov's bound in (2.57) and, thus, minimizes $\Pr\left(|e| \geq \delta\right)$ as explained in [Kay93b, Sec. 11.3].

Ziv-Zakai Bounds (ZZB)

The original work relating the estimation and detection problems was presented by Ziv and Zakai in [Ziv69]. However, they applied the Chebyshev's inequeality,

$$\Pr\left(|e| \geq \delta\right) \geq \frac{E_{\theta} E\left|e\right|^2}{\delta^2},$$

in lieu of the Kotelnivov's one (2.57), and the resulting bound was looser. The original idea was improved in [Cha75] [Bel74][Wei88a][Bel97] where the Kotelnikov's inequality is used to derive tight bounds on the (Bayesian) MSE. To do so, it is necessary to use the following relation between $\Pr\left(|e| \geq \delta\right)$ and the mean square error [Bel97, Eq. 2]:

$$E_{\theta} E\left|e\right|^2 = \int_0^{\infty} \Pr\left(|e| \geq \delta\right) \delta d\delta$$

where the Bayesian expectation is made explicit again. In the scalar case, the Ziv-Zakai bound is extended in [Bel97, Eq. 14] as follows

$$E_{\theta} E\left|e\right|^2 \geq \int_0^{\infty} \nu\left[D(\delta)\right] \delta d\delta \tag{2.58}$$

where $D(\delta)$ is the bound on $\Pr\left(|e| \geq \delta\right)$ introduced previously in (2.57) and $\nu\left[\cdot\right]$ is the "valley-filling" function introduced by Bellini and Tartara in [Bel74] and defined as

$$v\left[f\left(x\right)\right] \triangleq \max_{\xi \geq 0} f(x+\xi).$$

If the prior distribution is uniform on a finite interval, the above bound reduces to the Bellini-Tartara bound [Bel74].

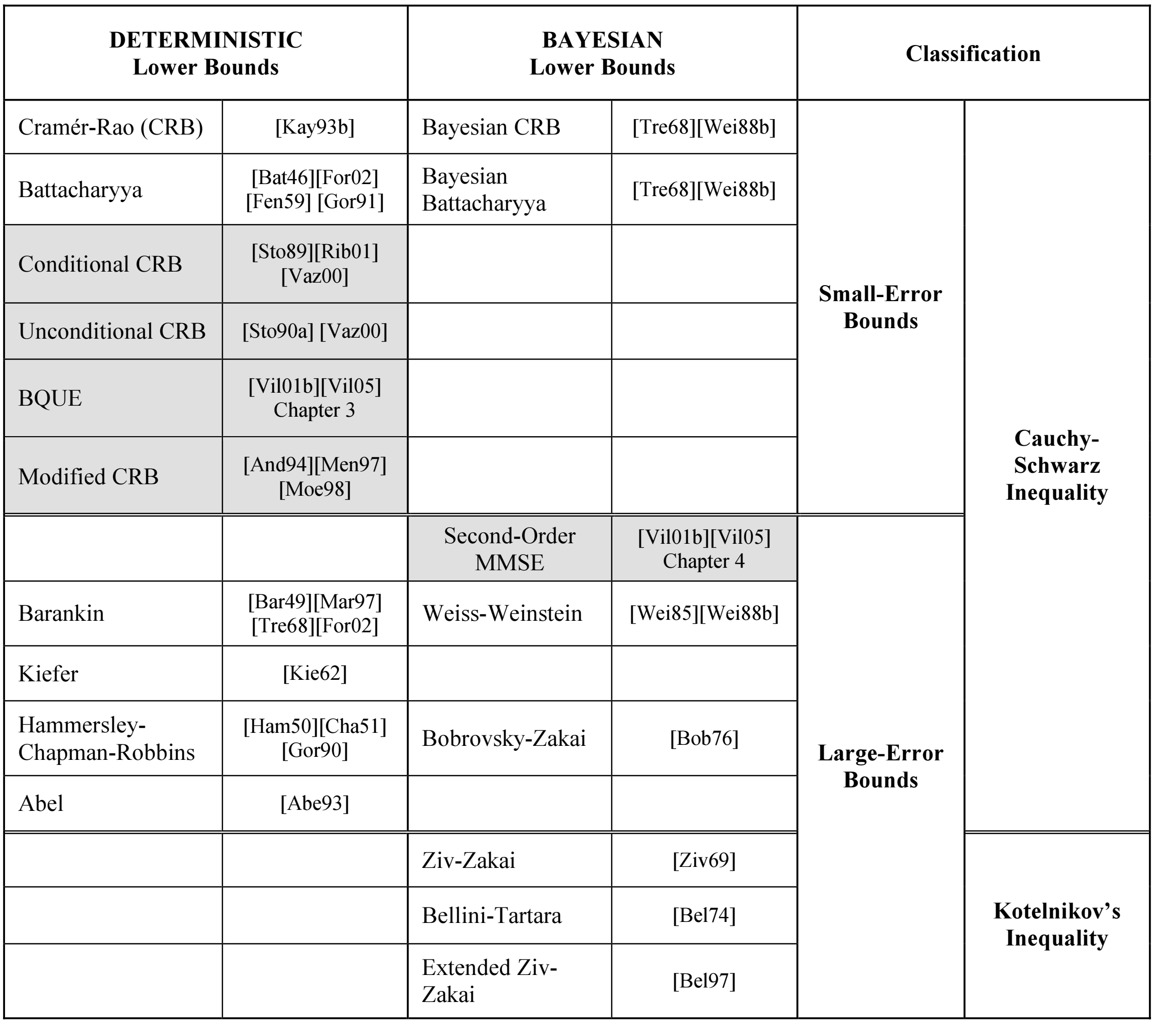
Finally, the Bellini-Tartara bound is generalized in [Bel97] to multivariate problems and arbitrary prior functions. In that case, the extended Ziv-Zakai bound is obtained projecting the estimation error $\mathbf{e} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ onto a given direction determined by the vector $\mathbf{v}$ [Bel97]. For a given $\mathbf{v}$, we have the same expression,

$$E_{\boldsymbol{\theta}} E\left|\mathbf{v}^H \mathbf{e}\right| \geq \int_0^{\infty} \nu\left[D_{\max}(\delta)\right] \delta d\delta,$$

where

$$D_{\max}(\delta) = \max_{\Delta : \mathbf{v}^H \Delta = \delta} \int_{-\infty}^{\infty} P_e\left(\boldsymbol{\theta} - \Delta, \boldsymbol{\theta} + \Delta\right)\left[f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} - \Delta\right) + f_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} + \Delta\right)\right] d\boldsymbol{\theta}.$$

In principle, the two hypothesis $\boldsymbol{\theta}^- \triangleq \boldsymbol{\theta} - \Delta$ and $\boldsymbol{\theta}^+ \triangleq \boldsymbol{\theta} + \Delta$ could be placed arbitrarily in the hyperplane $\mathbb{R}^P$ provided that the projection of the estimation error $\mathbf{v}^H \mathbf{e}$ is equal to δ in case of an erroneous detection or, in other words, Δ must hold that $\mathbf{v}^H \Delta = \delta$. Then, the tightest lower bound corresponds to the vector Δ, yielding the highest error probability. The reader is referred to the original work [Bel97] for further results, properties and examples. The utilization of the Ziv-Zakai bound (ZZB) in the problem of passive time delay estimation is carried out in detail in [Wei83][Wei84].



DETERMINISTIC Lower Bounds		BAYESIAN Lower Bounds		Classification	
Cramér-Rao (CRB)	[Kay93b]	Bayesian CRB	[Tre68][Wei88b]	Small-Error Bounds	Cauchy-Schwarz Inequality
Battacharyya	[Bat46][For02] [Fen59] [Gor91]	Bayesian Battacharyya	[Tre68][Wei88b]		
Conditional CRB	[Sto89][Rib01] [Vaz00]				
Unconditional CRB	[Sto90a] [Vaz00]				
BQUE	[Vil01b][Vil05] Chapter 3				
Modified CRB	[And94][Men97] [Moe98]				
		Second-Order MMSE	[Vil01b][Vil05] Chapter 4	Large-Error Bounds	
Barankin	[Bar49][Mar97] [Tre68][For02]	Weiss-Weinstein	[Wei85][Wei88b]		
Kiefer	[Kie62]				
Hammersley-Chapman-Robbins	[Ham50][Cha51] [Gor90]	Bobrovsky-Zakai	[Bob76]		
Abel	[Abe93]				
		Ziv-Zakai	[Ziv69]		Kotelnikov's Inequality
		Bellini-Tartara	[Bel74]		
		Extended Ziv-Zakai	[Bel97]		

Figure 2.5: Classification of the most important lower bounds in the literature. The lower bounds assuming a certain model for the nuisance parameters –or imposing the second-order constraint– are marked in gray.

Appendix 2.A UML for polyphase alphabets

Let us consider that the nuisance parameters belong to a polyphase alphabet of dimension I so that $x_k \in \left\{e^{j2\pi i/I}\right\}$ with $i = 0, ..., I-1$. In that case, it can be shown that the log-likelihood $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is the sum of a finite number of $\cosh(\cdot)$ functions, which are computed next

$$\begin{aligned}
E_x\left\{\exp\left(2\operatorname{Re}\left(x_k^*\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{y}\right)\right)\right\} &= \frac{1}{I}\sum_{i=0}^{I-1}\exp\left(2\operatorname{Re}\left(e^{-j2\pi i/I}\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{y}\right)\right) \\
&= \frac{2}{I}\sum_{i=0}^{I/2-1}\cosh\left(2\operatorname{Re}\left(e^{j2\pi i/I}\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{y}\right)\right) \\
\prod_{l>k}^{K} E_x\left\{\exp\left(2\operatorname{Re}\left(x_k^*x_l\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_l\right)\right)\right\} &= \prod_{l>k}^{K}\frac{2}{I^2}\sum_{i=0}^{I-1}(I-i)\cosh\left(2\operatorname{Re}\left(e^{j2\pi i/I}\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_l\right)\right) \\
E_x\left\{\exp\left(2\operatorname{Re}\left(|x_k|^2\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_k\right)\right)\right\} &= \frac{2}{I}\sum_{i=0}^{I/2-1}\cosh\left(\mathbf{a}_k^H\mathbf{R}_w^{-1}\mathbf{a}_k\right).
\end{aligned}$$

Notice that the term $E_{\mathbf{x}}\left\{\exp\left(\mathbf{x}^H\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\mathbf{x}\right)\right\}$ can be omitted if $\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}$ does not depend on the parameter. This situation is usual in digital communications [Men97, Sec. 5.7.3] because the noise is white (i.e., $\mathbf{R}_w = \sigma_w^2\mathbf{I}_M$) and $\mathbf{a}_k^H\mathbf{a}_l \cong E_s\delta(k,l)$ with E_s the energy of the received symbols. In that case, we have that

$$\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) \propto \sum_{k=1}^{K}\ln\sum_{i=0}^{I/2-1}\cosh\left(2\operatorname{Re}\left(e^{j2\pi i/I}\mathbf{a}_k^H(\boldsymbol{\theta})\mathbf{R}_w^{-1}\mathbf{y}\right)\right).$$

Appendix 2.B Low-SNR UML results

Unbiasedness Condition

It can be shown that the low-SNR UML estimator is unbiased for any positive SNR if $\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$. If this condition is verified, then the mean value of the log-likelihood gradient is null at $\boldsymbol{\theta}=\boldsymbol{\theta}_{o}$ for any value of the parameter. The proof is provided next. Let

$$E_{\mathbf{y}}\left\{\frac{\partial}{\partial\theta_{p}}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{o}}\right\}=\mathrm{Tr}\left(\mathbf{A}^{H}(\boldsymbol{\theta}_{o})\mathbf{R}_{w}^{-1}\mathbf{D}_{p}(\boldsymbol{\theta}_{o})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta}_{o})\right) \tag{2.59}$$

be the expected value of the log-likelihood gradient under the low SNR approximation with

$$\mathbf{D}_{p}(\boldsymbol{\theta})\triangleq\frac{\partial}{\partial\theta_{p}}\left[\mathbf{A}(\boldsymbol{\theta})\mathbf{A}^{H}(\boldsymbol{\theta})\right]=\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\theta_{p}}\mathbf{A}^{H}(\boldsymbol{\theta})+\mathbf{A}(\boldsymbol{\theta})\frac{\partial\mathbf{A}^{H}(\boldsymbol{\theta})}{\partial\theta_{p}}.$$

If we plug $\mathbf{D}_{p}(\boldsymbol{\theta})$ into (2.59), the argument of the trace can be written as

$$\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta})\frac{\partial}{\partial\theta_{p}}\left[\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta})\right]$$

using that $\mathrm{Tr}(\mathbf{AB})=\mathrm{Tr}(\mathbf{BA})$. Therefore, since $\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta})$ is positive definite, (2.59) vanishes iff $\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}\mathbf{A}(\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$. This condition implies that $\mathbf{R}_{w}^{-1}\mathbf{D}_{p}(\boldsymbol{\theta})\mathbf{R}_{w}^{-1}$ (2.59) must lie completely into the orthogonal subspace of $\mathbf{A}(\boldsymbol{\theta})$ for $p=1,...,P$.

Self-Noise Free Condition

If the gradient of the low-SNR UML log-likelihood function is not zero at $\boldsymbol{\theta}=\boldsymbol{\theta}_{o}$ as the noise variance goes to zero, the estimator variance exhibits a variance floor due to the randomness of the nuisance parameters. A sufficient condition to have self-noise free estimates at high SNR is that

$$\lim_{\sigma_{w}^{2}\to0}\frac{\partial}{\partial\theta_{p}}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{o}}=0$$

meaning that

$$\mathbf{x}^{H}\mathbf{A}^{H}(\boldsymbol{\theta})\mathbf{N}^{-1}\mathbf{D}_{p}(\boldsymbol{\theta})\mathbf{N}^{-1}\mathbf{A}(\boldsymbol{\theta})\mathbf{x}=0$$

for any value of $\boldsymbol{\theta}$ and $\mathbf{x}$. Notice that this requirement coincides with the unbiasedness condition if $\mathbf{A}(\boldsymbol{\theta})\mathbf{x}$ effectively spans all the signal subspace.

Appendix 2.C CML results

Unbiasedness

Plugging $\mathbf{B}(\boldsymbol{\theta}) \triangleq \mathbf{R}_w^{-1/2}\mathbf{A}(\boldsymbol{\theta})$ into $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta}))$ (2.14), we have that

$$\begin{aligned}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta})) &= C_4 + \operatorname{Tr}\left(\mathbf{R}_w^{-1/2}\mathbf{B}(\boldsymbol{\theta})\left[\mathbf{B}^H(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\right]^{-1}\mathbf{B}^H(\boldsymbol{\theta})\mathbf{R}_w^{-1/2}\widehat{\mathbf{R}}\right) = \\ &= C_4 + \operatorname{Tr}\left(\mathbf{R}_w^{-1/2}\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})\mathbf{R}_w^{-1/2}\widehat{\mathbf{R}}\right)\end{aligned}$$

with $\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta}) \triangleq \mathbf{B}(\boldsymbol{\theta})\left[\mathbf{B}^H(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\right]^{-1}\mathbf{B}^H(\boldsymbol{\theta})$ the orthogonal projector onto the subspace generated by the columns of $\mathbf{B}(\boldsymbol{\theta})$. Computing now the log-likelihood gradient, it is found that

$$\frac{\partial}{\partial\theta_p}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta})) = \operatorname{Tr}\left(\mathbf{R}_w^{-1/2}\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p}\mathbf{R}_w^{-1/2}\widehat{\mathbf{R}}\right)$$

where the derivative of the orthogonal projector is given by [Vib91, Eq. 33]

$$\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p} = \mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta})\frac{\partial\mathbf{B}(\boldsymbol{\theta})}{\partial\theta_p}\mathbf{B}^{\#}(\boldsymbol{\theta}) + \left[\mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta})\frac{\partial\mathbf{B}(\boldsymbol{\theta})}{\partial\theta_p}\mathbf{B}^{\#}(\boldsymbol{\theta})\right]^H \tag{2.60}$$

with $\mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta}) \triangleq \mathbf{I}_M - \mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})$. Therefore, the expected value of the gradient is

$$\begin{aligned}E_{\mathbf{y}}\left\{\left.\frac{\partial}{\partial\theta_p}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta}))\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right\} &= \operatorname{Tr}\left(\mathbf{R}_w^{-1/2}\left.\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\mathbf{R}_w^{-1/2}\left[\mathbf{A}(\boldsymbol{\theta}_o)\mathbf{A}^H(\boldsymbol{\theta}_o) + \mathbf{R}_w\right]\right) \\ &= \operatorname{Tr}\left(\left.\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} + \mathbf{B}^H(\boldsymbol{\theta}_o)\left.\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\mathbf{B}(\boldsymbol{\theta}_o)\right),\end{aligned}$$

that is equal to zero because

$$\begin{aligned}\mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}) &= \mathbf{0} \\ \mathbf{B}^H(\boldsymbol{\theta})\mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta}) &= \mathbf{0}.\end{aligned}$$

Self-Noise Free

If the gradient of the CML log-likelihood function is not zero at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ as the noise variance goes to zero, the estimator variance exhibits a variance floor due to the randomness of the nuisance parameters. A sufficient condition to have self-noise free estimates at high SNR is that

$$\lim_{\sigma_w^2\to 0}\left.\frac{\partial}{\partial\theta_p}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta}))\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = 0,$$

meaning that

$$\mathbf{x}^H\mathbf{B}^H(\boldsymbol{\theta})\left.\frac{\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})}{\partial\theta_p}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\mathbf{B}(\boldsymbol{\theta}_o)\mathbf{x} = 0$$

for any value of $\boldsymbol{\theta}$ and $\mathbf{x}$. Notice that the last equation is verified for any value of $\mathbf{x}$ due to (2.60). Actually, the CML is able to cancel out the self-noise as well as the bias because of the orthogonal projector $\mathbf{P}_{\mathbf{B}}^{\perp}(\boldsymbol{\theta})$ appearing in $\partial\mathbf{P}_{\mathbf{B}}(\boldsymbol{\theta})/\partial\theta_p$ (2.60).

Appendix 2.D GML asymptotic study

Using the inversion lemma [Kay93b, p. 571], we find that $\mathbf{R}^{-1}(\boldsymbol{\theta})$ has the following asymptotic expressions:

$$\begin{aligned}\lim_{\sigma_w^2\to\infty}\mathbf{R}^{-1} &= \mathbf{R}_w^{-1}\left(\mathbf{I}_M-\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\right)\\ \lim_{\sigma_w^2\to 0}\mathbf{R}^{-1} &= \mathbf{R}_w^{-1}\left(\mathbf{I}_M-\mathbf{A}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{R}_w^{-1}\right)\end{aligned}$$

with the operator lim meaning “asymptotically approximated to” in this appendix.

If we substitute these results into (2.24) and omit constant terms, we obtain the following asymptotic expressions for the GML cost function:

$$\begin{aligned}\lim_{\sigma_w^2\to\infty}\ln E_{\mathbf{x}}\left\{f_{\mathbf{y}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)\right\} &\propto \operatorname{Tr}\ln\left(\mathbf{I}_M-\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\right)+\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right)\\ &\simeq -\operatorname{Tr}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right)+\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right) \qquad (2.61)\\ \lim_{\sigma_w^2\to 0}\ln E_{\mathbf{x}}\left\{f_{\mathbf{y}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)\right\} &\propto -\operatorname{Tr}\ln\left(\mathbf{A}\mathbf{A}^H+\mathbf{R}_w\right)+\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right)\\ &\simeq \operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{A}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{R}_w^{-1}\widehat{\mathbf{R}}\right), \qquad (2.62)\end{aligned}$$

that correspond to the low-SNR UML and CML solutions obtained in (2.17) and (2.19), respectively.

The independent term $b(\boldsymbol{\theta})$ in (2.61) has been approximated using the Taylor expansion of the logarithm and the commutative property of the trace [Kay93b, p. 571], yielding

$$\lim_{\sigma_w^2\to\infty}\operatorname{Tr}\ln\left(\mathbf{I}_M-\sigma_w^{-2}\mathbf{A}\mathbf{A}^H\mathbf{N}^{-1}\right)=Tr\ln\left(\mathbf{I}_M\right)+\operatorname{Tr}\left(-\sigma_w^{-2}\mathbf{A}\mathbf{A}^H\mathbf{N}^{-1}\right)=-\operatorname{Tr}\left(\mathbf{A}^H\mathbf{R}_w^{-1}\mathbf{A}\right).$$

On the other hand, the independent term $b(\boldsymbol{\theta})$ in (2.62) is neglected at high SNR since it converges to the constant $-\operatorname{Tr}\ln\left(\mathbf{A}\mathbf{A}^H\right)$ whereas the second term is proportional to σ_w^{-2}.

Appendix 2.E Closed-loop estimation efficiency

Following the indications in [Kay93b, Appendix 7B], if the observation $\mathbf{y}$ is splitted into N statistically independent blocks, the log-likelihood function $\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is given by

$$\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})$$

and, thus, the corresponding gradient and Hessian are given by

$$\boldsymbol{\nabla}(\mathbf{y};\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\partial \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \triangleq \sum_{n=1}^{N} \boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta}) \tag{2.63}$$

$$\mathbf{H}(\mathbf{y};\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{2.64}$$

respectively. Therefore, the Newton-Raphson and scoring algorithms are updated in the k-th iteration adding the following term

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{D}_{\mathbf{g}}(\widehat{\boldsymbol{\theta}}_k)\mathbf{J}_{\mathbf{z}}^{-1}(\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\widehat{\boldsymbol{\theta}}_k), \tag{2.65}$$

in which we have taken into account that

$$\sum_{n=1}^{N} \frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \simeq N E_{\mathbf{z}} \left\{ \frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\} \triangleq -N\mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta}),$$

for N sufficiently large [Kay93b, Appendix 7B]. Notice that the last equation is approximatelly equal to the Fisher's information matrix:

$$\mathbf{J}(\boldsymbol{\theta}) = -E_{\mathbf{y}}\{\mathbf{H}(\mathbf{y};\boldsymbol{\theta})\} = -\sum_{n=1}^{N} E_{\mathbf{z}} \left\{ \frac{\partial^2 \ln f_{\mathbf{z}}(\mathbf{z}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\} = -N\mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta}).$$

Then, the averaging in (2.65) can be substituted by an exponential filtering such as

$$\boldsymbol{\varepsilon}_n = (1-\mu)\,\boldsymbol{\varepsilon}_{n-1} - \mu \mathbf{D}_{\mathbf{g}}(\widehat{\boldsymbol{\theta}}_k)\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\widehat{\boldsymbol{\theta}}_k), \tag{2.66}$$

with $\boldsymbol{\varepsilon}_0 = \mathbf{0}$. The step-size or forgetting factor μ is adjusted to yield the same noise equivalent bandwidth, which is defined as

$$B_n \triangleq \frac{\int_{-1/2T}^{1/2T} |H(f)|^2 \, df}{2T\,|H(0)|^2}$$

where T is the sampling period and $H(f)$ is the frequency response of the adopted filter [Men97, Sec. 3.5.5]. Using this formula, it follows that the noise equivalent bandwidth for the integrator (2.65) and the exponential filter (2.66) is $B_n = 0.5/N$ and $B_n = 0.5\mu/(2-\mu) \simeq \mu/4$, respectively, where the last approximation is verified for $\mu \ll 1$. Using this approximation, the step-size μ is approximatelly equal to $2/N$ (for $N \gg 1$) and (2.66) can be written as

$$\boldsymbol{\varepsilon}_n = \boldsymbol{\varepsilon}_{n-1} - \mu \mathbf{D}_{\mathbf{g}}(\widehat{\boldsymbol{\theta}}_k)\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}}_k)\boldsymbol{\nabla}_{\mathbf{z}}(\mathbf{z}_n;\widehat{\boldsymbol{\theta}}_k). \tag{2.67}$$

Finally, if (2.67) is integrated into the Newton-Raphson or scoring recursions, and the estimated parameter is updated after processing each block, we obtain the closed-loop estimator presented in (2.30). Notice that the obtained closed-loop estimator can also iterate the N blocks several times as the original iterative methods in (2.27) and (2.29).

Appendix 2.F Computation of $\Sigma_{\mathbf{sv}}(\boldsymbol{\theta})$ for the small-error bounds

The computation of $\Sigma_{\mathbf{sv}}(\boldsymbol{\theta}) = E\left\{\mathbf{s}(\mathbf{y};\boldsymbol{\theta})\mathbf{v}^T(\mathbf{y};\boldsymbol{\theta})\right\}$ for the Bhattacharyya and Cramér-Rao bounds requires to compute the following term

$$E_{\mathbf{y}}\left\{\frac{\partial^n \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\right\} = \int \frac{\partial^n f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\,d\mathbf{y} \tag{2.68}$$

The last term can be further manipulated taking into account that the estimator is unbiased, i.e., $E\{\mathbf{v}(\mathbf{y};\boldsymbol{\theta})\} = \mathbf{0}$. Then, if the chain rule is applied and the integral and derivative signs are swapped, we obtain that

$$\begin{aligned}\frac{\partial^n}{\partial\boldsymbol{\theta}^n}E_{\mathbf{y}}\left\{\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\right\} &= \frac{\partial^n}{\partial\boldsymbol{\theta}^n}\int f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\,\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\,d\mathbf{y}\\ &= \int \frac{\partial^n f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\,d\mathbf{y} + \int f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial^n\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}d\mathbf{y} = \mathbf{0},\end{aligned}$$

Then, using that $\mathbf{v}(\mathbf{y};\boldsymbol{\theta}) \triangleq \widehat{\boldsymbol{\alpha}}(\mathbf{y}) - \mathbf{g}(\boldsymbol{\theta})$, it follows that

$$\int f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial^n\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}d\mathbf{y} = -\int f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial^n\mathbf{g}^H(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}d\mathbf{y} = -\frac{\partial^n\mathbf{g}^H(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}$$

must be equal to (2.68) except for the minus sign. Thus, we conclude that

$$E_{\mathbf{y}}\left\{\frac{\partial^n \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}\mathbf{v}^H(\mathbf{y};\boldsymbol{\theta})\right\} = \frac{\partial^n\mathbf{g}^H(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^n}.$$

Appendix 2.G MCRB, CCRB and UCRB derivation

In this appendix, the derivation of the lower bounds introduced in Section 2.6.1 is sketched.

UCRB Derivation

The UCRB involves the computation of the following score function:

$$\begin{aligned}[\mathbf{s}_u(\mathbf{y};\boldsymbol{\theta})]_p &\triangleq \frac{\partial \ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_p} = -\frac{\partial}{\partial \theta_p}\left[\ln\det(\mathbf{R}(\boldsymbol{\theta})) + \operatorname{Tr}\left(\mathbf{R}^{-1}(\boldsymbol{\theta})\widehat{\mathbf{R}}\right)\right] \\ &= \operatorname{Tr}\left(\mathbf{R}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \theta_p}\mathbf{R}^{-1}(\boldsymbol{\theta})\left(\widehat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta})\right)\right),\end{aligned}$$

where $f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta})$ is the Gaussian p.d.f. introduced in (2.22) and the following two expressions from [Ott93, Eq. 4.57-58] have been applied:

$$\begin{aligned}\frac{\partial}{\partial \theta_p}\ln\det(\mathbf{R}(\boldsymbol{\theta})) &= \operatorname{Tr}\left(\mathbf{R}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \theta_p}\right) \\ \frac{\partial}{\partial \theta_p}\operatorname{Tr}\left(\mathbf{R}^{-1}(\boldsymbol{\theta})\widehat{\mathbf{R}}\right) &= -\operatorname{Tr}\left(\mathbf{R}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \theta_p}\mathbf{R}^{-1}(\boldsymbol{\theta})\widehat{\mathbf{R}}\right).\end{aligned}$$

Therefore, the score function $\mathbf{s}_u(\mathbf{y};\boldsymbol{\theta})$ can be written as follows

$$\mathbf{s}_u(\mathbf{y};\boldsymbol{\theta}) = \mathbf{D}_r^H(\boldsymbol{\theta})\left(\mathbf{R}^*(\boldsymbol{\theta}) \otimes \mathbf{R}(\boldsymbol{\theta})\right)^{-1}\left(\widehat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta})\right),$$

with the following definitions

$$\begin{aligned}[\mathbf{D}_r(\boldsymbol{\theta})]_p &\triangleq \operatorname{vec}\left(\partial \mathbf{R}(\boldsymbol{\theta})/\partial \theta_p\right) \\ \widehat{\mathbf{r}} &\triangleq \operatorname{vec}\left(\widehat{\mathbf{R}}\right) \\ \mathbf{r}(\boldsymbol{\theta}) &\triangleq \operatorname{vec}(\mathbf{R}(\boldsymbol{\theta})),\end{aligned} \tag{2.69}$$

and using the following relationships:

$$\begin{aligned}\operatorname{vec}\left(\mathbf{A}\mathbf{B}\mathbf{C}^H\right) &= (\mathbf{C}^* \otimes \mathbf{A})\operatorname{vec}(\mathbf{B}) \\ \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} &= (\mathbf{A} \otimes \mathbf{B})^{-1}.\end{aligned}$$

Finally, in the unconditional model, the Fisher's information matrix becomes

$$\mathbf{J}_u(\boldsymbol{\theta}) \triangleq E_{\mathbf{y}}\left\{\mathbf{s}_u(\mathbf{y};\boldsymbol{\theta})\mathbf{s}_u^H(\mathbf{y};\boldsymbol{\theta})\right\} = \mathbf{D}_r^H(\boldsymbol{\theta})\left(\mathbf{R}^*(\boldsymbol{\theta}) \otimes \mathbf{R}(\boldsymbol{\theta})\right)^{-1}\mathbf{D}_r(\boldsymbol{\theta})$$

using that the covariance matrix of $\widehat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta})$ is precisely $\mathbf{R}^*(\boldsymbol{\theta}) \otimes \mathbf{R}(\boldsymbol{\theta})$ under the Gaussian assumption [Li99, Eq. 20]. In Chapter 4, it will be shown that $\mathbf{J}_u$ can be obtained from $\mathbf{J}_2$ (2.52) when the nuisance parameters are Gaussian distributed.

CCRB Derivation

The CCRB was originally derived in [Sto89][Sto90a] for DOA estimation. A different derivation is given next based on the asymptotic performance of the CML estimator and the estimation bounds theory presented in Section 2.6.1.

In the conditional model, the CML estimator is formulated from the following score function:

$$\begin{aligned}[\mathbf{s}_c(\mathbf{y};\boldsymbol{\theta})]_p &\triangleq \frac{\partial}{\partial\theta_p}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta})) = 2\operatorname{Re}\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\theta_p}\mathbf{A}^{\#}(\boldsymbol{\theta})\widehat{\mathbf{R}}\right)\\ &= 2\operatorname{Re}\operatorname{Tr}\left(\mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\theta_p}\mathbf{A}^{\#}(\boldsymbol{\theta})\left(\widehat{\mathbf{R}}-\mathbf{R}(\boldsymbol{\theta})\right)\right)\end{aligned}$$

that is obtained using the results in Appendix 2.C. Using again (2.69) and

$$\operatorname{vec}\left(\mathbf{ABC}^H\right) = (\mathbf{C}^*\otimes\mathbf{A})\operatorname{vec}(\mathbf{B}),$$

it follows that

$$\mathbf{s}_c(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\theta},\widehat{\mathbf{x}}_{ML}(\boldsymbol{\theta})) = 2\operatorname{Re}\mathbf{D}_a^H(\boldsymbol{\theta})\left(\mathbf{A}^{\#*}(\boldsymbol{\theta})\otimes\mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\theta})\right)(\widehat{\mathbf{r}}-\mathbf{r}(\boldsymbol{\theta}))$$

with $[\mathbf{D}_a(\boldsymbol{\theta})]_p \triangleq \operatorname{vec}(\partial\mathbf{A}(\boldsymbol{\theta})/\partial\theta_p)$.

After some tedious simplifications that are omitted for the sake of brevity, in the conditional model, the Fisher's information matrix is given by

$$\begin{aligned}E_{\mathbf{y}}\left\{\mathbf{s}_c(\mathbf{y};\boldsymbol{\theta})\mathbf{s}_c^H(\mathbf{y};\boldsymbol{\theta})\right\} &= 2\operatorname{Re}\left\{\mathbf{D}_a^H(\boldsymbol{\theta})\left(\mathbf{x}\mathbf{x}^H\otimes\mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\theta})\right)\mathbf{D}_a(\boldsymbol{\theta})\right\}\\ &+2\operatorname{Re}\left\{\mathbf{D}_a^H(\boldsymbol{\theta})\left(\left(\mathbf{A}^H(\boldsymbol{\theta})\mathbf{R}_w^{-1}\mathbf{A}(\boldsymbol{\theta})\right)^{-1}\otimes\mathbf{R}_w^{-1}\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\theta})\right)\mathbf{D}_a(\boldsymbol{\theta})\right\},\end{aligned}$$

that is found to depend on the actual vector of nuisance parameters $\mathbf{x}$. It is shown in [Sto90a, Eq. 2.13] that the first term converges to its expected value as the observation size increases and, thus, $\mathbf{x}\mathbf{x}^H \rightarrow \mathbf{I}_K$. On the other hand, the second term can be neglected if the SNR or the observation length goes to infinity. Actually, this second term causes the CML degradation at low SNR when the observation is short [Sto90a, Eq. 2.15]. Bearing in mind these arguments, the *asymptotic* Fisher's information matrix apperaring in the CCRB expression (2.50) contains only the average of the first term. The resulting expression is known to bound the performance of the CML and GML estimators whatever the SNR or the observation size. However, the adopted CCRB becomes a loose bound for low SNRs in case of finite observations.

MCRB Derivation

A straightforward derivation of the multidimensional MCRB is provided next:

$$\begin{aligned}
\ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right) &= const - \left\|\mathbf{y}-\mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{x}\right\|^2_{\mathbf{R}_w^{-1}} \\
\frac{\partial \ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p} &= 2\,\mathrm{Re}\left\{\left(\mathbf{y}-\mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{x}\right)^H\mathbf{R}_w^{-1}\frac{\partial\mathbf{A}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p}\mathbf{x}\right\} \\
\frac{\partial^2 \ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_q} &= 2\,\mathrm{Re}\left\{\left(\mathbf{y}-\mathbf{A}\left(\boldsymbol{\theta}\right)\mathbf{x}\right)^H\mathbf{R}_w^{-1}\frac{\partial^2\mathbf{A}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_q}\mathbf{x}-\mathbf{x}^H\frac{\partial\mathbf{A}^H\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_q}\mathbf{R}_w^{-1}\frac{\partial\mathbf{A}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p}\mathbf{x}\right\} \\
E_{\mathbf{y}/\mathbf{x}}\left\{\frac{\partial^2 \ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_q}\right\} &= -2\,\mathrm{Re}\left\{\mathbf{x}^H\frac{\partial\mathbf{A}^H\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_q}\mathbf{R}_w^{-1}\frac{\partial\mathbf{A}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p}\mathbf{x}\right\}
\end{aligned}$$

and, therefore,

$$\begin{aligned}
\left[\mathbf{J}_m\right]_{p,q} &= -E_{\mathbf{x}}E_{\mathbf{y}/\mathbf{x}}\left\{\frac{\partial^2 \ln f_{\mathbf{y}/\mathbf{x}}\left(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_q}\right\} \\
&= 2\,\mathrm{Re}\left\{\mathrm{Tr}\left(\frac{\partial\mathbf{A}^H\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_q}\mathbf{R}_w^{-1}\frac{\partial\mathbf{A}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_p}\right)\right\}.
\end{aligned}$$

Finally, the elements of $\mathbf{J}_m$ can be arranged as in equation (2.54) using the following properties:

$$\begin{aligned}
\mathrm{Tr}\left(\mathbf{A}^H\mathbf{B}\right) &= \mathrm{vec}^H\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{B}\right) \\
\mathrm{vec}\left(\mathbf{ABC}\right) &= \left(\mathbf{C}^T\otimes\mathbf{A}\right)\mathrm{vec}\left(\mathbf{B}\right).
\end{aligned}$$