Signal Processing 182 (2021) 107917

Contents lists available at ScienceDirect

Signal Processing

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

A variational interpretation of the Cramér–Rao bound

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

[a] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
[b] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA

ARTICLE INFO

Article history:
Received 16 June 2020
Revised 24 November 2020
Accepted 27 November 2020
Available online 8 December 2020

MSC:
62F10
62F15
65K10

Keywords:
Cramér–Rao bound
Fisher information
Variational techniques

ABSTRACT

It is shown that both the classic and the Bayesian Cramér–Rao bounds can be obtained by minimizing the mean square error of an estimator while constraining the underlying distribution to be within a Fisher information ball. The presented results allow for some nonstandard interpretations of the Cramér–Rao bound and, more importantly, provide a template for novel bounds on the accuracy of estimators.

1. Introduction

The Cramér–Rao bound (CRB) is one of the most well-known lower bounds on the mean square error (MSE) of parametric estimators and has countless applications in statistics and related areas. It comes in two varieties, namely, the classic CRB [1,2], which provides a lower bound on the variance of an unbiased estimator, and the Bayesian CRB [3] (also known as the posterior CRB or van Trees inequality), which provides a lower bound on the expected MSE of an arbitrary estimator under a given prior.

Proofs for either variety of the CRB can be found in standard textbooks. Most of these proofs are based on the Cauchy–Schwarz inequality (CSI), the covariance inequality, or the Hammersley–Chapman–Robbins bound (HCRB). In contrast, the proof presented here is based on a variational approach and does not make use of other inequalities or bounds. But why is this alternative proof of interest, in particular, in a signal processing context?

First, we would like to emphasize that the aim of this paper is *not* to re-derive all known properties of the CRB and the estimators that attain it. Also, we do *not* want to claim that the presented proof is particularly elegant or in any way "superior" to the standard proofs given in textbooks. On the contrary, for many purposes, a proof via the CSI is more concise and transparent.

What we do want to show with this paper is the following: First, both the classic and the Bayesian CRB can be interpreted as solutions of an optimization problem, where the MSE is minimized with respect to both an estimator and a distribution, and the latter is subject to a constraint on its Fisher information. Based on this result, the second goal is to highlight that by varying the objective function (MSE) and/or the penalty term (Fisher information) a new family of bounds of the Cramér–Rao type can be obtained. In other words, the variational proof makes it possible to identify a family of bounds that otherwise cannot easily be identified as a generalization of the CRB.

Let us elaborate some more on the latter aspect. Existing generalizations of the CRB are typically obtained by using generalized versions of the CSI, such as Hölder's inequality, or by choosing different functions to which these inequalities are applied. Prominent examples of this type of bounds are the Bayesian Bhattacharyya bound, and the Bobrovsky-Zakai bound [4]. The generalization suggested by the variational proof, namely to replace the MSE or the Fisher information with other functions, is conceptually very different, and provides a new perspective on Bayesian and non-Bayesian bounds. Two novel bounds based on this perspective have already been studied in [5] and [6], and have been shown to be tighter than the CRB in some scenarios. This improvement is achieved by replacing the Fisher information with the Kullback–Leibler divergence (relative entropy), which is less

* Corresponding author.
E-mail addresses: mfauss@princeton.edu (M. Fauß), alex.dytso@njit.edu (A. Dytso), poor@princeton.edu (H.V. Poor).

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

https://doi.org/10.1016/j.sigpro.2020.107917
0165-1684/

sensitive to strong fluctuations in the density function. A more formal discussion of the proposed generalization is given in Section 4.

Note that establishing bounds on the accuracy of estimators by solving (constrained) optimization problems is not a novel approach in itself. For example, minimizing the MSE under unbiasedness constraints on the estimator leads to the well-known Barankin bound. Different approximations of the latter give rise to, among others, the Bhattacharyya bound, the HCRB, the McAulay–Seidman bound, and the classic CRB; see [7] and the references therein for more details. For the Bayesian case, a similar optimization problem is studied [8], where instead of the bias, the squared estimation error is subject to a constraint. In [9] and [10], a unifying framework for both Bayesian and non-Bayesian bounds is proposed, in which many existing results can be obtained by studying different integral transformations of functions of the likelihood-ratio type.

The most fundamental difference between this paper and the ones cited above is that in the problem studied here the minimum is taken *jointly* over the estimator and the distribution, with a constraint on the latter. By contrast, in the vast majority of works, the distribution is fixed, while the minimization is performed over the estimator or over a free function that is introduced in order to increase the degrees of freedom. Including the distribution in the free variables makes it possible to construct bounds that are tailored for certain properties of these distributions, such as having a bounded Fisher information or a bounded relative entropy with respect to some reference distribution. Of course, this approach is not the only way of incorporating prior knowledge of distributional properties into a bound. However, as mentioned before, it provides a systematic template and an intuitive interpretation that we conjecture to be conducive to future works on lower bounds in signal processing and other areas.

The paper is organized as follows: Some preliminary results that are used in the proofs are introduced in Section 2. In Section 3, a special case of the CRB is proven, namely, the case where a real scalar parameter is estimated from a real scalar observation. This case is the simplest and most instructive, in the sense that a variational proof of the corresponding CRB can be given in a perspicuous manner, without obscuring the main ideas by technical details. The result is then briefly discussed in Section 4, where also the outline of a variational proof of the CRB for vector parameters is given, and a possible generalization to obtain families of CRB-like bounds is proposed.

A note on notation: Random variables are denoted by upper case letters X and their realizations by the corresponding lower case letters x. Analogously, probability distributions are denoted by upper case letters P and their densities by the corresponding lower case letters p. The normal distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu,\sigma^2)$. The expected value of a random variable X under distribution P is written as $\mathrm{E}_P[X]$. The first and second (partial) derivatives of a function f with respect to the argument x are written as $\partial_x f(x)$ and $\partial_x^2 f(x)$, respectively. The difference of two functions $f(x)$ and $g(x)$ is written as $f(x)-g(x)=\Delta_{fg}(x)$. Vectors and matrices are indicated by boldface font. For matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ the notation $\boldsymbol{X}\preceq\boldsymbol{Y}$ is used to indicate that $\boldsymbol{Y}-\boldsymbol{X}$ is positive semidefinite. As is customary, the real line is denoted by $\mathbb{R}$ and the corresponding Borel σ-algebra by $\mathcal{B}$. All integrals in the paper are taken over either $\mathbb{R}$ or $\mathbb{R}\times\mathbb{R}$, and the domains of integration are omitted in the notation when they are clear from the context.

2. Preliminaries

Let X be a random variable with values in $(\mathbb{R},\mathcal{B})$. For the classic CRB, X is assumed to be distributed according to a distribution P_θ where $\theta\in\mathbb{R}$ is a deterministic but unknown parameter. The family of distributions $\{P_\theta\}_{\theta\in\mathbb{R}}$ is denoted by $\mathcal{P}$. Moreover, it is assumed that all $P_\theta\in\mathcal{P}$ are supported on $\mathbb{R}$, that is, $p_\theta(x)>0$ for all $x\in\mathbb{R}$, and that $\partial_\theta p_\theta(x)$ is well-defined and finite for all $x\in\mathbb{R}$; compare (5.12) in [11]. For the Bayesian Cramér–Rao bound, (X,Θ) is assumed to be a pair of random variables with values in $(\mathbb{R},\mathcal{B})\times(\mathbb{R},\mathcal{B})$ and joint distribution P. The marginal and conditional distributions are denoted by P_Θ, P_X, $P_{X|\Theta}$, and $P_{\Theta|X}$. Note that in the Bayesian scenario P_Θ denotes the distribution of Θ, whereas in the non-Bayesian scenario P_θ denotes a particular distribution from the family $\mathcal{P}$. Finally, it is assumed that differentiation and integration can be interchanged, that is,

$$\partial_\theta\int p_\theta(x)\,\mathrm{d}x=\int\partial_\theta p_\theta(x)\,\mathrm{d}x=0 \tag{1}$$

and

$$\partial_\theta\int p(x,\theta)\,\mathrm{d}x=\int\partial_\theta p(x,\theta)\,\mathrm{d}x=0 \tag{2}$$

for all $\theta\in\mathbb{R}$. See, for example, [12, Theorem 23] for conditions under which (1) and (2) hold.

The classic Fisher information is defined as

$$\mathcal{I}(P_\theta)=\mathrm{E}_{P_\theta}\Big[\big(\partial_\theta\log p_\theta(X)\big)^2\Big] \tag{3}$$

$$=\mathrm{E}_{P_\theta}\left[\left(\frac{\partial_\theta p_\theta(X)}{p_\theta(X)}\right)^2\right]=\int_{\mathbb{R}}\frac{\big(\partial_\theta p_\theta(x)\big)^2}{p_\theta(x)}\,\mathrm{d}x. \tag{4}$$

The Bayesian Fisher information is analogously defined as

$$\mathcal{I}(P)=\mathrm{E}_P\Big[\big(\partial_\theta\log p(X,\Theta)\big)^2\Big] \tag{5}$$

$$=\mathrm{E}_P\left[\left(\frac{\partial_\theta p(X,\Theta)}{p(X,\Theta)}\right)^2\right] \tag{6}$$

$$=\int_{\mathbb{R}\times\mathbb{R}}\frac{\big(\partial_\theta p(x,\theta)\big)^2}{p(x,\theta)}\,\mathrm{d}x\,\mathrm{d}\theta. \tag{7}$$

Note that $\mathcal{I}(P)$ can equivalently be written as

$$\mathcal{I}(P)=\mathrm{E}_{P_X}\big[\mathcal{I}(P_{\Theta|X})\big] \tag{8}$$

or

$$\mathcal{I}(P)=\mathrm{E}_{P_\Theta}\big[\mathcal{I}(P_{X|\Theta})\big]+\mathcal{I}(P_\Theta). \tag{9}$$

Using the arguments in [13], it can be shown that both the classic and Bayesian Fisher information are convex functions of P_θ and P, respectively.

Any measurable function f from $\mathbb{R}$ to $\mathbb{R}$ defines an estimator for θ by letting $\hat\theta=f(x)$. The MSE of this estimator under the distribution P is denoted by

$$\mathrm{mse}(f,P)=E_P\big[(f(X)-\Theta)^2\big]. \tag{10}$$

Analogously, under P_θ, the MSE is defined as

$$\mathrm{mse}(f,P_\theta)=E_{P_\theta}\big[(f(X)-\theta)^2\big]. \tag{11}$$

If f is unbiased, that is, if it holds that

$$\mathrm{E}_{P_\theta}\big[f(X)\big]=\theta \tag{12}$$

for all $\theta\in\mathbb{R}$, the MSE of an estimator coincides with its variance. The set of pairs $(f,\mathcal{P})$ for which f is unbiased is denoted by

$$\mathcal{U}=\Big\{(f,\mathcal{P}):\mathrm{E}_{P_\theta}\big[f(X)\big]=\theta\quad\forall P_\theta\in\mathcal{P}\Big\}. \tag{13}$$

A well-known property of unbiased estimators is that, under the assumption that (1) holds,

$$\int(f(x)-\theta)\,\partial_\theta p_\theta(x)\,\mathrm{d}x=1, \tag{14}$$

which can be shown by taking the derivative with respect to θ of $\mathrm{E}_{P_\theta}[f(X)-\theta]$ and interchanging differentiation and integration.

For the proofs presented in the next sections, the Gâteaux derivatives [14] of the Fisher information and the MSE are required. For the classic Fisher information, the Gâteaux derivative in the direction of a distribution Q_θ can be shown to be

$$\partial_{Q_\theta}\mathcal{I}(P_\theta)=\lim_{\varepsilon\to 0}\frac{\mathcal{I}((1-\varepsilon)P_\theta+\varepsilon Q_\theta)-\mathcal{I}(P_\theta)}{\varepsilon} \tag{15}$$

$$=\lim_{\varepsilon\to 0}\frac{\mathcal{I}(P_\theta+\varepsilon\Delta_{Q_\theta P_\theta})-\mathcal{I}(P_\theta)}{\varepsilon} \tag{16}$$

$$=2\int\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\,\partial_\theta\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x \tag{17}$$

$$-\int\left(\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\right)^2\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x. \tag{18}$$

Analogously, for the Bayesian Fisher information the Gâteaux derivative in the direction of a joint distribution Q can be shown to be

$$\partial_Q\mathcal{I}(P)=\lim_{\varepsilon\to 0}\frac{\mathcal{I}(P+\varepsilon\Delta_{QP})-\mathcal{I}(P)}{\varepsilon} \tag{19}$$

$$=2\int\frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\,\partial_\theta\Delta_{qp}(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta \tag{20}$$

$$-\int\left(\frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\right)^2\Delta_{qp}(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta. \tag{21}$$

As shown in [15], integration by parts and the fact that

$$\lim_{|x|\to\infty}\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\Delta_{q_\theta p_\theta}(x)=\lim_{|x|\to\infty}\frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\Delta_{qp}(x)=0. \tag{22}$$

for all $\theta\in\mathbb{R}$ whenever $\partial_Q\mathcal{I}(P)$ and $\partial_{Q_\theta}\mathcal{I}(P_\theta)$ are finite can be used to "shift" the derivatives of $\Delta_{q_\theta p_\theta}$ and Δ_{qp} from the first terms of (18) and (21) to the second terms so that

$$\partial_{Q_\theta}\mathcal{I}(P_\theta)=-2\int\frac{\partial_\theta^2 p_\theta(x)}{p_\theta(x)}\,\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x \tag{23}$$

$$+\int\left(\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\right)^2\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x. \tag{24}$$

and

$$\partial_Q\mathcal{I}(P)=-2\int\frac{\partial_\theta^2 p(x,\theta)}{p(x,\theta)}\,\Delta_{qp}(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta \tag{25}$$

$$+\int\left(\frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\right)^2\Delta_{qp}(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta. \tag{26}$$

Both versions of the Gâteaux derivative of the Fisher information are used in what follows.

The Gâteaux derivative of the MSE in the direction Q_θ or Q is given by

$$\partial_{Q_\theta}\operatorname{mse}(f,P_\theta)=\lim_{\varepsilon\to 0}\frac{\operatorname{mse}(f,P_\theta+\varepsilon\Delta_{Q_\theta P_\theta})-\operatorname{mse}(f,P_\theta)}{\varepsilon} \tag{27}$$

$$=\int(f(x)-\theta)^2\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x, \tag{28}$$

and

$$\partial_Q\operatorname{mse}(f,P)=\lim_{\varepsilon\to 0}\frac{\operatorname{mse}(f,P+\varepsilon\Delta_{QP})-\operatorname{mse}(f,P)}{\varepsilon} \tag{29}$$

$$=\int(f(x)-\theta)^2\Delta_{qp}(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta, \tag{30}$$

respectively.

3. Main results

In this section, a special case of the classic and the Bayesian CRB is proven using variational arguments. The proofs are intentionally stated in close analogy to each other in order to highlight that both versions of the CRB follow from the same optimization problem with different definitions of Fisher information. The Bayesian CRB is presented first since its proof is conceptually simpler.

3.1. Proof of the Bayesian Cramér–Rao Bound

Consider the following optimization problem:

$$\inf_{f,P}\operatorname{mse}(f,P)\quad\text{s.t.}\quad\mathcal{I}(P)\le\gamma. \tag{31}$$

In words, determine the minimal MSE of any estimator for Θ under the constraint that the Bayesian Fisher information of the joint distribution of X and Θ is bounded by γ.

Now, consider the auxiliary problem

$$\inf_{f,P}L_\lambda(f,P), \tag{32}$$

where λ is a positive scalar and

$$L_\lambda(f,P):=\operatorname{mse}(f,P)+\lambda^2\mathcal{I}(P). \tag{33}$$

The minimization in (32) can equivalently be written as

$$\inf_f\left\{\inf_P L_\lambda(f,P)\right\}. \tag{34}$$

By (26) and (30), the Gâteaux derivative of L_λ in the direction Q is given by

$$\partial_Q L_\lambda(f,P)=\partial_Q\operatorname{mse}(f,P)+\lambda^2\partial_Q\mathcal{I}(P) \tag{35}$$

$$=\int r_\lambda(x,\theta)\,\Delta_{qp}(x,\theta)\,\mathrm{d}\theta\,\mathrm{d}x, \tag{36}$$

where

$$r_\lambda(x,\theta)=(f(x)-\theta)^2-2\lambda^2\frac{\partial_\theta^2 p(x,\theta)}{p(x,\theta)}+\lambda^2\left(\frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\right)^2. \tag{37}$$

Since the MSE is linear in P and the Fisher information is convex in P, L_λ is also convex in P. Hence, a necessary and sufficient condition for a distribution P^* to solve the inner minimization in (34) is that

$$\partial_Q L_\lambda(f,P^*)\ge 0\quad\forall Q. \tag{38}$$

Since Q can be chosen arbitrarily, the condition in (38) can only be satisfied if r_λ is constant over $\mathbb{R}\times\mathbb{R}$. This yields a functional characterization of P^*.

Proposition 1. *A necessary and sufficient condition for P^* to solve the inner minimization in (34) is that*

$$2\frac{\partial_\theta^2 p^*(x,\theta)}{p^*(x,\theta)}-\left(\frac{\partial_\theta p^*(x,\theta)}{p^*(x,\theta)}\right)^2-\left(\frac{f(x)-\theta}{\lambda}\right)^2=c \tag{39}$$

for some measurable function f and some $c\in\mathbb{R}$.

From Proposition 1 an expression for the minimum in (32) in terms of P^* and the free parameter c can be obtained.

Proposition 2. *For any $\lambda>0$ it holds that*

$$\inf_{f,P}L_\lambda(f,P)=\lambda^2\left(2\int\partial_\theta^2 p^*(X,\Theta)\,\mathrm{d}x\,\mathrm{d}\theta-c\right), \tag{40}$$

where P^* satisfies the condition in (39) and f needs to be chosen such that P^* exists, but is otherwise arbitrary.

Proof. Taking the expected value with respect to P^* on the left- and right-hand side of (39) immediately yields the statement in Proposition 2. Note that existence of P^* means that the estimator f is chosen such that there exists a density p^*_θ that solves (39). □

In order to eliminate the dependence on P^* and c, it suffices to find a distribution that satisfies the condition in (39). In the next proposition it is shown that any distribution with a Gaussian posterior does the trick.

Proposition 3. *For any $\lambda > 0$ and any distribution P^* such that*

$$P^*_{\Theta|X} = \mathcal{N}(f^*(X), \lambda), \tag{41}$$

where

$$f^*(X) = E_{P^*_{\Theta|X}}[\Theta], \tag{42}$$

is optimal in the sense of (39).

Proof. For P^* as in Proposition 3, the density p^* is of the form

$$p^*(x,\theta) = \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(f^*(x)-\theta)^2}{2\lambda}} p_X(x), \tag{43}$$

where, the right hand side is a valid density by definition of $f^*(X)$ in (42). Consequently, it holds that

$$\frac{\partial_\theta p^*(x,\theta)}{p^*(x,\theta)} = \frac{f^*(x)-\theta}{\lambda} \tag{44}$$

and

$$\frac{\partial^2_\theta p^*(x,\theta)}{p^*(x,\theta)} = \frac{1}{\lambda} + \left(\frac{f^*(x)-\theta}{\lambda}\right)^2. \tag{45}$$

Therefore,

$$2\frac{\partial^2_\theta p^*(x,\theta)}{p^*(x,\theta)} - \left(\frac{\partial_\theta p^*(x,\theta)}{p^*(x,\theta)}\right)^2 - \left(\frac{f^*(x)-\theta}{\lambda}\right)^2 = \frac{2}{\lambda}, \tag{46}$$

which is the optimality condition in (39) with $c = \frac{2}{\lambda}$. □

With Proposition 3 at hand, the result in Proposition 2 can be made explicit.

Corollary 1. *For any $\lambda > 0$ it holds that*

$$\inf_{f,P} L_\lambda(f,P) = 2\lambda. \tag{47}$$

Proof. For P^* as in Proposition 3, it holds that

$$\int \partial^2_\theta p^*(X,\Theta)\,\mathrm{d}x\,\mathrm{d}\theta = \frac{1}{\lambda} + \int\left[\int\left(\frac{f^*(x)-\theta}{\lambda}\right)^2 p^*_{\Theta|X}(\theta)\mathrm{d}\theta\right] p_X(x)\mathrm{d}x \tag{48}$$

$$= \frac{1}{\lambda} + \int \frac{1}{\lambda}\, p_X(x)\,\mathrm{d}x = \frac{2}{\lambda}. \tag{49}$$

With $c = \frac{2}{\lambda}$ (compare the proof of Proposition 3), it follows from Proposition 2 that

$$\inf_{f,P} L_\lambda(f,P) = \lambda^2\left(\frac{4}{\lambda} - \frac{2}{\lambda}\right) = 2\lambda, \tag{50}$$

which is the statement in the corollary. □

Having solved the unconstrained problem (32), the solution of the constrained problem (31) falls into place.

Corollary 2 (Bayesian Cramér–Rao Bound)**.** *For any $\gamma > 0$ it holds that*

$$\inf_{\{f,P:\, I(P)\le\gamma\}} \mathrm{mse}(f,P) = \frac{1}{\gamma}. \tag{51}$$

Proof. For P^* and f^* as in Proposition 3 it holds that

$$\mathrm{mse}(f,P^*) = \lambda \tag{52}$$

and

$$\mathcal{I}(P) = E_{P_X}\left[\mathcal{I}(P_{\Theta|X})\right] = \frac{1}{\lambda}. \tag{53}$$

Hence, the MSE is minimized by choosing λ to be the smallest feasible value, which is $\lambda = 1/\gamma$. This completes the proof. □

3.2. Proof of the classic Cramér–Rao bound

Consider the following optimization problem:

$$\inf_{(f,\mathcal{P})\in\mathcal{U}} \mathrm{mse}(f,P_\theta) \quad \text{s.t.} \quad \mathcal{I}(P_\theta) \le \gamma_\theta. \tag{54}$$

This problem formulation has to be read in the sense that the MSE and the Fisher information are both evaluated at a particular parameter value θ, whereas the constraint that $f(X)$ is unbiased applies to the whole family of distributions $\mathcal{P}$. In words, determine the minimal MSE of an unbiased estimator for θ under the constraint that the Fisher information of the distribution P_θ is bounded by γ_θ.

Now, consider the auxiliary problem

$$\inf_{(f,\mathcal{P})\in\mathcal{U}} L_{\lambda_\theta}(f,P_\theta), \tag{55}$$

where λ_θ is a positive scalar and

$$L_{\lambda_\theta}(f,P_\theta) := \mathrm{mse}(f,P_\theta) + \lambda^2_\theta \mathcal{I}(P_\theta). \tag{56}$$

Here, λ is allowed to depend on θ since both the MSE and the Fisher information in (56) depend on θ. That is, the weighting of both terms is allowed to vary depending on the parameter value at which they are evaluated. The minimization in (55) can equivalently be written as

$$\inf_f \left\{ \inf_{\{\mathcal{P}:\,(f,\mathcal{P})\in\mathcal{U}\}} L_{\lambda_\theta}(f,P_\theta) \right\}, \tag{57}$$

where the inner minimization is performed over all distributions P_θ under which $f(X)$ is an unbiased estimator for θ. By (18) and (28), the Gâteaux derivative of L_{λ_θ} in the direction Q_θ is given by

$$\partial_{Q_\theta} L_\lambda(f,P_\theta) = \partial_{Q_\theta}\mathrm{mse}(f,P_\theta) + \lambda^2 \partial_{Q_\theta}\mathcal{I}(P_\theta) \tag{58}$$

$$= 2\lambda^2 \int s_{\theta,\lambda}(x)\,\partial_\theta \Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x + \int r_{\theta,\lambda}(x)\,\Delta_{q_\theta p_\theta}(x)\,\mathrm{d}x, \tag{59}$$

where

$$s_{\theta,\lambda}(x) = \frac{\partial_\theta p_\theta(x)}{p_\theta(x)} \tag{60}$$

and

$$r_{\theta,\lambda} = (f(x)-\theta)^2 - \left(\lambda\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\right)^2. \tag{61}$$

Since the MSE is linear in P_θ and the Fisher information is convex in P_θ, L_λ is also convex in P_θ. Hence, a necessary and sufficient condition for a distribution P^*_θ to solve the inner minimization in (57) is that

$$\partial_{Q_\theta} L_\lambda(f,P^*_\theta) \ge 0 \quad \forall Q_\theta \in \{P_\theta : (f,P_\theta)\in\mathcal{U}\}. \tag{62}$$

In the Bayesian case discussed before, this condition implied that the integrands need to be constant. Here, however, it follows from (1), (14) and the assumption that $f(X)$ is unbiased that functions

of the form $a_\theta\big(f(x)-\theta\big)+b_\theta$, where $a_\theta, b_\theta \in \mathbb{R}$ are scalars that are allowed to depend on θ, are "quasi-constant" in the sense that

$$\int \Big[a_\theta\big(f(x)-\theta\big)+b_\theta\Big] q_\theta(x)\,\mathrm{d}x = b_\theta \tag{63}$$

and

$$\int \Big[a_\theta\big(f(x)-\theta\big)+b_\theta\Big] \partial_\theta q_\theta(x)\,\mathrm{d}x = a_\theta \tag{64}$$

for all Q_θ such that $(f,\mathcal{Q}) \in \mathcal{U}$, where $\mathcal{Q} = \{Q_\theta\}_{\theta\in\mathbb{R}}$. This yields a functional characterization of the distributions that solve the inner problem in (57).

Proposition 4. *A necessary and sufficient condition for P_θ^* to solve the inner minimization in (57) is that*

$$\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)} = \frac{f(x)-\theta}{\lambda_\theta} + b_\theta \tag{65}$$

for some $b_\theta \in \mathbb{R}$.

Proof. For the optimality condition in (62) to be satisfied, the function $s_{\theta,\lambda}$ and $r_{\theta,\lambda}$ both need to be "quasi-constant" in the above sense, that is, it needs to hold that

$$\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)} = a_\theta\big(f(x)-\theta\big)+b_\theta \tag{66}$$

$$\left(\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)}\right)^2 - \left(\frac{f(x)-\theta}{\lambda_\theta}\right)^2 = c_\theta\big(f(x)-\theta\big)+d_\theta \tag{67}$$

for some constants $a_\theta, b_\theta, c_\theta, d_\theta \in \mathbb{R}$. Substituting (66) into (67) and comparing coefficients yields that

$$a_\theta\lambda_\theta = 1, \qquad d_\theta = b_\theta^2, \qquad c_\theta = 2a_\theta b_\theta, \tag{68}$$

so that (66) and (67) become

$$\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)} = \frac{f(x)-\theta}{\lambda_\theta} + b_\theta \tag{69}$$

$$\left(\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)}\right)^2 = \left(\frac{f(x)-\theta}{\lambda_\theta} + b_\theta\right)^2, \tag{70}$$

where the first condition clearly implies the second. Hence, P_θ^* in Proposition 4 satisfies the optimality condition in (62) with equality, meaning that it is a stationary point of L_λ. From the convexity of L_λ it immediately follows that P_θ^* is a global minimum, which concludes the proof. □

From the functional characterization of P_θ^*, one can obtain a solution of (55) in terms of λ_θ and the free parameter b_θ.

Proposition 5. *For any $\lambda_\theta > 0$ it holds that*

$$\inf_{(f,\mathcal{P})\in\mathcal{U}} L_\lambda(f,P_\theta) = L_\lambda(f,P_\theta^*) = \lambda_\theta(2-b_\theta), \tag{71}$$

where P^* satisfies the condition in (65) and f needs to be chosen such that P_θ^* exists and $f(X)$ is unbiased, but is otherwise arbitrary.

Proof. For P_θ^* as in (65) it holds that

$$\mathrm{mse}(f,P_\theta^*) = \int (f(x)-\theta)^2 p_\theta^*(x)\,\mathrm{d}x \tag{72}$$

$$= \lambda_\theta \int \big(f(x)-\theta\big)\left(\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)} - b_\theta\right) p_\theta^*(x)\mathrm{d}x \tag{73}$$

$$= \lambda_\theta \int \big(f(x)-\theta\big)\,\partial_\theta p_\theta^*(x)\mathrm{d}x - \lambda_\theta \int b_\theta\, p_\theta^*(x)\,\mathrm{d}x \tag{74}$$

$$= \lambda_\theta(1-b_\theta) \tag{75}$$

and

$$\lambda^2\mathcal{I}(P_\theta^*) = \int \left(\lambda\frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)\,\mathrm{d}x \tag{76}$$

$$= \lambda_\theta \int \Big(\big(f(x)-\theta\big)+\lambda_\theta b_\theta\Big)\,\partial_\theta p_\theta^*(x)\,\mathrm{d}x \tag{77}$$

$$= \lambda_\theta. \tag{78}$$

The statement in Proposition 5 follows. □

In order to eliminate the dependence on the free parameter b_θ, it suffices to find any distribution that satisfies the condition in Proposition 4. Again, a Gaussian distribution is a natural candidate.

Proposition 6. *For any $\lambda > 0$, the distribution*

$$P_\theta^* = \mathcal{N}(\theta,\lambda_\theta), \tag{79}$$

in combination with the estimator

$$f^*(x) = x, \tag{80}$$

is optimal in the sense of (65).

Proof. Clearly, $f^*(X)$ is an unbiased estimator under P_θ^* so that $(f^*,\mathcal{P}^*) \in \mathcal{U}$. A straightforward calculation yields

$$\frac{\partial_\theta p_\theta^*(x)}{p_\theta^*(x)} = \lambda_\theta\,\partial_\theta\left(\frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{(x-\theta)^2}{2\lambda_\theta}}\right)\frac{1}{p_\theta^*(x)} \tag{81}$$

$$= \frac{x-\theta}{\lambda_\theta}\frac{1}{\sqrt{2\pi\lambda_\theta}}e^{-\frac{(x-\theta)^2}{2\lambda_\theta}}\frac{1}{p_\theta^*(x)} \tag{82}$$

$$= \frac{f^*(x)-\theta}{\lambda_\theta}, \tag{83}$$

which is the optimality condition in (65) for $b_\theta = 0$. □

With Proposition 6 at hand, the result in Proposition 5 can be made explicit.

Corollary 3. *For any $\lambda > 0$ it holds that*

$$\inf_{(f,P_\theta)\in\mathcal{U}} L_{\lambda_\theta}(f,P_\theta) = 2\lambda_\theta. \tag{84}$$

Proof. For P_θ^* as in Proposition 6 it holds that $b_\theta = 0$ (compare the proof of Proposition 6). The corollary then follows immediately from Proposition 5. □

Having solved the unconstrained problem (55), the solution of the constrained problem (54) falls into place.

Corollary 4 (Classic Cramér–Rao Bound)**.** *For any $\gamma_\theta > 0$ it holds that*

$$\inf_{\substack{(f,\mathcal{P})\in\mathcal{U}\\ I(P_\theta)\le\gamma_\theta}} \mathrm{mse}(f,P_\theta) = \frac{1}{\gamma_\theta}. \tag{85}$$

Proof. For f^* and P_θ^* as in Proposition 6 it holds that (compare the proof of Proposition 5)

$$\mathrm{mse}(f^*,P_\theta^*) = \lambda_\theta \tag{86}$$

and

$$\mathcal{I}(P_\theta^*) = \frac{1}{\lambda_\theta}. \tag{87}$$

Hence, the MSE is minimized by choosing λ_θ to be the smallest feasible value, which is $\lambda_\theta = 1/\gamma_\theta$. This completes the proof. □

4. Discussion

In this section, some noteworthy aspects of the presented proofs are discussed and a brief outline is given of how they can be generalized to recover the vector versions of the CRB. The section concludes with an outlook on how the variational proof could provide a template for the derivation of novel bounds on the accuracy of estimators.

4.1. Remarks on the variational proofs

In light of the variational proof, the classic CRB is obtained as follows: Given the Fisher information of the true channel $\mathcal{I}(P_\theta)$, one searches for the most favorable channel within a Fisher ball of radius $\mathcal{I}(P_\theta)$. The variance of the best unbiased estimator when X is generated by transmitting θ over this optimal channel then provides a lower bound on the variance of any unbiased estimator and any channel within the Fisher information ball. The set of optimal channels is characterized by the differential Eq. (65). Since the additive Gaussian noise (AGN) channel is optimal in this sense, an intuitive interpretation of the classic CRB is as follows: If all that is known about an estimation problem is the Fisher information of the channel P_θ and that an unbiased estimator is used, then the best scenario one can hope for is that X is generated from θ by adding Gaussian noise with variance $1/\mathcal{I}(P_\theta)$.

It is well-known that the AGN channel is not the only channel that attains the CRB. A well-known necessary and sufficient condition for P_θ to attain the CRB is that the score function is affine in the estimation error, that is,

$$\frac{\partial_\theta p_\theta(x)}{p_\theta(x)} = \alpha_\theta\big(f(x) - \theta\big). \tag{88}$$

This condition is usually obtained by identifying the cases in which the CSI holds with equality. It is also implied by the variational proof: First, (65) establishes that the score function is an affine function of the estimation error with offset b_θ. However, from the results in Proposition 5 and Corollary 1, it follows that for $b_\theta > 0$, the variance of $f(X)$ would be below the CRB, and for $b_\theta < 0$ it would exceed the CRB, while simultaneous satisfying a sufficient condition for it to hold with equality. Consequently, it needs to hold that $b_\theta = 0$, which recovers (88).

The Bayesian CRB can be interpreted in close analogy to the classic CRB. It corresponds to the MSE of the best estimator when X and Θ are generated by an optimal joint distribution P^*, where the set of optimal distributions is characterized by the differential Eq. (39). Since any distribution with a Gaussian posterior is optimal in this sense, an intuitive interpretation of the Bayesian CRB is as follows: If all that is known about a Bayesian estimation problem is the Fisher information of the joint distribution P, than the best scenario one can hope for is that $\Theta|X$ follows a Gaussian distribution. In this sense, the Bayesian CRB can be seen as a reversed version of the classic CRB. While the latter bounds the variance of an estimator by assuming the *channel*, $P_{X|\Theta=\theta}$, to be of a specific type, the former bounds the MSE of an estimator by assuming the *posterior*, $P_{\Theta|X=x}$, to be of a specific type.

4.2. Extension to vector parameters

The vector case is deliberately excluded in Section 3 since it makes the proofs considerably more technical, running the risk of obscuring the underlying basic idea. However, in order to emphasize some aspects of the discussion in the previous subsection, it is useful to sketch a brief outline of how the proof extends to the vector case.

Consider the classic CRB, where $\boldsymbol{\theta} \in \mathbb{R}^K$, $K > 1$, is a deterministic but unknown parameter, $\boldsymbol{X}$ is a random variable with distribution $P_{\boldsymbol{\theta}}$, and $f(\boldsymbol{X})$ is constrained to be unbiased. First, the function L_{λ_θ} in (56) needs to be redefined as

$$L_{\boldsymbol{\Lambda}_{\boldsymbol{\theta}}}(\boldsymbol{f}, P_{\boldsymbol{\theta}}) = \boldsymbol{w}^{\mathrm{T}}\big(\mathbf{MSE}(\boldsymbol{f}, P_{\boldsymbol{\theta}}) + \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathcal{I}(P_{\boldsymbol{\theta}})\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\big)\boldsymbol{w}, \tag{89}$$

where $\mathbf{MSE}$ and $\mathcal{I}$ denote the MSE (covariance) matrix and the Fisher information matrix, respectively, and $\boldsymbol{f}$ is a vector-valued measurable function. The matrix $\boldsymbol{\Lambda}_{\boldsymbol{\theta}} \in \mathbb{R}^{K\times K}$ is positive definite and the vector $\boldsymbol{w} \in \mathbb{R}^K$ is arbitrary. It can be shown that $L_{\boldsymbol{\Lambda}_{\boldsymbol{\theta}}}$ is convex in $P_{\boldsymbol{\theta}}$ and that its optimality condition is of the form (65), where the derivative with respect to $\boldsymbol{\theta}$ is replaced by the gradient, and the devision by λ_θ is replaced by a matrix multiplication with $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}$. Moreover, this optimality condition is independent of $\boldsymbol{w}$ and satisfied by a Gaussian distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$. In analogy to Proposition 5, this leads to

$$\inf_{\boldsymbol{f}, P_{\boldsymbol{\theta}}} L_{\boldsymbol{\Lambda}_{\boldsymbol{\theta}}}(\boldsymbol{f}, P_{\boldsymbol{\theta}}) = L_{\boldsymbol{\Lambda}_{\boldsymbol{\theta}}}(\boldsymbol{f}^*, P^*_{\boldsymbol{\theta}}) = 2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{w} > 0 \tag{90}$$

and

$$\mathbf{MSE}(\boldsymbol{f}^*, P^*_{\boldsymbol{\theta}}) = \boldsymbol{\Lambda} = \mathcal{I}(P^*_{\boldsymbol{\theta}})^{-1}. \tag{91}$$

From (90) and (91), it can easily be shown that

$$\mathbf{MSE}(\boldsymbol{f}, P_{\boldsymbol{\theta}}) \preceq \mathcal{I}(P_{\boldsymbol{\theta}})^{-1}, \tag{92}$$

for all $(\boldsymbol{f}, P_{\boldsymbol{\theta}}) \in \mathcal{U}$, which is the classic vector CRB.

The vector version of the Bayesian CRB can be derived along the same lines. The alternative form of the Gâteaux derivative of the Fisher information in (26), which is required for the proof, can be obtained by means of the divergence theorem [16].

An interesting aspect of the CRB can be highlighted at this point. From a variational point of view, the classic CRB is obtained by minimizing over all feasible conditional distributions (channels) P_θ, i.e., over all feasible *ways of observing* θ. Analogously, the Bayesian CRB is obtained by minimizing over all feasible *ways of generating and observing* θ. In both cases it is not necessary to assume that x and θ are defined on the same space. That is, the CRB also holds for channels that map θ to much higher- or lower-dimensional spaces. For example, consider the case where a scalar parameter θ is observed via N parallel channels with scalar outputs, that is, θ is estimated from the vector $(X_1, \dots, X_N)$. The classic CRB states that, irrespective of how many and what kind of channels are used, the resulting estimator cannot be better than one that is based on a single scalar AGN channel. The reason for this is that the Fisher information constraint that needs to holds for the single AGN channel also applies to the *combined* Fisher information of the N channels in this example. In other words, the increase in Fisher information when using multiple channels is guaranteed to be at least large enough to nullify the reduction in MSE.

The same arguments apply to the case when multiple independent and identically distributed (IID) observations are observed via the same channel, which is statistically equivalent to the case where N observations are taken via N identical parallel channels. For this case, as discussed above, the CRB can again be obtained by constraining the joint distribution of $(X_1, \dots, X_N)$, which reduces to a product distribution under the IID assumption.

While these properties of the CRB are well-known and do by no means require the variational apparatus used in the previous section, following the variational path arguably forces one to think about interpretations and implications of the CRB in a more explicit manner.

4.3. Extension to a general class of bounds

A natural generalization of the optimization problems in (54) and (31) is to consider problems of the form

$$\inf_{f,P} \mathrm{E}_P\big[J(f(X), \Theta)\big] \quad \text{s.t.} \quad \mathrm{E}_P\big[\mathcal{F}\{p(X, \Theta)\}\big] \le \gamma, \tag{93}$$

or

$$\inf_{f,\mathcal{P}} \mathrm{E}_{P_\theta}\left[J_\theta(f(X))\right] \quad \text{s.t.} \quad \mathrm{E}_{P_\theta}\left[\mathcal{F}\{p_\theta(X)\}\right] \leq \gamma_\theta, \tag{94}$$

where J is suitable convex cost function and $\mathcal{F}$ is a convex operator acting on the density p_θ or p. As demonstrated in the previous section, the two variants of the CRB are obtained from this general setting by choosing

$$J_\theta(s) = J(s,\theta) = (s-\theta)^2 \tag{95}$$

and

$$\mathcal{F}\{\bullet\} = \left(\partial_\theta \log(\bullet)\right)^2. \tag{96}$$

Clearly, different choices for J and $\mathcal{F}$ lead to different bounds. In [5,6], for example, MSE bounds are derived under the assumption that the input distribution or the joint input-output distribution are close to a Gaussian reference distribution in terms of the Kullback–Leibler divergence (relative entropy). This corresponds to choosing J as in (95) and

$$\mathcal{F}\{\bullet\} = \log \frac{\bullet}{p_0}, \tag{97}$$

with p_0 denoting the density of the reference distributions. In fact, the work presented in this paper was partly motivated by the observation that the bound in [5], although derived from what appears to be a different starting point, is reminiscent of the CRB in many aspects. First results towards Cramér-Rao type bounds on Bregman risks based on Eqn 93 will be presented in a forthcoming publication [17].

With this paper, we hope to spark a more in-depth investigation of problems of the form (94) and (93), which we believe are likely to provide novel bounds on estimation risks and to contribute to a more unified treatment of existing results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Michael Fauß: Investigation, Writing - original draft. **Alex Dytso:** Investigation, Writing - original draft. **H. Vincent Poor:** Supervision, Resources, Funding acquisition, Writing - review & editing.

Acknowledgement

The authors would like to thank the associate editor and the anonymous reviewers for their constructive feedback which greatly helped to improve the quality of the paper.

References

[1] H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, NJ, USA, 1946.
[2] C.R. Rao, Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. 37 (1945) 81–91.
[3] H.L. van Trees, Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory, John Wiley, Hoboken, NJ, USA, 2004.
[4] Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking, H.L. van Trees, K.L. Bell (Eds.), John Wiley, Hoboken, NJ, USA, 2007.
[5] 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.
[6] M. Fauß, A. Dysto, H.V. Poor, MMSE Bounds under Kullback–Leibler divergence constraints on the joint input-output distribution, 2020, Preprint available online:arXiv:2006.03722.
[7] E. Chaumette, J. Galy, A. Quinlan, P. Larzabal, A new Barankin bound approximation for the prediction of the threshold region performance of maximum likelihood estimators, IEEE Trans. Signal Process. 56 (11) (2008) 5319–5333.
[8] A. Renaux, P. Forster, P. Larzabal, C.D. Richmond, A. Nehorai, A fresh look at the Bayesian bounds of the Weiss–Weinstein family, IEEE Trans. Signal Process. 56 (11) (2008) 5334–5352.
[9] K. Todros, J. Tabrikian, General classes of performance lower bounds for parameter estimation—Part I: non-Bayesian bounds for unbiased estimators, IEEE Trans. Inf. Theory 56 (10) (2010) 5045–5063.
[10] K. Todros, J. Tabrikian, General classes of performance lower bounds for parameter estimation—Part II: Bayesian bounds, IEEE Trans. Inf. Theory 56 (10) (2010) 5064–5082.
[11] E.L. Lehmann, G. Casella, Theory of Point Estimation, second edition, Springer, New York City, NY, USA, 1998.
[12] R.T. Rockafellar, Conjugate Duality and Optimization, SIAM, Philadelphia, PA, USA, pp. 1–74.
[13] M. Cohen, The fisher information and convexity (Corresp.), IEEE Trans. Inf. Theory 14 (4) (1968) 591–592.
[14] J. Bell, Fréchet derivatives and Gâteaux derivatives, 2014, Available online: http://individual.utoronto.ca/jordanbell/notes/frechetderivatives.pdf.
[15] P.J. Huber, Fisher information and spline interpolation, Ann. Stat. 2 (5) (1974) 1029–1033.
[16] Divergence Theorem, Encyclopedia of Mathematics. Available online: https://encyclopediaofmath.org/index.php?title=Divergence_theorem.
[17] Fauß M., Dytso A., Poor H. V., An Inequality for Bayesian Bregman Risks, Under review as a conference paper at ICASSP 2021.