A Bilateral Bound on the Mean-Square Error for Estimation in Model Mismatch

Amir Weiss⋆, Alejandro Lancho⋆, Yuheng Bu†, and Gregory W. Wornell⋆

⋆Massachusetts Institute of Technology   †University of Florida

{amirwei, lancho, gww}@mit.edu   buyuheng@ufl.edu

Abstract—A bilateral (i.e., upper and lower) bound on the mean-square error under a general model mismatch is developed. The bound, which is derived from the variational representation of the chi-square divergence, is applicable in the Bayesian and nonBayesian frameworks to biased and unbiased estimators. Unlike other classical MSE bounds that depend only on the model, our bound is also estimator-dependent. Thus, it is applicable as a tool for characterizing the MSE of a specific estimator. The proposed bounding technique has a variety of applications, one of which is a tool for proving the consistency of estimators for a class of models. Furthermore, it provides insight as to why certain estimators work well under general model mismatch conditions.

Index Terms—Parameter estimation, performance bounds, chi-square divergence, model mismatch.

I. INTRODUCTION

Classical bounds on the mean-square error (MSE) in parameter estimation traditionally assume that the statistical model, which describes the relation between the (random) observations and the parameter of interest, is fully known. Examples include the celebrated Cramér-Rao bound (CRB) for the nonBayesian framework [1], Van Trees (Bayesian CRB) [2], Barankin [3], Ziv-Zakai [4], Abel [5] and Weiss-Weinstein [6] bounds. For a more complete survey, see, e.g., [7], [8].

While these classical bounds are key to understanding the fundamental limitations in optimal parameter estimation, they all refer to the case where the statistical model is *exactly* faithful to the physics. A more realistic, and perhaps contemporary approach acknowledges that the mathematical model does not precisely describe the true underlying physics, and attempts to account for this inherent, almost inevitable mismatch. Of course, this is particularly important from an engineering perspective, where approximations are often made, preferring a mismatch for the benefit of simplicity in implementation.

In such cases, there is a need to understand the impact on performance. From this perspective, practical upper and lower bounds on the MSE under model mismatch are valuable, and as such have been recently receiving increasing attention [9]. More generally, aspects of this topic have a long history, for a variety of different forms of model mismatch and in a variety of different problems. For a selective list of representative examples, see, e.g., [10]–[22].

While much of the focus has been on lower bounds (e.g., the misspecified CRB (MCRB) [23]), upper bounds are also important. For instance, it is often unclear when the MSE of an estimator derived under a different model from the true one will be bounded from above. While upper bounds on MSE have been considered for some (mismatch-free) special cases—including, e.g., [24]–[30]—we are not aware of tools for upper bounding the MSE of a given estimator under general model mismatch conditions. Such an upper bound would provide a guarantee for the actual performance, even if with a gap from the exact performance, which is at any rate unknown since the true model is unknown.

In this paper, we develop a useful bilateral (i.e., upper and lower) bound on MSE under a general model mismatch. The bound is applicable to both biased and unbiased estimators, and in both Bayesian and nonBayesian frameworks. It can be used for establishing the consistency of estimators for a class of models rather than for a single model. Furthermore, it has the potential to be useful in understanding why certain estimators are robust—i.e., work well for various models that deviate from the true underlying (and unknown) model.

II. PRELIMINARIES AND BACKGROUND

Let $\Theta \subseteq \mathbb{R}^{K \times 1}$ be a parameter space, and $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ be a complete probability space, where $\mathcal{X}, \mathcal{F}$ and $\mathcal{P} = \{P_\theta : \boldsymbol{\theta} \in \Theta\}$ denote an observation (or sample) space, a σ-algebra on $\mathcal{X}$, and a collection of probability distributions indexed by $\boldsymbol{\theta} \in \Theta$ over the common measurable space $(\mathcal{X}, \mathcal{F})$, respectively. We assume that all the distributions $\{P_\theta \in \mathcal{P}\}$ are absolutely continuous with respect to a measure μ, which is assumed to be the Lebesgue measure unless stated otherwise. For brevity, we write P_θ simply as P, except where emphasis is required.

We now provide the necessary background for our results.

A. f-divergence and Variational Representation

An f-divergence is the following measure of discrepancy between distributions, defined over a measurable space [31].

Definition 1 (f-divergence): Let $P, Q \in \mathcal{P}$ be two probability distributions on $\mathcal{X}$, such that $P \ll Q$, namely, P is absolutely continuous with respect to Q. Then, for any convex function $f : (0, \infty) \to \mathbb{R}$ that: (i) is strictly convex at 1; and (ii) $f(1) = 0$, the f-divergence of Q from P is defined as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right], \tag{1}$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{X} \to [0,\infty)$ denotes the Radon-Nikodym derivative of P with respect to Q.

An equivalent and useful form of (1) is by the variational representation of f-divergence (e.g., [32]), which uses the notion of convex conjugation, defined as follows.

Definition 2 (Convex conjugate): Let $f : (0,\infty) \to \mathbb{R}$ be a convex function. The convex conjugate f^* of f is defined by

$$f^*(x) \triangleq \sup_{\lambda \in \mathbb{R}} \left[\lambda x - f(\lambda) \right] \triangleq \sup_{\lambda \in \mathbb{R}} \widetilde{f}(x, \lambda).$$

Using the fact that a convex conjugate of a convex function is also convex, and that $(f^*)^* = f$, we obtain the variational representation of f-divergence [32, Ch. 6.1.1] in terms of f^*,

$$D_f(P||Q) = \sup_{g:\mathcal{X} \to \mathbb{R}} \left[\mathbb{E}_P[g(\mathsf{x})] - \mathbb{E}_Q[f^*(g(\mathsf{x}))] \right], \quad (2)$$

where g is such that the expectations in (2) are finite.

B. Chi-square Divergence

Specializing (1) with the function $f(x) = (x-1)^2$, we obtain the chi-square divergence (CSD, e.g, [33]),

$$\chi^2(P||Q) \triangleq \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right)^2 \right] = \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^2 \right] - 1.$$

Using the fact that the convex conjugate of f is given by

$$f^*(x) = \sup_{\lambda \in \mathbb{R}} \left[\lambda x - (x-1)^2 \right],$$

and after a relatively simple change of (the maximization) variable (see, e.g., [32, Ch. 6.1]), one obtains

$$\chi^2(P||Q) = \sup_{g:\mathcal{X} \to \mathbb{R}} \frac{\left(\mathbb{E}_P[g(\mathsf{x})] - \mathbb{E}_Q[g(\mathsf{x})] \right)^2}{\mathbb{V}\mathrm{ar}_Q(g(\mathsf{x}))}, \quad (3)$$

from which the Hammersley-Chapman-Robbins (HCR) [34], [35] bound is readily derived [32, Ch. 6.2]. While the HCR pertains to a mismatch-free setting, we consider the fundamentally different model mismatch setting, as described next.

III. ESTIMATION IN MODEL MISMATCH

Assume that the observations $\mathsf{x}_1, \ldots, \mathsf{x}_N$, denoted collectively as $\mathbf{x} \triangleq [\mathsf{x}_1 \cdots \mathsf{x}_N]^\mathsf{T} \in \mathcal{X}$,[1] are available for estimation of a vector of unknown parameters $\boldsymbol{\theta} \in \Theta$, and that $\mathbf{x}$ and $\boldsymbol{\theta}$ are related via the model $P_{\mathrm{data}} \in \mathcal{P}$. In a standard setting, P_{data} is assumed to be fully known, and given a proper criterion,[2] $\boldsymbol{\theta}$ can be estimated from $\mathbf{x}$. Here, we consider the case where P_{data}, the *true* underlying relation between $\mathbf{x}$ and $\boldsymbol{\theta}$, is (possibly partially) unknown. This setting is realistic in engineering problems where a physical model is unknown, or simply too complicated to describe analytically.

Since P_{data}, the true relation between $\mathbf{x}$ and $\boldsymbol{\theta}$, is in general unknown, the system designer chooses $Q_{\mathrm{data}} \in \mathcal{P}$ to describe this relation. This choice is possibly based on some partial knowledge and/or simplifying approximations. Having chosen

[1] As in traditional notation of classical estimation, with a slight abuse of notation, we assume henceforth that $\mathbf{x}$ (rather than x) is an element in $\mathcal{X}$.

[2] Be it for the frequentist or the Bayesian approach.

Q_{data}, the system designer devises an estimator of $\boldsymbol{\theta}$ based on $\mathbf{x}$, denoted by $\widehat{\boldsymbol{\theta}}(\mathbf{x})$, with an estimation error $\boldsymbol{\varepsilon} \triangleq \widehat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}$.

Note that while the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ is designed based on Q_{data}, its performance is affected by the true underlying model P_{data}. Specifically, if we denote the resulting distribution of the squared estimation error as $P \in \mathcal{P}$, namely $\|\boldsymbol{\varepsilon}\|_2^2 \sim P$, then the actual MSE of this estimator is given by

$$\mathrm{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) \triangleq \mathbb{E}_P\left[\|\boldsymbol{\varepsilon}\|_2^2 \right] \in \mathbb{R}_+. \quad (4)$$

However, since P_{data} is unknown, the resulting P is unknown, and therefore (4) cannot be evaluated. Thus, the following natural (informal) questions arise: (i) Is it still possible to provide some performance guarantees in such (common) model mismatch situations? (ii) Is partial knowledge enough for some strong guarantees (e.g., consistency)? If so, (iii) how much knowledge about P (or P_{data}) is required to this end?

Our main result below shows that the answer to questions (i) and (ii) is yes. As for (iii), the more quantitative question, our results provide one possible answer that offers a trade-off between partial (reasonably available) knowledge and performance guarantees in terms of MSE (e.g., Corollary 2).

A. Main Results

We now state our main result—a bilateral bound on the MSE in model mismatch.

Theorem 1: Let $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ be an estimator of $\boldsymbol{\theta}$, and $\|\boldsymbol{\varepsilon}\|_2^2 = \|\widehat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}\|_2^2$ its associated squared estimation error, distributed according to P, stemming from the data distribution P_{data}. Further, denote by $Q \in \mathcal{P}$ the distribution of $\|\boldsymbol{\varepsilon}\|_2^2$, stemming from the chosen (possibly mismatched) data distribution Q_{data}. Then, the true MSE (4) is lower- and upper-bounded by

$$\mathrm{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) \geq \mathrm{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) - \Delta\left(P, Q, \widehat{\boldsymbol{\theta}}(\mathbf{x}) \right), \quad (5a)$$

$$\mathrm{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) \leq \mathrm{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) + \Delta\left(P, Q, \widehat{\boldsymbol{\theta}}(\mathbf{x}) \right), \quad (5b)$$

where $\Delta\left(P, Q, \widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) \triangleq \sqrt{\mathbb{V}\mathrm{ar}_Q(\|\boldsymbol{\varepsilon}\|_2^2) \cdot \chi^2(P||Q)} \in \mathbb{R}_+$.

Proof: It follows from (3) that

$$\chi^2(P||Q) \geq \frac{\left(\mathbb{E}_P[g(\mathsf{x})] - \mathbb{E}_Q[g(\mathsf{x})] \right)^2}{\mathbb{V}\mathrm{ar}_Q(g(\mathsf{x}))}, \quad (6)$$

for any $g(\cdot)$ for which the right-hand side in (6) is finite.

Now, choose $g(t) = t$ and $\mathsf{x} = \|\boldsymbol{\varepsilon}\|_2^2$ such that P and Q are the distributions of $\|\boldsymbol{\varepsilon}\|_2^2$, induced by the data distributions P_{data} and Q_{data}, respectively. For this particular choice, using the same definition (4) for Q as well, (6) specializes to

$$\chi^2(P||Q) \geq \frac{\left(\mathrm{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) - \mathrm{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right) \right)^2}{\mathbb{V}\mathrm{ar}_Q(\|\boldsymbol{\varepsilon}\|_2^2)}.$$

Multiplying both sides by (the positive) $\mathbb{V}\mathrm{ar}_Q\left(\|\boldsymbol{\varepsilon}\|_2^2 \right)$, taking the square root, writing the two resulting inequalities, and isolating $\mathrm{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right)$, readily gives (5a)–(5b). ∎

Remark 1: In some cases, the terms that depend on Q only (and not on P), namely $\mathrm{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x}) \right)$ and $\mathbb{V}\mathrm{ar}_Q\left(\|\boldsymbol{\varepsilon}\|_2^2 \right)$, can be computed analytically, as they result from the chosen and

known data distribution Q_{data}. Although Q_{data} may be different from the true data distribution P_{data}, this can be very useful, e.g., for proving consistency (see Corollary 2 below).

Remark 2: A prior distribution may or may not be assigned to $\boldsymbol{\theta}$, and (5) still holds (see Section III-C). Hence, the bound is applicable for both the Bayesian and nonBayesian frameworks.

Remark 3: In the absence of model mismatch, i.e., when the chosen data distribution accurately describes the data such that $Q_{\text{data}} = P_{\text{data}}$, the upper and lower bounds coincide, hence $\text{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) = \text{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right)$, as expected. While this property is trivial, it verifies that the bound (5) is sensible. More generally, as long as the variance of the squared error is bounded under Q, the accuracy of the bound improves as the deviation (in the CSD sense) of Q_{data} from P_{data}, and therefore of Q from P, decreases. A desirable property, indeed.

Remark 4: The presumed, chosen model Q_{data} may be such that $\text{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) \leq \text{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right)$ (or even with strict inequality). However, (5) shows that for such "overly optimistic" choices of Q, a sufficiently high penalty must be incurred to the bound in the form of an increased $\Delta\left(P, Q, \widehat{\boldsymbol{\theta}}(\mathbf{x})\right)$ term. This is demonstrated via a simple example in Section III-C.

Remark 5: Recall that the MSE under P cannot be computed, since P is assumed to be partly or fully unknown. In this case, any nontrivial information regarding the actual MSE performance is valuable. As shown in Corollary 2 below, in some cases it is enough to (only) bound the $\chi^2(P||Q)$ term in order to attain nontrivial analytical performance guarantees.

While the expectation and variance under Q do not change when computed under Q_{data} (due to the "law of the unconscious statistician", e.g., [36]), the CSD between the data distributions may be easier to compute or bound. This is owing to the fact that Q and P are transformed probability distributions, which are determined by the estimation rule $\widehat{\boldsymbol{\theta}}(\mathbf{x})$, and Q_{data} and P_{data}. This motivates the following corollary.

Corollary 1: Consider the setting of Theorem 1. Then, for (the true and presumed) distributions $P_{\text{data}}, Q_{\text{data}}$ (respectively) of any function of the raw data, from which the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ may be computed, we have

$$\text{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) \geq \text{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) - \Delta\left(P_{\text{data}}, Q_{\text{data}}, \widehat{\boldsymbol{\theta}}(\mathbf{x})\right), \text{(7a)}$$

$$\text{MSE}_P\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) \leq \text{MSE}_Q\left(\widehat{\boldsymbol{\theta}}(\mathbf{x})\right) + \Delta\left(P_{\text{data}}, Q_{\text{data}}, \widehat{\boldsymbol{\theta}}(\mathbf{x})\right), \text{(7b)}$$

Proof: By the data processing inequality (e.g., [31], [37]),

$$\chi^2(P||Q) \leq \chi^2(P_{\text{data}}||Q_{\text{data}})$$
$$\implies \Delta\left(P, Q, \widehat{\boldsymbol{\theta}}(\mathbf{x})\right) \leq \Delta\left(P_{\text{data}}, Q_{\text{data}}, \widehat{\boldsymbol{\theta}}(\mathbf{x})\right).$$

Since $\mathbb{V}\text{ar}_Q(\|\boldsymbol{\varepsilon}\|_2^2) = \mathbb{V}\text{ar}_{Q_{\text{data}}}(\|\boldsymbol{\varepsilon}\|_2^2)$, (7a) and (7b) bound from below and above (5a) and (5b), respectively. ∎

While (7) is weaker than (5), as explained in Remark 5, it may be more convenient to work with; approximating the estimation error distribution is less trivial than approximating the distribution of the data, since in general we have access to data. Further, (7) may already provide a satisfactory bound.

B. The Gaussian Signal Model

A particular case of high interest is when Q_{data} is chosen as the Gaussian distribution. Indeed, in many applications, such as communication, localization and image denoising, the Gaussian signal model is used in order to derive different estimators for various purposes, even though the actual measured signals clearly do not follow a Gaussian distribution. However, despite the model mismatch, many of these methods work well on real data. This suggests that there exists a more fundamental justification for this fact than simply a good empirical fit.

Theorem 1 provides such a justification, and in particular, an accurate analytical description to the class of models for which the (mismatched) Gaussian model still yields "good" (e.g., consistent) estimators. Consider, for example, the general, ubiquitous signal model (e.g., as in [38])

$$\mathbf{x}_n = \mathbf{h}(\boldsymbol{\theta}) + \mathbf{v}_n \in \mathbb{R}^{M \times 1}, \quad n \in \{1, \ldots, N\}, \tag{8}$$

where $\mathbf{h} : \mathbb{R}^{K \times 1} \to \mathbb{R}^{M \times 1}$ is a known (possibly random) function, $\boldsymbol{\theta}$ is deterministic and unknown, and $\{\mathbf{v}_n\}$ are zero-mean additive noise vectors with an unknown distribution. If we nonetheless choose to assume that $\mathbf{x}_n \overset{\text{iid}}{\sim} Q_{\text{data}} = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Lambda}(\boldsymbol{\theta}))$,[3] we may derive $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$, the maximum-likelihood estimator (MLE) of $\boldsymbol{\theta}$ for this chosen model. However, the performance of $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$, which is not necessarily the MLE for data P_{data} when $P_{\text{data}} \neq Q_{\text{data}}$, is generally no longer necessarily appealing. We emphasize that $\{\mathbf{x}_n\}$ are generally *not* even iid under P_{data}. The following corollary describes the data distributions for which $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$ is consistent despite model mismatch.

Corollary 2: Consider the signal model (8) and assume that $\mathbf{h}$ is such that $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$ is consistent under $\mathbf{x}_n \overset{\text{iid}}{\sim} Q_{\text{data}} = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Lambda}(\boldsymbol{\theta}))$, i.e., $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}} \overset{p}{\longrightarrow} \boldsymbol{\theta}$ as $N \to \infty$, where $\overset{p}{\longrightarrow}$ denotes convergence in probability. Define $\bar{\mathbf{x}} \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$, and denote $\bar{\mathbf{x}} \sim \bar{P}_{\text{data}}$ and $\bar{Q}_{\text{data}} \triangleq \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), N^{-1} \cdot \boldsymbol{\Lambda}(\boldsymbol{\theta}))$. If $\bar{P}_{\text{data}}, \bar{Q}_{\text{data}}$ are "not too far", i.e.,[4]

$$\chi^2\left(\bar{P}_{\text{data}}||\bar{Q}_{\text{data}}\right) = o(N^2), \tag{9}$$

then the estimator $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$ is MSE-consistent under P_{data}, namely,

$$\lim_{N \to \infty} \text{MSE}_P\left(\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}\right) = 0. \tag{10}$$

Proof: For the model (8), under $\mathbf{x}_n \overset{\text{iid}}{\sim} Q_{\text{data}}$, a sufficient statistic is $\bar{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), N^{-1} \cdot \boldsymbol{\Lambda}(\boldsymbol{\theta})) \triangleq \bar{Q}_{\text{data}}$. Therefore, $\bar{\mathbf{x}}$ can be treated as the observed data, with $\bar{Q}_{\text{data}}$ and $\bar{P}_{\text{data}}$ as its hypothesized (possibly mismatched) and true distributions, respectively. From Corollary 1, we have

$$\text{MSE}_P\left(\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}\right) \leq \text{MSE}_Q\left(\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}\right) + \Delta\left(\bar{P}_{\text{data}}, \bar{Q}_{\text{data}}, \widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}\right).$$

While we use $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$ for brevity, we emphasize that in this case we have $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}(\bar{\mathbf{x}})$, since $\bar{\mathbf{x}}$ is a sufficient statistic. Now, since $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}}$ is asymptotically efficient under Q_{data},

$$\boldsymbol{\varepsilon} = \widehat{\boldsymbol{\theta}}_{\text{ML}}^{\mathcal{N}} - \boldsymbol{\theta} \overset{d}{\longrightarrow} \mathcal{N}\left(\mathbf{0}, N^{-1}\mathbf{I}^{-1}(\boldsymbol{\theta})\right), \tag{11}$$

---

[3]The symbol $\overset{\text{iid}}{\sim}$ stands for independent, identically distributed (iid).
[4]The "little-o" notation $a_n = o(b_n)$ means that $\lim_{n \to \infty} a_n \cdot b_n^{-1} = 0$.

where $\xrightarrow{d}$ denotes convergence in distribution, $\mathbf{0}$ is the all-zeros vector (with proper dimensions), and $\mathbf{I}(\boldsymbol{\theta}) \in \mathbb{R}^{K \times K}$ is the Fisher information matrix (see, e.g., [39, Ch. 11.10]) of a single observation $\mathbf{x}_n$, hence

$$\mathrm{MSE}_Q\left(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}\right) = \frac{1}{N} \sum_{k=1}^{K} \underbrace{\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{kk}}_{\triangleq \sigma_{\mathrm{CRB},k}^2} \xrightarrow{N \to \infty} 0.$$

It remains to show that $\Delta\left(\bar{P}_{\mathrm{data}}, \bar{Q}_{\mathrm{data}}, \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}\right) \xrightarrow{N \to \infty} 0$, where

$$\Delta\left(\bar{P}_{\mathrm{data}}, \bar{Q}_{\mathrm{data}}, \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}\right) = \sqrt{\mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}(\|\boldsymbol{\varepsilon}\|_2^2) \cdot \chi^2\left(\bar{P}_{\mathrm{data}} || \bar{Q}_{\mathrm{data}}\right)}.$$

Focusing on the variance of the squared (norm of the) error,

$$\mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\|\boldsymbol{\varepsilon}\|_2^2\right) = \sum_{k=1}^{K} \mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_k^2\right) + \sum_{\substack{k,\ell=1 \\ k \neq \ell}}^{K} \mathrm{COV}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_k^2, \varepsilon_\ell^2\right), \tag{12}$$

where $\mathrm{COV}_{\bar{Q}_{\mathrm{data}}}(\mathsf{a}, \mathsf{b})$ denotes the covariance of a and b, evaluated under $\bar{Q}_{\mathrm{data}}$. Due to (11), we have, asymptotically,

$$\mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_k^2\right) = \underbrace{\mathbb{E}\left[\varepsilon_k^4\right]}_{3 \frac{\sigma_{\mathrm{CRB},k}^4}{N^2}} - \underbrace{\mathbb{E}^2\left[\varepsilon_k^2\right]}_{\left(\frac{\sigma_{\mathrm{CRB},k}^2}{N}\right)^2} = \frac{2\sigma_{\mathrm{CRB},k}^4}{N^2}, \tag{13}$$

where we have used Isserlis' theorem [40] to compute $\mathbb{E}\left[\varepsilon_k^4\right]$. As for the covariance terms in (12), applying the Cauchy-Schwarz inequality and (13), we obtain

$$\mathrm{COV}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_k^2, \varepsilon_\ell^2\right) \leq \sqrt{\mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_k^2\right) \mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\varepsilon_\ell^2\right)}$$
$$= \frac{2\sigma_{\mathrm{CRB},k}^2 \sigma_{\mathrm{CRB},\ell}^2}{N^2}.$$

Therefore, an upper bound on (12) is

$$\mathbb{V}\mathrm{ar}_{\bar{Q}_{\mathrm{data}}}\left(\|\boldsymbol{\varepsilon}\|_2^2\right) \leq \frac{2}{N^2} \underbrace{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}_{\text{independent of } N}, \tag{14}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and the entries of the auxiliary matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$ are defined as

$$[\boldsymbol{\Sigma}]_{k\ell} \triangleq \begin{cases} \sigma_{\mathrm{CRB},k}^2, & k = \ell, \\ \sigma_{\mathrm{CRB},k} \sigma_{\mathrm{CRB},\ell}, & k \neq \ell. \end{cases} \tag{15}$$

Using the bound (14) and assumption (9), we conclude that

$$\Delta\left(\bar{P}_{\mathrm{data}}, \bar{Q}_{\mathrm{data}}, \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}\right) \leq \sqrt{\frac{2\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}{N^2} \cdot \chi^2\left(\bar{P}_{\mathrm{data}} || \bar{Q}_{\mathrm{data}}\right)} \xrightarrow{N \to \infty} 0,$$

and (10) follows. ∎

Remark 6: For simplicity, in the proof of Corollary 2 we assume that the "standard" regularity conditions hold (e.g., [41]), such that $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}$ under Q_{data} is asymptotically efficient, and the Fisher information matrix exists. However, consistency (rather than asymptotic efficiency) of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}^{\mathcal{N}}$ can suffice for (10). It is interesting to note that, similarly to the requirement (9), the Barankin bound also requires a (finiteness) condition on the CSD [42, Eq. 6] (though different in nature).
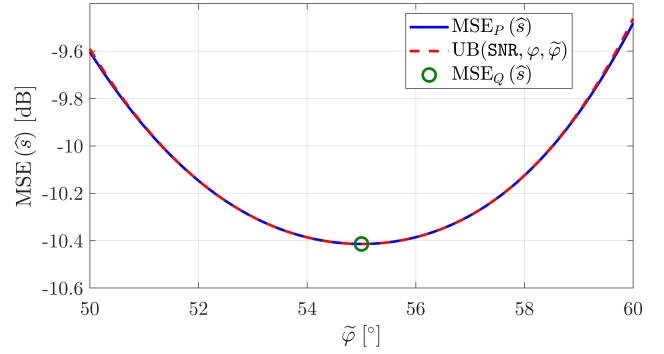

Fig. 1. MSE and the upper bound (16) vs. $\widetilde{\varphi}$ in "DOA-mismatched" estimation for $\varphi = 55°$, $\mathrm{SNR} = 10$. Here, $\mathrm{MSE}_Q(\widehat{\mathsf{s}})$ is independent of $(\varphi, \widetilde{\varphi})$.

The importance of Corollary 2 is that it provides an analytical characterization, in the form of a sufficient condition, for the success of the Gaussian quasi-ML approach [11] in terms of MSE-consistency. Put simply, the condition (9) means that as long as the data distribution is not "too far" from the Gaussian distribution, using the Gaussian MLE is a reasonable approach for estimation, at least asymptotically.

Note further that the proof can be generalized for a non-Gaussian Q_{data}, as long as a similar condition as (9) holds. Moreover, theoretically, it is possible to take the infimum of (7b) over the parameters of the chosen Q_{data}—for example, in the Gaussian case, over the mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and the positive-definite covariance matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$—to get the tightest upper bound of this type.

C. Example Calculation

The following example, which permits an analytical calculation of the bound, demonstrates aspects of the bound behavior. In particular, we consider the optimal multi-sensor receiver with angular mismatch. More specifically, consider a Bayesian version of (8), with $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{a}(\varphi)\mathsf{s}$, where $\theta = \mathsf{s} \sim \mathcal{N}(0, 1)$ is the estimand (i.e., $K = 1$ and $N = 1$) and $\mathbf{a}(\varphi) \in \mathbb{R}^{M \times 1}$ is a (unit-norm) steering vector as a function of the direction-of-arrival (DOA) $\varphi \in [0°, 180°)$, and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\mathrm{SNR}}\mathbf{I}_M)$.[5] In this case, the minimum MSE (MMSE) estimator of s is given by $\widehat{\mathsf{s}}(\varphi, \mathrm{SNR}) \triangleq \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\mathbf{a}(\varphi)^{\mathsf{T}}\mathbf{x}$, when φ is known. In practice, however, φ is known only up to some accuracy level. Consequently, if some $\widetilde{\varphi}$ is used in place of φ, then[6]

$$\varepsilon = \widehat{\mathsf{s}}(\widetilde{\varphi}, \mathrm{SNR}) - \mathsf{s} \sim \begin{cases} P = \mathcal{N}\left(0, \sigma^2\right), & \text{for } P_{\mathrm{data}} \leftrightarrow \widetilde{\varphi} \neq \varphi, \\ Q = \mathcal{N}\left(0, \frac{1}{1+\mathrm{SNR}}\right), & \text{for } Q_{\mathrm{data}} \leftrightarrow \widetilde{\varphi} = \varphi, \end{cases}$$

$$\underbrace{\sigma^2}_{\mathrm{MSE}_P(\widehat{\mathsf{s}})} \triangleq \underbrace{\frac{1}{1+\mathrm{SNR}}}_{\mathrm{MSE}_Q(\widehat{\mathsf{s}})} \cdot \underbrace{\left[3 - 2\mathbf{a}(\varphi)^{\mathsf{T}}\mathbf{a}(\widetilde{\varphi}) + \frac{\left(\mathrm{SNR}(1-\mathbf{a}(\varphi)^{\mathsf{T}}\mathbf{a}(\widetilde{\varphi}))\right)^2}{1+\mathrm{SNR}}\right]}_{\triangleq 1/\gamma^2(\mathrm{SNR}, \varphi, \widetilde{\varphi}) \geq 1 \text{ ("inflation factor")}}.$$

Moreover, in this case the CSD is given by

$$\chi^2(P||Q) = \frac{\gamma^2}{\sqrt{2\gamma^2 - 1}} - 1 \triangleq c^2(\mathrm{SNR}, \varphi, \widetilde{\varphi}), \quad \forall \gamma > \frac{1}{\sqrt{2}},$$

---

[5]This example in fact deals with the linear minimum MSE (LMMSE) receiver [43] of a single-input multiple-output (SIMO) communication system.

[6]Here, P and Q are the error (rather than the squared error) distributions. Thus, by virtue of Corollary 1, we consider a slight variant of Theorem 1.

where γ is shorthand for $\gamma(\text{SNR}, \varphi, \widetilde{\varphi})$. Thus, using (5b),

$$\text{MSE}_P(\widehat{\mathsf{s}}) \leq \frac{(1 + c(\text{SNR}, \varphi, \widetilde{\varphi})\sqrt{2})}{1 + \text{SNR}} \triangleq \text{UB}(\text{SNR}, \varphi, \widetilde{\varphi}), \quad (16)$$

where an obvious lower bound is $\text{MSE}_Q(\widehat{\mathsf{s}})$, since the MSE is minimized for the true DOA. Fig. 1 shows the MSE under P and the upper bound (16) (i.e., (5b)) vs. the mismatched DOA $\widetilde{\varphi}$ for an angular uncertainty of $10°$ centered at the true DOA $\varphi = 55°$ for $\text{SNR} = 10$ dB; the tightness of (5b) is evident.

An intuitive interpretation of the above is that Q_{data} represents an "overly optimistic" point of view ($\gamma \leq 1$), where the assumed DOA is exact: $\widetilde{\varphi} = \varphi$. Thus, the upper bound is closer to the true performance curve, and the lower bound becomes less informative. Clearly, here $\text{MSE}_Q(\widehat{\mathsf{s}}) \leq \text{MSE}_P(\widehat{\mathsf{s}})$ for any (DOA) mismatch. Furthermore, $\text{MSE}_Q(\widehat{\mathsf{s}})$ is greater than the lower bound (5a). Evidently, this example reveals that the bounds (5a) and (5b) are not always *simultaneously* informative. However, perhaps surprisingly, in some situations they are, as demonstrated in the next section.

IV. Representative Application of the Bound

We now consider a representative application of the bound in a nonBayesian framework (in contrast to that of Section III-C), which showcases bilateral tightness, as well as an improvement over the MCRB in the non-asymptotic regime.

Consider the time-of-arrival (TOA) estimation problem, which is instrumental in a host of engineering applications, e.g., [44]–[46]. Specifically, we focus on the case study of a mismatched waveform considered in [21], which is a special case of (8), with $\theta = \tau$ (i.e., $K = 1$), $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_M)$, and

$$h_m(\tau) = e^{-\left(\frac{t-\tau}{T_p}\right)^2}\bigg|_{t=mT_s} \triangleq h_m(\tau, T_p), \ m \in \{1, \ldots, M\},$$

where τ is the unknown TOA, $h_m(\tau, T_p)$ is the m-th sample of a τ-shifted Gaussian pulse with pulse width T_p, and T_s is the sampling period, hence MT_s is the observation interval.

While the true model of the observation $\mathbf{x}$ is $P_{\text{data}} = \mathcal{N}(\mathbf{h}(\tau, T_p), \sigma^2 \mathbf{I}_M)$, if there is imprecise knowledge of the pulse width,[7] and it is assumed to be T_Q, we have $Q_{\text{data}} = \mathcal{N}(\mathbf{h}(\tau, T_Q), \sigma^2 \mathbf{I}_M)$. When the estimator $\widehat{\tau}$ is designed based on T_Q, and the system designer is aware of the potential mismatch due to some inherent physical uncertainty, performance guarantees (upper bounds) and fundamental limitation (lower bounds) of the actual MSE can be of high practical value.

Remark 7: Although we do not use the looser version of the bound (Corollary 1), we now show that in this problem, the CSD of the data distributions can be computed in closed-form. It is known that for two multivariate normal distributions $P = \mathcal{N}(\boldsymbol{\mu}_P, \boldsymbol{\Lambda}), Q = \mathcal{N}(\boldsymbol{\mu}_Q, \boldsymbol{\Lambda})$ [48],

$$D_f(P||Q) = D_f(\mathcal{N}(0,1)||\mathcal{N}(\delta_{\boldsymbol{\Lambda}}(\boldsymbol{\mu}_P, \boldsymbol{\mu}_Q), 1)), \quad (17)$$

where $\delta_{\boldsymbol{\Lambda}}^2(\boldsymbol{\mu}_P, \boldsymbol{\mu}_Q) \triangleq (\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)^{\mathsf{T}} \boldsymbol{\Lambda}^{-1}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)$ is the Mahalanobis generalized distance. Hence, in our TOA estimation problem, we have (e.g., [26])

$$\chi^2(P_{\text{data}}||Q_{\text{data}}) = \exp\left\{\frac{\|\mathbf{h}(\tau, T_P) - \mathbf{h}(\tau, T_Q)\|_2^2}{\sigma^2}\right\} - 1. \quad (18)$$

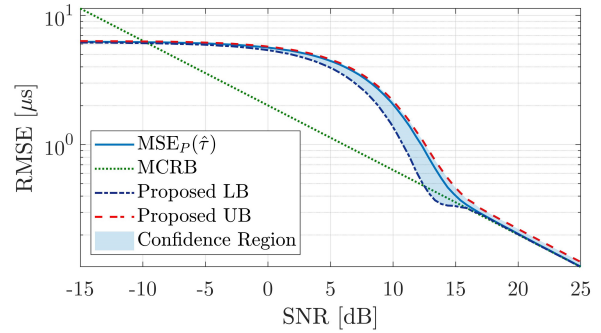[7]Such a mismatch can occur, for example, in ultrasound [47].

Fig. 2. Root MSE (RMSE) vs. SNR in TOA estimation, for $T_s = 10^{-2}\mu s, T_P = 2\mu s, T_Q = 1.1 \cdot T_P, M = 2 \cdot 10^3$ and $\tau \sim \text{Unif}(-5, 5)\mu s$.

However, as explained in [21], for Gaussian pulses, the squared-norm in (18) can be approximated (with exponentially vanishing approximation errors) by an integral over the whole real line, after which a trivial change of integration variable shows that the squared norm in (18) is independent of τ.

It is well-known that in this problem, at low signal-to-noise ratios (SNRs), the estimation error (of any reasonable estimator) is uniformly distributed on the uncertainty time-interval. Furthermore, at high SNRs, a cross-correlation-based estimator (CCE) is normally distributed around the true TOA, where the optimal CCE attains the minimal attainable variance. Thus, we expect that the CSD between the *error* distributions of two slightly different CCEs will not differ significantly.

To compute the CSD between the distributions of the squared errors resulting from P_{data} and Q_{data} (Theorem 1), we use the data-dependent partition divergence estimator [49, Eq. 6]. For the lower bound, we also use the (trivial) fact that the MSE is a monotonic nonincreasing function of the SNR, namely that $\text{MSE}(\text{SNR}_1) \leq \text{MSE}(\text{SNR}_2)$ for all $\text{SNR}_1 \geq \text{SNR}_2$. Hence, a refinement of the lower bound (5a) in this case is

$$\text{LB}(\text{SNR}) \triangleq \max\{\text{LB}_{\text{Thm1}}(\varrho^2) : \varrho^2 \geq \text{SNR}\}, \quad (19)$$

where $\text{LB}_{\text{Thm1}}(\varrho^2)$ denotes (19) at an SNR level of ϱ^2.

Fig. 2 presents the MSE vs. the SNR of the CCE, designed based on $T_Q = 1.1 \cdot T_P$ (namely, erroneously assuming $\mathbf{x} \sim Q_{\text{data}}$), the MCRB [21, Eq. 14], and the proposed (refined) lower (19) and upper (5b) bounds (LB and UB, respectively). Evidently, the proposed LB is tighter than the MCRB at the low SNR regime, where the MCRB is not only uninformative, but no longer serves as a legitimate lower bound. In the transition region, the LB and UB satisfactorily capture the threshold phenomenon (e.g., [50], [51]). Moreover, the UB (5b) provides the guaranteed accuracy despite the mismatch.

V. Concluding Remarks

We develop a bilateral bound on MSE that is applicable to a general (not necessarily unbiased) estimator derived under a mismatched model. The bound provides performance guarantees for the operation of estimators that are designed to operate in one setting, but are then applied in a different one.

An interesting direction for future research is the potential applications of the bound in the context of machine learning, such as supervised regression, where the presence and assumption of model mismatch is ubiquitous.

REFERENCES

[1] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[2] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[3] E. W. Barankin, "Locally best unbiased estimates," *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 477–501, 1949.

[4] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inf. Theory*, vol. 15, no. 3, pp. 386–391, 1969.

[5] J. S. Abel, "A bound on mean-square-estimate error," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1675–1680, 1993.

[6] A. J. Weiss and E. Weinstein, "A lower bound on the mean-square error in random parameter estimation (corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, no. 5, pp. 680–682, 1985.

[7] K. Todros and J. Tabrikian, "General classes of performance lower bounds for parameter estimation—Part I: Non-Bayesian bounds for unbiased estimators," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5045–5063, 2010.

[8] ——, "General classes of performance lower bounds for parameter estimation—Part II: Bayesian bounds," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5064–5082, 2010.

[9] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, "Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 142–157, 2017.

[10] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather Modification; University of California Press: Berkeley, CA, USA*, 1967, p. 221.

[11] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the econometric society*, pp. 1–25, 1982.

[12] Q. H. Vuong, "Cramér-Rao bounds for misspecified models," *Div. of the Humanities and Social Sci., California Inst. of Technol., Pasadena, CA, USA*, 1986.

[13] W. Xu, A. B. Baggeroer, and K. L. Bell, "A bound on mean-square estimation error with background parameter mismatch," *IEEE Trans. Inf. Theory*, vol. 50, no. 4, pp. 621–632, 2004.

[14] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3712–3720, 2010.

[15] M. Fozunbal, "On regret of parametric mismatch in minimum mean square error estimation," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2010, pp. 1408–1412.

[16] C. Fritsche, U. Orguner, E. Özkan, and F. Gustafsson, "On the Cramér-Rao lower bound under model mismatch," in *Proc. of ICASSP*, 2015, pp. 3986–3990.

[17] S. Fortunati, M. S. Greco, and F. Gini, "A lower bound for the mismatched maximum likelihood estimator," in *IEEE Radar Conference (RadarCon)*, 2015, pp. 0180–0185.

[18] C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2263–2278, 2015.

[19] M. L. Diong, E. Chaumette, and F. Vincent, "Generalized Barankin-type lower bounds for misspecified models," in *Proc. of ICASSP*, 2017, pp. 4466–4470.

[20] M. Pajovic, "Misspecified Bayesian cramér-rao bound for sparse Bayesian," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 2018, pp. 263–267.

[21] F. Roemer, "Misspecified Cramer-Rao bound for delay estimation with a mismatched waveform: A case study," in *Proc. of ICASSP*, 2020, pp. 5994–5998.

[22] L. T. Thanh, K. Abed-Meraim, and N. L. Trung, "Misspecified Cramer–Rao bounds for blind channel estimation under channel order misspecification," *IEEE Trans. Signal Process.*, vol. 69, pp. 5372–5385, 2021.

[23] S. Fortunati, F. Gini, and M. S. Greco, "The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2387–2399, 2016.

[24] L. P. Seidman, "An upper bound on average estimation error in nonlinear systems," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 243–250, 1968.

[25] U. Timor, "An upper bound on the estimation error in the threshold region," *IEEE Trans. Inf. Theory*, vol. 16, no. 6, pp. 692–699, 1970.

[26] M. Zakai and J. Ziv, "Lower and upper bounds on the optimal filtering error of certain diffusion processes," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 325–331, 1972.

[27] R. M. Hawkes and J. B. Moore, "An upper bound on the mean-square error for Bayesian parameter estimators (corresp.)," *IEEE Trans. Inf. Theory*, vol. 22, no. 5, pp. 610–615, 1976.

[28] Y. Ephraim and N. Merhav, "Lower and upper bounds on the minimum mean-square error in composite source signal estimation," *IEEE Trans. Inf. Theory*, vol. 38, no. 6, pp. 1709–1724, 1992.

[29] P. Schniter and C. R. Johnson, Jr., "Bounds for the MSE performance of constant modulus estimators," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2544–2560, 2000.

[30] F. Belliardo and V. Giovannetti, "Achieving Heisenberg scaling with maximally entangled states: An analytic upper bound for the attainable root-mean-square error," *Physical Review A*, vol. 102, no. 4, p. 042613, 2020.

[31] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Sci. Math. Hungar.*, vol. 2, pp. 229–318, 1967.

[32] Y. Wu, "Lecture notes on information-theoretic methods for high-dimensional statistics," *Lecture Notes for ECE598YW (UIUC)*, vol. 16, 2017.

[33] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 10–13, 2013.

[34] J. M. Hammersley, "On estimating restricted parameters," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950.

[35] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *The Annals of Mathematical Statistics*, pp. 581–586, 1951.

[36] S. Ross, *A first course in probability*. Pearson, 2010.

[37] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, no. 1-4, pp. 191–213, 1972.

[38] A. Yeredor, A. Weiss, and A. J. Weiss, "High-order analysis of the efficiency gap for maximum likelihood estimation in nonlinear Gaussian models," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4782–4795, 2018.

[39] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

[40] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918.

[41] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[42] F. E. Glave, "A new look at the Barankin lower bound," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 349–356, 1972.

[43] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[44] J. D. Taylor, *Ultra-Wideband Radar Systems*. CRC press, 2001.

[45] C. Falsi, D. Dardari, L. Mucchi, and M. Z. Win, "Time of arrival estimation for UWB localizers in realistic environments," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–13, 2006.

[46] C. Jansen, S. Wietzke, O. Peters, M. Scheller, N. Vieweg, M. Salhi, N. Krumbholz, C. Jördens, T. Hochrein, and M. Koch, "Terahertz imaging: applications and perspectives," *Applied optics*, vol. 49, no. 19, pp. E48–E57, 2010.

[47] J. A. Jensen and S. Leeman, "Nonparametric estimation of ultrasound pulses," *IEEE Trans. Biomed. Eng.*, vol. 41, no. 10, pp. 929–936, 1994.

[48] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc., Ser. B*, vol. 28, pp. 131–142, 1966.

[49] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[50] A. J. Weiss and E. Weinstein, "Fundamental limitations in passive time delay estimation–Part I: Narrow-band systems," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 31, no. 2, pp. 472–486, 1983.

[51] E. Weinstein and A. J. Weiss, "Fundamental limitations in passive time-delay estimation–Part II: Wide-band systems," *IEEE Trans. Acoust., Speech Signal Process.,*, vol. 32, no. 5, pp. 1064–1078, 1984.