Signal Processing 207 (2023) 108933

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A Kullback-Leibler Divergence Variant of the Bayesian Cramér-Rao Bound

Michael Fauß[a,1,*], Alex Dytso[b,2], H. Vincent Poor[a,2]

[a] Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA
[b] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA

ARTICLE INFO

Article history:
Received 14 April 2022
Revised 1 November 2022
Accepted 6 January 2023
Available online 20 January 2023

MSC:
0000
1111

PACS:
0000
1111

Keywords:
MMSE bounds
information inequalities
Cramér–Rao bound
Kullback–Leibler divergence

ABSTRACT

This paper proposes a Bayesian Cramér-Rao type lower bound on the minimum mean square error. The key idea is to minimize the latter subject to the constraint that the joint distribution of the input-output statistics lies in a Kullback–Leibler divergence ball centered at a Gaussian reference distribution. The bound is tight and is attained by a Gaussian distribution whose mean is identical to that of the reference distribution and whose covariance matrix is determined by a scalar parameter that can be obtained by finding the unique root of a simple function. Examples of applications in signal processing and information theory illustrate the usefulness of the proposed bound in practice.

1. Introduction

The mean square error (MSE) is a natural and commonly used measure for the accuracy of an estimator. The minimum MSE (MMSE) plays a central role in statistics [1,2], information theory [3,4], and signal processing [5–7] and has been shown to have close connections to entropy and mutual information [8,9].

However, often the MMSE is difficult to compute so that bounds have to be considered instead. Generally, MMSE lower bounds can roughly be broken into three families. The first family, termed Ziv–Zakai bounds, works by connecting estimation and binary hypothesis testing [10]. The second family, termed the genie approach, works by providing side information and, thus, reducing the MMSE [11]. Finally, the third family, termed Weiss–Weinstein, works by using the Cauchy-Schwarz inequality [12]; the ubiquitous Cramér-Rao bound (CRB) is an example of this family. The bound proposed here falls in neither of these families, yet it can be argued to be of the Cramér–Rao type.

In [13], we showed that instead of using the Cauchy–Schwarz inequality, both the conventional and the Bayesian CRB can also be proven via variational arguments. Seen from this angle, the Bayesian CRB corresponds to the minimum MMSE that can be attained under a constraint on the Fisher information of the joint input-output distribution. The bound proposed in this paper is defined in analogy to this interpretation, but with the Fisher information replaced by the Kullback–Leibler (KL) divergence to a Gaussian reference distribution.

A specialized version of the bound presented in this paper has previously been shown in [14], where a similar approach was used to derive upper and lower bounds on the MMSE of additive Gaussian noise channels. The bound presented here extends the one in [14] in three ways. First, it applies to a significantly larger class of channels, namely, to all joint input-output distributions whose KL divergence from a Gaussian reference distribution is finite. Second, it is tight, which is not the case for the lower bound in [14]. Finally, the connection to the variational interpretation of the CRB, which positions the proposed bound in the larger con-

* Corresponding author.
E-mail address: mfauss@princeton.edu (M. Fauß).

[1] The work of M. Fauß was supported by the German Research Foundation (DFG) under Grant 424522268.
[2] The work of H. V. Poor was in part supported by the U.S. National Science Foundation (NSF) under Grant CCF-1908308.

https://doi.org/10.1016/j.sigpro.2023.108933
0165-1684/

text of lower bounds in estimation theory, was not identified in [14].

It is interesting to note that a complementary result to the bound in this paper, namely, an *upper bound* on the MMSE based on the KL divergence to a Gaussian reference distribution, is well-known in the literature and has been studied repeatedly in the context of robust estimation [15–17]. While both bounds are closely related and formally similar, their underlying motivations are markedly different: In robust estimation, the goal is to design estimators that are insensitive against small, random model mismatches. Consequently, the KL divergence is used to model the deviation of the true, unknown distribution from an ideal Gaussian distribution. The minimax robust estimator is then defined as the one that minimizes the worst-case MSE over this neighborhood. The motivation for introducing a KL divergence constraint in the derivation of the lower bound is different. Here, it is assumed that the true distribution is *known exactly*, but that the corresponding MMSE is prohibitively hard to calculate. Hence, the idea is to look for a "surrogate property" that is easier to evaluate. A bound can then be obtained by minimizing the MMSE over all distributions for which the value of the surrogate property does not exceed that of the true distribution. For the CRB, this surrogate property is the Fisher information; for the bound proposed in this paper, it is the KL divergence to a Gaussian reference distribution. In a nutshell, while in robust MMSE estimation the KL divergence constraint captures *unknown* and potentially *harmful* aspects of the distribution, in Cramér–Rao type lower bounds it captures a *known* and *useful* property.

This conceptual difference between the lower and upper bound might explain why the two have not been derived as a pair. Another reason for why the upper bound was established independently and much earlier could be that the underlying minimax optimization problem is strictly convex and, therefore, guaranteed to have a unique solution. The minimization problem underlying the lower bound, on the other hand, is non-convex, so that it is not clear whether a global minimum can be identified in the first place. This aspect will be discussed in more detail later in the paper. Finally, as can be seen in [15–17], the focus in robustness is naturally on the minimax robust estimator that attains the upper bound, while the upper bound itself is of little interest. More precisely, if the aim is to upper bound the MMSE, the bounds obtained via the best linear estimator are simpler and more accurate; compare the discussion in [14, Section V.B].

Clearly, the idea of bounding the MMSE by minimizing it over a suitable class of distributions and estimators is applicable in a much more general setting. The MMSE, the Fisher information, and the KL divergence are by no means the only candidates for useful bounds of this type. In the course of the paper, it should become clear that the Bayesian CRB and its Kullback–Leibler variant are just two members of a much larger family of bounds. We conjecture that further investigating the properties and members of this family is a promising avenue for future research. Preliminary results exist for risks induced by Bregman divergences and exponential family reference distributions [18]. Nevertheless, the specific bound presented in this paper stands out among the members of the proposed family since it is an *exact equivalent of the Bayesian CRB*, which makes it relevant from a theoretical point of view, and since it is easy to evaluate numerically, which makes it relevant from a practical point of view.

The remainder of the paper is organized as follows: A formal problem formulation is given in Section 2. The main result of the paper, a Cramér–Rao type lower bound on the MMSE, is stated in Section 3, proved in Section 4, and discussed in Section 5. Some illustrative numerical examples are presented in Section 6. Section 7 concludes the paper.

2. Problem formulation

Let $(\mathbb{R}^K, \mathcal{B}^K)$ denote the K-dimensional Borel space, and let $X \in (\mathbb{R}^K, \mathcal{B}^K)$ and $Y \in (\mathbb{R}^M, \mathcal{B}^M)$ be two random variables with joint distribution P. The MSE when estimating X from Y is defined as a function of the joint distribution P and an estimator f, that is,

$$mse_{X|Y}(f,P) := E_P\Big[\|X - f(Y)\|^2\Big], \tag{1}$$

where E_P denotes the expectation taken with respect to P (the subscript P will occasionally be dropped when the distribution is clear from the context), and f denotes a measurable function mapping from $(\mathbb{R}^K, \mathcal{B}^K)$ to $(\mathbb{R}^M, \mathcal{B}^M)$. The set of all estimators[3] is denoted by $\mathcal{F}$, and the MMSE is defined as

$$mmse_{X|Y}(P) := \inf_{f\in\mathcal{F}} mse_{X|Y}(f,P). \tag{2}$$

The estimator attaining the MMSE is $f(Y) = E\big[X|Y\big]$, the latter denoting the expected value of X given Y.

The problem investigated in this paper is

$$\inf_P mmse_{X|Y}(P) \quad \text{s.t.} \quad P \in \mathcal{P}_\varepsilon(P_0), \tag{3}$$

where $\mathcal{P}_\varepsilon(P_0)$ is a KL divergence ball of radius ε centered at P_0, that is,

$$\mathcal{P}_\varepsilon(P_0) := \big\{P : D_{\mathrm{KL}}\big(P\|P_0\big) \le \varepsilon\big\}. \tag{4}$$

Moreover, P_0 is assumed to be a Gaussian distribution,

$$P_0 = \mathcal{N}(\mu_0, \Sigma_0), \tag{5}$$

with mean vector

$$\mu_0 = \begin{bmatrix}\mu_{X_0}\\ \mu_{Y_0}\end{bmatrix} \in \mathbb{R}^{K+M}, \tag{6}$$

$\mu_{X_0} \in \mathbb{R}^K$, $\mu_{X_0} \in \mathbb{R}^M$, and covariance matrix

$$\Sigma_0 = \begin{bmatrix}A_0 & B_0\\ B_0^{\mathrm{T}} & C_0\end{bmatrix} \in \mathbb{S}^{K+M}, \tag{7}$$

where $A_0 \in \mathbb{S}^K$, $B_0 \in \mathbb{R}^{K\times M}$ and $C_0 \in \mathbb{S}_+^M$. Here $\mathbb{S}_+^K$ ($\mathbb{S}^K$) denotes the sets of real positive (semi)definite matrices of size $K \times K$.

3. Main result

Before stating the solution of (3), it is useful to briefly summarize the Gaussian case

$$P = P_0 = \mathcal{N}(\mu_0, \Sigma_0), \tag{8}$$

which corresponds to $\varepsilon = 0$. In this case, the MMSE estimator is given by

$$f_0(Y) := E_{P_0}[X|Y] = \mu_{X_0} + B_0 C_0^{-1}(Y - \mu_{Y_0}) \tag{9}$$

and the MMSE calculates to

$$mmse_{X|Y}(P_0) = tr\big(\Xi_0\big) = \sum_{k=1}^{K} \xi_{0,k}, \tag{10}$$

where

$$\Xi_0 := \Sigma_0/C_0 = A_0 - B_0 C_0^{-1} B_0^{\mathrm{T}} \tag{11}$$

denotes the Schur complement of Σ_0 in C_0 and $\xi_{0,1} \ge \xi_{0,2} \ge \ldots \ge \xi_{0,K}$ denote the ordered eigenvalues of Ξ_0. It is now possible to state the main results of this paper.

[3] More precisely, $\mathcal{F}$ denotes a quotient set, where two estimators are equivalent if they differ only on an Lebesgue null set.

Theorem 1. *For all distributions $P \in \mathcal{P}_\varepsilon(P_0)$, with P_0 of the form Eq. (5), it holds that*

$$mmse_{X|Y}(P) \geq \sum_{k=1}^{K} \frac{\xi_{0,k}}{1+\gamma^* \xi_{0,k}} \tag{12}$$

where γ^* is the unique positive solution of

$$\sum_{k=1}^{K} \phi(\gamma\, \xi_{0,k}) = 2\varepsilon, \tag{13}$$

and

$$\phi(t) := \log(1+t) - \frac{t}{1+t}. \tag{14}$$

The bound is attained by the estimator f_0 in Eq. (9) and the Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_{\gamma^*})$, where

$$\Sigma_\gamma = \Sigma_0 - \gamma \begin{bmatrix} \Xi_0 (I_K + \gamma\, \Xi_0)^{-1} \Xi_0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{15}$$

and I_K denotes the identity matrix of size $K \times K$.

The theorem is proved in the next section.

4. Proof of the main result

The proof of the bounds given in the previous section is based on the Lagrange function

$$L_\lambda(f, P) := mse_{X|Y}(f, P) + \lambda D_{\mathrm{KL}}\big(P \| P_0\big), \tag{16}$$

with $\lambda > 0$. Some useful properties of L_λ are stated in the following two Lemmas.

Lemma 1. *The function $L_\lambda(f, \bullet)$ is strictly convex for all $f \in \mathcal{F}$.*

Lemma 2. *Let $f \in \mathcal{F}$ be given, and let η denote the standard Lebesgue measure on $(\mathbb{R}^K, \mathcal{B}^K)$. If some $c_f > 0$ exists such that*

$$p_f^*(x,y) = c_f\, p_0(x,y)\, e^{-\frac{1}{\lambda}\|x - f(y)\|_2^2} \tag{17}$$

is a valid density w.r.t. η, then the corresponding distribution P_f^* solves

$$\inf_P L_\lambda(f, P). \tag{18}$$

The proofs of both Lemmas follow in close analogy to the proofs of Lemma 1 and Lemma 2 in [14] and are omitted for brevity. A proof of Lemma 2, based on the Donsker–Varadhan representation of the KL divergence can be found in Appendix A of [19].

Another result that will be useful in what follows is the uniqueness of the MMSE estimator, which is fixed in the following lemma.

Lemma 3. *Suppose that $E\big[\|f(Y)\|_2^2\big] < \infty$. Then,*

$$E\big[\langle X - f(Y), g(Y)\rangle\big] \leq 0 \quad \forall g \text{ s.t. } E\big[\|g(Y)\|_2^2\big] < \infty \tag{19}$$

where $\langle \bullet, \bullet \rangle$ denotes the inner product, if and only if

$$f(Y) = E[X|Y] \quad \text{a.s.} \tag{20}$$

Moreover, $E[X|Y]$ attains Eq. (19) with equality.

Proof. First, Eq. (19) can be re-written as

$$E[\langle X - f(Y), g(Y)\rangle]$$

$$= E[\langle X - E[X|Y], g(Y)\rangle] + E[\langle E[X|Y] - f(Y), g(Y)\rangle] \tag{21}$$

$$= E\big[\langle E[X|Y] - f(Y), g(Y)\rangle\big], \tag{22}$$

where the last equality follows from the orthogonality principle. Therefore, Eq. (19) can equivalently be written as

$$E\big[\langle E[X|Y] - f(Y), g(Y)\rangle\big] \leq 0 \quad \forall g \text{ s.t. } E\big[\|g(Y)\|_2^2\big] < \infty. \tag{23}$$

Clearly, the estimator in Eq. (20) forces Eq. (23) to be zero. In order to see the other direction, choose $g(Y) = E[X|Y] - f(Y)$, which results in

$$E\big[\|E[X|Y] - f(Y)\|_2^2\big] \leq 0. \tag{24}$$

Hence, the only function that satisfies Eq. (24) is given by $f(Y) = E[X|Y]$ a.s.. This concludes the proof. □

4.1. Proof of the main result

Consider the auxiliary problem

$$\inf_f \inf_P L_\lambda(f, P), \tag{25}$$

with

$$\lambda = \frac{2}{\gamma}, \quad \gamma > 0. \tag{26}$$

The inner minimization in Eq. (25) can be solved via Lemma 2:

$$\inf_P L_\lambda(f, P) = E_{P_f^*}\left[\|X - f(Y)\|_2^2 + \frac{2}{\gamma} \log \frac{p_f^*(X,Y)}{p_0(X,Y)}\right] \tag{27}$$

$$= E_{P_f^*}\left[\|X - f(Y)\|_2^2 + \frac{2}{\gamma} \log c_f - \|X - f(Y)\|_2^2\right] \tag{28}$$

$$= \frac{2}{\gamma} \log c_f \tag{29}$$

$$= -\frac{2}{\gamma} \log E_{P_0}\left[e^{-\frac{\gamma}{2}\|X - f(Y)\|_2^2}\right], \tag{30}$$

where the last equality follows from c_f having to be chosen such that p_f^* is a valid density function, that is

$$\int p_f^*(x,y)\, \mathrm{d}x\, \mathrm{d}y = \int c_f\, p_0(x,y)\, e^{-\frac{1}{\lambda}\|x - f(y)\|_2^2}\, \mathrm{d}x\, \mathrm{d}y \tag{31}$$

$$= c_f E_{P_0}\left[e^{-\frac{1}{\lambda}\|x - f(y)\|_2^2}\right] = 1. \tag{32}$$

The optimal estimator in Eq. (25) can hence be characterized by the problem

$$\sup_f E_{P_0}\left[e^{-\frac{\gamma}{2}\|X - f(Y)\|_2^2}\right]. \tag{33}$$

This problem is in general nonconvex. However, using the uniquess of the MMSE estimator, we will show that Eq. (33) admits a single, unique local maximum, which in turn implies that the local maximum is a global maximum.

A necessary condition for an estimator to be a local maximum of Eq. (33) is that the Gâteaux derivative of the objective function is nonpositive in the direction of every estimator $g \in \mathcal{F}$ [20, Ch. 7.4, Thm. 2]. This derivative calculates to

$$-\gamma\, E_{P_0}\left[\langle X - f(Y)\,,\, g(Y)\rangle e^{-\frac{\gamma}{2}\|X - f(Y)\|_2^2}\right]$$

$$= -\frac{\gamma}{c_f} E_{P_f^*}[\langle X - f(Y)\,,\, g(Y)\rangle], \tag{34}$$

This yields the necessary optimality condition

$$E_{P_f^*}\Big[\langle X - f(Y)\,,\, g(Y)\rangle\Big] \geq 0 \tag{35}$$

for all $g \in \mathcal{F}$. In view of Lemma 3, the only function that satisfies this condition is the conditional mean estimator under P_f^*, that is

$$f^*(Y) = E_{P_{f^*}^*}\big[X|Y\big] \quad \text{a.s.} \tag{36}$$

Moreover, since the conditional expectation is the only function that satisfies Eq. (35), there exist no other local maxima. Since the global maximum is a local maximum, this implies that Eq. (36) attains the global maximum.

Next, it is shown that the condition in Eq. (36) is satisfied by the estimator f_0 in Eq. (9). In order to see this, note that for $f = f_0$, $P^*_{f_0}$ in Eq. (17) is a Gaussian distribution with mean $\mu_{f_0} = \mu_0$ and precision matrix

$$\Sigma_{f_0}^{-1} = \Sigma_0^{-1} - \gamma U_0^{\mathrm{T}} U_0 \tag{37}$$

where

$$U_0 = \begin{bmatrix} I_K & -B_0 C_0^{-1} \end{bmatrix}. \tag{38}$$

From Woodbury's matrix identity it follows that the corresponding covariance matrix is of the form

$$\Sigma_{f_0} = \Sigma_0 - \left(U_0 \Sigma_0\right)^{\mathrm{T}} \left(\tfrac{1}{\gamma} I + U_0 \Sigma_0 U_0^{\mathrm{T}}\right)^{-1} U_0 \Sigma_0. \tag{39}$$

Using (7),

$$U_0 \Sigma_0 = \begin{bmatrix} I_K & -B_0 C_0^{-1} \end{bmatrix} \begin{bmatrix} A & B \\ B^{\mathrm{T}} & C \end{bmatrix} \tag{40}$$

$$= \begin{bmatrix} A_0 - B_0 C_0^{-1} B_0^{\mathrm{T}} & B_0 - B_0 C_0^{-1} C_0 \end{bmatrix} \tag{41}$$

$$= \begin{bmatrix} \Xi_0 & 0 \end{bmatrix}, \tag{42}$$

so that Σ_{f_0} calculates to

$$\Sigma_{f_0} = \Sigma_0 - \begin{bmatrix} \Xi_0 \left(\gamma^{-1} I_K + \Xi_0\right)^{-1} \Xi_0 & 0 \\ 0 & 0 \end{bmatrix} \tag{43}$$

$$= \Sigma_0 - \gamma \begin{bmatrix} \Xi_0 (I_K + \gamma \Xi_0)^{-1} \Xi_0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{44}$$

Since the Gaussian MMSE estimator only depends on the mean vector and the M right most columns of the covariance matrix, compare (9), it immediately follows that

$$E_{P^*_{f_0}}\left[X|Y=y\right] = E_{\mathcal{N}(\mu_{f_0}, \Sigma_{f_0})}\left[X|Y=y\right] \tag{45}$$

$$= E_{\mathcal{N}(\mu_0, \Sigma_0)}\left[X|Y=y\right] \tag{46}$$

$$= f_0(y), \tag{47}$$

which is the optimality condition in (36).

Using this result, it holds that

$$\sup_f E_{P_0}\left[e^{-\frac{\gamma}{2}\|X - f(Y)\|_2^2}\right] = E_{P_0}\left[e^{-\frac{\gamma}{2}\|(X-\mu_X) - BC^{-1}(Y-\mu_Y)\|_2^2}\right] \tag{48}$$

$$= E_{\mathcal{N}(0, I_K)}\left[e^{-\frac{\gamma}{2}\|\Xi_0^{1/2} Z\|^2}\right] \tag{49}$$

$$= E_{\mathcal{N}(0, I_K)}\left[e^{-\frac{\gamma}{2}\sum_{k=1}^K \xi_{0,k} Z_k^2}\right] \tag{50}$$

where $Z = [Z_1, \ldots, Z_K]$ is a vector of standard normally distributed random variables. The expression in (50) is the product of K moment generating functions of χ^2 distributed random variables evaluated at $-\frac{\gamma}{2}\xi_{0,k}$, hence, it evaluates to

$$\prod_{k=1}^K E_{\mathcal{N}(0,1)}\left[e^{-\frac{\gamma}{2}\xi_{0,k} Z_k^2}\right] = \prod_{k=1}^K (1 + \gamma \xi_{0,k})^{-\frac{1}{2}}. \tag{51}$$

Inserting this result back into (30) yields

$$\inf_f \inf_P L_\lambda(f, P) = \frac{1}{\gamma} \sum_{k=1}^K \log(1 + \gamma \xi_{0,k}) \tag{52}$$

for all $\gamma > 0$.

In order to establish the connection to the original problem (3), let $(f^\dagger, P^\dagger)$ denote the solution of the latter. For all $\gamma > 0$ it holds that

$$mmse_{X|Y}(P^\dagger) \geq mse_{X|Y}(f^\dagger, P^\dagger) + \frac{2}{\gamma}\left(D_{\mathrm{KL}}\left(P^\dagger \| P_0\right) - \varepsilon\right) \tag{53}$$

$$\geq \inf_{P,f}\left(mse_{X|Y}(f, P) + \frac{2}{\gamma} D_{\mathrm{KL}}\left(P \| P_0\right)\right) - \frac{2}{\gamma}\varepsilon \tag{54}$$

$$\geq \frac{1}{\gamma}\left(\sum_{k=1}^K \log(1 + \gamma \xi_{0,k}) - 2\varepsilon\right) =: \rho(\gamma). \tag{55}$$

In order to maximize this bound with respect to γ, note that

$$\rho'(\gamma) = \frac{1}{\gamma}\left(\sum_{k=1}^K \frac{\xi_{0,k}}{1 + \gamma \xi_{0,k}} - \rho(\gamma)\right) \tag{56}$$

$$=: \frac{1}{\gamma}\left(\tilde{\rho}(\gamma) - \rho(\gamma)\right), \tag{57}$$

where ρ' denotes the derivative of ρ and $\tilde{\rho}(\gamma)$ is defined implicitly. Since ρ is concave by construction, every stationary point is a global maximum, which yields the optimality condition

$$\rho(\gamma) - \tilde{\rho}(\gamma) = 0 \tag{58}$$

$$\sum_{k=1}^K \left(\log(1 + \gamma \xi_{0,k}) - \frac{\gamma \xi_{0,k}}{1 + \gamma \xi_{0,k}}\right) = 2\varepsilon \tag{59}$$

$$\sum_{k=1}^K \phi(\gamma\, \xi_{0,k}) = 2\varepsilon. \tag{60}$$

Since $\phi : [0, \infty) \to [0, \infty)$ is continuous and increasing, the left-hand side of (60) is continuous and increasing in γ, so that γ^* is unique. Finally, by definition of γ^*,

$$\rho(\gamma^*) = \tilde{\rho}(\gamma^*) = \sum_{k=1}^K \frac{\xi_{0,k}}{1 + \gamma^* \xi_{0,k}}. \tag{61}$$

Since the estimator/distribution pair in Theorem 1 attains this bound, it is tight. This completes the proof.

5. Discussion

In this section, some notewothy properties and special cases of the proposed bound are discussed.

5.1. Connection to the Cramér–Rao bound

As explained in the introduction, the bound in Theorem 1 can be interpreted as a Bayesian Cramér–Rao bound that is based on the KL divergence instead of the Fisher information. More formally, the CRB can be obtained by solving (3) with $\mathcal{P}_\varepsilon$ redefined as

$$\mathcal{P}_\varepsilon := \left\{P : \mathcal{I}(P) \leq \varepsilon\right\}, \tag{62}$$

where $\mathcal{I}(P)$ denotes the Bayesian Fisher information; see [13] for more details.

In view of this interpretation, the question which version of the CRB is more suitable for a given estimation problem translates to

the question which property, the Fisher information or the KL divergence from P_0, leads to a tighter constraint in (3). Both properties are measures for the deviation from Gaussianity, but they quantify the deviation differently. Since the requirement of having a finite KL divergence to a Gaussian reference distribution requires no additional regularity conditions, the KL divergence based bound is defined for a larger class of distributions than the conventional Bayesian CRB. As a rule of thumb, the less smooth a distribution function, the more likely it is that the KL divergence based CRB outperforms the conventional CRB. This aspect will be illustrated with an example in Section 6.

5.2. Additive noise channels

The bounds presented in [14] hold for additive noise channels in which noise and input are independent and at least one of them is Gaussian distributed. With the bound in Theorem 1 at hand, these assumptions can be relaxed. Moreover, if the additive noise channel is approximated by an additive Gaussian noise (AGN) channel, the KL divergence of the joint input-output distributions simplifies to the sum of the KL divergences of the input and the noise distributions.

In general, the KL divergence between two distributions $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X Q_{Y|X}$ can be decomposed into

$$D_{\mathrm{KL}}\big(P_{XY}\|Q_{XY}\big) = D_{\mathrm{KL}}\big(P_X\|Q_X\big) + E_{P_X}\big[D_{\mathrm{KL}}\big(P_{Y|X}\|Q_{Y|X}\big)\big]. \tag{63}$$

Now, consider an additive channel

$$Y = X + N, \tag{64}$$

where $X \sim P_X$ and $N \sim P_N$ are independent. In this case it holds that

$$\begin{aligned} E_{P_X}\big[D_{\mathrm{KL}}\big(P_{Y|X}\|Q_{Y|X}\big)\big] &= E_{P_X}\big[D_{\mathrm{KL}}\big(P_{Y-X}\|Q_{Y-X}\big)\big] \\ &= E_{P_N}\big[D_{\mathrm{KL}}\big(P_N\|Q_N\big)\big] \qquad (65)\\ &= D_{\mathrm{KL}}\big(P_N\|Q_N\big), \qquad (66)\end{aligned}$$

so that

$$D_{\mathrm{KL}}\big(P_{XY}\|Q_{XY}\big) = D_{\mathrm{KL}}\big(P_X\|Q_X\big) + D_{\mathrm{KL}}\big(P_N\|Q_N\big). \tag{67}$$

Moreover, if the reference distribution P_0 is chosen such that it corresponds to an additive Gaussian noise channel

$$Y_0 = X_0 + N_0, \tag{68}$$

where $X_0 \sim \mathcal{N}(\mu_{X_0}, \Sigma_{X_0})$ and $N_0 \sim \mathcal{N}(\mu_{N_0}, \Sigma_{N_0})$, then the MMSE matrix Ξ_0 in (11) simplifies to

$$\Xi_0 = \Sigma_{X_0}(\Sigma_{X_0} + \Sigma_{N_0})^{-1}\Sigma_{N_0}. \tag{69}$$

Using these results, MMSE bounds for additive noise channels can be obtained by adding the non-Gaussianity parameters of the input and noise distributions instead of considering their joint non-Gaussianity. This is a natural extension of the bounds in [14] and will be illustrated with an example in the next section. However, this simplicity comes at the cost of less tight bounds, since only allowing reference distributions that correspond to AGN channels reduces the degrees of freedom.

5.3. A special case with explicit bounds

For the special case that the covariance matrix of the reference distribution, Σ_0, is chosen such that its Schur complement admits a flat spectrum, that is, if $\xi_{0,1} = \ldots = \xi_{0,K} = \xi_0$, the solution of (13) can be expressed explicitly, namely

$$\gamma^*\xi_0 = \frac{1-\omega_0(\varepsilon/K)}{\omega_0(\varepsilon/K)} \tag{70}$$

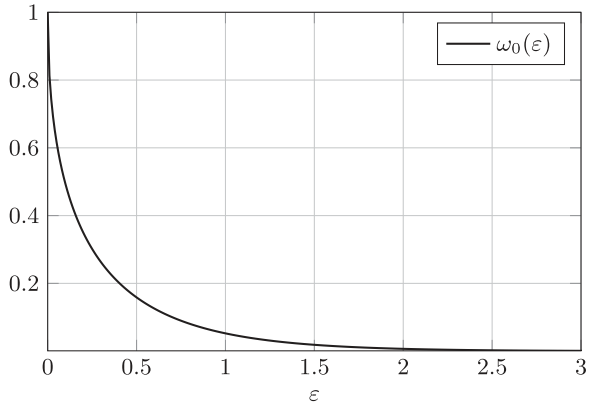

Fig. 1. Graph of ω_0 defined in (71).

where

$$\omega_0(t) = -W_0\big(-e^{-(2t+1)}\big) \tag{71}$$

and W_0 denotes the main branch of the Lambert W function [21]. Inserting (71) back into (12) yields bounds of the simple form

$$\frac{mmse_{X|Y}(P)}{K} \geq \omega_0(\varepsilon/K)\xi_0 \tag{72}$$

For illustration purposes, the function ω_0 is plotted in Fig. 1. Moreover, (72) indicates that the proposed bound is asymptotically exact for $K \to \infty$, whenever the KL divergence of P from P_0 grows sublinearly in K, that is $D_{\mathrm{KL}}\big(P\|P_0\big) \in o(K)$. This is in line with the results in [14], where this behavior was demonstrated for uniform input distributions on balls in $\mathbb{R}^K$.

5.4. Bounds on γ^*

In order to solve (13) for γ, it is useful to be able to bound γ^* from above and below, so that the problem can be reduced to finding the root of a monotonic function on a finite interval. The following corollary provides such bounds

Corollary 1. *For γ^* as in Theorem 1 it holds that*

$$\frac{1-\omega_0(\varepsilon/K)}{\omega_0(\varepsilon/K)} \leq \gamma^*\xi_{0,1} \leq \frac{1-\omega_0(\varepsilon)}{\omega_0(\varepsilon)}, \tag{73}$$

with ω_0 defined in (71).

The corollary follows from the monotonicity of ϕ and the bounds

$$\phi(\gamma\,\xi_{0,1}) \leq \sum_{k=1}^{K}\phi(\gamma\,\xi_{0,k}) \leq K\phi(\gamma\,\xi_{0,1}). \tag{74}$$

Bounding the sum on the left-hand side of (60) via (74) and solving for γ yields the desired result.

5.5. Connection to the AWGN channel

A possibly helpful interpretation of the bound in Theorem 1 is via the AWGN channel:

$$V = \sqrt{\gamma}X + N, \tag{75}$$

where $N \sim \mathcal{N}(0, I_K)$, $X \sim \mathcal{N}(\mu_X, \Xi_0)$, and γ denotes the signal-to-noise ratio (SNR). The MMSE of the channel in (75) is given by

$$mmse_{X|V}(\gamma) = \sum_{k=1}^{K}\frac{\xi_{0,k}}{1+\gamma\xi_{0,k}}, \tag{76}$$

so that the bounds in (12) can be written as

$$mmse_{X|Y}(P) \geq mmse_{X|U}(\gamma^*). \tag{77}$$

That is, the parameter γ^* can be interpreted as the SNR of an equivalent AWGN channel.

Interestingly, it can be shown that the minimax robust MMSE estimator studied in [15–17] leads to an MMSE upper bound of the form

$$mmse_{X|Y}(P) \leq mmse_{X|U}(\tilde{\gamma}), \tag{78}$$

where $\tilde{\gamma}$ is the unique *negative* solution of (13). However, this negative SNR clearly does not admit a physical interpretation comparable to that of γ^*.

6. Examples

In this section, three examples are given that illustrate how the proposed bound

A. can be applied to additive channels when both the input signal and the noise are non-Gaussian distributed.
B. can be used to obtain (asymptotically tight) MMSE bounds in a case where the Bayesian Cramér–Rao bound does not apply.
C. can be used to arrive at useful asymptotic expressions, with a particular emphasize on the fact that it is guaranteed to outperform the Baysian Cramér–Rao lower bound at low SNRs.

Python code for all examples can be found in a public Git repository [22].

In all but the first example the result in Theorem 1 is not used directly to bound the MMSE over a KL divergence ball, but to bound the MMSE for a given distribution P_{XY}. Such bounds can be obtained as follows: For any given P_0, it trivially holds that P_{XY} is within a KL divergence ball with radius $D_{\mathrm{KL}}\big(P_{XY}\|P_0\big)$ centered at P_0. Hence, $mmse_{X|Y}(P_{XY})$ is bounded by (12) with $\varepsilon = D_{\mathrm{KL}}\big(P_{XY}\|P_0\big)$. Since this is true for all Gaussian reference distributions, the lower bound can further be maximized with respect to P_0, that is,

$$mmse_{X|Y}(P_{XY}) \geq \sup_{P_0} \inf_{P} \Big\{ mmse_{X|Y}(P) : P \in \mathcal{P}_{D_{\mathrm{KL}}(P_{XY}\|P_0)}(P_0) \Big\}, \tag{79}$$

where P_0 is restricted to be Gaussian. We refer to a reference distribution that solves (79) as *best Gaussian approximation* of P_{XY}.

6.1. Generalized-Gaussian signal in additive generalized-Gaussian noise

There are many applications in signal processing and communications for which the noise can be assumed to be additive, but it cannot be assumed to be normally distributed [23,24]. Given this prominent role of additive channels, it is useful to illustrate how to apply the bound in Theorem 1 to this particular model.

Using the channel model in (64), let $X \sim \mathcal{G}(a,p)$ and $N \sim \mathcal{G}(b,q)$, where $\mathcal{G}(a,p)$ denotes a generalized Gaussian (GG) distribution with density function

$$g(x|a,p) = \frac{p}{2a\Gamma(1/p)} e^{-\left(\frac{|x|}{a}\right)^p}, \tag{80}$$

where Γ denotes the gamma function [25], $a>0$ is a scale parameter, and $p>0$ determines the type of decay of the tails [26]. In [14], it is shown that the best Gaussian approximation of a zero-mean GG distribution, in terms of the KL divergence, is attained by choosing the variance of the reference distribution as

$$\sigma_0^2 = a\sqrt{\frac{\Gamma(3/p)}{\Gamma(1/p)}}, \tag{81}$$

so that

$$d_{\mathcal{G}}(p) := \min_{\sigma_0^2 \geq 0} D_{\mathrm{KL}}\Big(\mathcal{G}(a,p)\|\mathcal{N}(0,\sigma_0^2)\Big) \tag{82}$$

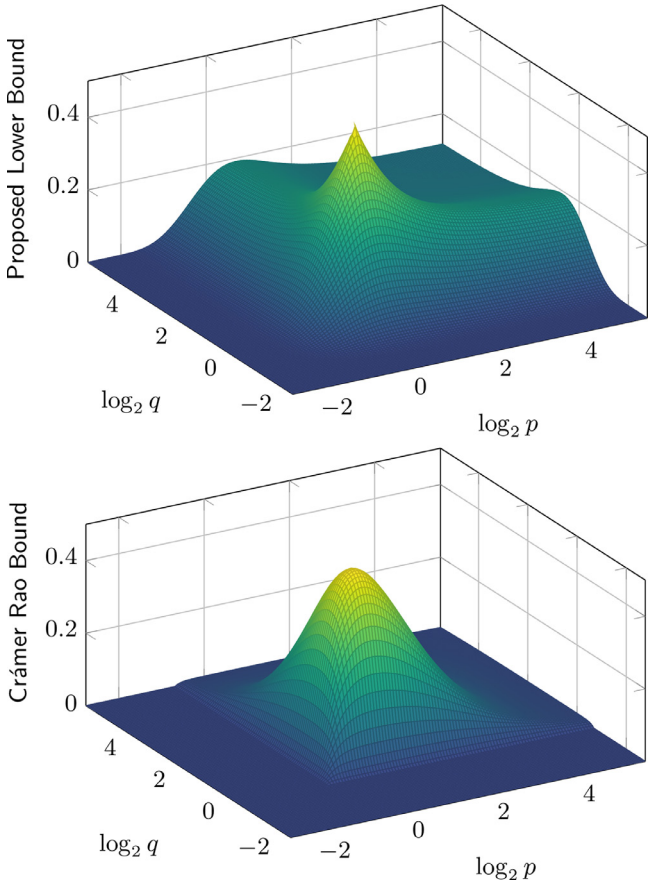

Fig. 2. Proposed MMSE lower bound (top) and Cramér–Rao bound (bottom) for an additive channel with generalized Gaussian noise and input.

$$= \log \frac{p}{\sqrt{2}} \sqrt{\frac{\Gamma(3/p)}{\Gamma(1/p)}} \frac{\Gamma(1/2)}{\Gamma(1/p)} + \frac{1}{2} - \frac{1}{p}. \tag{83}$$

See Fig. 4 in [14] for a plot of the graph of $d_{\mathcal{G}}$. From (67) it follows that the KL divergence of the true input-output distribution[4] and its best (additive) Gaussian approximation is given by

$$d_{XY}(p,q) := d_{\mathcal{G}}(p) + d_{\mathcal{G}}(q). \tag{84}$$

Combining (84), (69), and (72) yields the bound

$$mmse_{Y|X}(P_{\mathcal{G}\mathcal{G}}) \geq \omega_0\big(d_{X,Y}(p,q)\big) mmse_{Y|X}(P_0) \tag{85}$$

$$= \omega_0\big(d_{X,Y}(p,q)\big) \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}, \tag{86}$$

where σ_X^2 and σ_N^2 denote the signal and noise power, respectively.

Examples of the lower bound in (86) are shown in the upper plot of Fig. 2 for $p,q \in [2^{-2}, 2^5]$ and at an SNR of 0dB ($\sigma_X^2 = \sigma_N^2 = 1$). For comparison, the Crámer–Rao bound is depicted in the lower plot. The latter can be shown to be given by

$$mmse_{Y|X}(P_{\mathcal{G}\mathcal{G}}) \geq \frac{1}{I\big(\mathcal{G}(a,p)\big) + I\big(\mathcal{G}(b,q)\big)}, \tag{87}$$

where

$$I\big(\mathcal{G}(a,p)\big) = \begin{cases} \frac{p^2}{a^2}\frac{\Gamma(2-1/p)}{\Gamma(1/p)}, & 1/2 < p < \infty \\ \infty, & 0 < p \leq 1/2 \end{cases} \tag{88}$$

[4] Note that unless both the input and the noise are Gaussian distributed, the joint input-output distribution is not a multivariate GG distribution itself.

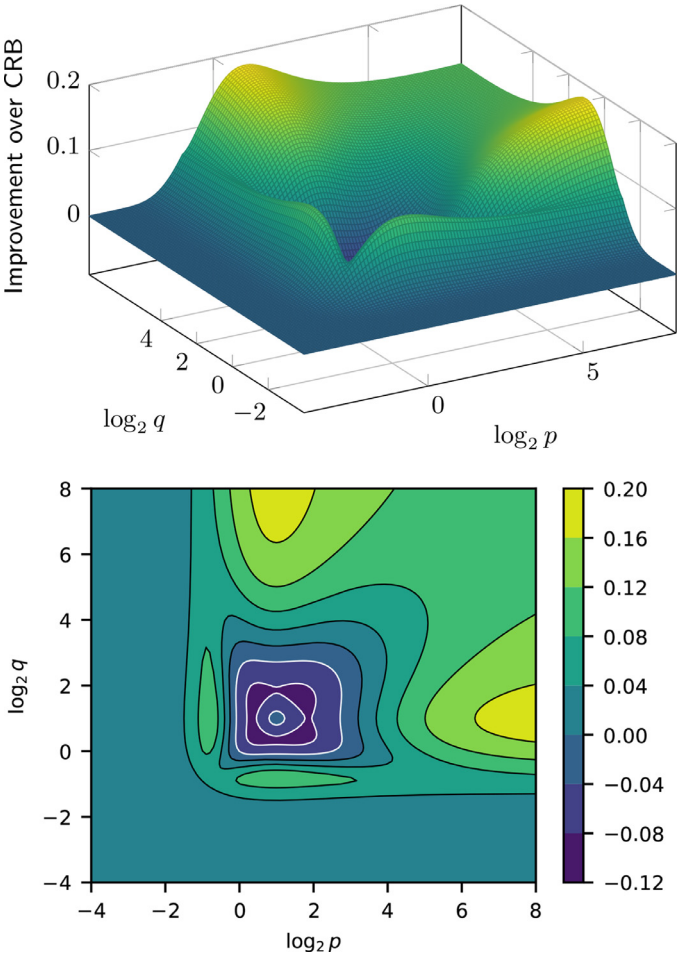

Fig. 3. Difference between the proposed MMSE lower bound and the Cramér–Rao bound for an additive channel with generalized Gaussian noise and input.

denotes the Fisher information of the zero-mean generalized Gaussian distribution [27, Chapter 3.2.1].

By inspection, the KL divergence based CRB is an improvement over the Fisher information based CRB for a variety of combinations of p and q. In particular, the proposed bound is significantly tighter as long as one of the distributions is close to Gaussian ($p, q \approx 2$), while the other distribution is more concentrated ($q, p > 2$). In contrast, it can be seen that the conventional CRB only performs well if both distributions are approximately Gaussian, with a pronounced peak around $p = q = 2$.

This improvement becomes more obvious when considering the difference between the two bounds, which is plotted in Fig. 3. Again, the proposed bound is notably tighter, with the exception of a small region around the Gaussian case. Since this region is difficult to recognize in the surface plot, it is shown separately in the plot below, where it is indicated by the white contour lines. Finally, for comparison, note that for all $p, q > 0$ in this example the MMSE is upper bounded by 0.5, which is the MSE obtained by the best linear estimator.

6.2. Multiplicative channel with uniform input distribution

The example in the previous section allowed for a comparison between the KL divergence based CRB and the conventional CRB. In this example, we would like to highlight a scenario in which the latter cannot be applied at all, namely that of uniform input distributions on K-dimensional balls, K-balls for short. A K-ball is defined as

$$\mathcal{B}_K(c, r) = \left\{ x \in \mathbb{R}^K : \sum_{k=1}^{K} \frac{(x_k - c_k)^2}{r^2} \leq 1 \right\}, \tag{89}$$

where $r > 0$ denotes the radius of the K-ball and $c \in \mathbb{R}^K$ denotes its center. Uniform distributions over balls are used, for example, in communication and signal processing to model parameter tolerances or estimation errors [28].

Now, consider the multiplicative channel

$$Y = X \cdot N, \tag{90}$$

where $\cdot$ denotes the elementwise product, $N \in \mathbb{R}^K$ is standard normally distributed, $N \sim \mathcal{N}(0, I_K)$, and X is uniformly distributed on $\mathcal{B}_K(c, r)$, here denoted by $X \sim \mathcal{U}_{\mathcal{B}_K}(c, r)$. For simplicity, it is assumed that $\mathcal{B}_K(c, r) \subset \mathbb{R}^K_+$, that is, $c_k > r$ for all $k = 1, \ldots, K$. In light of the discussion above, this scenario corresponds to measuring a parameter X with known tolerance r via a multiplicative channel with random channel gain N.

The joint distribution of X and Y in (90) can be approximated by jointly Gaussian random variables X_0 and Y_0 as follows. First, it is shown in [14] that the best Gaussian approximation for $\mathcal{B}_K(c, r) \subset \mathbb{R}^K_+$ is obtained by moment matching, that is,

$$\mu_{X_0} = c \quad \text{and} \quad \Sigma_{X_0} = I_k \frac{r^2}{K+2}. \tag{91}$$

The KL divergence between $P_X = \mathcal{U}_{\mathcal{B}_K}(c, r)$ and $P_{X_0} = \mathcal{N}(\mu_{X_0}, \Sigma_{X_0})$ can be shown to be given by

$$D_{\mathrm{KL}}\big(P_X \| P_{X_0}\big) = \frac{K}{2} - \frac{K}{2} \log \frac{K+2}{2} + \log \Gamma \frac{K+2}{2} \tag{92}$$

$$=: d_{\mathcal{U}}(K). \tag{93}$$

Since in a jointly Gaussian channel the conditional variance of $Y|X$ is independent of X, the conditional distribution $P_{Y|X} = \mathcal{N}(0, I_K X^2)$ can only be approximated by a Gaussian distribution with fixed, diagonal covariance matrix, so that $P_{Y_0|X_0} = P_{Y_0} = \mathcal{N}(0, I_K \sigma^2_{Y_0})$. The corresponding conditional KL divergence is given by

$$D_{\mathrm{KL}}\big(P_{Y|X} \| P_{Y_0}\big) = \sum_{k=1}^{K} D_{\mathrm{KL}}\big(P_{Y_k|X_k} \| P_{Y_{0,k}}\big), \tag{94}$$

where

$$D_{\mathrm{KL}}\big(P_{Y_k|X_k} \| P_{Y_{0,k}}\big) = \frac{1}{2}\left(\frac{X_k^2}{\sigma^2_{Y_{0,k}}} - 1 - \log \frac{X_k^2}{\sigma^2_{Y_{0,k}}} \right).$$

In order to evaluate $D_{\mathrm{KL}}\big(P_{XY} \| P_{X_0 Y_0}\big)$ via (63), the expected value of $D_{\mathrm{KL}}\big(P_{Y|X} \| P_{Y_0}\big)$ with respect to P_X is required, which is given by

$$E_{P_X}\Big[D_{\mathrm{KL}}\big(P_{Y|X} \| P_{Y_0}\big)\Big] = \sum_{k=1}^{K} E_{P_{X_k}}\Big[D_{\mathrm{KL}}\big(P_{Y_k|X_k} \| P_{Y_{0,k}}\big)\Big] \tag{95}$$

$$= \frac{1}{2} \sum_{k=1}^{K} \left(\frac{E[X_k^2]}{\sigma^2_{Y_{0,k}}} - 1 - E\big[\log X_k^2\big] + \log \sigma^2_{Y_{0,k}} \right). \tag{96}$$

Minimizing with respect to $\sigma^2_{Y_0|X_0}$ yields the best Gaussian approximation $\sigma^2_{Y_{0,k}} = E[X_k^2]$, so that

$$E_{P_X}\Big[D_{\mathrm{KL}}\big(P_{Y|X} \| P_{Y_0}\big)\Big] = \frac{1}{2} \sum_{k=1}^{K} \Big(\log E\big[X_k^2\big] - E\big[\log X_k^2\big] \Big). \tag{97}$$

It is not difficult to show[5] that

$$p_{X_k}(x) = \frac{1}{\sqrt{\pi}\, r} \frac{\Gamma(\frac{K+2}{2})}{\Gamma(\frac{K+3}{2})} \beta_{1,\frac{K+1}{2}}\Big(\frac{(x-c_k)^2}{r^2}\Big), \tag{98}$$

where $\beta_{a,b}$ denotes the PDF of the beta distribution with parameters a and b. From (98) it follows that

$$E\big[X_k^2\big] = c_k^2 + \frac{r^2}{K+2} \tag{99}$$

and

$$E\big[\log X_k^2\big] = \log c_k^2 + \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{K+2}{2})}{\Gamma(\frac{K+1}{2})} H\Big(\frac{r}{c_k}, \frac{K+1}{2}\Big), \tag{100}$$

where $H : (0,1] \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$H(a,b) := \int_{-1}^{1} \log(1+ax)(1-x^2)^{b-1}\, \mathrm{d}x. \tag{101}$$

If evaluating the right hand side of (100) is too costly, the bound

$$E\big[\log X_k^2\big] > \frac{1}{2}\big(\log(c_k - r)^2 + \log(c_k + r)^2\big) \tag{102}$$

$$= \log(c_k^2 - r^2) \tag{103}$$

can be used instead, which is obtained by lower bounding the logarithmic function by an affine function on the interval $[c_k - r, c_k + r]$ and is a good approximation when $c_k \gg r$.

Given this Gaussian approximation and using the fact that Σ_0 in (91) admits a flat spectrum, the lower MMSE bound in Theorem 1 evaluates to

$$mmse_{X|Y}(P_{XY}) \geq K\omega_0\Big(\frac{\varepsilon_K(c,r)}{K}\Big)\sigma^2_{X_{0,1}} \tag{104}$$

$$= \omega_0\Big(\frac{\varepsilon_K(c,r)}{K}\Big)\frac{K}{K+2}r^2, \tag{105}$$

where

$$\varepsilon_K(c,r) = D_{\mathrm{KL}}\big(P_{XY} \| P_{X_0Y_0}\big) \tag{106}$$

$$= d_{\mathcal{U}}(K) + \frac{1}{2}\sum_{k=1}^{K} d_\beta(c_k, r, K), \tag{107}$$

with $d_{\mathcal{U}}$ defined in (93) and

$$d_\beta(c_k, r, K) = \log E\big[X_k^2\big] - E\big[\log X_k^2\big] \tag{108}$$

$$= c_k^2 + \frac{r^2}{K+2} - \log c_k^2 - \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{K+2}{2})}{\Gamma(\frac{K+1}{2})} H\Big(\frac{r}{c_k}, \frac{K+1}{2}\Big) \tag{109}$$

$$< c_k^2 + \frac{r^2}{K+2} - \log(c_k^2 - r^2). \tag{110}$$

An example of the bound in (105) is shown in Fig. 4. Here the center point is chosen to be $c_1 = \ldots = c_K = 10$, the radius of the K-ball is set to $r = 2$, and K varies between 1 and 100. Clearly, the lower bound becomes tighter for large K. In fact, it is not hard to show that

$$\frac{\varepsilon_K(c,r)}{K} \to 0 \quad \text{for} \quad K \to \infty, \tag{111}$$

meaning the lower bound is asymptotically tight and

$$\lim_{K\to\infty} mmse_{X|Y}(P_{XY}) = r^2. \tag{112}$$

[5] For a unit K-ball centered at the origin, the probability of the event $\{X_k \leq x\}$, $x \in [0,1]$, corresponds to the ratio of the volume of the spherical cap [29] of height x to the volume of the entire K-ball.

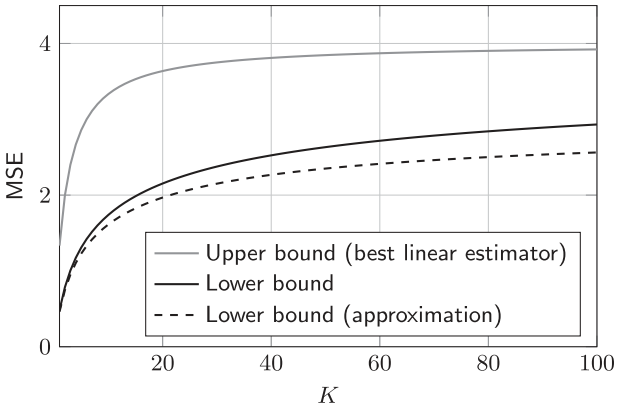

Fig. 4. MMSE bounds for the model in (90), where X is distributed uniformly on a K-ball with center point $c_1 = \ldots = c_K = 10$ and radius $r = 2$. The exact lower bound is given in (105), and the approximate version is obtained by using the inequality in (103) to bound ε_K in (107).

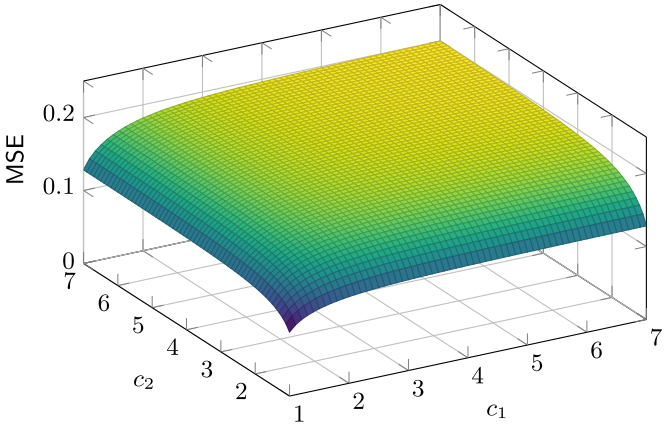

Fig. 5. MMSE lower bound in (105) with $K = 2$ and $r = 1$ for different center points $c = (c_1, c_2)$.

While this result could have been obtained in a more straightforward manner, it allows for some interesting insights. The limit in (112) implies that, asymptotically, the MMSE estimator for the model in (90) is a constant, namely $f^*(y) = \mu_X = c$. Interestingly, the aspect that the observations contain a vanishingly small amount of information is captured by the Gaussian approximation model, where X_0 and Y_0 are entirely independent. Nevertheless, the distribution of Y is of importance since it contributes to the distance between the approximated and the true joint distribution. Hence, the proposed bound capture the asymptotic independence of input and output, while using the Gaussian approximation to bound the impact of ignoring this dependence for finite K.

The influence of the center point c on the lower bound is illustrated in Fig. 5 for the case $K = 2$ and $r = 1$. The bound is lower towards the axis, where small values of X lead to small variances of Y, which in turn makes estimating X from Y easier. This effect becomes less and less pronounced as c moves away from the origin, thus increasing the variance of Y. Asymptotically, for $c \to \infty$, the lower bound again approaches the upper bound ($\sigma^2_{X_0} = 1/4$), meaning that Y becomes increasingly uninformative.

6.3. Low SNR behavior

Next, we show that the proposed bounds can be used to study the SNR behavior of the MMSE. More precisely, we show that the proposed lower bound performs better than the Cramér-Rao bound in the low SNR regime. This result is fixed in the next Lemma.

Lemma 4. *For additive channels, the KL divergence based CRB is tighter than the Fisher information based CRB in the low SNR regime.*

Proof. Consider the additive channel in (64), with $N \sim \mathcal{N}(0, \sigma_N^2 I)$ and choose $\Sigma_{N_0} = \sigma_N^2 I$ and $\Sigma_{X_0} = \sigma_0^2 I$. Using the lower bound in (72), we have that

$$\lim_{\sigma_N^2 \to \infty} mmse_{X|Y}(P_X) \geq \omega_0\left(\frac{\varepsilon}{K}\right)\sigma_0^2 K, \tag{113}$$

where σ_0^2 is arbitrary and

$$\varepsilon = D_{\mathrm{KL}}\left(P_X \| Q_{X_0}\right) = -h(X) + \frac{K}{2}\log(2\pi\sigma_0^2) + \frac{tr(\Sigma_X)}{2\sigma_0^2}. \tag{114}$$

Taking $\sigma_0^2 \to \infty$ on the right side of (113) leads to

$$\lim_{\sigma_N^2 \to \infty} mmse_{X|Y}(P_X) \geq \frac{1}{2\pi e} e^{\frac{2}{K}h(X)}. \tag{115}$$

The above procedure can now be compared to the Cramér-Rao bound, which leads to the following limit:

$$\lim_{\sigma_N^2 \to \infty} mmse_{X|Y}(P_X) \geq \lim_{\sigma_N^2 \to \infty} tr\left(\left(\frac{1}{\sigma_N^2} I + I(P_X)\right)^{-1}\right) \tag{116}$$

$$= tr\left(I^{-1}(P_X)\right). \tag{117}$$

Next, invoking Stam's inequality [30] we have that

$$\frac{1}{2\pi e} e^{\frac{2}{K}h(X)} \geq tr\left(I^{-1}(P_X)\right). \tag{118}$$

This completes the proof. □

7. Conclusion

This work has considered the problem of minimizing the mean square error when estimating a random vector $X \in \mathbb{R}^K$ from a random vector $Y \in \mathbb{R}^M$, subject to the constraint that their joint distribution P_{XY} lies in a KL divergence ball of radius ε centered at a Gaussian reference distribution. It has been shown that the minimum is attained by a jointly Gaussian distribution whose mean is identical to that of the reference distribution and whose covariance matrix can be determined by finding a scalar root of a simple function. This bound has been identified as a variant of the Bayesian Cramér–Rao bound, where the Fisher information is replaced by the Kullback–Leibler divergence.

Credit author statement

All authors contributed equally to the manuscript

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Michael Faus reports financial support was provided by the German Research Foundation. H. Vincent Poor reports financial support was provided by the U.S. National Science Foundation.

Data availability

No data was used for the research described in the article.

References

[1] E.L. Lehmann, G. Casella, Theory of Point Estimation, 2nd ed., Springer, New York City, New York, USA, 1998.

[2] Y. Dodge, in: The Concise Encyclopedia of Statistics, Springer, New York City, New York, USA, 2008, pp. 141–144, doi:10.1002/0471667196.ess1587.pub2.

[3] D. Guo, Y. Wu, S. Shamai (Shitz), S. Verdú, Estimation in Gaussian noise: properties of the minimum mean-square error, IEEE Trans. Inf. Theory 57 (4) (2011) 2371–2385, doi:10.1109/TIT.2011.2111010.

[4] A. Dytso, R. Bustin, D. Tuninetti, N. Devroye, H.V. Poor, S. Shamai (Shitz), On communication through a Gaussian channel with an MMSE disturbance constraint, IEEE Trans. Inf. Theory 64 (1) (2018) 513–530, doi:10.1109/TIT.2017.2747556.

[5] S.M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Upper Saddle River, NJ, USA, 1993.

[6] L.A. Dalton, E.R. Dougherty, Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error—part I: representation, IEEE Trans. Signal Process. 60 (5) (2012) 2575–2587, doi:10.1109/TSP.2012.2184101.

[7] L.A. Dalton, E.R. Dougherty, Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error—part II: consistency and performance analysis, IEEE Trans. Signal Process. 60 (5) (2012) 2588–2603, doi:10.1109/TSP.2012.2184102.

[8] D. Guo, S. Shamai (Shitz), S. Verdú, Mutual information and minimum mean-square error in Gaussian channels, IEEE Trans. Inf. Theory 51 (4) (2005) 1261–1282, doi:10.1109/TIT.2005.844072.

[9] S. Verdú, D. Guo, A simple proof of the entropy-power inequality, IEEE Trans. Inf. Theory 52 (5) (2006) 2165–2166, doi:10.1109/TIT.2006.872978.

[10] J. Ziv, M. Zakai, Some lower bounds on signal parameter estimation, IEEE Trans. Inf. Theory 15 (3) (1969) 386–391.

[11] J.T. Flam, S. Chatterjee, K. Kansanen, T. Ekman, On MMSE estimation: s linear model under Gaussian mixture statistics, IEEE Trans. Signal Process. 60 (7) (2012) 3840–3845.

[12] E. Weinstein, A.J. Weiss, A general class of lower bounds in parameter estimation, IEEE Trans. Inf. Theory 34 (2) (1988) 338–342.

[13] M. Fauß, A. Dytso, H.V. Poor, A variational interpretation of the Cramér–Rao bound, Signal Process. 182 (2021) 107917, doi:10.1016/j.sigpro.2020.107917.

[14] A. Dytso, M. Fauß, A.M. Zoubir, H.V. Poor, MMSE bounds for additive noise channels under Kullback-Leibler divergence constraints on the input distribution, IEEE Trans. Signal Process. 67 (24) (2019) 6352–6367.

[15] B.C. Levy, R. Nikoukhah, Robust least-squares estimation with a relative entropy constraint, IEEE Trans. Inf. Theory 50 (1) (2004) 89–104, doi:10.1109/TIT.2003.821992.

[16] Y. Guo, B.C. Levy, Robust MSE equalizer design for MIMO communication systems in the presence of model uncertainties, IEEE Trans. Signal Process. 54 (5) (2006) 1840–1852, doi:10.1109/TSP.2006.872322.

[17] M. Zorzi, On the robustness of the Bayes and Wiener estimators under model uncertainty, Automatica 83 (2017) 133–140, doi:10.1016/j.automatica.2017.06.005.

[18] M. Fauß, A. Dytso, H.V. Poor, An inequality for Bayesian Bregman risks with applications in directional estimation, in: Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2021, pp. 1–6, doi:10.1109/MFI52462.2021.9591193.

[19] V. Anantharam, A variational characterization of Rényi divergences, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT), 2017, pp. 893–897.

[20] D.G. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons, 1997.

[21] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, D.E. Knuth, On the Lambert W function, Adv. Comput. Math. 5 (1) (1996) 329–359, doi:10.1007/BF02124750.

[22] Git Repository, https://github.com/mifauss/KL-Divergence-MMSE-Bounds.

[23] D. Middleton, Non-Gaussian noise models in signal processing for telecommunications: new methods and results for class A and class B noise models, IEEE Trans. Inf. Theory 45 (4) (1999) 1129–1149, doi:10.1109/18.761256.

[24] M. Nassar, J. Lin, Y. Mortazavi, A. Dabak, I.H. Kim, B.L. Evans, Local utility power line communications in the 3–500 kHz band: channel impairments, noise, and standards, IEEE Signal Process. Mag. 29 (5) (2012) 116–127, doi:10.1109/MSP.2012.2187038.

[25] P.J. Davis, Leonhard Euler's integral: a historical profile of the Gamma function, Am. Math. Mon. 66 (10) (1959) 849–869.

[26] A. Dytso, R. Bustin, H.V. Poor, S. Shlomo (Shitz), On additive channels with generalized Gaussian noise, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT), 2017, pp. 426–430, doi:10.1109/ISIT.2017.8006563.

[27] S.A. Kassam, J.B. Thomas, Signal Detection in Non-Gaussian Noise, Springer Texts in Electrical Engineering, Springer, New York City, New York, USA, 2012.

[28] B.R. Barmish, C.M. Lagoa, The uniform distribution: rigorous justification for its use in robustness analysis, in: Proceedings of the of 35th IEEE Conference on Decision and Control (CDC), volume 3, 1996, pp. 3418–3423, doi:10.1109/CDC.1996.573689.

[29] S. Li, Concise formulas for the area and volume of a hyperspherical cap, Asian J. Math. Stat. 4 (1) (2011) 66–70.

[30] M. Raginsky, I. Sason, Concentration of measure inequalities in information theory, communications and coding, arXiv preprint arXiv:1212.4663 (2012).