Lower-bounds on the Bayesian Risk in estimation procedures via Sibson's α-Mutual Information

Amedeo Roberto Esposito, Michael Gastpar
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{amedeo.esposito, michael.gastpar}@epfl.ch

***Abstract*—In this work, we consider the problem of parameter estimation in a Bayesian setting. We propose a new approach to *lower-bounding* the Bayesian risk, based on Sibson's α-Mutual Information. The results are then applied to specific settings of interest. As an example, we provide a lower-bound on the risk of the so-called "Hide-and-Seek" problem. Generalisations of the results and alternative directions are also briefly presented.**

***Index Terms*—Bayesian Risk, Distributed Estimation, Sibson's α-Mutual Information, Maximal Leakage**

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517954

I. Introduction

In this work we consider the problem of parameter estimation in a Bayesian setting. More precisely, we propose a new approach to *lower-bounding* the Bayesian risk based on Sibson's α-Mutual Information. Through a classical reduction from estimation to testing [1], [2] we can shift the focus from the Bayesian risk to the computation of two objects:

- a measure of information (Sibson's α-MI);
- a small-ball probability [3];

We will look at the problem through an information-theoretic lens, similarly to [3]. We will thus treat the parameter to be estimated as a message sent through a channel. This allows us to include frameworks where, in a distributed fashion, m processors observe noisy samples of this parameter. The processors will then send a version of their observations to a central node. The central node will then proceed to estimate the parameter. The main advantage of using this type of bounds is that the small-ball probability does not depend on the specific estimator. Similarly, the information measure can also be rendered independent of the estimator via strong/classical DPIs (Data Processing Inequalities). Therefore, these lower-bounds can be applied to any estimation framework that matches this one, regardless of the choice of the estimator. It is important to note that, although the problem can be interpreted as a transmission problem, a fundamental difference is that the size of the quantised messages may not grow with the number of samples. This might render the reconstruction of the samples impossible but the estimation of the parameter may remain feasible [3]. Our main focus will not be on asymptotic results but rather on finite samples lower-bounds.

A. Sibson's α-Mutual Information

Introduced by Rényi as a generalization of KL-divergence, α-divergence has found many applications ranging from hypothesis testing to guessing and several other statistical inference problems [4]. It can be defined as follows [5].

Definition 1. Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real number different from 1. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$ (such a measure always exists, e.g. $\mu = (\mathcal{P}+\mathcal{Q})/2)$) and denote with p, q the densities of $\mathcal{P}, \mathcal{Q}$ with respect to μ. The α-Divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha} d\mu. \tag{1}$$

For an extensive treatment of α-divergences and their properties we refer the reader to [5]. Starting from Rényi's Divergence and the geometric averaging that it involves, Sibson built the notion of Information Radius [6]. A deconstructed and generalised version of the Information Radius leads us to the following definition of a generalisation of Shannon's Mutual Information [4]:

Definition 2. Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$, and with marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$, respectively. For $\alpha > 0$, the Sibson's mutual information of order α between X and Y is defined as:

$$I_\alpha(X,Y) = \min_{Q_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X Q_Y). \tag{2}$$

One can prove that $\lim_{\alpha\to 1} I_\alpha(X,Y) = I(X;Y)$. On the other hand when $\alpha \to \infty$, we get: $I_\infty(X,Y) = \log \mathbb{E}_{\mathcal{P}_Y}\left[\sup_{x:\mathcal{P}_X(x)>0} \frac{\mathcal{P}_{XY}(\{x,Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})}\right] = \mathcal{L}(X\to Y)$, where $\mathcal{L}(X\to Y)$ denotes the Maximal Leakage from X to Y, a recently defined information measure with an operational meaning in the context of privacy and security [7]. For more details on Sibson's α-MI we refer the reader to [4], as for Maximal Leakage the reader is referred to [7].

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 $\phi : \mathcal{X} \to \mathcal{W}$ one can then estimate W from $X \sim \mathcal{P}_{X|W}$ via $\phi(X) = \hat{W}$. Let us denote with $\ell : \mathcal{W}\times\mathcal{W} \to \mathbb{R}^+$ a loss function, the Bayesian risk is defined as:

$$R = \inf_\phi \mathbb{E}[\ell(W, \phi(X)] = \inf_\phi \mathbb{E}[\ell(W,\hat{W})]. \tag{3}$$

Our purpose will be to lower-bound R using the tools described in the previous section. In particular, using a simple Markov's inequality approach, for every estimator ϕ and $\rho \geq 0$, one can do the following:

$$\mathbb{E}[\ell(W,\hat{W})] \geq \rho \left(\mathbb{P}(\ell(W,\hat{W}) \geq \rho) \right). \tag{4}$$

Computing the deviation probability in (4) is not necessarily easier than computing the expectation itself as it still depends on a specific choice of ϕ. However, with further manipulations we can actually relate $\mathbb{P}(\ell(W,\hat{W}) \geq \rho)$ to our information-measure term $I_\alpha(W,X)$ and the small-ball probability

$$L_W(\rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} \mathbb{P}(\ell(W,w) \leq \rho). \tag{5}$$

Both these terms are now independent of ϕ granting us the tools to provide general lower-bounds on 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], [10]. Closely connected to this work is [3]. The approach is quite similar, with the main difference that we employ a family of bounds involving Sibson's $\alpha-$Mutual Information [11], [12] while [3] relies solely on Mutual Information.

II. The lower bounds

Our main contribution is the connection between (4), $I_\alpha(W,X)$ and (5) and is summarised in the following result.

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

$$R \geq \rho \left(1 - \exp\left(\frac{\alpha-1}{\alpha} \left(I_\alpha(W,X) + \log(L_W(\rho)) \right) \right) \right). \tag{6}$$

Proof. We have that

$$P_{W\hat{W}}(\ell(W,\hat{W}) \leq \rho) \tag{7}$$

$$\leq \left(\sup_{\hat{w} \in \hat{\mathcal{W}}} P_W(\ell(W,\hat{w}) \leq \rho) \right)^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,\hat{W}) \right) \tag{8}$$

$$= \exp\left(\frac{\alpha-1}{\alpha} \left(I_\alpha(W,\hat{W}) + \log(L_W(\rho)) \right) \right) \tag{9}$$

$$\leq \exp\left(\frac{\alpha-1}{\alpha} \left(I_\alpha(W,X) + \log(L_W(\rho)) \right) \right). \tag{10}$$

(8) follows from [11, Corollary 1]. (10) follows from the Data Processing Inequality for I_α [4] and the Markov Chain $W - X - Y - \hat{W}$. Moreover, starting from (4) and using Markov's inequality one has that

$$R \geq \rho \cdot P_{W\hat{W}}(\ell(W,\hat{W}) \geq \rho) \tag{11}$$

$$= \rho \cdot (1 - P_{W\hat{W}}(\ell(W,\hat{W}) \leq \rho)). \tag{12}$$

The statement follows from upper-bounding (12) with (10). □

Two remarks are in order:

- It is important to notice that the behaviour of (6) is fundamentally different from [3, Theorem 1]. In [3, Theorem 1] the dependence is linear with respect to the Mutual Information and logarithmic in $L_W(\rho)$ while in our Theorem 1, we have an exponential dependence in I_α and linear in $L_W(\rho)$.
- The theorem introduces a new parameter $\alpha > 1$ to optimise over. α introduces a trade-off between the two quantities for a given ρ, $I_\alpha(W,X)$ and $L_W(\rho)$: $\frac{\alpha-1}{\alpha} I_\alpha(W,X)$ will increase with α whereas $L_W(\rho)^{\frac{\alpha-1}{\alpha}}$ will decrease.

Taking the limit of $\alpha \to \infty$ in (6) allows us to get rid of the parameter α while maintaining the same type of behaviour. This also brings in Sibson's α-MI of order ∞ also known in the literature as Maximal Leakage [7]:

Corollary 1. *Consider the Bayesian framework described in Sec. I-B.*

$$R \geq \sup_{\rho > 0} \rho \left(1 - \exp\left(\mathcal{L}\left(W \to X\right) + \log(L_W(\rho)) \right) \right). \tag{13}$$

An interesting characteristic of Corollary 1 is that $\mathcal{L}\left(W \to X\right)$ depends on W only through the support. This allows us to provide, essentially for free, an even more general lower-bound on R. Indeed, ignoring $L_W(\rho)$ for a moment (it can be upper-bounded regardless of P_W as it is done in Sec. III-A), for a fixed family of $\mathcal{P}_{X|W}$, $\mathcal{L}\left(W \to X\right)$ has the same value regardless of $\mathcal{P}_W$, as long as the support of $\mathcal{P}_W$ remains the same.

III. Examples

In this section we will apply Theorem 1 and Corollary 1 to different estimation settings. The first example will show a setting where a simple application of Corollary 1 can already provide an interesting lower-bound. Example 2 shows the limitations of Corollary 1 and proposes a setting where the more general Theorem 1 needs to be applied. To conclude, Example 3 shows how one can provide a meaningful lower-bound for the "Hide-and-seek" problem presented in [10].

A. Bernoulli Bias

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

It is easy to see that using the sample mean estimator, *i.e.* $\hat{W} = \frac{1}{n}\sum_{i=1}^n X_i$ one has that $R \leq \frac{1}{\sqrt{6n}}$. Let us now lower-bound the risk in this setting. Without making any further assumptions on W, we can only trivially upper-bound $L_W(\rho)$ with 2ρ. In settings where it is difficult to directly estimate $L_W(\rho)$ but only an upper-bound is available it is useful to follow the technique described in the Appendix. Thus, a simple application of (35) allows us to get a result tight in n. Given that $L_W(\rho) \leq 2\rho = g(\rho)$, we have that g is clearly an increasing function and thus invertible. Moreover, $g^{-1}(k) = k/2$. Given the definition of $P_{X_i|W=w}$ one also has that

$$\mathcal{L}\left(W \to X^n\right) \leq \log\left(2 + \sqrt{\frac{\pi n}{2}} \right). \tag{14}$$

Using these considerations in (35) we get:

$$R \geq \sup_{s\in(0,1)} \frac{s(1-s)}{2\left(2+\sqrt{\frac{\pi n}{2}}\right)} \tag{15}$$

$$= \frac{1}{8\left(2+\sqrt{\frac{\pi n}{2}}\right)}, \tag{16}$$

which is a constant away from the upper-bound. For n large enough (*i.e.*, $n \geq 160$) we can actually further lower bound the quantity as follows:

$$R \geq \frac{1}{\sqrt{2\pi n}}. \tag{17}$$

Surprisingly, the maximal leakage offers already a good result without having to add any extra machinery. Corollary 1 provides a better lower-bound than the Mutual Information result in this simple setting (c.f., the bound proved in [3, Cor 2] via conditioning wrt an independent copy of X^n provides a lower bound of $1/(16\sqrt{\pi n})$.) Moreover, if one has access to an upper-bound on $L_W(\rho)$ that does not employ any knowledge of the measure $\mathcal{P}_W$, except for the support (e.g., if W were to be discrete, an upper-bound of 1 over the pmf would suffice) the lower-bound on the risk would apply to any W whose support is the interval $[0,1]$. This shows the potential of employing Maximal Leakage in these settings.

B. Gaussian prior with Gaussian noise

However simple and effective, Corollary 1 has a disadvantage: $\mathcal{L}(X\to Y) = +\infty$ whenever $Y = f(X)$, X has an infinite support and f is deterministic. The same applies if $Y = X + Z$ where X is a real valued random variable and Z is an independent random variable and they both have infinite support [7]. In these settings, one must resort to $I_\alpha(X,Y)$ with $\alpha < +\infty$. The following example presents one such setting.

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

Using the sample mean estimator one has that $R \leq \sqrt{\frac{\sigma_W^2}{1+n\sigma_W^2/\sigma^2}}$. In this setting, $\mathcal{L}(W\to X^n)$ is actually infinite. However, $I_\alpha(W,X^n)$ is finite for every $\alpha < +\infty$. Thus, let us provide a lower bound to the risk R through (36). Since the empirical mean is a sufficient statistic for W we have that

$$I_\alpha(W,X^n) = I_\alpha\left(W, \frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{2}\log\left(1+\alpha n\frac{\sigma_W^2}{\sigma^2}\right). \tag{18}$$

Moreover, we have that for $\ell(w,\hat{w}) = |w-\hat{w}|$, $L_W(\rho) \leq \rho\sqrt{\frac{2}{\sigma_W^2\pi}}$. This implies that $g^{-1}(k) = k\sqrt{\frac{\sigma_W^2\pi}{2}}$ and:

$$R \geq \sup_s \sup_{\alpha\in(1,+\infty)} s(1-s)^\gamma \sqrt{\frac{\sigma_W^2\pi}{2\left(1+n\alpha\frac{\sigma_w^2}{\sigma^2}\right)}} \tag{19}$$

$$\geq \sup_{\alpha\in(1,+\infty)} \frac{1}{2}^{\gamma+1} \sqrt{\frac{\sigma_W^2\pi}{2\left(1+n\alpha\frac{\sigma_w^2}{\sigma^2}\right)}} \tag{20}$$

$$\geq \frac{1}{2}^{7/2} \sqrt{\frac{\sigma_W^2\pi}{\left(1+2n\frac{\sigma_w^2}{\sigma^2}\right)}}. \tag{21}$$

Here (20) follows from choosing $s = \frac{1}{2}$ and (21) follows from choosing $\alpha = 2$ which in turn implies $\gamma = 2$. Similarly to the previous example, (21) is a constant away from the upper-bound. One could also take α arbitrarily close to 1 but that would imply that γ is approaching $+\infty$, as $\frac{1}{\alpha}+\frac{1}{\gamma} = 1$, resulting in a worse multiplicative constant. To conclude, let us consider next a d-dimensional distributed estimation problem, known as the "Hide and seek" problem. It has been first presented in [10] and also studied in [3].

C. Hide-and-seek problem

Example 3. Consider a family of distributions $\mathcal{P} = \{\mathcal{P}_w : w = 1,\dots,d\}$ on $\{0,1\}^d$. Under $\mathcal{P}_w$, the $w-$th coordinate of the random vector $X \in \{0,1\}^d$ has bias $\frac{1}{2}+\rho$ while the other coordinates of X are independently drawn from Ber$(1/2)$. For $i = 1,\dots,m$, the i-th processor observes n samples X_i^n drawn independently from $\mathcal{P}_W$, and sends a b-bits message $Y_i = \varphi(X_i^n, Y^{i-1})$ to the estimator. The estimator computes $\hat{W} = \psi(Y^m)$ from the received messages. The risk in this example is defined as:

$$R_M = \inf_{\varphi^m,\psi} \max_{w\in[d]} \mathbb{P}[W \neq \hat{W}]. \tag{22}$$

A lower-bound for R_M derived in [10] is as follows:

$$R_M \geq 1 - \left(\frac{3}{d} + 5\sqrt{\min\left\{\frac{10\rho nmb}{d}, mn\rho^2\right\}}\right) \tag{23}$$

and only holds for $0 \leq \rho \leq 1/(4n)$. On the other hand, in [3], a quite different lower-bound has been proposed:

$$R_M \geq 1 - \frac{1}{\log d}\min\left\{\left[1-\left(\frac{1-2\rho}{1+2\rho}\right)^n\right]mb + 1, \min(4mn\rho^2, \log d) + 1\right\}, \tag{24}$$

and it holds for $0 \leq \rho \leq 1/2$. Let us now use a naïve approach with Maximal Leakage. We have that $W - X^{n\times m} - Y^m - \hat{W}$ forms a Markov Chain. Thus, $\mathcal{L}\left(W\to\hat{W}\right) \leq \min(\mathcal{L}(W\to X^{n\times m}), \mathcal{L}(W\to Y^m))$. We also have that $\mathcal{L}(W\to Y^m) \leq mb$ and that:

$$\mathcal{L}\left(W\to X^{n\times m}\right) \leq \min(nm\mathcal{L}(W\to X), \log d) \tag{25}$$

$$= \min(nm\log(1+2\rho), \log d). \tag{26}$$

Hence:

$$\mathcal{L}\left(W\to\hat{W}\right) \leq \min(nm\log(1+2\rho), \log d, mb). \tag{27}$$

Using (27) in Corollary 1 we get the following:

$$\mathbb{P}(\{\hat{W} \neq W\}) \geq 1 - \frac{\exp(\min\{mb, \log d, nm\log(1+2\rho)\})}{d}. \tag{28}$$

Notice that (28) is such that the right-hand side is always greater or equal to 0. Indeed, assuming d to be fixed, and letting n and m grow, we have that the minimum is achieved by $\log d$ and in that case we have $\mathbb{P}(\{\hat{W} \neq W\})] \geq 0$. Here, the difference in behaviour of Corollary 1 with respect to [3, Theorem 1] is pivotal. Let us now compare the results on a common setting. The setting chosen in [3], where $d = 512, b = 3d, m = 10$ and $\rho = 1/(4n)$ does not represent a choice of parameters where (28) is interesting. Indeed, for large enough n, $nm\log(1+2\rho) = nm\log(1+1/2n) \approx m/2$ and as a consequence, the expression will converge to a constant determined by the minimum between $mb, \log d, m/2$. On the other hand, both (23) and (24) have a term that depends on $mn\rho^2$ which, for $\rho = 1/(4n)$, will decay with n. Choosing instead $\rho \sim n^{-p}$ with $p > 1$ represents an interesting setting for our bound, as the plots in Fig. 1a and 1b show. Thanks to the different behaviour of (28) (reaching 1 exponentially fast) we can see a much sharper jump towards 1 with respect to (24), which instead plateaus below 1, and with respect to (23) that reaches 1 more slowly. The jump to 1 of (28) becomes even sharper with larger p and converges towards a specific behaviour at $p \approx 2$. Increasing p any further does not alter the behaviour of the bound meaningfully. As for the behaviour for fixed ρ, for $\rho = 0.01$, (28) provides a larger lower-bound only for $n < 25$. If we lower the parameter down to $\rho = 0.0001$ then (28) is larger than (24) for all n but only larger than (23) for $n < 1850$. Regardless of the considerations related to the specific settings, it is interesting how a very simple application of Corollary 1 can provide a tighter lower-bound without having to compute any SDPI constant. Moreover, in the proof of (24) in [3], in order to compute $I(W;X)$ an assumption on the distribution of W was necessary and the choice fell on W uniform on $[d]$. With Maximal Leakage, $\mathcal{L}(W \to X)$ does not depend on the specific distribution over W, rendering the bound potentially more general.

IV. Further generalisations

A. Inverting the roles

α-mutual information is an asymmetric quantity. Theorem 1 involves $I_\alpha(W,X)$. A natural follow-up question is: Can one give a similar bound involving $I_\alpha(X,W)$ instead? Indeed, by inverting the roles of W and $\hat{W}$, such a bound can be given, as we now show. The bound involves the small ball probability for $\hat{W}$, that is,

$$L_{\hat{W}}(\rho) = \sup_w \mathcal{P}_{\hat{W}}(d(w,\hat{W}) \geq \rho). \tag{29}$$

This quantity hinges on the marginal distribution of $\hat{W}$, which, in turn, depends on the estimator used. In terms of $L_{\hat{W}}(\rho)$, we can give the following general bound:

Lemma 1. *Consider the Bayesian framework described in Sec. I-B. The following holds for every $\alpha > 1$ and $\rho > 0$:*

$$R \geq \rho\left(1 - \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(X,W) + \log(L_{\hat{W}}(\rho))\right)\right)\right). \tag{30}$$

Moreover, taking the limit of $\alpha \to \infty$ one has:

$$R \geq \rho\left(1 - \exp\left(\mathcal{L}(X \to W) + \log(\mathbb{E}[L_{\hat{W}}(\rho)])\right)\right). \tag{31}$$

To apply this lemma in concrete cases, one needs to compute or upper bound the small ball probability $L_{\hat{W}}(\rho)$. Leveraging basic properties of the estimator, one can sometimes bound it. For example, if the estimator is a linear function of the noisy observations one can leverage results related to Lévy's concentration functions of sums of independent random variables. E.g., if $Y_1, \ldots, Y_m$ are uncorrelated and have log-concave distributions, then for every $\rho \geq 0$ [13, Thm 1.1],

$$L_{\sum_i^m Y_i}(\rho) \leq \frac{2\rho}{\sqrt{\mathrm{Var}(\sum_{i=1}^m Y_i) + \rho^2/3}} = \frac{2\rho}{\sqrt{m\mathrm{Var}(Y_1) + \rho^2/3}}.$$

More general statements can be made, assuming $\psi(Y^m) = \sum_{i=1}^m a_i Y_i$ under different constraints over a_i [14]. To appreciate the promise of this lemma, let us also discuss the behaviors of $I_\alpha(W,X)$ and $I_\alpha(X,W)$, respectively. Specifically, let us consider again the "Hide and Seek" problem. Assuming, as in [3, Ex. 12], that $\mathcal{P}_W$ is uniform over $[d]$, we have that $\mathcal{L}(X^{n\times m} \to W) = \log \frac{d(1/2+\rho)}{(d-1)(1/2-\rho)+(1/2+\rho)} = \log \kappa(d,\rho) < \log d$. In the case of a constant ρ and d and a linear estimator ψ, using (IV-A) in Lemma 1 one would get $R \geq \rho\left(1 - \frac{\kappa(d,\rho)2\rho}{\sqrt{m\mathrm{Var}(Y_i)+\rho^2/3}}\right)$. This lower bound approaches ρ as m grows, rather than providing the trivial lower bound of 0 as in (28). The assumptions required, along with the need of specifying a prior over W, clearly restrict the domain of applicability of Lemma 1 with respect to Theorem 1 and Corollary 1. However, this approach can provide results in settings where Theorem 1 and Corollary 1 become vacuous.

B. Conditioning

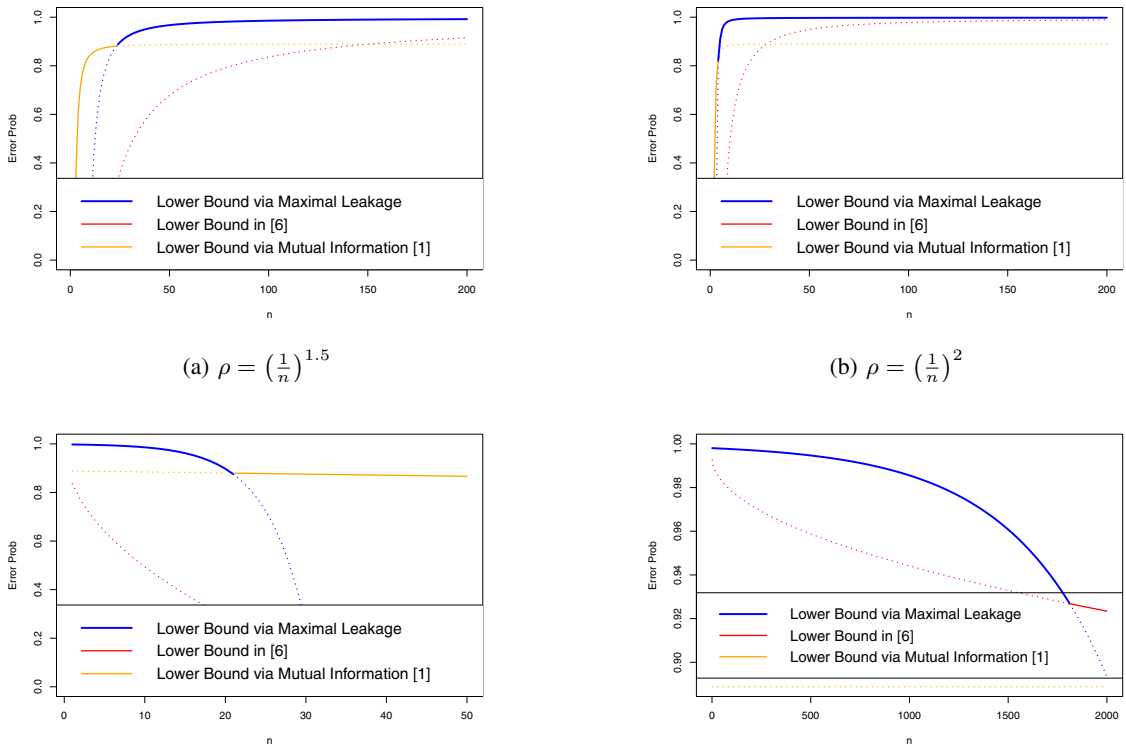
Following the approach undertaken in [3], it is also possible to propose a conditional version of Theorem 1 and Corollary 1, in order to get tighter bounds. In this spirit we will consider the following definition:

Definition 3. Let X, Y, Z be three random variables jointly distributed according to $\mathcal{P}_{XYZ}$. For $\alpha > 0$, a conditional Sibson's mutual information of order α between X and Y given Z is defined as:

$$I_\alpha(X,Y|Z) = \min_{Q_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} Q_{Y|Z} \mathcal{P}_Z). \tag{32}$$

Remark 1. This definition has already appeared in [15, Section IV.C.2] where a closed-form expression has been computed alongside with an operational meaning through a hypothesis testing problem.

It has been shown in [12] that several such definitions can be advanced, depending on the operational meaning and corresponding probability bound one needs. The choice of this specific definition has been done in order to be able to use [12, Theorem 1, Corollary 1]. These results are in turn necessary to provide a conditional version of Theorem 1 and Corollary 1 similar to [3, Theorem 1 eq. (5)]. Leveraging Definition 3 and the fact that when $\alpha \to \infty$ then $I_\alpha(X,Y|Z) \to \mathcal{L}(X \to Y|Z)$ [12], we can now give a conditional version for Theorem 1 and Corollary 1.



(a) $\rho = \left(\frac{1}{n}\right)^{1.5}$

(b) $\rho = \left(\frac{1}{n}\right)^{2}$

(c) $\rho = 0.01$

(d) $\rho = 0.0001$

Fig. 1: Behaviour of (28) for various values of ρ. Solid lines mean that the corresponding lower-bound is the largest.

Theorem 2. *Consider the Bayesian framework described in Sec. I-B.*

$$R \geq \sup_{\mathcal{P}_{U|W,X}} \sup_{\rho>0,\alpha\geq 1} \rho\Bigg(1-\exp\left(\frac{\alpha-1}{\alpha}(I_\alpha(W,X|U) + \log(\mathbb{E}[L_{W|U}(\rho)])\right)\Bigg).$$

Moreover, taking the limit of $\alpha \to \infty$ one has:

$$R \geq \sup_{\mathcal{P}_{U|W,X}} \sup_{\rho>0} \rho\left(1-\exp\left(\mathcal{L}\left(W\to X|U\right)+\log(\mathbb{E}[L_{W|U}(\rho)])\right)\right).$$

The main idea lying behind the conditional version for Mutual Information is that, choosing an appropriate U, it is possible to control the growth of $I(W;X|U)$ and obtain tighter bounds. In particular, let us assume we have n samples X^n. If the family $\mathcal{P} = \{\mathcal{P}_{X|W=w} : w \in \mathcal{W}\}$ is a subset of a finite-dimensional exponential family and W has a density supported on a compact subset of $\mathbb{R}^d$, if we choose U to be a conditionally independent copy $\hat{X}^n$ of X^n (given W) the mutual information $I(W;X^n|\hat{X}^n)$ will converge to a constant as n grows (rather than grow with n) [3]. However, in the specific examples discussed earlier, there does not appear to be a suitable U that tightens the bounds further. Nonetheless we state the result as it may be of interest in other settings.

ACKNOWLEDGMENT

The work in this paper was supported in part by the Swiss National Science Foundation under Grants 169294 and 200364.

APPENDIX

The quantity $L_W(\rho)$ can be hard to estimate precisely. In cases when only an upper-bound is available, the following is a sometimes useful technique that can be used to lower-bound the risk [3, Remark 1]: assume that we can upper-bound $L_W(\rho)$ with an increasing function $g(\rho)$. Let us also assume that g admits a generalised inverse $g^{-1}(k) = \sup\{\rho \geq 0 : g(\rho) \leq k\}$. Let $s \in (0,1)$ be fixed. If for some $\hat{\rho}$

$$g(\hat{\rho}) \leq (1-s)\exp(-\mathcal{L}(W \to X)) \tag{33}$$

then we can guarantee that

$$1 - \exp(\mathcal{L}(W \to X))L_W(\hat{\rho}) \geq s. \tag{34}$$

Using this in Corollary 1 one gets the following:

$$R \geq \sup_{s\in(0,1)} s(g^{-1}((1-s)\exp(-\mathcal{L}(W \to X))). \tag{35}$$

Similarly, for $I_\alpha(W,X)$ one can state the following:

$$R \geq \sup_{s\in(0,1)} s(g^{-1}\left((1-s)^\gamma \exp(-I_\alpha(W,X))\right)), \tag{36}$$

where $\gamma = \alpha/(\alpha-1)$.

REFERENCES

[1] B. Yu, *Assouad, Fano, and Le Cam*. New York, NY: Springer New York, 1997, pp. 423–435.

[2] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," *CoRR*, vol. abs/1405.0782, 2014.

[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. Verdú, "α-mutual information," in *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, 2015, pp. 1–6.

[5] T. van Erven and P. Harremoës, "Rényi divergence and kullback-keibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[6] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitstheorie verw Gebiete 14*, pp. 149–160, 1969.

[7] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.

[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] 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.

[10] 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.

[11] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via rényi-, f-divergences and maximal leakage," *Accepted for Publication in IEEE Transactions on Information Theory*, 2021. [Online]. Available: http://arxiv.org/abs/1912.01439

[12] A. R. Esposito, D. Wu, and M. Gastpar, "On conditional Sibson's α-Mutual Information," 2021. [Online]. Available: http://arxiv.org/abs/2102.00720

[13] S. G. Bobkov and G. P. Chistyakov, "On concentration functions of random variables," *Journal of Theoretical Probability volume*, vol. 28, 2015.

[14] H. H. Nguyen and V. H. Vu, *Small Ball Probability, Inverse Theorems, and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 409–463.

[15] M. Tomamichel and M. Hayashi, "Operational interpretation of rényi information measures via composite hypothesis testing against product and markov distributions," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1064–1082, 2018.