

# Lower-bounds on the Bayesian Risk in Estimation Procedures via $f$ -Divergences

Adrien Vandenbroucq, Amedeo Roberto Esposito, Michael Gastpar

School of Computer and Communication Sciences

EPFL, Lausanne, Switzerland

adrien.vandenbroucq@alumni.epfl.ch, {amedeo.esposito, michael.gastpar}@epfl.ch

**Abstract**—We consider the problem of parameter estimation in a Bayesian setting and propose a general lower-bound that includes part of the family of  $f$ -Divergences. The results are then applied to specific settings of interest and compared to other notable results in the literature. In particular, we show that the known bounds using Mutual Information can be improved by using, for example, Maximal Leakage, Hellinger divergence, or generalizations of the Hockey-Stick divergence.

**Index Terms**—Bayesian Risk, Parameter Estimation, Information Measures,  $f$ -Divergences, Mutual Information, Hockey-Stick Divergence

## I. INTRODUCTION

In this work<sup>1</sup> we consider the problem of parameter estimation in a Bayesian setting. The connection between said problem and information measures has been established multiple times over the years [2]–[4]. Here we further develop the perspective undertaken in [3] and in [5]. Similarly to [3] and [5] we will look at the problem through an information-theoretic lens and we will thus treat the parameter to be estimated as a message sent through a channel. The family of bounds one can derive in this framework generally give rise to two objects:

- a measure of information (Shannon’s Mutual Information was employed in [3], Sibson’s  $\alpha$ -Mutual Information in [5], Hockey-Stick Divergence in [4], etc.);
- a small-ball probability;

The main advantage of this is that both terms can be rendered independent of the specific choice of the estimator, which in turns renders these lower-bounds quite general. Our main focus will not be on asymptotic results but rather on finite sample lower-bounds. In particular, we will expand upon [5], utilizing the same approach but focusing on  $f$ -Divergences rather than on Sibson’s Mutual Information.

## II. BACKGROUND AND DEFINITIONS

**Definition 1.** Given a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , the Legendre-Fenchel transform of  $f$  is defined as

$$f^*(x^*) = \sup_{x \in \mathcal{X}} \langle x^*, x \rangle - f(x), \quad (1)$$

where  $\langle x^*, x \rangle$  denotes the natural pairing between a space  $\mathcal{X}$  and its topological dual  $\mathcal{X}^*$ , i.e.,  $\langle x^*, x \rangle = x^*(x)$ . Given

<sup>1</sup>A more detailed version of this work including all the proofs can be found in [1]. Moreover, the code for the experiments can be found at [https://github.com/Adirlou/f-divergences\\_lower\\_bounds](https://github.com/Adirlou/f-divergences_lower_bounds)

a function  $f$ ,  $f^*$  is guaranteed to be lower semi-continuous and convex. If  $f$  is convex and lower semi-continuous then  $f = f^{**}|_{\mathcal{X}}$  (the restriction of  $f^{**}$  on  $\mathcal{X}$  agrees with  $f$ ).

### A. $f$ -Divergences

A straightforward generalization of the KL-Divergence can be obtained by considering a generic convex function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , usually with the simple constraint that  $f(1) = 0$ .

**Definition 2.** Let  $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$  be two probability spaces. Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function such that  $f(1) = 0$ . Consider a measure  $\mu$  such that  $\mathcal{P} \ll \mu$  and  $\mathcal{Q} \ll \mu$  (i.e.,  $\mathcal{P}$  and  $\mathcal{Q}$  are absolutely continuous with respect to  $\mu$ ). Denoting with  $p, q$  the densities of the measures with respect to  $\mu$ , the  $f$ -Divergence of  $\mathcal{P}$  from  $\mathcal{Q}$  is defined as follows:

$$D_f(\mathcal{P} \parallel \mathcal{Q}) = \int q f\left(\frac{p}{q}\right) d\mu. \quad (2)$$

Note that  $f$ -divergences are independent from the choice of the dominating measure  $\mu$  [6]. When absolute continuity between  $\mathcal{P}, \mathcal{Q}$  holds, denoted with  $\mathcal{P} \ll \mathcal{Q}$  one retrieves the following [6]:

$$D_f(\mathcal{P} \parallel \mathcal{Q}) = \int f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q}. \quad (3)$$

This generalization includes the KL divergence (by simply setting  $f(t) = t \log(t)$ ), but it also includes:

- Total Variation distance, with  $f(t) = \frac{1}{2}|t - 1|$ ;
- Hellinger distance, with  $f(t) = (\sqrt{t} - 1)^2$ ;
- Pearson  $\chi^2$ -divergence, with  $f(t) = (t - 1)^2$ .

In particular, in this paper, we will be interested in two families of divergences. The first family, also known as Hellinger Divergences, is typically characterized by a parameter  $p > 0$ . More precisely, we are referring to the  $f$ -Divergences that stem from  $f_p(t) = \frac{t^p - 1}{p - 1}$  and that will be denoted as follows:

$$\mathcal{H}_p(\mathcal{P} \parallel \mathcal{Q}) = D_{f_p}(\mathcal{P} \parallel \mathcal{Q}). \quad (4)$$

The second family we consider is characterized by two parameters, namely  $\beta > 0$  and  $\gamma \geq \beta$ , and arise from the parametric family of functions  $f_{\beta, \gamma}(t) = \max\{0, \beta t - \gamma\}$ . We denote it as:

$$E_{\beta, \gamma}(\mathcal{P} \parallel \mathcal{Q}) = D_{f_{\beta, \gamma}}(\mathcal{P} \parallel \mathcal{Q}). \quad (5)$$

For the case  $\beta = 1$ , one retrieves the family of so-called  $E_\gamma$ -Divergences [7, Eq. (47)].

Much like  $f$ -Divergences, a generalization of Shannon's Mutual Information, denoted in the literature as  $f$ -Mutual Information, can be defined starting from  $f$ -Divergences as follows:

**Definition 3.** Let  $X$  and  $Y$  be two random variables jointly distributed according to  $\mathcal{P}_{XY}$  over a measurable space  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{XY})$ . Let  $(\mathcal{X}, \mathcal{F}_X, \mathcal{P}_X), (\mathcal{Y}, \mathcal{F}_Y, \mathcal{P}_Y)$  be the corresponding probability spaces induced by the marginals. Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function such that  $f(1) = 0$ . The  $f$ -Mutual Information between  $X$  and  $Y$  is defined as:

$$I_f(X, Y) = D_f(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y). \quad (6)$$

If  $f$  is strictly convex at 1 and satisfies  $f(1) = 0$ , then  $I_f(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent [6, Theorem 5]. Choosing  $f(t) = t \log t$ , one recovers the Mutual Information. With a slight abuse of notation, we will denote  $f$ -Mutual Informations with the same symbols used to characterize the corresponding divergences, e.g.,  $\mathcal{H}_p(X, Y) = \mathcal{H}_p(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$  will represent the  $f_p$ -Mutual Information, while  $E_{\beta, \gamma}(X, Y) = E_{\beta, \gamma}(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$  will represent the  $f_{\beta, \gamma}$ -Mutual Information.

### B. Problem Setting - the Bayesian framework

Let  $\mathcal{W}$  denote the parameter space and assume that we have access to a prior distribution over this space  $\mathcal{P}_W$ . Suppose then that we observe  $W$  through the family of distributions  $\mathcal{P} = \{\mathcal{P}_{X|W=w} : w \in \mathcal{W}\}$ . Given a function  $\phi : \mathcal{X} \rightarrow \mathcal{W}$  one can then estimate  $W$  from  $X \sim \mathcal{P}_{X|W}$  via  $\phi(X) = \hat{W}$ . Let us denote with  $\ell : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{R}^+$  a loss function, the Bayesian risk is defined as:

$$R = \inf_{\phi} \mathbb{E}[\ell(W, \phi(X))] = \inf_{\phi} \mathbb{E}[\ell(W, \hat{W})]. \quad (7)$$

Our purpose will be to lower-bound  $R$  using the tools described in the previous section. To this end, we will be using a simple Markov's inequality approach: i.e., for every estimator  $\phi$  and  $\rho \geq 0$ , one can do the following

$$\mathbb{E}[\ell(W, \hat{W})] \geq \rho \left( P_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho) \right). \quad (8)$$

With further manipulations we can actually relate  $\mathbb{P}(\ell(W, \hat{W}) \geq \rho)$  to the information-measures described before and some function  $\psi$  of  $P_W P_{\hat{W}}(\ell(W, \hat{W}) \geq \rho)$  (the measure of  $\{\ell(W, \hat{W}) \geq \rho\}$  under the product of the marginals  $P_W P_{\hat{W}}$ ). Let us denote  $P_W P_{\hat{W}}(\ell(W, \hat{W}) \leq \rho) = L_W(\hat{W}, \rho)$ . In some cases, this  $\psi$  will lead us to considering the so-called small-ball probability

$$L_W(\rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} L_W(\hat{w}, \rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} \mathbb{P}(\ell(W, \hat{w}) \leq \rho). \quad (9)$$

The purpose is to render both of these quantities independent of  $\phi$ , granting us the tools to provide general lower-bounds on the risk  $R$ .

### C. Related Works

A survey of early works in this area, mainly focusing on asymptotic settings, can be found in [8]. More recent but important advances are instead due to [2], [9]. Closely connected to this work is [3]. The approach is quite similar, with the main difference that we employ a family of bounds involving a variety of divergences while [3] relies solely on Mutual Information and the Kullback-Leibler Divergence. [5] focuses on Sibson's  $\alpha$ -Mutual Information, and [4] uses the  $E_\gamma$ -Divergence. A similar approach was also undertaken in [10]. The authors focused on the notion of  $f$ -informativity (cf. [11]) and leveraged the data processing inequality similarly to [12, Theorem 3]. In particular,  $f$ -informativities are more general than the  $f$ -Mutual Informations considered in this work (cf. Definition 3) and they can potentially lead to tighter results. The technique used to provide lower-bounds on the Bayesian risk for general non-negative losses (cf. [10, Section 4]) is, however, different. It is unclear whether the results provided in this work are equivalent (or weaker) with respect to those obtained in [10].

### III. THE LOWER BOUNDS

Let us start with our main result and then show how it is connected to the Bayesian Risk.

**Theorem 1.** Consider the Bayesian framework described in Sec. II-B. Let  $f : [0, +\infty) \rightarrow \mathbb{R}$  be an increasing convex function such that  $f(1) = 0$  and suppose that the generalized inverse, defined as  $f^{-1}(y) = \inf\{t \geq 0 : f(t) > y\}$ , exists. Then the following must hold for every  $\rho > 0$  and every estimator  $\hat{W}$ :

$$\mathbb{E}[\ell(W, \hat{W})] \geq \rho \left( 1 - L_W(\hat{W}, \rho) \cdot f^{-1} \left( \frac{I_f(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))f^*(0)}{L_W(\hat{W}, \rho)} \right) \right). \quad (10)$$

Moreover, if  $f^*(0) \leq 0$ , the bound simplifies to

$$\mathbb{E}[\ell(W, \hat{W})] \geq \rho \left( 1 - L_W(\hat{W}, \rho) \cdot f^{-1} \left( \frac{I_f(W, \hat{W})}{L_W(\hat{W}, \rho)} \right) \right). \quad (11)$$

*Proof.* To prove the statement we use [12, Theorem 3]. In our notation, it states that for every function  $f$  with the desired properties, we have

$$P_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho) \leq L_W(\hat{W}, \rho). \quad (12)$$

$$f^{-1} \left( \frac{I_f(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))f^*(0)}{L_W(\hat{W}, \rho)} \right). \quad (13)$$

In particular when  $f^*(0) \leq 0$ , the bound reduces to

$$P_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho) \leq L_W(\hat{W}, \rho) \cdot f^{-1} \left( \frac{I_f(W, \hat{W})}{L_W(\hat{W}, \rho)} \right). \quad (14)$$

Rewriting  $P_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho)$  as  $1 - P_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho)$  and combining this with Equations (8) and (13) concludes the proof.  $\square$

In order to provide a lower-bound on the Bayesian Risk, one needs to render the right-hand side of Equations (10) (or (11)) independent of  $\hat{W} = \phi(X)$  and, in order to do that, one needs to render independent of  $\hat{W}$ :

- 1) The information-measure, *e.g.*, through the data-processing inequality  $I_f(W, \hat{W}) \leq I_f(W, X)$ ;
- 2) The quantity  $L_W(\hat{W}, \rho)$ , that can be easily upper-bounded in the following way:  $L_W(\hat{W}, \rho) \leq \sup_{\hat{w}} L_W(\hat{w}, \rho) = L_W(\rho)$ .

For simplicity, consider Equation (11) and introduce the following object

$$G_f(I_f, L_W) := L_W(\hat{W}, \rho) \cdot f^{-1} \left( \frac{I_f(W, \hat{W})}{L_W(\hat{W}, \rho)} \right). \quad (15)$$

To use the two inequalities just stated above in items 1) and 2), one thus needs that for a given choice of  $f$ ,  $G_f(I_f, L_W)$  is increasing in  $I_f$  for a given value of  $L_W$  and vice-versa. This allows us to further lower-bound (11) and render the quantity independent of the specific choice of  $\phi$ . Hence, starting from (7) one can provide a lower-bound on the risk  $R$  that is independent of  $\phi$ . Let us now look at some specific choices of  $f$  such that  $G_f$  satisfies the desired properties and for which a bound on the Bayesian risk can indeed be retrieved.

**Corollary 1.** *Consider the Bayesian framework described in Sec. II-B. The following must hold for every  $p > 1$  and  $\rho > 0$ :*

$$R \geq \rho \left( 1 - L_W(\rho)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(W, X) + 1)^{\frac{1}{p}} \right). \quad (16)$$

*Proof.* Since  $f(x) = \frac{x^{p-1}}{p-1}$ , we have that  $f^*(0) = \sup_{x \geq 0} (-f(x)) = \frac{1}{p-1}$  and  $f^{-1}(t) = ((p-1)t + 1)^{\frac{1}{p}}$ .

For every estimator  $\hat{W}$ ,

$$L_W(\hat{W}, \rho) \cdot f^{-1} \left( \frac{I_f(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))f^*(0)}{L_W(\hat{W}, \rho)} \right) \quad (17)$$

$$= L_W(\hat{W}, \rho) \left( \frac{(p-1)\mathcal{H}_p(W, \hat{W}) + 1}{L_W(\hat{W}, \rho)} \right)^{\frac{1}{p}} \quad (18)$$

$$= L_W(\hat{W}, \rho)^{\frac{p-1}{p}} \left( (p-1)\mathcal{H}_p(W, \hat{W}) + 1 \right)^{\frac{1}{p}} \quad (19)$$

$$\leq L_W(\rho)^{\frac{p-1}{p}} ((p-1)\mathcal{H}_p(W, X) + 1)^{\frac{1}{p}}, \quad (20)$$

where in (20) we used the data-processing inequality for  $f$ -divergences. Using (20) with Theorem 1, we retrieve that for every estimator  $\hat{W}$

$$\mathbb{E}[\ell(W, \hat{W})] \geq \rho \left( 1 - L_W(\rho)^{\frac{p-1}{p}} ((p-1)\mathcal{H}_p(W, X) + 1)^{\frac{1}{p}} \right). \quad (21)$$

Since the right-hand side of (21) is independent of  $\hat{W} = \phi(X)$  one can use it to lower-bound the risk  $R$ .  $\square$

Restricting the choice of  $f$  to this family of polynomials we can thus state the following lower-bound on the risk:

$$R \geq \sup_{\rho > 0} \sup_{p > 1} \rho \left( 1 - L_W(\rho)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(W, \hat{W}) + 1)^{\frac{1}{p}} \right). \quad (22)$$

*Remark 1.* Using the one-to-one mapping connecting Hellinger divergences and Rényi's  $\alpha$ -Divergence [7, Eq. (30)], the bound above can be re-written as follows:

$$R \geq \sup_{\rho > 0} \sup_{\alpha > 1} \rho \left( 1 - L_W(\rho)^{\frac{\alpha-1}{\alpha}} \cdot \exp \left( \frac{\alpha-1}{\alpha} D_\alpha(P_{W\hat{W}} \| P_W P_{\hat{W}}) \right) \right). \quad (23)$$

In addition, given the generality of Theorem 1 we can also recover other notable results present in the literature (cf. [4, Remark 1]) through the following:

**Corollary 2.** *Consider the Bayesian framework described in Sec. II-B. The following must hold for every  $\beta > 0$ ,  $\gamma \geq \beta$ , and  $\rho > 0$ :*

$$R \geq \rho \left( 1 - \frac{E_{\beta, \gamma}(W, \hat{W}) + \gamma L_W(\rho)}{\beta} \right). \quad (24)$$

*Proof.* We take the same approach as in Corollary 1. Let  $f(x) = \max\{0, \beta x - \gamma\}$ , consequently one has that  $f^*(0) = \sup_{x \geq 0} (-f(x)) = 0$  and that the generalized inverse corresponds to  $f^{-1}(t) = \frac{t+\gamma}{\beta}$ . Using Theorem 1, along with the fact that  $f^*(0) \leq 0$  we have that for every estimator  $\hat{W}$ ,

$$\mathbb{E}[\ell(W, \hat{W})] \geq \rho \left( 1 - \frac{E_{\beta, \gamma}(W, \hat{W}) + \gamma L_W(\hat{W}, \rho)}{\beta} \right) \quad (25)$$

$$\geq \rho \left( 1 - \frac{E_{\beta, \gamma}(W, X) + \gamma L_W(\rho)}{\beta} \right). \quad (26)$$

Since (26) is independent of  $\hat{W} = \phi(X)$  one can use it to lower-bound the risk  $R$ .  $\square$

We thus retrieve the following lower-bound on the risk:

$$R \geq \sup_{\rho > 0} \sup_{\beta > 0, \gamma \geq \beta} \rho \left( 1 - \frac{E_{\beta, \gamma}(W, \hat{W}) + \gamma L_W(\rho)}{\beta} \right). \quad (27)$$

*Remark 2.* Note that setting  $\beta = 1$  (24) recovers the result in [4, Remark 1]. In fact, by introducing an additional degree of freedom through the  $\beta$  parameter in Equation (27), the resulting lower-bound can only be tighter than [4, Remark 1].

#### IV. EXAMPLES

In this section we apply Corollaries 1 and 2 to two classical estimation settings. The resulting lower-bounds are then compared with those obtained in [5] involving Sibson's  $\alpha$ -Mutual Information and Maximal Leakage and with those in [3] involving Shannon's Mutual Information and Maximal Leakage.

Ultimately, for each example, we would like to compare the tightest versions of our bounds, which are given by Equation

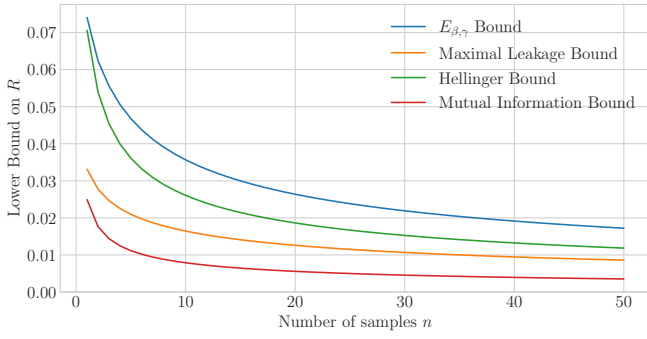


Fig. 1: Setting: Example 1. Comparison between the largest lower-bounds one can retrieve for different information measures in Example 1: that is between (27), (28), [5, Eq. (16)] and [3, Corollary 2, Eq. (19)]. The quantities are analytically maximized over  $\rho$  (cf. [1, Appendix A]) and numerically optimized over, respectively,  $p > 1$ ,  $\beta > 0$ , and  $\gamma \geq \beta$ .

(22) for the  $\mathcal{H}_p$ -Divergence and (27) for the  $E_{\beta,\gamma}$ -Divergence. However, since their computations involve a maximization problem over some parameters ( $p$  or  $\beta, \gamma$ ) that we cannot analytically solve, we compute these lower-bounds only for specific values of the parameters. The choice of parameters we use might seem arbitrary but it correctly captures the behavior of the bounds. Indeed, experiments show that when solving the maximization over  $p$  or  $\beta, \gamma$  (e.g., through the `scipy.optimize.minimize` function from the Python library SciPy) the same behaviors are observed, like Figure 1 shows in the context of Example 1.

#### A. Example 1: Bernoulli Bias Estimation

*Example 1.* Suppose that  $W \sim U[0, 1]$  and that for each  $i \in [n]$ ,  $X_i | \{W = w\} \sim \text{Ber}(w)$ . Also, assume that  $\ell(w, \hat{w}) = |w - \hat{w}|$ .

We first provide a closed-form expression of the lower-bound resulting from Corollary 1 for a specific choice of  $p$  which enables to match the upper-bound up to a constant factor. In fact in general, the tightest bound in this family comes from Equation (22) and can, in this example, be stated as follows:

$$R \geq \sup_{\rho > 0} \sup_{p > 1} \rho \left( 1 - (2\rho)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(W, X^n) + 1)^{\frac{1}{p}} \right). \quad (28)$$

The value of  $\mathcal{H}_p(W, X^n)$  for this setting is expressed in the following Lemma.

**Lemma 1.** Consider the setting described in Example 1. Then for every  $p > 1$ ,

$$\begin{aligned} & (p-1)(\mathcal{H}_p(W, X^n) + 1) \\ &= (n+1)^{p-1} \sum_{k=0}^n \binom{n}{k}^p \frac{\Gamma(kp+1)\Gamma((n-k)p+1)}{\Gamma(np+2)}. \end{aligned} \quad (29)$$

In particular with  $p = 2$ , one recovers:

$$\chi^2(W, X^n) + 1 = \frac{n+1}{2n+1} \cdot \frac{4^n}{\binom{2n}{n}} \leq \frac{16\sqrt{\pi n}}{21}. \quad (30)$$

*Proof.* The proof is omitted due to space constraints but can be found in [1, Appendix B].  $\square$

**Corollary 3.** Consider the setting described in Example 1. The Bayesian risk is lower-bounded by

$$R \geq \frac{7}{72\sqrt{\pi n}}. \quad (31)$$

*Proof.* Let  $p = 2$  in Corollary 1 along with  $L_W(\rho) \leq 2\rho$ , one has that

$$R \geq \sup_{\rho > 0} \rho \left( 1 - \sqrt{2\rho(\chi^2(W, X^n) + 1)} \right). \quad (32)$$

Solving the maximization over  $\rho$  (cf. [1, Appendix A]) and using (30) we conclude that

$$R \geq \frac{2}{27} \cdot \frac{1}{\chi^2(W, X^n) + 1} \geq \frac{7}{72\sqrt{\pi n}}. \quad (33)$$

$\square$

Notice that (31) matches the upper-bound up to a constant, and tightens the result in [3, Corollary 2] while not requiring that  $n \rightarrow \infty$ .

*Remark 3.* As mentioned in previous proof, Stirling's approximation yields  $(\chi^2(W, X^n) + 1) \sim \frac{\sqrt{\pi n}}{2}$  when  $n$  is large. This implies that for  $n$  large one can show that  $R \gtrsim \frac{4}{27\sqrt{\pi n}}$ , thus leading to a slight improvement over (31).

Similarly, one can do the same steps used to retrieve Corollary 3, but this time using the  $E_{\beta,\gamma}$ -Divergence instead of the  $\mathcal{H}_p$ -Divergence. In particular, for the case  $\beta = 0.75$  and  $\gamma = 2.2$ , Eq. (24) in this example can be expressed as

$$R \geq \sup_{\rho > 0} \rho \left( 1 - \frac{4}{3} (E_{0.75, 2.2}(W, X^n) + 4.4\rho) \right) \quad (34)$$

$$= \frac{5(0.75 - E_{0.75, 2.2}(W, X^n))^2}{66}. \quad (35)$$

A direct comparison between the bounds we provide and those already present in the literature can be seen in Figure 2. The lower-bounds are computed as a function of the number of samples  $n$ , which we consider to be in the range  $\{1, \dots, 50\}$ . The figure shows that all the divergences we considered in this work provide a larger (and thus, better) lower-bound on the Bayesian risk when compared with results that stem from using Shannon's Mutual Information (cf. [3, Corollary 2]). In particular, the lower-bound involving the  $E_{\beta,\gamma}$ -Mutual Information represents the largest among the ones we consider. Given the lack of a closed-form expression for  $E_{\beta,\gamma}$  in this example the quantities (35) along with ([3, Corollary 2, Eq. (19)] and [5, Eq. (16)]) and (31) are computed numerically.

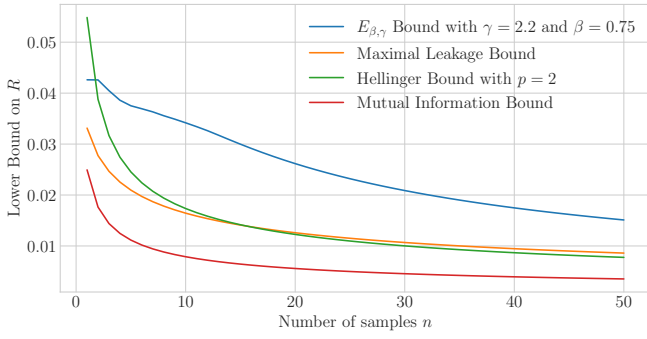


Fig. 2: Setting: Example 1. The picture shows the behaviour of (31), (35), [5, Eq. (16)], and [3, Corollary 2, Eq. (19)] as a function of  $n$ . The values of  $E_{0.75,2.2}(W, X^n)$  for each  $n$  are computed numerically. Here, unlike in Figure 1 where parameters are optimized, the values are fixed to  $\gamma = 2.2, \beta = 0.75$  and  $p = 2$ .

### B. Gaussian prior with Gaussian noise in $d$ dimensions

**Example 2.** Assume that  $W \sim \mathcal{N}(0, \sigma_W^2)$  and that for  $i \in [n]$ ,  $X_i = W + Z_i$  where  $Z_i \sim \mathcal{N}(0, \sigma^2)$ . Assume also that the loss is s.t.  $\ell(w, \hat{w}) = |w - \hat{w}|$ .

Using the estimator  $\hat{W} = \mathbb{E}[W|\bar{X}]$  with  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ , one has that  $R \leq \sqrt{\sigma_W^2 / (1 + n \frac{\sigma_W^2}{\sigma^2})}$ . Moreover, the small-ball probability can be upper-bounded as follows

$$L_W(\rho) \leq \left( \sup_{w \in \mathbb{R}} P_W(w) \right) \left( \int_{-\rho}^{\rho} 1 du \right) = \frac{2\rho}{\sqrt{2\pi\sigma_W^2}}. \quad (36)$$

Once again the largest lower bound on the risk, in the family of bounds provided by Corollary 1, can be expressed as follows

$$R \geq \sup_{\rho > 0} \sup_{p > 1} \rho \left( 1 - \left( \frac{2\rho}{\sqrt{2\pi\sigma_W^2}} \right)^{\frac{p-1}{p}} ((p-1)\mathcal{H}_p(W, X^n) + 1)^{\frac{1}{p}} \right). \quad (37)$$

To compute the Hellinger information, we make use of the following lemma:

**Lemma 2.** Let  $W \sim \mathcal{N}(0, \sigma_W^2 I_d)$  and  $Z \sim \mathcal{N}(0, \sigma^2 I_d)$  be two Gaussian random variables, where  $I_d$  denotes the  $d \times d$  identity matrix. Moreover, let  $X = W + Z$  and  $p > 1$ . Then

$$(p-1)(\mathcal{H}_p(W, X) + 1) = \left( \frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^p}{1 + (2-p)p \frac{\sigma_W^2}{\sigma^2}} \right)^{\frac{d}{2}}. \quad (38)$$

In particular, with  $p = 3/2$  and  $d = 1$ , one recovers:

$$\frac{1}{2}(\mathcal{H}_{3/2}(W, X) + 1) = \sqrt{\frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^{\frac{3}{2}}}{1 + \frac{3\sigma_W^2}{4\sigma^2}}}. \quad (39)$$

*Proof.* The proof is omitted due to space constraints but can be found in [1, Appendix C].  $\square$

Setting  $p = 3/2$  in (37) leads to the following result:

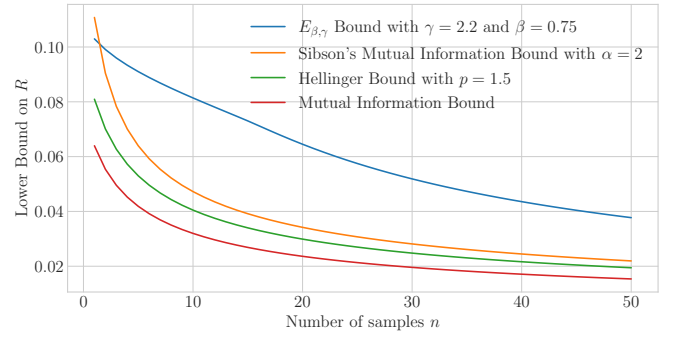


Fig. 3: Setting: Example 2 with  $\sigma_W^2 = 1$  and  $\sigma^2 = 2$ . The picture shows the behaviour of (40), (42), [5, Eq. (21)], and [3, Corollary 1, Eq. (16)] as a function of  $n$ . The values of  $E_{0.75,2.2}(W, X^n)$  for each  $n$  are computed numerically. Here, the values of the parameters are fixed to  $\gamma = 2.2, \beta = 0.75, \alpha = 2$  and  $p = 1.5$ .

**Corollary 4.** Consider the setting described in Example 2. The Bayesian risk is lower-bounded by

$$R \geq \frac{81\sqrt{2\pi}}{2048} \sqrt{\frac{\sigma_W^2}{1 + n \frac{\sigma_W^2}{\sigma^2}}}. \quad (40)$$

*Proof.* Given that  $\bar{X}$  is a sufficient statistic we have that  $\mathcal{H}_p(W, X^n) = \mathcal{H}_p(W, \bar{X})$ . Plugging this choice of  $\bar{X}$  in (39), substituting in (37), and then optimizing over  $\rho$  (cf. [1, Eq. (45)]), yields the statement after some algebraic manipulations.  $\square$

Note that (40) matches the upper-bound up to a constant factor, and provides a strengthening of the bounds obtained in [3, Corollary 1]. One can, as in Example 1, repeat the analysis with the  $f_{\beta,\gamma}$ -Divergence instead of the  $f_p$ -Divergence. In particular for the case  $\beta = 0.75$  and  $\gamma = 2.2$ , Equation (24) in this example can be expressed as

$$R \geq \sup_{\rho > 0} \rho \left( 1 - \frac{4}{3} \left( E_{0.75,2.2}(W, X^n) + \frac{4.4\rho}{\sqrt{2\pi\sigma_W^2}} \right) \right) \quad (41)$$

$$= \frac{5\sqrt{2\pi\sigma_W^2}(0.75 - E_{0.75,2.2}(W, X^n))^2}{66}, \quad (42)$$

where the optimization over  $\rho$  stems from [1, Appendix A].

Similarly to Example 1, we numerically evaluate (42) and compare it with [3, Corollary 1, Eq. (16)], [5, Eq. (21)] (with  $\alpha = 2$ ), and (40). Figure 3 shows the resulting lower-bounds as a function of the number of samples  $n$ . One can observe similar behaviors when comparing with the results from previous example: the bounds retrieved through the  $\mathcal{H}_p$ - and  $E_{\beta,\gamma}$ -Divergences are able to both improve on the lower-bound relying on Shannon's Mutual Information. Once again, Equation (27) gives the largest lower-bound in this example, while Sibson's  $\alpha$ -Mutual Information is still able to provide a stronger result than (22).

## REFERENCES

- [1] A. Vandenbroucq, A. R. Esposito, and M. Gastpar, "Lower-bounds on the bayesian risk in estimation procedures via  $f$ -divergences," *CoRR*, vol. abs/2202.02557, 2022. [Online]. Available: <https://arxiv.org/abs/2202.02557>
- [2] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013, pp. 2328–2336.
- [3] A. Xu and M. Raginsky, "Information-theoretic lower bounds on bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.
- [4] S. Asodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an  $f$ -divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 545–550.
- [5] A. R. Esposito and M. Gastpar, "Lower-bounds on the bayesian risk in estimation procedures via Sibson's  $\alpha$ -mutual information," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 748–753.
- [6] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theor.*, vol. 52, no. 10, pp. 4394–4412, 2006. [Online]. Available: <http://dx.doi.org/10.1109/TIT.2006.881731>
- [7] I. Sason, "On  $f$ -divergences: Integral representations, local behavior, and inequalities," *Entropy*, vol. 20, no. 5, 2018. [Online]. Available: <https://www.mdpi.com/1099-4300/20/5/383>
- [8] Te Sun Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.
- [9] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014, pp. 163–171.
- [10] X. Chen, A. Guntuboyina, and Y. Zhang, "On bayes risk lower bounds," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 7687–7744, jan 2016.
- [11] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, pp. 191–213, 1972.
- [12] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-,  $f$ -divergences and Maximal Leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.