Journal of Machine Learning Research 25 (2024) 1-45 Submitted 3/23; Revised 2/24; Published 11/24

Lower Bounds on the Bayesian Risk via Information Measures

Amedeo Roberto Esposito AMEDEO.ESPOSITO@OIST.JP
Okinawa Institute of Science and Technology, Okinawa

Adrien Vandenbroucque ADRIEN.VANDENBROUCQUE@EPFL.CH
École Polytechnique Fédérale de Lausanne, Switzerland

Michael Gastpar MICHAEL.GASTPAR@EPFL.CH
École Polytechnique Fédérale de Lausanne, Switzerland

Editor: Daniel Roy

Abstract

This paper focuses on parameter estimation and introduces a new method for lower bounding the Bayesian risk. The method allows for the use of virtually *any* information measure, including Rényi's α, φ-divergences, and Sibson's α-Mutual Information. The approach considers divergences as functionals of measures and exploits the duality between spaces of measures and spaces of functions. In particular, we show that one can lower bound the risk with any information measure by upper bounding its dual via Markov's inequality. We are thus able to provide estimator-independent impossibility results thanks to the Data-Processing Inequalities that divergences satisfy. The results are then applied to settings of interest involving both discrete and continuous parameters, including the "Hide-and-Seek" problem, and compared to the state-of-the-art techniques. An important observation is that the behaviour of the lower bound in the number of samples is influenced by the choice of the information measure. We leverage this by introducing a new divergence inspired by the "Hockey-Stick" divergence, which is demonstrated empirically to provide the largest lower bound across all considered settings. If the observations are subject to privatisation, stronger impossibility results can be obtained via Strong Data-Processing Inequalities. The paper also discusses some generalisations and alternative directions.

Keywords: Bayesian Risk, Estimation, Divergences, Duality, Hölder's Inequality

1 Introduction

In this work,[1] we consider the problem of parameter estimation in a Bayesian setting. In this problem, an underlying parameter is modelled as a random variable. Noisy observations are made according to a given conditional probability distribution, conditioned on the realisation of the underlying parameter. Based on these observations, the parameter needs to be estimated. Estimation quality is assessed through a fidelity criterion, expressed in terms of a loss function. The average incurred loss is referred to as the Bayesian risk. This problem has a rich history, dating back to Bayes (1764). Given the characteristics of the observation process, it is of interest to characterise the performance of the optimal estimator. This is

[1] This article was presented in part at the 2021 and 2022 IEEE International Symposia on Information Theory

well known to be prohibitive. Short of such a characterisation, it is therefore relevant to develop fundamental lower bounds on the performance of any estimator. We propose an approach to *lower bounding* the Bayesian risk leveraging most information measures present in the literature. We look at the problem through an information-theoretic lens, similarly to Xu and Raginsky (2017). We 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. We thus shift the focus from the estimation problem to the computation of two main quantities (which we render as independent of the estimator as possible):

1. an information measure (*e.g.*, Sibsons's α-Mutual Information, φ-Mutual Information, etc.);

2. a functional of the probability of some event under independence (*e.g.*, a small-ball probability (Li and Shao, 2001), like it happens in (Xu and Raginsky, 2017)).

The main tools utilised rely on Legendre-Fenchel duality and they allow us to introduce bounds involving Rényi's, φ-divergences and Sibson's α-Mutual Information. An advantage of using this type of bounds is that one can render the functional in Item 2 (*e.g.*, the small-ball probability) independent of the specific estimator. Similarly, the information measure can also be rendered independent of the estimator via Data-Processing Inequalities. Therefore, these lower bounds can be applied to any standard estimation framework regardless of the specific choice of the estimator. More details on the formal framework that we adopted can be found in Section 1.3.

It is important to notice 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 (Xu and Raginsky, 2017). Our main focus will not be on asymptotic results but rather on finite number of samples lower bounds.

1.1 Overview of the document

Following the Introduction, the paper will be broken into four main sections:

- Section 2: Preliminaries, in which we will define the information measures of interest as well as describe the theoretical framework leveraged to provide the bounds;

- Section 4: Main Bounds, in which (making use of the framework described in Section 2) we propose a variety of lower bounds on the Bayesian risk involving a variety of information-measures, in particular:

  - Sibson's α-Mutual Information and Maximal Leakage (Theorem 8);
  - φ-Mutual Information (Theorem 9). In particular:
    - Hellinger p-divergence (Corollary 11);
    - Rényi's α-divergence (Remark 12);

  - ∗ a generalisation of the "Hockey-Stick" divergence, E_γ (Corollary 14);

- Section 5: Examples, in which we apply the bounds proposed in Section 4 to a variety of classical and less classical settings:

  - estimation of the bias of a Bernoulli random variable (see Section 5.1);
  - estimation of the bias of a Bernoulli random variable after injection of additional noise (*e.g.*, observing privatised samples, see Section 5.2);
  - estimation of the mean of a Gaussian random variable (with Gaussian prior, see Section 5.3);
  - lower bound on the minimax risk for the "Hide-and-Seek" problem (Shamir, 2014) (see Section 5.4).

  For each of the problems we derive bounds involving a variety of information measures and we compare said bounds among themselves and with respect to relevant bounds in the literature as well.

- Throughout the document we also consider further generalisations, in which we propose a variety of ways of extending/tightening/altering the results we proposed in Section 4. In particular, one can provide new bounds:

  - conditioning on an additional random variable (see Appendix E.1);
  - leveraging the asymmetry of some information measures (see Appendix E.2);
  - lower bounding the expected risk directly (*i.e.*, without using Markov's inequality, see Appendix E.3).

1.2 Related Work

The problem of parameter estimation has been extensively studied over the years, with many contributions coming from a variety of fields. Relevant literature, mostly leveraging the Van Trees Inequality (and the quadratic risk) can be found in Van Trees (2001); Sato and Akahira (1996); Brown and Gajek (1990); Van Trees and Bell (2007); Brown and Liu (1993). Moreover, a survey of early work in this area (mainly focusing on asymptotic settings) can be found in Te Sun Han and Amari (1998). More recent but important advances are instead due to Zhang et al. (2013); Duchi and Wainwright (2013); Shamir (2014). Closely connected to this work is Xu and Raginsky (2017). The approach is quite similar, with the main difference that we employ a family of bounds involving a variety of divergences while Xu and Raginsky (2017) relies on Mutual Information and on the information density. Related is also Asoodeh et al. (2021), where the authors use the so-called E_γ-divergence to provide a lower bound on the Bayesian Risk. A similar approach was also undertaken in Chen et al. (2016). The authors focused on the notion of φ-informativity (Csiszár, 1972)) and leveraged the Data-Processing inequality similarly to (Esposito et al., 2021a, Theorem 3). A more thorough discussion of the differences between this work and Chen et al. (2016) can be found in Appendix B.

1.3 Problem Setting

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

$$R_B = \inf_\phi \mathcal{P}_{W\hat{W}}(\ell(W, \phi(X))) = \inf_\phi \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})). \tag{1}$$

Our purpose is to lower bound R_B using a variety of information measures. With this drive and leveraging various tools that stem from Legendre-Fenchel duality, one can connect the expected value of ℓ under the joint $\mathcal{P}_{W\hat{W}}$ to

- the expected value of the same function under the product of the marginals ($\mathcal{P}_W \mathcal{P}_{\hat{W}}$) or a "small-ball probability";
- an information measure (quantifying how "far" the joint is from the product of the marginals).

This will allow us to render the lower bound *as independent as possible* from the specific choice of the estimator ϕ. More precisely, our desideratum will be a lower bound of the following form:

$$R_B \geq \varpi \left(\frac{d\mathcal{P}_{W\hat{W}}}{d\mathcal{P}_W \mathcal{P}_{\hat{W}}} \right) \vartheta(\mathcal{P}_W \mathcal{P}_{\hat{W}}, \ell), \tag{2}$$

with, once again, the purpose of then rendering the right-hand side of Equation (2) as independent as possible of the estimator ϕ. Let us denote with $L_W(\hat{W}, \rho) = \mathcal{P}_W \mathcal{P}_{\hat{W}}(\ell(W, \hat{W}) < \rho)$, a functional ϑ of particular interest to us is the one that leads to (a function of) the so-called small-ball probability

$$L_W(\rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} L_W(\hat{w}, \rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} \mathcal{P}_W(\ell(W, \hat{w}) < \rho). \tag{3}$$

More generally, the choice of ϑ will depend on the choice of ϖ and vice versa. The discussion just above represents a generalisation of the approach undertaken in Xu and Raginsky (2017). In particular, in Xu and Raginsky (2017) the authors also target a bound similar to Equation (2). However, the choice of ϖ is essentially made so that the lower bound involves either Shannon's Mutual Information (simple or conditional) (Xu and Raginsky, 2017, Theorem 1) or the information density (simple or conditional) (Xu and Raginsky, 2017, Theorem 2). As a consequence, this also fixes the choice of ϑ. Our purpose is to allow for a wider range of functionals to provide stronger results. In the following sections, we will explore different choices of said functionals that lead to interesting results in the field.

2 Preliminaries

In this section, we will define the main objects utilised throughout the document and define the relevant notation. We will mainly adopt a measure-theoretic framework. Given a

measurable space $(\mathcal{X}, \mathcal{F})$ and two measures μ, ν which render it a measure space, if ν is absolutely continuous with respect to μ (denoted with $\nu \ll \mu$) then we will represent with $\frac{d\nu}{d\mu}$ the Radon-Nikodym derivative of ν with respect to μ.
Given a (measurable) function $f : \mathcal{X} \to \mathbb{R}$ and a measure μ, adopting the De Finetti's notation, we denote the Lebesgue integral of f with respect to the measure μ as follows

$$\mu(f) = \langle f, \mu \rangle = \int f \, \mathrm{d}\mu.$$

Consequently, given the bijection between events and indicators functions, with a slight abuse of notation, we also denote measures of events as follows

$$\mu(\mathbb{1}_E) = \int \mathbb{1}_E \, \mathrm{d}\mu = \mu(E).$$

The object $\mu(f)$ represents a bilinear inner product which will characterise a pairing between a (properly defined) space of functions and a (properly defined) space of measures. Once such a pairing is set, one can then proceed to define the Legendre-Fenchel transform connecting functionals acting on measures to functionals acting on functions. More formally, let $C_b(\mathcal{X})$ denote the space of continuous and bounded functions defined on $\mathcal{X}$ and $\mathcal{M}(\mathcal{X})$ the set of Radon measures defined on the same space, then one has that $\mathcal{M}(\mathcal{X})$ and $C_b(\mathcal{X})$ are in separating duality through the bilinear mapping $\langle \cdot, \cdot \rangle$ defined above (see Rassoul-Agha and Seppäläinen (2015)). Thus, given a functional $\psi : C_b(\mathcal{X}) \to \mathbb{R}$ one can define its Legendre-Fenchel dual as follows:

$$\psi^\star(\mu) = \sup_{f \in C_b(\mathcal{X})} \langle f, \mu \rangle - \psi(f). \tag{4}$$

Another connection between the spaces of interest comes from considering a norm on a space and the corresponding dual norm on the dual space, *i.e.,* given a norm acting on $\mathcal{X}$, $\|\cdot\|$ and a pairing between two spaces $(\mathcal{X}, \mathcal{Y})$, one can construct a norm on $\mathcal{Y}$ as follows:

$$\|h\|_\star = \sup_{f : \|f\| \leq 1} |\langle h, f \rangle|. \tag{5}$$

For this paper, we will essentially interpret the expected value $\mathcal{P}_{W\hat{W}}(\ell)$ as $\langle \mathcal{P}_{W\hat{W}}, \ell \rangle$. Once this simple observation is made, these tools will allow us to connect functionals of measure (*e.g.*, information-measures) to functionals of the loss (*e.g.*, small-ball probabilities). More details about this connection can be found in Appendix A.

3 Information Measures

In this section, we will introduce information measures and the necessary tools utilised to provide the main results of this work.

3.1 Rényi's Information Measures

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. It can be defined as follows (van Erven and Harremoës, 2014).

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:*

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu. \tag{6}$$

The definition can be proved to be independent of the chosen measure μ, see (van Erven and Harremoës, 2014). Moreover, one has that if $\alpha > 1$ then $D_\alpha(\mathcal{P}\|\mathcal{Q}) \geq D(\mathcal{P}\|\mathcal{Q})$, where $D(\mathcal{P}\|\mathcal{Q})$ denotes the Kullback-Leibler divergence between $\mathcal{P}$ and $\mathcal{Q}$. Under mild additional conditions one can also prove that $D_\alpha(\mathcal{P}\|\mathcal{Q}) \xrightarrow{\alpha\to1} D(\mathcal{P}\|\mathcal{Q})$. For an extensive treatment of Rényi's α-divergences, we refer the reader to (van Erven and Harremoës, 2014). Starting from Rényi's divergence, Sibson built the notion of Information Radius (Sibson, 1969). A deconstructed and generalised version of the Information Radius leads us to the following definition of a generalisation of Shannon's Mutual Information (Verdú, 2015):

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:*

$$I_\alpha(X,Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y). \tag{7}$$

In particular, Equation (7) admits the following closed-form expression:

$$I_\alpha(X,Y) = \frac{\alpha}{\alpha-1}\log\left\|\left\|\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right\|_{L^\alpha(\mathcal{P}_X)}\right\|_{L^1(\mathcal{P}_Y)} \tag{8}$$

Notice that $I_\alpha(X,Y) \neq I_\alpha(Y,X)$ *i.e.*, differently from Shannon's Mutual Information, I_α is not symmetric in its arguments. Similarly to Rényi's α-divergences one has that if $\alpha > 1$ then $I_\alpha(X,Y) \geq I(X;Y)$, where $I(X;Y)$ denotes the Shannon's Mutual Information. Under mild additional conditions one can also prove that $I_\alpha(X,Y) \xrightarrow{\alpha\to1} I(X;Y)$. Moreover, the limit of $\alpha \to \infty$ is also meaningful. Indeed, $I_\infty(X,Y)$ has gained independent interest in Privacy and Security. It goes under the name of Maximal Leakage, is denoted by $\mathcal{L}(X{\to}Y)$ and it has been endowed with an operational meaning (Issa et al., 2020). In particular one has that:

$$\mathcal{L}(X{\to}Y) = \log\mathcal{P}_Y\left(\operatorname*{ess\,sup}_{\mathcal{P}_X}\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right) \overset{X,Y\,\text{discrete}}{=} \log\sum_y \max_x P_{Y|X=x}(y). \tag{9}$$

Maximal Leakage also admits a conditional version that, for discrete random variables X, Y, Z, has the following form (Issa et al., 2020, Definition 6 & Theorem 6):

$$\mathcal{L}(X{\to}Y|Z) = \log\max_{z:\mathcal{P}_Z(z)>0}\sum_y\max_{x:\mathcal{P}_{X|Z}(x|z)>0}\mathcal{P}_{Y|X,Z}(y|x,z), \tag{10}$$

with an associated chain-rule-like inequality

$$\mathcal{L}(X{\to}(Y,Z)) \leq \mathcal{L}(X{\to}Y) + \mathcal{L}(X{\to}Z|Y). \tag{11}$$

For an extensive treatment of Sibson's α-Mutual Information we refer the reader to Verdú (2015) while for Maximal Leakage the reader is referred to Issa et al. (2020).

3.2 φ-Mutual Information

Another generalisation of the KL-divergence can be obtained by considering a generic convex function $\varphi : \mathbb{R}^+ \to \mathbb{R}$, usually with the simple constraint that $\varphi(1) = 0$. The constraint can be ignored as long as $\varphi(1) < +\infty$ by simply considering a new mapping $\tilde{\varphi}(x) = \varphi(x) - \varphi(1)$.

Definition 3. *Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\varphi : \mathbb{R}^+ \to \mathbb{R}$ be a convex function. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$. Denoting with p, q the densities of the measures with respect to μ, the φ-divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:*

$$D_\varphi(\mathcal{P}\|\mathcal{Q}) = \int q\varphi\left(\frac{p}{q}\right) \mathrm{d}\mu. \tag{12}$$

It is possible to show that φ-divergences are independent of the dominating measure (Liese and Vajda, 2006). Indeed, when absolute continuity between $\mathcal{P}, \mathcal{Q}$ holds *i.e.*, $\mathcal{P} \ll \mathcal{Q}$, an assumption we will often use, we retrieve the following (Liese and Vajda, 2006):

$$D_\varphi(\mathcal{P}\|\mathcal{Q}) = \int \varphi\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) \mathrm{d}\mathcal{Q}. \tag{13}$$

Examples of divergences included in this generalisation are:

- the Kullback-Leibler divergence by setting $\varphi(t) = t\log(t)$;
- the Total Variation distance with $\varphi(t) = \frac{1}{2}|t-1|$;
- the Hellinger distance with $\varphi(t) = (\sqrt{t}-1)^2$.

Denoting with $\mathcal{F}_X$ the σ-field generated from the random variable X, (i.e., $\sigma(X)$), φ-Mutual Information is defined as follows:

Definition 4. *Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$ over the 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 $\varphi : \mathbb{R}^+ \to \mathbb{R}$ be a convex function such that $\varphi(1) = 0$. The φ-Mutual Information between X and Y is defined as:*

$$I_\varphi(X, Y) = D_\varphi(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y). \tag{14}$$

If $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$ we have that:

$$I_\varphi(X, Y) = \int \varphi\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right) \mathrm{d}\mathcal{P}_X\mathcal{P}_Y. \tag{15}$$

It is possible to see that, if φ satisfies $\varphi(1) = 0$ and it is strictly convex at 1, then $I_\varphi(X, Y) = 0$ if and only if X and Y are independent (Liese and Vajda, 2006). Notice that, in general, $I_\varphi(X, Y) \neq I_\varphi(Y, X)$ *i.e.*, the information measure is not symmetric. Relevant to this work will be the family of Hellinger divergences (Liese and Vajda, 1987, Definition 2.10), that stem from the parametrised family of functions $\varphi_p(x) = (x^p - 1)/(p-1)$ with $p \geq 1$ and denoted in the following way:

$$\mathcal{H}_p(X, Y) = I_{\varphi_p}(X, Y). \tag{16}$$

Exploiting a bound involving $I_\varphi(X,Y)$ for a broad enough set of functions φ allows us to differently measure the dependence between X and Y. This allows us to provide bounds that are tailored to the specific problem at hand and, as we will see, to improve over bounds leveraging Shannon's Mutual Information.

3.3 Data-Processing Inequality

An important property that φ-divergences share is the Data-Processing Inequality (DPI). *I.e.*, given two measures μ,ν and a Markov Kernel K, one has that for every convex φ

$$D_\varphi(\nu K\|\mu K) \leq D_\varphi(\nu\|\mu). \tag{17}$$

This property holds as well for Rényi's α-divergences, despite them not being φ-divergences (van Erven and Harremoës, 2014, Theorem 9). An analogous version also holds for generalisation of Mutual Information and can be formulated as follows: assume that $X-Y-Z$ forms a Markov chain, then for every convex φ and every $\alpha\in(0,+\infty]$:

$$I_\varphi(X,Z) \leq \min\{I_\varphi(X,Y),I_\varphi(Y,Z)\} \quad \text{and} \quad I_\alpha(X,Z) \leq \min\{I_\alpha(X,Y),I_\alpha(Y,Z)\}. \tag{18}$$

In many cases, the inequality can be proved to be strict, and a large body of literature has been devoted to computing or bounding the so-called Strong Data-Processing Inequalities, see *e.g.*, Cohen et al. (1993); Ahlswede and Gács (1976); Polyanskiy and Wu (2017); Raginsky (2016); Makur and Zheng (2020). More details on the subject and how it can be leveraged in an estimation setting can be found in Section 4.1.

3.4 Functional Inequalities and Divergences

One of the key technical and conceptual tools is delineated in Appendix A. The key take-away is the following: interpreting $D_\varphi(\cdot\|\mu)=\psi_\mu(\cdot)$ as a convex and lower semi-continuous mapping, it is possible to characterise its variational representation (Theorem 22), allowing us to link divergences, expected values of functions and a corresponding functional. This allows us to retrieve Theorem 9. The other main tool that will be utilised is Hölder's inequality. In particular, there is a connection between Rényi's α-information measure and L^α-norms of the Radon-Nikodym derivative. Some of the results we are about to provide are a consequence of one or multiple applications of Hölder's inequality, in particular, one can prove the following:

Theorem 5. *Let $(\mathcal{X}\times\mathcal{Y},\mathcal{F},\mathcal{P}_{XY}),(\mathcal{X}\times\mathcal{Y},\mathcal{F},\mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY}\ll\mathcal{P}_X\mathcal{P}_Y$. Given an $\mathcal{F}$-measurable function $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}^+$, then,*

$$\mathcal{P}_{XY}(f) \leq \left\| \|f\|_{L^\beta(\mathcal{P}_X)} \right\|_{L^{\beta'}(\mathcal{P}_Y)} \cdot \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{L^\alpha(\mathcal{P}_X)} \right\|_{L^{\alpha'}(\mathcal{P}_Y)} \tag{19}$$

where $\beta,\alpha,\beta',\alpha'$ are such that $1=\frac{1}{\alpha}+\frac{1}{\beta}=\frac{1}{\alpha'}+\frac{1}{\beta'}$. Given a measurable function g, $\|g\|_{L^\alpha(\mu)}$ denotes the α-Norm of g under μ i.e., $\left(\int g^\alpha d\mu\right)^{\frac{1}{\alpha}}$.

Theorem 5 (and corresponding generalisations including Orlicz and Amemiya norms) has already appeared in Esposito et al. (2021a) in a slightly less general form, and in Esposito

(2022) in a variety of forms. It has been re-stated here for ease of reference. Moreover, Theorem 5 provides multiple degrees of freedom:

- the parameters characterising the norms: α, α';
- the (positive-valued) function f.

Three choices of the above are meaningful to us:

1. $\alpha' = \alpha$, which makes Rényi's divergence of order α appear on the right-hand side of Equation (19) (as a norm of the Radon-Nikodym derivative);
2. $\alpha' \to 1$, which makes Sibson's Mutual Information of order α appear on the right-hand side of Equation (19);
3. $f = \mathbb{1}_E$, which allows us to relate the probability of the same event under the joint and a function of the product of the marginals (and an information measure).

Selecting $\alpha' = \alpha$ and $f = \mathbb{1}_E$ gives rise to Esposito et al. (2021a, Corollary 6), while letting $\alpha' \to 1$ and selecting again $f = \mathbb{1}_E$ give rise to the following corollary:

Corollary 6 ((Esposito et al., 2021a, Corollary 1))**.** *Given $E \in \mathcal{F}$, we have that:*

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right)^{1/\beta} \cdot \mathcal{P}_Y\left(\mathcal{P}_X^{1/\alpha}\left(\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right)\right) \tag{20}$$

$$= \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right)^{1/\beta} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right), \tag{21}$$

where $I_\alpha(X,Y)$ is the Sibson Mutual Information of order α, (Verdú, 2015). Moreover, α and β are such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

Corollary 6 is pivotal in providing a family of lower bounds on the Bayesian risk *i.e.*, Theorem 8.

4 Main Results: lower bounds on the Risk

To provide our main results, let us state a fundamental lemma which is a simple consequence of Markov's inequality and allows us to provide lower bounds on the expected risk bounding small-ball probabilities instead:

Lemma 7. *Let W and $\hat{W}$ be two random variables jointly distributed according to $\mathcal{P}_{W,\hat{W}}$ and let $\ell : \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$. For every $\rho > 0$ the following holds true:*

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho(1 - \mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}) < \rho)). \tag{22}$$

Proof. One has that for every $\rho > 0$, due to Markov's inequality the following steps hold:

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho \cdot \mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}) \geq \rho) \tag{23}$$

$$= \rho \cdot (1 - \mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}) < \rho)). \tag{24}$$

□

An application of Corollary 6 in conjunction with Lemma 7 yields our first main result:

Theorem 8. *Consider the Bayesian framework described in Section 1.3. The following holds for every $\alpha > 1$ and $\rho > 0$:*

$$R_B \geq \rho\left(1 - \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(W,X) + \log(L_W(\rho))\right)\right)\right). \tag{25}$$

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

$$R_B \geq \sup_{\rho>0} \rho\left(1 - \exp\left(\mathcal{L}\left(W \to X\right) + \log(L_W(\rho))\right)\right). \tag{26}$$

The proof can be found in Appendix C.1. Two remarks are in order:

- It is important to notice that the behaviour of Equation (25) is fundamentally different from Xu and Raginsky (2017, Theorem 1). In Xu and Raginsky (2017, Theorem 1) the dependence is linear with respect to the Mutual Information and logarithmic in $L_W(\rho)$ while in Theorem 8 there is an exponential dependence on I_α and linear in $L_W(\rho)$.

- Theorem 8 introduces a new parameter $\alpha > 1$ to optimise over. The presence of α leads to 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 with α.

An interesting characteristic of Equation (26) 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 the risk. Indeed, ignoring $L_W(\rho)$ for a moment, 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 W remains the same). We can also walk the same path undertaken in Esposito et al. (2021a) and derive a variety of lower bounds involving a variety of information measures.

Theorem 9. *Consider the Bayesian framework described in Section 1.3. Let $\varphi : [0, +\infty) \to \mathbb{R}$ be a monotone, strictly convex function and suppose that the generalised inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$, exists. Then for every $\rho > 0$ and every estimator $\hat{W}$, if φ is non-decreasing one has the following*

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - L_W(\hat{W},\rho)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W}) + (1 - L_W(\hat{W},\rho))\cdot\varphi^\star(0)}{L_W(\hat{W},\rho)}\right)\right), \tag{27}$$

while if φ is non-increasing one recovers the following:

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - L_W(\hat{W},\rho)\right)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W}) + L_W(\hat{W},\rho)\cdot\varphi^\star(0)}{1 - L_W(\hat{W},\rho)}\right). \tag{28}$$

The proof can be found in Appendix C.2.

Remark 10 (Recovering Mutual Information). *A natural question is whether Theorem 9 also includes Shannon's Mutual Information (and, consequently, the results in (Xu and Raginsky, 2017)). Selecting $\varphi(x) = x\log x$ is problematic as the function is non-monotonic and its inverse would not have a closed-form expression one could leverage in Equation* (27) *and* (28)*. However, following the same steps undertaken in Appendix C.2, with $\varphi(x) = x\log(x)$, but with a different choice of $f = -\tilde{\lambda}\ell - \log \mathcal{P}_W\mathcal{P}_{\hat{W}}(\exp(-\tilde{\lambda}\ell)) + 1$ and $\tilde{\lambda} = -\frac{1}{\rho}\log\left(\mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell < \rho\})\right)$, Equation* (110) *does lead to (Xu and Raginsky, 2017, Theorem 1).*

Whenever $\varphi^\star(0) \leq 0$, the expressions in Equations (27) and (28) can be respectively simplified as follows: if φ is non-decreasing, Equation (27) specialises to:

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - L_W(\hat{W},\rho)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W})}{L_W(\hat{W},\rho)}\right)\right), \tag{29}$$

while if φ is non-increasing, Equation (28) specialises to:

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - L_W(\hat{W},\rho)\right)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W})}{1 - L_W(\hat{W},\rho)}\right). \tag{30}$$

The assumption that $\varphi^\star(0) \leq 0$ holds indeed true in a variety of cases (cf. Corollary 11).

Although Theorem 9 represents a quite general result, in order to apply it to the Bayesian risk setting (and provide an estimator-independent lower bound) one has to select φ carefully. In particular, one has to render the right-hand side of Equation (27) (or Equation (29)) independent of $\hat{W} = \phi(X)$. In order to do that, the following two quantities need to be rendered independent of $\hat{W}$:

1. The information-measure (*e.g.*, through the data-processing inequality $I_\varphi(W,\hat{W}) \leq I_\varphi(W,X)$);
2. The quantity $L_W(\hat{W},\rho)$ (which 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 (29) and introduce the following object

$$G_\varphi(I_\varphi, L_W) = L_W(\hat{W},\rho)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W})}{L_W(\hat{W},\rho)}\right). \tag{31}$$

To use the two inequalities just stated in Item 1 and Item 2, one needs that for a given choice of φ, $G_\varphi(I_\varphi, L_W)$ is increasing in I_φ for a given value of L_W and increasing in L_W for a given value of I_φ. This allows us to further lower bound Equation (29) and render the quantity independent of the specific choice of ϕ. Analogously, one can state similar assumptions in order to apply the same reasoning to Equation (30). Hence, starting from Equation (1) one can provide a lower bound on the risk R_B that is independent of ϕ.

Let us now look at some specific choices of φ such that G_φ satisfies the desired properties and, thus, for which a bound on the Bayesian risk can be retrieved.

Corollary 11. *Consider the Bayesian framework described in Section 1.3. The following holds for every $p > 1$ and $\rho > 0$:*

$$R_B \geq \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(W,X)+1)^{\frac{1}{p}}\right). \tag{32}$$

The proof of Corollary 11 is in Appendix C.3. Restricting the choice of φ to the family of polynomials φ_p that gives rise to the Hellinger divergences (see Equation (16)) one can thus state the following lower bound on the risk:

$$R_B \geq \sup_{\rho>0} \sup_{p>1} \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \cdot \left((p-1)\mathcal{H}_p(W,\hat{W})+1\right)^{\frac{1}{p}}\right). \tag{33}$$

Remark 12. *Using the one-to-one mapping connecting Hellinger divergences and Rényi's α-divergence (Sason and Verdú, 2016, Eq. (80)) one can rewrite Equation* (32) *for a given p as follows:*

$$R_B \geq \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \cdot \exp\left(\frac{p-1}{p} D_p(\mathcal{P}_{W\hat{W}} \| \mathcal{P}_W \mathcal{P}_{\hat{W}})\right)\right). \tag{34}$$

Moreover, given the definition of Sibson's α-Mutual Information one has that for a given $\alpha > 1$:

$$\begin{aligned}
\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,X)\right) &= \exp\left(\frac{\alpha-1}{\alpha} \inf_{\mathcal{Q}_X} D_\alpha(\mathcal{P}_{WX} \| \mathcal{P}_W \mathcal{Q}_X)\right) &(35)\\
&\leq \exp\left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{WX} \| \mathcal{P}_W \mathcal{P}_X)\right) &(36)\\
&= ((\alpha-1)\mathcal{H}_\alpha(W,X)+1)^{\frac{1}{\alpha}}. &(37)
\end{aligned}$$

Consequently, one has that:

$$\begin{aligned}
R_B &\geq \left(1 - L_W(\rho)^{\frac{\alpha-1}{\alpha}} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,X)\right)\right) &(38)\\
&\geq \left(1 - L_W(\rho)^{\frac{\alpha-1}{\alpha}} \cdot ((\alpha-1)\mathcal{H}_\alpha(W,X)+1)^{\frac{1}{\alpha}}\right). &(39)
\end{aligned}$$

Hence, Equation (25) *always improves over Equation* (32)*. However, as we will see, in a variety of settings, the Hellinger α-divergence can be computed explicitly while I_α cannot. For this reason, we will often leverage the Hellinger divergence to provide closed-form lower bounds on the risk even though they do not yield the best lower bound.*

Several results can be derived from Theorem 9. Each of them has potentially interesting applications in specific Bayesian Estimation settings. In this work, we will mostly focus on Sibson's α-Mutual Information and Hellinger p-divergences. In the spirit of leveraging the generality of Theorem 9, we also provide a bound involving a novel information measure $E_{\gamma,\zeta}$, strongly inspired by the so-called E_γ-divergence (Sason and Verdú, 2016, Equation (66)), (Polyanskiy et al., 2010, page 2314), also known in the literature as the Hockey-stick divergence. Applications of the Hockey-Stick divergence in this framework have been explored in (Asoodeh et al., 2021). Its definition is the following:

Definition 13. *Let $(\Omega, \mathcal{F})$ be a measurable space and let μ and ν be two probability measures defined on the space. Denote with $\varphi_{\gamma,\zeta}(x) = (\zeta x - \gamma)_+ - (\zeta - \gamma)_+$ with $\zeta > 0$, $\gamma \geq 0$, and where $(x)_+ = \max\{0, x\}$. The function $\varphi_{\gamma,\zeta}(x)$ is convex, increasing and is such that $\varphi_{\gamma,\zeta}(1) = 0$. Assume that $\nu \ll \mu$, then define the following object:*

$$E_{\gamma,\zeta}(\nu\|\mu) = D_{\varphi_{\gamma,\zeta}}(\nu\|\mu). \tag{40}$$

Moreover, whenever $\nu = \mathcal{P}_{XY}$ and $\mu = \mathcal{P}_X\mathcal{P}_Y$ we denote (with a slight abuse of notation) $E_{\gamma,\zeta}(\mathcal{P}_{XY}\|\mathcal{P}_{XY})$ with $E_{\gamma,\zeta}(X, Y)$. If $\zeta = 1$ then one recovers the usual E_γ-divergence.

Leveraging it, one can provide the following result in this framework:

Corollary 14. *Consider the Bayesian framework described in Section 1.3. The following holds for every $\zeta > 0$, $\gamma \geq 0$, and $\rho > 0$:*

$$R_B \geq \rho \left(1 - \frac{E_{\gamma,\zeta}(W, \hat{W}) + \gamma L_W(\rho) + (\zeta - \gamma)_+}{\zeta}\right). \tag{41}$$

One can thus retrieve the following lower bound on the risk:

$$R_B \geq \sup_{\rho>0} \sup_{\zeta>0,\gamma\geq 0} \rho \left(1 - \frac{E_{\gamma,\zeta}(W, \hat{W}) + \gamma L_W(\rho) + (\zeta - \gamma)_+}{\zeta}\right). \tag{42}$$

The proof is in Appendix C.4.

Remark 15. *Setting $\zeta = 1$ in Equation (41) one recovers (Asoodeh et al., 2021, Remark 1). In fact, by introducing an additional degree of freedom through the ζ parameter in Equation (42), the resulting lower bound can only be tighter than (Asoodeh et al., 2021, Remark 1).*

Using these results one can provide meaningful lower bounds on the risk in a variety of settings of interest, as we will see in Section 5. Some natural extensions over the framework introduced are presented in Appendix E. They consider either a slight change of perspective or a slight alteration of the observation model. We will now see how our bounds can be improved if one has more information on the type of noise present in the observation channels or if the samples used for estimation are privatised.

4.1 Leveraging Strong Data-Processing Inequalities

A key step in the results proved here consists of leveraging the Markov Chain $W - X - \hat{W}$ along with the Data-Processing Inequality as follows: $I_\varphi(W, \hat{W}) \leq I_\varphi(W, X)$. For more details on DPI see Section 3.3. In case more information is available concerning the kernel linking X and $\hat{W}$ then one can leverage the so-called Strong Data-Processing Inequality (SDPI), a tightening of the classical DPI. In particular, in many settings of interest, one can show that $I_\varphi(W, \hat{W})$ is strictly smaller than $I_\varphi(W, X)$ unless $\hat{W} = X$ and the characterisation of the ratio between these two quantities for Markov Kernels can be formalised via SDPIs:

Definition 16 (Raginsky 2016, Definition 3.1)**.** *Given a probability measure μ, a Markov Kernel K and a convex function φ, we say that K satisfies a φ-type strong data-processing inequality at μ with constant $c \in [0,1)$ if, for all $\nu \ll \mu$ one has that*

$$D_\varphi(\nu K \| \mu K) \leq c \cdot D_\varphi(\nu \| \mu). \tag{43}$$

In order to characterise the tightest such constant c, let us define the following objects:

$$\eta_\varphi(\mu, K) = \sup_{\nu \neq \mu} \frac{D_\varphi(\nu K \| \mu K)}{D_\varphi(\nu \| \mu)} \quad \textit{and} \quad \eta_\varphi(K) = \sup_\mu \eta_\varphi(\mu, K).$$

Moreover, under mild condition on φ one can also prove that if $U - X - Y$ forms a Markov chain then:

$$\sup_{P_{U|X}} \frac{I_\varphi(U,Y)}{I_\varphi(U,X)} = \eta_\varphi(P_Y, P_{Y|X}). \tag{44}$$

Said quantities are generally hard to compute for a given Markov Kernel K and functional φ, however, a variety of bounds is present in the literature (see Raginsky (2016)). In particular, given any convex φ it is possible to show the following result:

Lemma 17 (Del Moral et al. 2003, Proposition 1.1.)**.** *Let $K : \mathcal{F} \times \Omega \to [0,1]$ be a Markov Kernel, and let φ be a convex functional such that $\varphi(1) = 0$, one has that*

$$\eta_\varphi(K) \leq \eta_{\mathrm{TV}}(K) = \sup_{x, \hat{x} \in \Omega} \mathrm{TV}(K(\cdot|x), K(\cdot|\hat{x})), \tag{45}$$

where TV *denotes the Total Variation distance i.e., the φ divergence that stems from $\varphi(t) = \frac{1}{2}|t-1|$.*

The bound in Equation (45) does not, however, hold for Rényi's divergences:

Example 1. *Let $\mu = (1/2, 1/2)$ and $K = BSC(\lambda)$ with $\lambda < \frac{1}{2}$. Then $\eta_{\mathrm{TV}}(K) = (1-2\lambda)$. Consider now $D_\alpha(K(\cdot|0)\|\mu) = D_\alpha(\delta_0 K \| \mu K) = \frac{1}{\alpha-1}\log(2^{1-\alpha}(\lambda^\alpha + (1-\lambda)^\alpha))$. Moreover, $D_\alpha(\delta_0\|\mu) = -\log(2)$. If $\lambda = 0.2$ and $\alpha = 6$ one has that*

$$\eta_{D_\alpha}(K) > \frac{D_\alpha(\delta_0 K \| \mu K)}{D_\alpha(\delta_0 \| \mu)} = 0.6138 > \vartheta(K) = 0.6, \tag{46}$$

and the gap increases with α.

In the specific context of estimation problems, one can leverage SDPI coefficients in the following way:

$$I_\varphi(W, \hat{W}) \leq I_\varphi(W, X)\eta_\varphi(\mathcal{P}_X, \mathcal{P}_{\hat{W}|X}) \leq I_\varphi(W, X)\eta_\varphi(\mathcal{P}_{\hat{W}|X}),$$

This can potentially provide a refinement of the results presented so far. Moreover, the same technique can be employed in settings where one does not have direct access to samples but has rather access to noisy copies or privatised versions. In this case, one can provide lower bounds tailored to the type of noise that has been injected to privatise the data. Consider the following setting in which one has access to n independent samples $X^n = (X_1, \dots, X_n)$

generated from W. Moreover, assume the samples are not directly observed but rather one has access to a sequence Z^n of noisy/privatised version of X^n (obtained through the sequence of Markov kernels $K_1, \dots, K_n$, where for every $i \geq 1$, $K_i : \mathcal{F} \times \mathcal{X} \to [0,1]$ and $\mathcal{P}_{Z_i} = \mathcal{P}_{X_i} K_i$). The goal is to estimate W from the sequence Z^n via an estimator $\psi : \mathcal{Z}^n \to \mathcal{W}$. Given a loss function $\ell : \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$, the noisy Bayesian risk is thus defined as

$$R_B^{\text{noisy}} = \inf_\psi \mathcal{P}_{WZ}(\ell(W, \psi(Z^n))) = \inf_\psi \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})), \tag{47}$$

and it can be lower bounded similarly to the non-private/noisy case via Theorem 9. For simplicity of exposition, we will consider a single kernel K and, consequently, one has that Z^n is obtained from X^n through the tensor-product $K^{\otimes n}$. In this case, one has the Markov chain $W - X^n - Z^n - \hat{W}$ and can leverage SDPI twice as follows:

$$I_\varphi(W, \hat{W}) \leq \eta_\varphi(\mathcal{P}_{W|Z^n}) I_\varphi(W, Z^n) \leq \eta_\varphi(\mathcal{P}_{W|Z^n}) \eta_\varphi(\mathcal{P}_{X^n}, K^{\otimes n}) I_\varphi(W, X^n), \tag{48}$$

and consequently, state the following result:

Corollary 18. *Consider the private Bayesian framework considered above. Denote with Z^n the private samples obtained from X^n through the tensor product of a kernel K. Let $\varphi : [0, +\infty) \to \mathbb{R}$ be a monotone convex function and suppose that the generalised inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$, exists. Assume as well that the function G_φ defined in Equation (31) is non-decreasing in both arguments. Then, for every estimator $\hat{W}$, if φ is non-decreasing one has the following*

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \sup_{\rho > 0} \rho \Bigg(1 - L_W(\hat{W}, \rho) \\ \cdot \varphi^{-1} \left(\frac{\eta_\varphi(\mathcal{P}_{\hat{W}|Z^n}) \eta_\varphi(\mathcal{P}_{X^n}, K^{\otimes n}) I_\varphi(W, X^n) + (1 - L_W(\rho)) \varphi^\star(0)}{L_W(\rho)} \right) \Bigg), \tag{49}$$

whereas if φ is non-increasing one recovers the following:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \sup_{\rho > 0} \rho \left(1 - L_W(\hat{W}, \rho) \right) \\ \cdot \varphi^{-1} \left(\frac{\eta_\varphi(\mathcal{P}_{\hat{W}|Z^n}) \eta_\varphi(\mathcal{P}_{X^n}, K^{\otimes n}) I_\varphi(W, X^n) + L_W(\rho) \cdot \varphi^\star(0)}{1 - L_W(\rho)} \right). \tag{50}$$

Notice that in case one does *not* estimate from noisy observations Z^n then one can still consider the same setting as in Corollary 18 with K representing the kernel associated to the identity mapping *i.e.*, $K(y|x) = \delta_x(y)$. In this case one has that $\mathcal{P}_{X^n} = \mathcal{P}_{Z^n}$, $\eta_\varphi(\mathcal{P}_{X^n}, K^{\otimes n}) = 1$ and, consequently, $\eta_\varphi(\mathcal{P}_{\hat{W}|Z^n}) = \eta_\varphi(\mathcal{P}_{\hat{W}|X^n})$. Hence, Corollary 18 boils down to a refinement of Theorem 9 without any additional noise injection.

Remark 19 (Connection to Local-Differential Privacy)**.** *In case one is considering $E_{\gamma,\zeta}$ with $\zeta = 1$, then the corresponding contraction parameter has been analysed in Asoodeh* *et al. (2021), where contraction has been shown to be equivalent to Local Differential-Privacy (LDP). I.e., a kernel K is said to be (ϵ,δ) "Locally-Differentially Private" (LDP), if:*

$$\max_{E\in\mathcal{F},x,\hat{x}} |K(E|x) - e^{\epsilon}K(E|\hat{x})| \leq \delta. \tag{51}$$

In this case, one has that $\eta_{\varphi_{e^{\epsilon}}}(K) \leq \delta$. Moreover, due to (Asoodeh et al., 2021, Lemma 2) one has that if K is (ϵ,δ)-LDP then for every φ one has that $\eta_{\varphi}(K) \leq (1-(1-\delta)e^{-\epsilon})$. It is unclear, however, whether there are settings in which said upper-bound is more convenient than others. Indeed, one also has that for every convex function φ

$$\eta_{\varphi}(K) \leq \eta_{\mathrm{TV}}(K) < (1-(1-\delta)e^{-\epsilon}), \tag{52}$$

and, for many channels, $\eta_{\mathrm{TV}}(K)$ is relatively easy to compute. Moreover, even $\eta_{\mathrm{TV}}(K)$ tends to be quite larger than the effective contraction coefficient of the divergence at hand: e.g., if $K = BSC(\lambda_{\epsilon})$ with $\lambda_{\epsilon} = \frac{1}{1+e^{\epsilon}}$ then K is $(\epsilon,0)$-LDP (see (Asoodeh et al., 2021, Example 1)) and if φ is operator-convex (e.g., $\varphi(x) = x\log x$ or $\frac{x^p-1}{p-1}$with $1 < p \leq 2$, etc.) then (Raginsky, 2016, Corollary 3.1):

$$\eta_{\varphi}(K) = \left(1-\frac{2}{1+e^{\epsilon}}\right)^2 \ll \left|1-\frac{2}{1+e^{\epsilon}}\right| < (1-e^{-\epsilon}). \tag{53}$$

An important feature of Corollary 18 is that for a subclass of functions φ one can leverage tensorisation properties of η_{φ}. Indeed, if φ satisfies the conditions of (Raginsky, 2016, Theorem 3.9) then

$$\eta_{\varphi}(\mu^{\otimes n}, K^{\otimes n}) = \eta_{\varphi}(\mu, K). \tag{54}$$

Moreover, if φ is operator convex, then one can say the following:

$$\eta_{\varphi}(K^{\otimes n}) = \eta_{\chi^2}(K^{\otimes n}) \leq 1-(1-\eta_{\varphi}(K))^n. \tag{55}$$

Both the results are true for instance, for $\varphi(x) = (x^p-1)/(p-1)$ with $1 \leq p \leq 2$ (but the assumptions of (Raginsky, 2016, Theorem 3.9) are violated if $p > 2$). This means that one can leverage Equations (54) to (55) for the Hellinger divergence $\mathcal{H}_p$ with $1 \leq p \leq 2$.

5 Examples of application

In this section, we apply the results presented in the previous section to four estimation settings. The first three are classical settings, while the fourth comes from a distributed estimation setting:

- estimation of the mean of a Bernoulli random variable with parameter W, where W is assumed to be uniform between $(0,1)$;
- the same setting as above with the difference that one does not observe the samples X^n directly but rather a noisy/privatised version Z^n, where each Z_i is assumed to be the outcome of X_i after being passed through a Binary Symmetric Channel with parameter λ (BSC(λ));
- estimation of the mean of a Gaussian random variable with Gaussian prior;

- identification of the biased random variable in a d-dimensional vector in a distributed fashion (cf., the "Hide-and-seek" problem advanced in Shamir (2014)).

The loss function for the first three cases will be the L^1-distance while for the fourth one, we will consider the $0-1$ loss. For the first three cases, the maximisation over ρ in the lower bounds is carried out analytically (details in Appendix D.1).

5.1 Bernoulli Bias

Example 2. *Suppose that $W \sim \mathcal{U}([0,1])$ and that for each $i \in [n]$, the random variables $X_i|W = w$ are distributed according to a* Ber(w)*, i.e., $P(X_i = 1|W = w) = w$ and $P(X_i = 0|W = w) = 1 - w$. Also, assume that $\ell(w, \hat{w}) = |w - \hat{w}|$.*

Using the sample mean estimator, *i.e.*, $\hat{W} = \frac{1}{n}\sum_{i=1}^{n} X_i$, one has that (see (Xu and Raginsky, 2017, Equation (20))):

$$R_B \leq \frac{1}{\sqrt{6n}}. \tag{56}$$

Let us now lower bound the risk in this setting. First, we find that

$$L_W(\rho) = \sup_{\hat{w}} \mathcal{P}_W(|W - \hat{w}| < \rho) = 2\rho. \tag{57}$$

To obtain a lower bound involving Maximal Leakage, one can see that (details in Section D.2.1)

$$\mathcal{L}\left(W \to X^n\right) \leq \log\left(2 + \sqrt{\frac{\pi n}{2}}\right). \tag{58}$$

Substituting Equation (58) in Equation (26), along with $L_W(\rho) = 2\rho$, provides us with the following lower bound on the risk:

$$R_B \geq \sup_{\rho > 0} \rho \left(1 - \exp\left(\mathcal{L}\left(W \to X^n\right)\right) L_W(\rho)\right) \tag{59}$$

$$\geq \sup_{\rho > 0} \rho \left(1 - \left(2 + \sqrt{\frac{\pi n}{2}}\right) 2\rho\right). \tag{60}$$

The quantity in Equation (60) is a concave function of ρ and thus we can maximise it. In particular, the maximiser is $\hat{\rho} = \frac{1}{4\left(2+\sqrt{\frac{\pi n}{2}}\right)}$ and plugging it in Equation (60) one gets the following:

$$R_B \geq \frac{1}{8\left(2 + \sqrt{\frac{\pi n}{2}}\right)}, \tag{61}$$

which, for n large enough (*i.e.*, $n \geq 127/\pi \approx 41$), can be further lower bounded as follows

$$R_B \geq \frac{1}{5\sqrt{2\pi n}}. \tag{62}$$

Surprisingly, Maximal Leakage already offers a lower bound that matches the upper bound up to a constant (see Equation (56)) without any extra machinery. Equation (61) provides a larger lower bound than the one provided using Mutual Information (see Xu and Raginsky

(2017, Corollary 2)) for $n \geq 1$. Moreover, the proof in Xu and Raginsky (2017) needs a more complicated setting involving a conditioning with respect to an independent copy of X^n and can only provide an *asymptotic* lower bound on the risk of $1/(16\sqrt{2\pi n})$ while Equation (61) holds for every n.
On the contrary, given the closed-form expression, Maximal Leakage can be quite easy to compute or upper-bound. Moreover, the information measure depends on $\mathcal{P}_W$ only through the support. This means that if one has access to an upper-bound on $L_W(\rho)$ that does not employ any knowledge of $\mathcal{P}_W$ except for the support (*e.g.*, if W were to be discrete, an upper-bound of 1 over the probability mass function could suffice) the resulting lower bound on the risk (in this example), would apply to any W whose support is the interval $[0,1]$.
One can also provide a more general lower bound involving I_α. Indeed, one has that (details in Section D.2.2), in this setting:

$$\exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W,X^n)\right) = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{\Gamma(k\alpha+1)\Gamma((n-k)\alpha+1)}{\Gamma(n\alpha+2)}\right)^{\frac{1}{\alpha}}. \tag{63}$$

Plugging Equation (63) in Equation (25) one obtains the following lower bound on the risk:

$$R_B \geq \sup_{\rho>0}\sup_{\alpha>1}\rho\left(1-(2\rho)^{\frac{\alpha-1}{\alpha}}\exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W,X^n)\right)\right). \tag{64}$$

The lower bound in Equation (64) can clearly only improve the one provided in Equation (60), as $\mathcal{L}(W \!\to\! X^n) = I_\infty(W,X^n) > I_\alpha(W,X^n)$ for every $\alpha < \infty$. However, differently from Equation (60), it does not admit a closed-form expression and needs to be computed numerically in order to assess how far it is from the upper bound. Similarly, one could try to employ a lower bound that includes Hellinger$-p$ divergences. The lower bound on the risk induced by Corollary 11 is given by

$$R_B \geq \sup_{\rho>0}\sup_{p>1}\rho\left(1-(2\rho)^{\frac{p-1}{p}}\cdot(\mathcal{H}_p(W,X^n))^{\frac{1}{p}}\right). \tag{65}$$

Via the argument delineated in Remark 12 one can see that Equation (64) always improves over Equation (65). However, for some values of p, one can provide a closed-form expression for the lower bound provided by Equation (65) while this is not possible for Equation (64). For this reason, we decided to explicitly state both results. Indeed, in general, one has that (details in Section D.2.3):

$$((p-1)\mathcal{H}_p(W,X^n)+1) = (n+1)^{p-1}\sum_{k=0}^{n}\binom{n}{k}^p\frac{\Gamma(kp+1)\Gamma((n-k)p+1)}{\Gamma(np+2)}, \tag{66}$$

Then, with $p=2$ one recovers (details in Section D.2.3):

$$\mathcal{H}_2(W,X^n)+1 = \chi^2(W,X^n) = \frac{n+1}{2n+1}\cdot\frac{4^n}{\binom{2n}{n}} \leq \frac{16\sqrt{\pi n}}{21}. \tag{67}$$

Hence, specialising Equation (65) to $p=2$ leads us to:

$$R_B \geq \sup_{\rho>0}\rho\left(1-\sqrt{2\rho(\chi^2(W,X^n)+1)}\right). \tag{68}$$

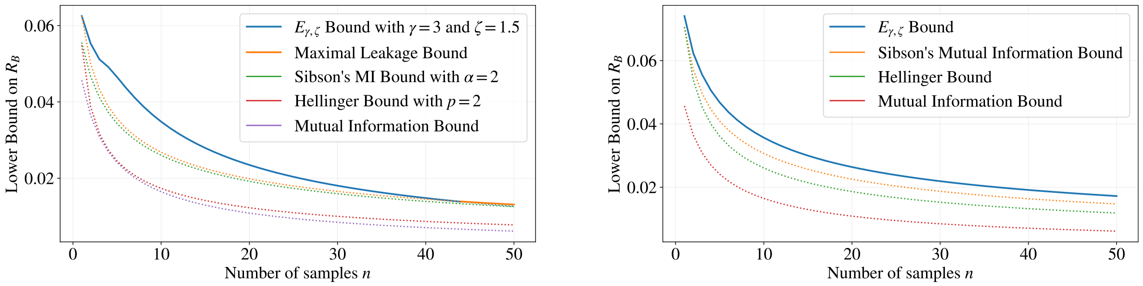
Solving then the maximisation over ρ and using Equation (67) one can conclude that:

$$R_B \geq \frac{2}{27} \cdot \frac{1}{\chi^2(W, X^n) + 1} \geq \frac{7}{72\sqrt{\pi n}}. \tag{69}$$

Notice that Equation (69) also matches the upper-bound up to a constant and, similarly to Maximal Leakage, improves over Xu and Raginsky (2017, Corollary 2) while not requiring that $n \to \infty$. 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 $R_B \gtrsim \frac{4}{27\sqrt{\pi n}}$, thus leading to a slight improvement over Equation (69). To conclude, one can apply the same steps with the $E_{\gamma,\zeta}$-divergence. The lower bound on the risk one can retrieve via Corollary 14 in this example can thus be expressed as

$$\begin{aligned} R_B &\geq \sup_{\zeta,\gamma} \sup_{\rho>0} \rho \left(1 - \frac{(E_{\gamma,\zeta}(W, X^n) + 2\rho\gamma)}{\zeta}\right) && (70) \\ &= \sup_{\zeta,\gamma} \frac{(\zeta - E_{\gamma,\zeta}(W, X^n))^2}{8\gamma\zeta}. && (71) \end{aligned}$$

The lower bound in Equation (71) can be empirically seen to be the best among the ones presented so far (thus beating Hellinger, I_α and, consequently, Maximal Leakage and Mutual Information). A direct comparison between the bounds provided here and those already present in the literature can be seen in Figure 1a and Figure 1b. 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, tighter) lower bound on the Bayesian risk when compared with results that stem from using Shannon's Mutual Information (see Xu and Raginsky (2017, Corollary 2)). In particular, the lower bound involving the $E_{\gamma,\zeta}$-Mutual Information represents the largest among the ones we consider. Given the lack of a closed-form expression for $E_{\gamma,\zeta}$ in this example, the quantity in Equation (71) was computed numerically (see Section D.2.4). Moreover, in order to verify whether the behaviour (and ordering) of the lower bounds in Figure 1a was determined by the specific choices of the parameters p, γ, ζ and α, in Figure 1b the lower bounds on the risk have also been numerically optimised over the respective parameters p, γ, ζ, α. As Figure 1b shows, the lower bound provided by $E_{\gamma,\zeta}$ remains the best. Notice that the lower bound involving Mutual Information has no parameter to optimise over (other than ρ). Maximal Leakage does not provide the best bound, but it possesses the interesting characteristic of depending on $\mathcal{P}_W$ only through the support, thus leading to potential applicability in a variety of settings in which $\mathcal{P}_W$ is not accessible. In contrast, Mutual Information, the Hellinger divergence and the $E_{\gamma,\zeta}$-divergence all require to know $\mathcal{P}_W$. The lower bounds on the risk in this Example can thus be summarised as follows:



(a) The picture shows the behaviour of Equation (60), Equation (64) with $\alpha = 2$, Equation (69), Equation (71) with $\gamma = 3$ and $\zeta = 1.5$ and (Xu and Raginsky, 2017, Equation (19)) as a function of n. The values of $E_{3,1.5}(W, X^n)$ for each n are computed numerically. A solid line means that the corresponding lower bound is the largest.

(b) Comparison between the largest lower bounds one can retrieve for different information measures in Example 2: that is between Equation (64), Equation (65), Equation (70) and (Xu and Raginsky, 2017, Equation (19)). The quantities are analytically maximized over ρ and numerically optimized over, respectively, $\alpha > 1, p > 1$, $\zeta > 0$, and $\gamma \geq 0$. A solid line means that the corresponding lower bound is the largest.

Figure 1: Comparison of various bounds for Example 2 with and without (numerical) optimisation of parameters.

Corollary 20. *In the setting described in Example 2 one has the following lower bound on the Bayesian risk:*

$$
\begin{aligned}
R_B \geq \max \Bigg\{ & \max_{\zeta>0,\gamma\geq 0} \left\{ \frac{(\zeta - E_{\gamma,\zeta}(W, X^n))^2}{8\gamma\zeta} \right\}, \\
& \max_{\alpha>1} \left\{ \left(\frac{(2\alpha-1)}{2\alpha} \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W, X^n) \right) \right)^{-\frac{\alpha}{\alpha-1}} \frac{(\alpha-1)}{(2\alpha-1)} \right\} \Bigg\}.
\end{aligned}
\tag{72}
$$

5.2 Noisy Bernoulli Bias

Assume, like in Section 5.1, that W is uniform on the $[0, 1]$ interval, $X_i \sim \text{Ber}(W)$. In line with the discussion above, suppose that one observes noisy copies of X_i's denoted with Z_i's, where Z_i is the outcome of X_i after being passed through a Binary Symmetric Channel with parameter λ (K=BSC(λ)). The purpose is to estimate W through a function of $Z_1, \ldots, Z_n$ *i.e.*, $\hat{W} = \psi(Z^n)$. One thus has the following Markov Chain $W - X^n - Z^n - \hat{W}$. In order to lower bound the Bayesian risk in this setting, one can use Corollary 18. In particular, given the additional injection of noise, it is to be expected that one has a stronger impossibility result with respect to the non-noisy version. This is reflected in the computations below. Let us restrict ourselves to φ-divergences as one can then leverage the results in the literature on SDPI constants for the channel considered here. In particular, if one considers Hellinger p-divergences then the following can be said about their associated

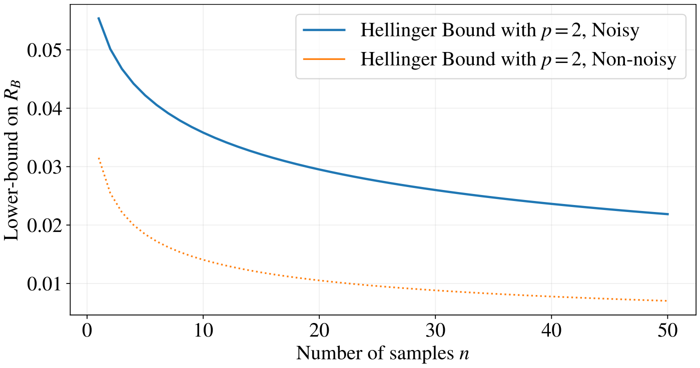


Figure 2: Comparison of the lower bounds in Equation (69) and Equation (78) for the noisy Bernoulli bias setting described in Section 5.2 with $\lambda = 0.25$. A solid line means that the corresponding lower bound is the largest.

SDPI-coefficients η_p (Raginsky, 2016, Corollary 3.1), (Cohen et al., 1993):

$$\eta_p(K) = (1-2\lambda)^2 \text{ if } 1 \le p \le 2 \tag{73}$$

$$\eta_p(K) \le |1-2\lambda| \quad \text{if } p > 2. \tag{74}$$

Moreover, if $p \le 2$ then one can leverage tensorisation properties of SDPI-coefficients (see Raginsky (2016, Section 3.5)) and the following can be said

$$\eta_p(\mathcal{P}_{X^n}, K^{\otimes n}) = \eta_p(\mathcal{P}_X, K) \le \eta_p(K) = (1-2\lambda)^2 \text{ if } 1 \le p \le 2 \tag{75}$$

In this setting, one has that the Hellinger divergence $\mathcal{H}_p(W, X^n)$ is given in Equation (66). With $p = 2$ and without making any assumption on $\mathcal{P}_{\hat{W}|Z^n}$, one can leverage Corollary 18 and retrieve the following closed-form expression for the risk in this setting:

$$R_B^{\text{noisy}} \ge \sup_{\rho>0} \rho \left(1 - \sqrt{2\rho((1-2\lambda)^2\chi^2(W,X^n)+1)}\right) \tag{76}$$

$$= \frac{2}{27} \frac{1}{(1-2\lambda)^2\chi^2(W,X^n)+1} \tag{77}$$

$$\ge \frac{2}{27} \frac{1}{(1-2\lambda)^2 \frac{16\sqrt{\pi n}}{21} + 1}. \tag{78}$$

Clearly, the denominator in Equation (78) is smaller than the one in Equation (69) thus yielding a larger lower bound on the risk, this is depicted in Figure 2 for the case $\lambda = 0.25$.

5.3 Gaussian prior with Gaussian noise (and absolute error)

Another classical and interesting setting is given by the following example:

Example 3. *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:

$$R_B \leq \sqrt{\frac{\sigma_W^2}{1+n\sigma_W^2/\sigma^2}}. \tag{79}$$

Moreover, given that $\ell(w,\hat{w}) = |w-\hat{w}|$ it is also possible to show that:

$$L_W(\rho) \leq \left(\sup_{w\in\mathbb{R}} \mathcal{P}_W(w)\right)\left(\int_{-\rho}^{\rho} 1\,\mathrm{d}u\right) \leq \rho\sqrt{\frac{2}{\sigma_W^2\pi}}. \tag{80}$$

In this setting, $\mathcal{L}\left(W\!\to\!X^n\right)$ is infinite. However, $I_\alpha(W,X^n)$ is finite for every $\alpha < +\infty$. One can thus provide a lower bound on the risk, resorting to I_α via Equation (25). Given that the empirical mean is a sufficient statistic for W in this case, one has that (Verdú, 2015, Example 5):

$$I_\alpha(W,X^n) = I_\alpha\left(W, \frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{2}\log\left(1+\alpha n\frac{\sigma_W^2}{\sigma^2}\right). \tag{81}$$

These considerations imply that :

$$R_B \geq \sup_{\alpha>1}\sup_{\rho>0} \rho\left(1-\exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W,X^n)\right)\left(\rho\sqrt{\frac{2}{\sigma_W^2\pi}}\right)^{\frac{\alpha-1}{\alpha}}\right) \tag{82}$$

$$= \sup_{\alpha>1}\sup_{\rho>0} \rho\left(1-\left(\rho\sqrt{\left(1+\alpha n\frac{\sigma_W^2}{\sigma^2}\right)\frac{2}{\sigma_W^2\pi}}\right)^{\frac{\alpha-1}{\alpha}}\right) \tag{83}$$

$$= \sup_{\alpha>1} \frac{1}{(\beta+1)}\left(\frac{\beta}{\beta+1}\right)^{\beta}\left(\sqrt{\left(1+\alpha n\frac{\sigma_W^2}{\sigma^2}\right)\frac{2}{\sigma_W^2\pi}}\right)^{-\frac{1}{\beta}}, \tag{84}$$

remembering that $\beta = \frac{\alpha}{\alpha-1}$.

Stepping away from Sibson's α-Mutual Information one can look at Hellinger p-divergences and $E_{\gamma,\zeta}$ once again. In particular, one has that for $p>1$ (details in Section D.3.1):

$$\mathcal{H}_p(W,X) = \left(\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^p}{1+(2-p)p\frac{\sigma_W^2}{\sigma^2}}\right)^{\frac{1}{2}}. \tag{85}$$

Thus, the family of bounds provided by Corollary 11 can be expressed as follows

$$R \geq \sup_{p>1}\sup_{\rho>0} \rho\left(1-\left(\frac{2\rho}{\sqrt{2\pi\sigma_W^2}}\right)^{\frac{p-1}{p}} \mathcal{H}_p^{\frac{1}{p}}(W,X^n)\right) \tag{86}$$

$$= \sup_{p>1}\sup_{\rho>0} \rho\left(1-\left(\frac{2\rho}{\sqrt{2\pi\sigma_W^2}}\right)^{\frac{p-1}{p}} \left(\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^p}{1+(2-p)p\frac{\sigma_W^2}{\sigma^2}}\right)^{\frac{1}{2p}}\right) \tag{87}$$

$$= \sup_{p>1} \frac{1}{q+1}\left(\frac{q}{q+1}\right)^q \left(\left(\frac{2}{\sqrt{2\pi\sigma_W^2}}\right)^{\frac{p-1}{p}} \left(\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^p}{1+(2-p)p\frac{\sigma_W^2}{\sigma^2}}\right)^{\frac{1}{2p}}\right)^{-\frac{1}{q}}, \tag{88}$$

where q represents the Hölder's conjugate with respect to p, *i.e.*, $q=\frac{p}{p-1}$.

In particular, setting $p=3/2$ one obtains:

$$\mathcal{H}_{3/2}(W,X) = \sqrt{\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{3}{2}}}{1+\frac{3\sigma_W^2}{4\sigma^2}}}, \tag{89}$$

leading us to a lower bound on the Bayesian risk given by:

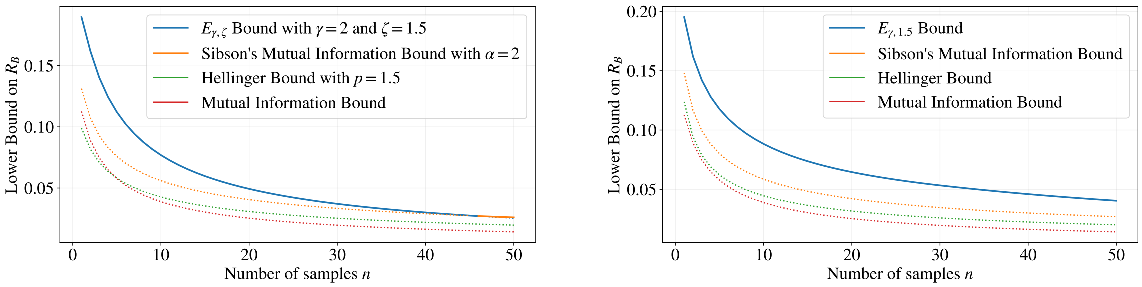
$$R_B \geq \frac{81\sqrt{2\pi}}{2048}\sqrt{\frac{\sigma_W^2}{1+n\frac{\sigma_W^2}{\sigma^2}}}. \tag{90}$$

Similarly to the previous example, one has that Equation (90) matches the upper-bound up to a constant factor and provides a strengthening of the bounds obtained in (Xu and Raginsky, 2017, Corollary 1). Repeating the analysis with the $E_{\gamma,\zeta}$-divergence, one obtains the following:

$$R_B \geq \sup_{\rho>0}\rho\left(1-\frac{\left(E_{\gamma,\zeta}(W,X^n)+\frac{2\rho\gamma}{\sqrt{2\sigma_W^2\pi}}\right)}{\zeta}\right) \tag{91}$$

$$= \frac{\sqrt{2\sigma_W^2\pi}\,(\zeta-E_{\gamma,\zeta}(W,X^n))^2}{8\gamma\zeta}. \tag{92}$$

Like in Example 2, one can numerically evaluate Equation (92) and compare it with Equation (84), Equation (90) and Xu and Raginsky (2017, Equation (16)). Figure 3a and Figure 3b show the resulting lower bounds as a function of the number of samples n. One can observe similar behaviors when comparing with the results from the previous example: the bounds retrieved through the $\mathcal{H}_p$- and $E_{\gamma,\zeta}$-divergences are both able to improve on the lower bound relying on Shannon's Mutual Information. Once again, $E_{\gamma,\zeta}$, (cf. Equation (92)) provides the largest lower bound, while Sibson's α-Mutual Information is still able



(a) Setting: Example 3 with $\sigma_W^2 = 1$ and $\sigma^2 = 2$. The picture shows the behaviour of Equation (84) with $\alpha = 2$, Equation (90), Equation (92) with $\gamma = 2$ and $\zeta = 1.5$ and (Xu and Raginsky, 2017, Equation (16)) as a function of n. The values of $E_{2,1.5}(W, X^n)$ for each n are computed numerically. A solid line means that the corresponding lower bound is the largest.

(b) Comparison between the largest lower bounds one can retrieve for different information measures in Example 3: that is between, Equation (84), Equation (88), Equation (92) with $\zeta = 1.5$, and (Xu and Raginsky, 2017, Equation (16)). The quantities are numerically optimised over, respectively, $\gamma \geq 1$, $p > 1$ and $\alpha > 1$. The numerical optimisation over the parameter ζ is not carried out for computational reasons. A solid line means that the corresponding lower bound is the largest.

Figure 3: Comparison of various bounds for Example 3 with and without (numerical) optimisation of parameters.

to provide a stronger result than Equation (88). Similarly to before, one can also numerically optimise the bounds with respect to the corresponding parameters $\alpha > 1, p > 1, \zeta > 0$ and $\gamma \geq 0$ and the resulting comparison is depicted in Figure 3b.

5.4 “Hide-and-seek” problem

To conclude, let us consider next a d-dimensional distributed estimation problem, known as the “Hide-and-seek” problem. It was first presented in Shamir (2014) and also studied in Xu and Raginsky (2017).

Example 4. *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 $\frac{1}{2} + \theta$ 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:*

$$R_M = \inf_{\varphi^m, \psi} \max_{w \in [d]} \mathbb{P}[W \neq \hat{W}]. \tag{93}$$

A lower bound for R_M derived in (Shamir, 2014) is as follows:

$$R_M \geq 1 - \left(\frac{3}{d} + 5\sqrt{\min\left\{ \frac{10\theta nmb}{d}, mn\theta^2 \right\}} \right) \tag{94}$$

and only holds for $0 \leq \theta \leq 1/(4n)$. Additionally, in (Xu and Raginsky, 2017) a quite different lower bound has been proposed:

$$R_M \geq 1 - \frac{1}{\log d} \min\left\{\left[1 - \left(\frac{1-2\theta}{1+2\theta}\right)^n\right] mb + 1, \min(4mn\theta^2, \log d) + 1\right\}, \tag{95}$$

and it holds for $0 \leq \theta \leq 1/2$. 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,

$$\mathcal{L}\left(W{\to}\hat{W}\right) \leq \min\left\{\mathcal{L}\left(W{\to}X^{n\times m}\right), \mathcal{L}\left(W{\to}Y^m\right)\right\}.$$

We also have that $\mathcal{L}\left(W{\to}Y^m\right) \leq mb$ and that:

$$\begin{aligned}
\mathcal{L}(W \to X^{n\times m}) &\leq nm\mathcal{L}(W \to X) &(96)\\
&= nm \log \sum_x \max_w \mathcal{P}_{X|W=w}(x) &(97)\\
&\leq nm \log \sum_x \left(\frac{1}{2}\right)^{d-1}\left(\frac{1}{2}+\theta\right) &(98)\\
&= nm \log(2^d(2^{-d} + 2^{-d+1}\theta)) &(99)\\
&= nm\log(1+2\theta), &(100)
\end{aligned}$$

Hence:

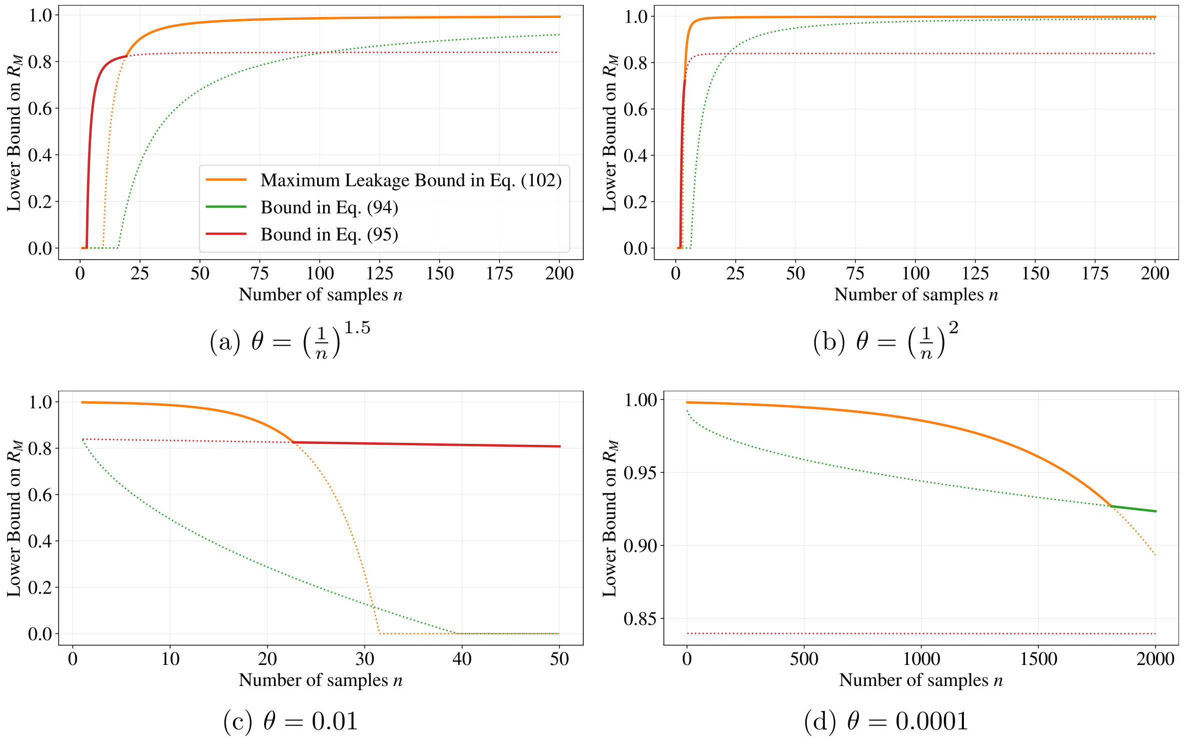
$$\mathcal{L}\left(W{\to}\hat{W}\right) \leq \min(nm\log(1+2\theta), \log d, mb). \tag{101}$$

Using Equation (101) in Equation (26) we get the following:

$$\mathbb{P}(\{\hat{W} \neq W\}) \geq 1 - \frac{\exp(\min\{mb, \log d, nm\log(1+2\theta)\})}{d}. \tag{102}$$

Notice that Equation (102) 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 Equation (26) with respect to (Xu and Raginsky, 2017, Theorem 1) is pivotal. Let us now compare the results in a common setting. The setting chosen in (Xu and Raginsky, 2017), where $d = 512, b = 3d, m = 10$ and $\theta = 1/(4n)$ does not represent a choice of parameters where Equation (102) is interesting. Indeed, for large enough n, $nm\log(1+2\theta) = nm\log(1+1/2n) \approx m/2$ and, as a consequence, the expression will converge to a constant determined by the minimum between $mb, \log d, m/2$. Furthermore, both Equation (94) and Equation (95) have a term that depends on $mn\theta^2$ which, for $\theta = 1/(4n)$, will decay with n. Thus, choosing $\theta \sim n^{-q}$ with $q > 1$ represents an interesting setting for the bound in Equation (102), as the plots in Figure 4a and Figure 4b show.

Thanks to the different behaviour of Equation (102) (reaching 1 exponentially fast) we can see a much sharper jump towards 1 with respect to Equation (95), which instead reaches a plateau strictly below 1, and with respect to Equation (94) that reaches 1 more slowly. The growth towards 1 of Equation (102) becomes even sharper with faster q and converges towards a specific behaviour at $q \approx 2$. Increasing q any further does not alter the behaviour



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

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

(c) $\theta = 0.01$

(d) $\theta = 0.0001$

Figure 4: Setting: Example 4 with various values of θ. Behaviour of Equation (102) and its comparison with bounds in Equation (95) and Equation (94). A solid line means that the corresponding lower bound is the largest.

of the bound meaningfully. As for the behaviour of the bound for fixed θ, if $\theta = 0.01$. then Equation (102) provides a larger lower bound only for $n < 25$. If the parameter is brought down to $\theta = 0.0001$ then Equation (102) is larger than Equation (95) for all n but only larger than Equation (94) for $n < 1850$. Regardless of the considerations related to the specific settings, it is interesting how a very simple application of Equation (26) can provide a tighter lower bound. Moreover, in Xu and Raginsky (2017), to compute $I(W;X)$ an assumption on the distribution of W was necessary, and the choice fell on W uniform on $[d]$. In contrast, $\mathcal{L}(W \to X)$ does not depend on the specific distribution over W, rendering the bound more general. Other divergences could be explored in this setting as well. However, one, in general, does not have a chain rule for any other φ-divergence (or Sibson's α-Mutual Information with $\alpha < +\infty$) which is a fundamental step in the proof for Maximal Leakage (see Equation (96)). Moreover, some assumption (or maximisation over) $\mathcal{P}_W$ would be necessary. In general, some additional machinery would be required to employ them in this setting. These approaches will not be explored in this document.

6 Conclusions

Inf. Measure	Special case	Lower Bound on R_B for $\rho > 0$
$I_\varphi(W,X)$	φ non-decreasing	$\rho\left(1 - L_W(\rho)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,X)+(1-L_W(\rho))\cdot\varphi^\star(0)}{L_W(\rho)}\right)\right)$
	φ non-increasing	$\rho\left(1 - L_W(\rho)\right)\cdot\varphi^{-1}\left(\frac{I_\varphi(W,X)+L_W(\rho)\cdot\varphi^\star(0)}{1-L_W(\rho)}\right)$
	$\varphi(x) = \frac{x^p-1}{p-1}$	$\rho\left(1 - L_W(\rho)^{\frac{p-1}{p}}\cdot\left((p-1)\mathcal{H}_p(W,X)+1\right)^{\frac{1}{p}}\right)$
	$\varphi(x) = (\zeta x-\gamma)_+ - (\zeta-\gamma)_+$	$\rho\left(1 - \frac{E_{\gamma,\zeta}(W,X)+\gamma L_W(\rho)+(\zeta-\gamma)_+}{\zeta}\right)$
$I_\alpha(W,X)$	$\alpha > 0$	$\rho\left(1 - \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(W,X)+\log(L_W(\rho))\right)\right)\right)$
	$\alpha \to \infty$	$\rho\left(1 - \exp\left(\mathcal{L}(W \to X)+\log(L_W(\rho))\right)\right)$

Table 1: Summary of the bounds derived in Section 4 and their special cases.

We have introduced a methodological framework to provide lower bounds on the Bayesian risk leveraging virtually any information measure. The lower bound encapsulates the intuition that if the observations X^n do not share enough "information" with the parameter W, then estimation of W is impossible regardless of the number of observations n. However, "information" can be measured in a variety of ways: via Sibson's Mutual Information, Rényi's Divergences, or φ-Mutual Information. Different choices yield different lower bounds. One can thus select the one that provides the best result in a specific setting of interest. The difficulty in computing the risk is relayed to the computation of an information measure which depends explicitly on the observation channel. The lower bounds are characterised by being estimator-independent and by the fact that one can explicitly take into account the information loss that a specific privacy-enforcing kernel can induce (see Section 4.1 and

Section 5.2). Given a function φ, all the bounds are characterised by a simple expression that involves the computation of two objects:

- the functional $L_W(\cdot,\cdot)$ (see Equation (3));
- and an information measure $I_{\{\cdot\}}(W, X^n)$.

Although no clear algorithm to create the largest lower bound for an estimation problem via information measures can be provided, we can highlight some observations. Consider the parametrised families of Sibson's, Rényi's, and the Hellinger α-Information, the shape of the bound is the same and the only difference lies in the information measure. Thus, the smallest information measure in the group will yield the largest lower bound and one can give a clear ordering (see Remark 12). For a given α:

1. I_α will provide the best lower bound between the three, however, it may be harder to calculate to provide a closed-form expression;
2. the Hellinger p-divergence (with $p = \alpha$) will provide worse lower bounds. It can, however, lead to closed-form expressions for the bound (that match the upper bound up to a constant, see Section 5.1).

For a given family of information measures (*e.g.*, Sibson's α-Mutual Information), comparisons between different choices of the free parameters (*e.g.*, α) are also not straightforward:

1. choosing a larger α implies that the information measure will be larger;
2. however, the multiplicative term $L_W^{\frac{\alpha-1}{\alpha}}$ will be smaller;
3. taking $\alpha \to \infty$ for I_α (Maximal Leakage $\mathcal{L}(W \to X^n)$) is interesting despite item 1.:
   (a) the information measure is "easier" to compute as it depends only on the observation channel;
   (b) it leads to results that hold for every W with the same support;
   (c) it satisfies a chain rule which allows us to employ in settings like the one described in Section 5.4.

Bounds induced by information measures not in the I_α family are also useful for other reasons. For instance, the SDPI constant of I_α has not been characterised yet while there exist universal upper and lower bounds on the SDPI constant of every divergence D_φ. This allows us to tighten the bounds, as well as to quantify the information loss that privacy-enforcing mechanisms induce and how much harder the estimation problem becomes as a consequence of this (see Sections 4.1 and 5.2). Table 1 summarises the bounds provided in this work.

Acknowledgments and Disclosure of Funding

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

Appendices

A Variational Representations of Divergences

A re-interpretation of the comments stated in Section 2 and tailored to divergences leads us to the main technical tools that will be used through the document: “variational representations” and functional inequalities. The main starting point will be looking at divergences as functionals acting on the first measure *i.e.*, $D_\varphi(\cdot\|\mu) = \psi_\mu(\cdot)$. Once this is established, most variational representations are instances of Legendre-Fenchel duality as stated in Equation (4). The most well-known is certainly the Donsker-Varadhan representation of the Kullback-Leibler divergence, which states the following (Varadhan, 1984):

$$D(\nu\|\mu) = \sup_{f\in B(\mathcal{X})} \langle \nu, f\rangle - \log\left(\mu(\exp(f))\right), \tag{103}$$

where $B(\mathcal{X})$ denotes the space of bounded and measurable real-valued functions defined on $\mathcal{X}$. Equation (103) characterises the Kullback-Leibler divergence as the Legendre-Fenchel dual of the functional $\log(\mu(\exp(f))) = \vartheta_\mu(f)$ *i.e.*, $D(\cdot\|\mu) = \vartheta^\star_\mu(\cdot)$. Similar variational representation can be found for large families of divergences, like Rényi’s divergences (Anantharam, 2018; Birrell et al., 2021) and φ-divergences (Broniatowski and Keziou, 2006). We will now lay the groundwork to state the variational representation for φ-divergences as it represents a meaningful tool for the scope of this work. In particular, let $F(\mathcal{X})$ be an arbitrary family of real-valued functions defined on $\mathcal{X}$ and denote with $\mathcal{M}_1(\mathcal{X})$ the space of probability measures over $\mathcal{X}$. Denote with $\langle F(\mathcal{X}) \cup B(\mathcal{X})\rangle$ the linear span of $F(\mathcal{X}) \cup B(\mathcal{X})$. Moreover, denoting with $|\nu|$ denotes the total variation of the measure ν, consider the following sets:

$$\mathcal{M}_1^F(\mathcal{X}) = \left\{\nu \in \mathcal{M}_1(\mathcal{X}) : \int |f|\,\mathrm{d}\nu < \infty \text{ for } f \in F(\mathcal{X})\right\},$$

and

$$\mathcal{M}^F(\mathcal{X}) = \left\{\nu \in \mathcal{M}(\mathcal{X}) : \int |f|\,\mathrm{d}|\nu| < \infty \text{ for } f \in F(\mathcal{X})\right\}.$$

If $F(\mathcal{X}) = B(\mathcal{X})$ then $\mathcal{M}_1^F(\mathcal{X}) = \mathcal{M}_1(\mathcal{X})$ and $\mathcal{M}^F(\mathcal{X}) = \mathcal{M}(\mathcal{X})$. Denote with τ_F the weakest topology on $\mathcal{M}^F(\mathcal{X})$ such that all mappings $\nu \to \nu(f)$ are continuous when $f \in \langle F(\mathcal{X}) \cup B(\mathcal{X})\rangle$ and with τ_M the weakest topology on $\langle F(\mathcal{X}) \cup B(\mathcal{X})\rangle$ such that all mappings $f \to \nu(f)$ are continuous when $\nu \in \mathcal{M}^F(\mathcal{X})$. One can then show the following result

Proposition 21 ((Broniatowski and Keziou, 2006, Proposition 2.1)). *The space of measures $\mathcal{M}^F(\mathcal{X})$ equipped with the τ_F-topology and the space of functions $\langle F(\mathcal{X}) \cup B(\mathcal{X})\rangle$ equipped with the τ_M are locally convex topological vector spaces and are the topological dual of each other.*

$D_\varphi(\cdot\|\mu) = \psi_\mu(\cdot)$ is thus a convex and lower semi-continuous mapping with respect to τ_F (Broniatowski and Keziou, 2006, Proposition 2.2) and it is possible to characterise its variational representation, bridging us between the two spaces $\mathcal{M}^F(\mathcal{X})$ and $\langle F(\mathcal{X}) \cup B(\mathcal{X})\rangle$. For the additional technical condition required on φ (*i.e*, guaranteeing the uniqueness of the dual optimal solution the reader is referred to Broniatowski and Keziou (2006)).

Theorem 22 (Broniatowski and Keziou 2006, and Theorem 4.3)**.**
Let φ be a strictly convex functional and let $\mu \in \mathcal{M}(\mathcal{X})$. One has that for every $\nu \in \mathcal{M}^F(\mathcal{X})$:

$$D_\varphi(\nu\|\mu) = \sup_{f\in\langle F(\mathcal{X})\cup B(\mathcal{X})\rangle} \nu(f) - \mu(\varphi^\star(f)), \tag{104}$$

where $\varphi^\star$ denotes the Legendre-Fenchel dual of φ. Moreover, one has that for a given $f \in \langle F(\mathcal{X})\cup B(\mathcal{X})\rangle$:

$$\mu(\varphi^\star(f)) = \sup_{\nu\in\mathcal{M}^F(\mathcal{X})} \nu(f) - D_\varphi(\nu\|\mu). \tag{105}$$

Through Equation (104), given a measure μ, one can connect the expected value of any function $f \in \langle F(\mathcal{X})\cup B(\mathcal{X})\rangle$ under any measure $\nu \ll \mu$ (*i.e.*, $\nu(f)$) to the divergence $D_\varphi(\nu\|\mu)$. The behavior of the third actor in Equation (104), the dual of $D_\varphi(\cdot\|\mu)$, is crucial in order to obtain bounds. For instance, when f is the indicator function of an event, one can explicitly compute the dual (and then retrieve a family of Fano-like inequalities involving arbitrary divergences. For more details see Esposito et al. (2021a) and Esposito (2022, Chapter 3)). When f is *not* an indicator function, one cannot typically compute the dual explicitly and has to upper-bound it leveraging properties of μ and f. In this work, to provide such an upper bound on the dual, we will make use of Markov's inequality. This takes us back to indicator functions for which we can completely characterise the dual. For technical details see Appendix C.2. This pattern is fundamental whenever one is trying to relate (via upper or lower bounds) the expected value of a function to some divergence/entropy (see Esposito (2022, Chapter 2)).

B Comparison with similar approaches

An approach closely connected to the one proposed in here is in Chen et al. (2016). The authors therein focused on the notion of φ-informativity (Csiszár, 1972) and leveraged the Data-Processing inequality of the information measure. In particular, φ-informativities can potentially lead to tighter results than the φ-Mutual Information considered in this work. Similarly to Sibson's α-Mutual Information, they are defined as follows:

$$\hat{I}_\varphi(X,Y) = \inf_{\mathcal{Q}_Y} D_\varphi(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) \leq I_\varphi(X,Y). \tag{106}$$

Given that the minimum-achieving distribution, $\mathcal{Q}^\star_Y$, is guaranteed to exist in Equation (106) (see (Csiszár, 1972)), one can see that $\hat{I}_\varphi(X,Y) = D_\varphi(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}^\star_Y)$. Consequently, the same steps followed in the proof of Theorem 9 can be undertaken in order to reach a similar result involving $\hat{I}_\varphi$ and $\mathcal{P}_X\mathcal{Q}^\star_Y$ rather than I_φ and $\mathcal{P}_X\mathcal{P}_Y$. However, except in some specific settings, the minimum-achieving distribution in Equation (106) does not necessarily admit a closed-form expression (Csiszár, 1972). As a consequence, the corresponding φ-Informativity does not admit a closed-form expression. Moreover, another step the authors leveraged to achieve Chen et al. (2016, Theorem 3.2), is the inversion of the resulting binary divergence, leading to a bound which can rarely be expressed in closed form and can only be computed numerically. While a direct comparison between the two approaches would be hard, some similarities are present and hint at the fact that Chen et al. (2016, Theorem 3.2) is tighter than Theorem 9. Indeed, an alternative proof for Theorem 9 also stems from leveraging the DPI of I_φ (see Esposito et al. (2021a, Theorem 3)). However, additional steps are introduced in order to get a closed-form lower bound. Our analysis is designed to retrieve a large family of results which are amenable to analysis and interpretable. This allows us to retrieve lower bounds in closed-form expressions that can be seen to match the upper bounds, up to a constant, in a variety of settings. From a more conceptual standpoint, one could see Esposito et al. (2021a, Theorem 3) (and, consequently, Theorem 9) as a generalisation of Hölder's[2] inequality to arbitrary convex functionals. This generalisation, which in turn can be seen as a generalisation of Fano's inequality for φ-Mutual Information, allows us to also encompass divergences from the Rényi's family and Sibson's α-Mutual Information, which are not φ-divergences and are thus excluded from Chen et al. (2016). To conclude, let us highlight that our approach, which leverages duality, allows us to provide a single analysis for every type of loss and does not require a separate treatise for $0-1$ losses and more general losses. Consequently, the two approaches for general losses are different and hard to compare.

C Proof of Section 4

C.1 Proof of Theorem 8

Proof. We have that

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})<\rho) \leq \left(\sup_{\hat{w}\in\hat{\mathcal{W}}} \mathcal{P}_W(\ell(W,\hat{w})<\rho)\right)^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,\hat{W})\right) \tag{107}$$

$$= \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(W,\hat{W}) + \log(L_W(\rho))\right)\right) \tag{108}$$

$$\leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(W,X) + \log(L_W(\rho))\right)\right). \tag{109}$$

Equation (107) follows from Corollary 6, Equation (109) follows from the Data-Processing Inequality for I_α and the Markov Chain $W - X - Y - \hat{W}$. The statement follows from lower bounding Equation (24) using Equation (109). □

C.2 Proof of Theorem 9

Proof. From the variational representation for φ-divergences (see Equation (104)), given $\mathcal{P}_{W\hat{W}}$, for every function f in the respective space (defined in Theorem 22) one has that:

$$I_\varphi(W,\hat{W}) = D_\varphi(\mathcal{P}_{W\hat{W}}\|\mathcal{P}_W\mathcal{P}_{\hat{W}}) \geq \mathcal{P}_{W\hat{W}}(f) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\varphi^\star(f)). \tag{110}$$

Equation (110) allows us to relate the expected value of any function $f : \mathcal{W}\times\hat{\mathcal{W}}$ under the joint with $I_\varphi(W,\hat{W})$ and the corresponding Legendre-Fenchel dual. Our purpose is to provide a lower bound on the expected loss ℓ. Hence, we will select $f = \tilde{\lambda}(\rho-\ell)$ with $\rho, \tilde{\lambda} > 0$. Moreover, given the non-negativity of ℓ one can also see that $\ell \geq \rho\mathbb{1}_{\{\ell\geq\rho\}}$ (*i.e.*,

[2] Selecting $\varphi(x) = x^p$ specialises Theorem 9 to Corollary 11 which can also be proven as an application of Hölder's inequality followed by Markov's inequality, cf. (Esposito et al., 2021a, Corollary 6).

Markov's Inequality in its functional form). Thus, plugging our choice of f in Equation (110) the following chain of inequalities follows:

$$\begin{aligned}\tilde{\lambda}\mathcal{P}_{W\hat{W}}(\ell) &\geq \tilde{\lambda}\rho - I_\varphi(W,\hat{W}) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\varphi^\star(\tilde{\lambda}\rho - \tilde{\lambda}\ell)) &(111)\\ &\geq \tilde{\lambda}\rho - I_\varphi(W,\hat{W}) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\varphi^\star(\tilde{\lambda}\rho(1-\mathbb{1}_{\{\ell\geq\rho\}}))) &(112)\\ &= \tilde{\lambda}\rho - I_\varphi(W,\hat{W}) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\varphi^\star(\tilde{\lambda}\rho\mathbb{1}_{\{\ell<\rho\}})) &(113)\\ &= \tilde{\lambda}\rho - I_\varphi(W,\hat{W}) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})\cdot\varphi^\star(\tilde{\lambda}\rho) - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell\geq\rho\})\cdot\varphi^\star(0), &(114)\end{aligned}$$

where Equation (112) follows by the monotonicity of $\varphi^\star$ which can be seen as stemming from the strict convexity and the monotonicity of φ. Indeed, if φ is strictly convex then $\varphi^{\star\prime}(t) = \varphi'^{-1}(t)$ for every $t \in \mathrm{Im}(\varphi')$ (Rockafellar, 1970, Theorem 26.5). Since φ is monotone non-decreasing on the positive axis, one has that $\varphi'(t) \geq 0$ on $[0,+\infty]$. Accordingly, the inverse of φ' will also be non-negative on $[0,+\infty]$, which implies the non-negativity of $\varphi^{\star\prime}$ and, therefore, the monotonicity of $\varphi^\star$. A similar argument shows the monotonicity of $\varphi^\star$ when φ is monotone non-increasing. Then, dividing both sides by $\tilde{\lambda}$ and selecting $\tilde{\lambda} = \frac{1}{\rho}\lambda$ with $\lambda > 0$ one recovers the following:

$$\begin{aligned}\mathcal{P}_{W\hat{W}}(\ell) &\geq \sup_{\lambda>0}\rho\left(1 - \frac{I_\varphi(W,\hat{W}) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell\geq\rho\})\cdot\varphi^\star(0) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})\cdot\varphi^\star(\lambda)}{\lambda}\right) &(115)\\ &= \rho\left(1 - \inf_{\lambda>0}\frac{I_\varphi(W,\hat{W}) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell\geq\rho\})\cdot\varphi^\star(0) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})\cdot\varphi^\star(\lambda)}{\lambda}\right) &(116)\\ &= \rho\left(1 - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})\inf_{\lambda>0}\frac{\frac{I_\varphi(W,\hat{W}) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell\geq\rho\})\cdot\varphi^\star(0)}{\mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})} + \varphi^\star(\lambda)}{\lambda}\right) &(117)\\ &= \rho\left(1 - \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})\varphi^{-1}\left(\frac{I_\varphi(W,\hat{W}) + \mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell\geq\rho\})\cdot\varphi^\star(0)}{\mathcal{P}_W\mathcal{P}_{\hat{W}}(\{\ell<\rho\})}\right)\right), &(118)\end{aligned}$$

where Equation (118) follows from the same argument as in (Esposito, 2022, Theorem 13). Equation (112) is the step of the proof which is relevant to the discussion at the end of Appendix A. In particular, the choice of f, along with the non-decreasability of φ allowed us to leverage the functional form of Markov's inequality and, consequently, to upper-bound the dual of $D_\varphi(\cdot\|\mathcal{P}_W\mathcal{P}_{\hat{W}})$. Upper-bounding the dual is crucial in order to achieve a bound of the form of Equation (118). In order to prove the result for φ non-increasing one has to select $f = -\tilde{\lambda}\ell$ leverage Markov's inequality and select $\tilde{\lambda} = -\frac{1}{\rho}\lambda$ with $\lambda < 0$. The result then follows from the same argument as in Esposito (2022, Theorem 13) *i.e.*, from selecting $\lambda = \varphi'(\varphi^{-1}(c))$ with $c = \frac{I_\varphi(W,\hat{W}) + \varphi^\star(0)\mathcal{P}_W\mathcal{P}_{\hat{W}}(E)}{\mathcal{P}_W\mathcal{P}_{\hat{W}}(E^c)}$ and $E = \{\ell < \rho\}$. □

C.3 Proof of Corollary 11

Proof. The statement follows from Theorem 9 with $\varphi_p(x) = \frac{x^p-1}{p-1}$ for $p \geq 1$. Hence, for every estimator $\hat{W} = \phi(X^n)$,

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}) \geq \rho) \leq L_W(\hat{W},\rho) \cdot \varphi_p^{-1}\left(\frac{I_{\varphi_p}(W,\hat{W}) + (1 - L_W(\hat{W},\rho)) \cdot \varphi_p^{\star}(0)}{L_W(\hat{W},\rho)}\right) \tag{119}$$

$$= L_W(\hat{W},\rho)^{1-1/p}((p-1)\mathcal{H}_p(W,\hat{W}) + 1)^{\frac{1}{p}} \tag{120}$$

$$\leq L_W(\rho)^{\frac{p-1}{p}}\left((p-1)\mathcal{H}_p(W,X) + 1\right)^{\frac{1}{p}}, \tag{121}$$

where Equation (121) follows from the fact that in this case the functional G (see Equation (31)) is increasing in $L_W(\hat{W},\rho)$ for a given value of $\mathcal{H}_p(W,\hat{W})$ and increasing in $\mathcal{H}_p(W,\hat{W})$ for a given value of $L_W(\hat{W},\rho)$. Hence one can use both these inequalities: $L_W(\hat{W}\rho) \leq L_W(\rho)$ and $\mathcal{H}_p(W,\hat{W}) \leq \mathcal{H}_p(W,X)$. One thus retrieves that for every estimator $\hat{W}$

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - L_W(\rho)^{\frac{p-1}{p}}\left((p-1)\mathcal{H}_p(W,X) + 1\right)^{\frac{1}{p}}\right). \tag{122}$$

Since the right-hand side of Equation (122) is independent of $\hat{W} = \phi(X)$ one can use it to lower bound the risk R. □

C.4 Proof of Corollary 14

Proof. Let $\varphi(x) = (\zeta x - \gamma)_+ - (\zeta - \gamma)_+$ in Theorem 9, along with the fact that $\varphi^{-1}(y) = \frac{y+(\zeta-\gamma)_+ + \gamma}{\zeta}$ for $y > 0$ and $\varphi^{\star}(0) = (\zeta - \gamma)_+$ one has that for every estimator $\hat{W} = \phi(X^n)$,

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \rho\left(1 - \frac{E_{\gamma,\zeta}(W,\hat{W}) + \gamma L_W(\hat{W},\rho) + (\zeta-\gamma)_+}{\zeta}\right) \tag{123}$$

$$\geq \rho\left(1 - \frac{E_{\gamma,\zeta}(W,X) + \gamma L_W(\rho) + (\zeta-\gamma)_+}{\zeta}\right). \tag{124}$$

Since Equation (124) is independent of $\hat{W} = \phi(X)$ one can use it to lower bound the risk R. □

D Computations for Section 5

D.1 Maximisation over ρ

The bounds considered in the first three examples have the following form

$$\sup_{\rho>0} \rho(1 - c\rho^t - b), \tag{125}$$

for some $c, t, b \geq 0$. Letting $h(\rho) := \rho(1 - c\rho^t - b)$, the optimal value $\rho_\star$ is found by setting $h'(\rho_\star) = 0$, which yields

$$1 - (t+1)c\rho_\star^t - b = 0 \iff \rho_\star = \left(\frac{1-b}{(t+1)c}\right)^{\frac{1}{t}}. \tag{126}$$

Since $h''(\rho_\star) = -t(t+1)c\rho_\star^{t-1} \leq 0$, this ensures $\rho_\star$ is a maximum. Substituting $\rho_\star$ back in Equation (125), we can express the lower bound as

$$\sup_{\rho>0} \rho(1 - c\rho^t - b) = \frac{t}{c^{\frac{1}{t}}} \left(\frac{1-b}{t+1}\right)^{1+\frac{1}{t}}. \tag{127}$$

D.2 Section 5.1

D.2.1 Maximal Leakage

In this setting one has that

$$\mathcal{P}_{X^n|W=w}(x^n) = w^k(1-w)^{n-k}$$

where $k = \sum_{i=1}^n x_i$ is the hamming weight of x^n. As per assumption, $\mathcal{P}_W(w) = 1$ if $0 \leq w \leq 1$ and, consequently, one has that

$$\mathcal{P}_{W|X^n=x^n}(w) = (n+1)\binom{n}{k}(1-w)^{n-k}w^k.$$

One can thus compute Maximal Leakage in this setting:

$$\begin{aligned}
\mathcal{L}\left(W \to X^n\right) &= \log \sum_{x^n} \max_w \mathcal{P}_{X^n|W=w}(x^n) && (128)\\
&= \log \sum_{k=0}^{n} \binom{n}{k} \max_w w^k(1-w)^{n-k} && (129)\\
&= \log \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} && (130)\\
&\leq \log\left(2 + \sum_{k=1}^{n-1} \sqrt{\frac{n}{2\pi k(n-k)}}\right) && (131)\\
&\leq \log\left(2 + \sqrt{\frac{\pi n}{2}}\right), && (132)
\end{aligned}$$

where Equation (131) follows from Stirling's approximation (see Feller (1968, Page 54)), while the last bound follows from upper bounding the sum by an integral (and elementary integration rules).

D.2.2 Sibson's α-Mutual Information

For Sibson's α-Mutual Information with $\alpha > 1$, one has that:

$$\begin{aligned}
\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W, X^n))\right) &= \mathbb{E}\left[\mathbb{E}^{\frac{1}{\alpha}}\left[\left(\frac{\mathcal{P}_{X^n|W}}{\mathcal{P}_{X^n}}\right)^\alpha \middle| X^n\right]\right] && (133)\\
&= \sum_{x^n} \mathcal{P}_{X^n}(x^n)\left(\int_0^1 \mathcal{P}_W(w)\left(\frac{\mathcal{P}_{W|X^n=x^n}(w)}{\mathcal{P}_W(w)}\right)^\alpha \mathrm{d}w\right)^{\frac{1}{\alpha}} && (134)\\
&= \sum_{x^n} \mathcal{P}_{X^n}(x^n)\left(\int_0^1 \left(\mathcal{P}_{W|X^n=x^n}(w)\right)^\alpha \mathrm{d}w\right)^{\frac{1}{\alpha}} && (135)\\
&= \sum_{k=0}^n \binom{n}{k}\frac{1}{(n+1)\binom{n}{k}}\left(\int_0^1 \left((n+1)\binom{n}{k} w^k(1-w)^{n-k}\right)^\alpha \mathrm{d}w\right)^{\frac{1}{\alpha}} && (136)\\
&= \sum_{k=0}^n \binom{n}{k}\left(\int_0^1 \left(w^k(1-w)^{n-k}\right)^\alpha \mathrm{d}w\right)^{\frac{1}{\alpha}} && (137)\\
&= \sum_{k=0}^n \binom{n}{k}\left(\frac{\Gamma(k\alpha+1)\Gamma((n-k)\alpha+1)}{\Gamma(n\alpha+2)}\right)^{\frac{1}{\alpha}}, && (138)
\end{aligned}$$

where Equation (138) uses the identity relating the Beta function with the Gamma function *i.e.*,

$$\mathrm{Beta}(x,y) = \int_0^1 w^{x-1}(1-w)^{y-1}\,\mathrm{d}w = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{139}$$

so that

$$\int_0^1 w^{k\alpha}(1-w)^{(n-k)\alpha}\,\mathrm{d}w = \frac{\Gamma(k\alpha+1)\Gamma((n-k)\alpha+1)}{\Gamma(n\alpha+2)}. \tag{140}$$

D.2.3 Hellinger p-Divergence

For the Hellinger p-divergence with $p > 1$, one has that:

$$\begin{aligned}
((p-1)\mathcal{H}_p(W,X^n)+1) &= \left\|\frac{d\mathcal{P}_{WX^n}}{d\mathcal{P}_W\mathcal{P}_{X^n}}\right\|^p_{L_p(\mathcal{P}_W\mathcal{P}_{X^n})} && (141)\\
&= \sum_{x^n}\mathcal{P}_{X^n}(x^n)\int_0^1 \mathcal{P}_W(w)\left(\frac{\mathcal{P}_{W|X^n=x^n}(w)}{\mathcal{P}_W(w)}\right)^p \mathrm{d}w && (142)\\
&= \sum_{x^n}\mathcal{P}_{X^n}(x^n)\int_0^1 \left(\mathcal{P}_{W|X^n=x^n}(w)\right)^p \mathrm{d}w && (143)\\
&= \sum_{k=0}^n\binom{n}{k}\frac{1}{(n+1)\binom{n}{k}}\int_0^1\left((n+1)\binom{n}{k}w^k(1-w)^{n-k}\right)^p\mathrm{d}w && (144)\\
&= (n+1)^{p-1}\sum_{k=0}^n\binom{n}{k}^p\int_0^1\left(w^k(1-w)^{(n-k)}\right)^p\mathrm{d}w && (145)\\
&= (n+1)^{p-1}\sum_{k=0}^n\binom{n}{k}^p\frac{\Gamma(kp+1)\Gamma((n-k)p+1)}{\Gamma(np+2)}, && (146)
\end{aligned}$$

where Equation (146) follows from Equation (140). For the special case $p = 2$, we get

$$\begin{aligned}
\chi^2(W,X^n)+1 &= (n+1)\sum_{k=0}^n\binom{n}{k}^2\frac{(2k)!(2(n-k))!}{(2n+1)!} && (147)\\
&= \frac{n+1}{(2n+1)}\sum_{k=0}^n\frac{(n!)^2(2k)!(2(n-k))!}{(k!)^2((n-k)!)^2(2n)!} && (148)\\
&= \frac{n+1}{(2n+1)\binom{2n}{n}}\sum_{k=0}^n\binom{2k}{k}\binom{2(n-k)}{n-k} && (149)\\
&= \frac{n+1}{2n+1}\cdot\frac{4^n}{\binom{2n}{n}}, && (150)
\end{aligned}$$

where in Equation (150) we use the result in (Graham et al., 1989, Eq. (5.39), p.187) stating that $\sum_{k=0}^n\binom{2k}{k}\binom{2(n-k)}{n-k} = 4^n$.

D.2.4 Modified Hockey-Stick Divergence $E_{\gamma,\zeta}$

For the $E_{\gamma,\zeta}$ divergence with $\zeta > 0, \gamma \geq 0$, one has that:

$$\begin{aligned}
E_{\gamma,\zeta}(W, X^n) &= \sum_{x^n} \mathcal{P}_{X^n}(x^n) \int_0^1 \mathcal{P}_W(w) E_{\gamma,\zeta}(\mathcal{P}_{WX^n} \| \mathcal{P}_W \mathcal{P}_{X^n}) \,\mathrm{d}w \qquad (151)\\
&= \sum_{x^n} \mathcal{P}_{X^n}(x^n) \int_0^1 \mathcal{P}_W(w) \left[\left(\zeta \frac{\mathcal{P}_{W|X^n=x^n}(w)}{\mathcal{P}_W(w)} - \gamma\right)_+ - (\zeta-\gamma)_+\right] \mathrm{d}w \qquad (152)\\
&= \sum_{x^n} \mathcal{P}_{X^n}(x^n) \int_0^1 \left[\left(\zeta \mathcal{P}_{W|X^n=x^n}(w) - \gamma\right)_+ - (\zeta-\gamma)_+\right] \mathrm{d}w \qquad (153)\\
&= \frac{1}{n+1} \sum_{k=0}^{n} \int_0^1 \left[\left(\zeta(n+1)\binom{n}{k} w^k (1-w)^{n-k} - \gamma\right)_+ - (\zeta-\gamma)_+\right] \mathrm{d}w \qquad (154)\\
&= \frac{1}{n+1} \sum_{k=0}^{n} \int_0^1 \left[\left(\zeta(n+1)\binom{n}{k} w^k (1-w)^{n-k} - \gamma\right)_+\right] \mathrm{d}w - (\zeta-\gamma)_+. \qquad (155)
\end{aligned}$$

Since there is no closed-form formula for the integral, we compute the integration numerically in order to evaluate $E_{\gamma,\zeta}(W, X^n)$ in our experiments.

D.3 Section 5.3

D.3.1 Hellinger p-Divergence

For the Hellinger p-divergence with $p > 1$, one has that:

$$\begin{aligned}
((p-1)\mathcal{H}_p(W, X^n) + 1) &= \left\| \frac{d\mathcal{P}_{WX^n}}{d\mathcal{P}_W \mathcal{P}_{X^n}} \right\|^p_{L_p(\mathcal{P}_W \mathcal{P}_{X^n})} \qquad (156)\\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathcal{P}_W(w) \mathcal{P}_X(x) \left(\frac{\mathcal{P}_{X|W=w}(x)}{\mathcal{P}_X(x)}\right)^p \mathrm{d}w \,\mathrm{d}x \qquad (157)\\
&= \int_{\mathbb{R}} \mathcal{P}_X(x)^{1-p} \int_{\mathbb{R}} \mathcal{P}_W(w) \mathcal{P}_{X|W=w}(x)^p \,\mathrm{d}w \,\mathrm{d}x. \qquad (158)
\end{aligned}$$

Focusing on the innermost integral (which we denote as $I_p(x)$), one has

$$\begin{aligned}
I_p(x) &:= \int_{\mathbb{R}} \mathcal{P}_W(w) \mathcal{P}_{X|W=w}(x)^p \,\mathrm{d}w \qquad (159)\\
&= \left(\frac{(2\pi\sigma^2)^{-p}}{2\pi\sigma_W^2}\right)^{\frac{1}{2}} \int_{\mathbb{R}} e^{-\frac{w^2}{2\sigma_W^2} - \frac{p(w-x)^2}{2\sigma^2}} \,\mathrm{d}w \qquad (160)\\
&= \left(\frac{(2\pi\sigma^2)^{-p}}{2\pi\sigma_W^2}\right)^{\frac{1}{2}} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}\left(px^2 - 2pxw + \left(\frac{\sigma^2}{\sigma_W^2} + p\right) w^2\right)} \,\mathrm{d}w \qquad (161)\\
&= \left(\frac{(2\pi\sigma^2)^{-p}}{2\pi\sigma_W^2}\right)^{\frac{1}{2}} e^{\frac{-p \cdot x^2}{2\sigma^2}} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}\left(-2pxw + \left(\frac{\sigma^2}{\sigma_W^2} + p\right) w^2\right)} \,\mathrm{d}w. \qquad (162)
\end{aligned}$$

Adding and subtracting cx^2 with $c = -p\left(1 + p\frac{\sigma_W^2}{\sigma^2}\right)^{-1}$ in the exponent inside the integral in Equation (162) leads to

$$
\begin{aligned}
I_p(x) &= \left(\frac{(2\pi\sigma^2)^{-p}}{2\pi\sigma_W^2}\right)^{\frac{1}{2}} e^{\frac{cx^2}{2\sigma^2}} \int_{\mathbb{R}} e^{-\frac{\frac{\sigma^2}{\sigma_W^2}+p}{2\sigma^2}\left(w - \sqrt{\frac{p+c}{\frac{\sigma^2}{\sigma_W^2}+p}}x\right)^2} \mathrm{d}w & (163)\\
&= \left(\frac{(2\pi\sigma^2)^{-p}}{2\pi\sigma_W^2}\right)^{\frac{1}{2}} \exp\left(-\frac{px^2}{2\sigma^2\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)}\right)\left(2\pi\frac{\sigma^2}{\frac{\sigma^2}{\sigma_W^2}+p}\right)^{\frac{1}{2}} & (164)\\
&= \left(2\pi\sigma^2\right)^{-\frac{p}{2}}\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{px^2}{2\sigma^2\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)}\right). & (165)
\end{aligned}
$$

Finally, plugging the value of I_p back in (158), we retrieve that:

$$
\begin{aligned}
(p-1)\mathcal{H}_p(W,X)+1 &= \int_{\mathbb{R}} \mathcal{P}_X(x)^{1-p}\frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} e^{-\frac{px^2}{2\left(\sigma^2+p\sigma_W^2\right)}}\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)^{-\frac{1}{2}} \mathrm{d}x & (166)\\
&= \frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{d(p-1)}{2}}}{(2\pi\sigma^2)^{\frac{1}{2}}\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{1}{2}}}\int_{\mathbb{R}} e^{\frac{(p-1)x^2}{2\left(\sigma^2+\sigma_W^2\right)}-\frac{px^2}{2\left(\sigma^2+p\sigma_W^2\right)}} \mathrm{d}x & (167)\\
&= \frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{(p-1)}{2}}}{(2\pi\sigma^2)^{\frac{1}{2}}\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{1}{2}}}\int_{\mathbb{R}} e^{-\frac{x^2}{2}\left(\frac{1-p}{\sigma^2+\sigma_W^2}+\frac{p}{\sigma^2+p\sigma_W^2}\right)} \mathrm{d}x & (168)\\
&= \frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{(p-1)}{2}}}{(2\pi\sigma^2)^{\frac{1}{2}}\left(1+p\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{1}{2}}}\left(\frac{2\pi}{\frac{1-p}{\sigma^2+\sigma_W^2}+\frac{p}{\sigma^2+p\sigma_W^2}}\right)^{\frac{1}{2}} & (169)\\
&= \frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{\frac{(p-1)}{2}}}{\left(\sigma^2+p\sigma_W^2\right)^{\frac{1}{2}}}\left(\frac{1}{\frac{1-p}{\sigma^2+\sigma_W^2}+\frac{p}{\sigma^2+p\sigma_W^2}}\right)^{\frac{1}{2}} & (170)\\
&= \left(\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{p-1}}{\frac{(1-p)\left(\sigma^2+p\sigma_W^2\right)}{\sigma^2+\sigma_W^2}+p}\right)^{\frac{1}{2}} & (171)\\
&= \left(\frac{\left(1+\frac{\sigma_W^2}{\sigma^2}\right)^{p}}{1+(2-p)p\frac{\sigma_W^2}{\sigma^2}}\right)^{\frac{1}{2}}. & (172)
\end{aligned}
$$

E Other approaches

E.1 Conditioning

Following the approach undertaken in Xu and Raginsky (2017), it is also possible to propose a conditional version of the theorems proposed above. For this to happen one needs a definition of conditional information measures. For φ–divergences the choice would typically fall on objects of the following form

$$I_\varphi(X,Y|Z) = D_\varphi(\mathcal{P}_{XYZ}\|\mathcal{P}_Z\mathcal{P}_{X|Z}\mathcal{P}_{Y|Z}). \tag{173}$$

As for Sibson's I_α, the matter becomes slightly more complicated since one has that $I_\alpha(X,Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y)$. In the case of three random variables, it is unclear which factorisation of the joint and which minimisation to consider. Indeed, it has been shown in Esposito et al. (2021b) that several definitions of conditional I_α can be proposed, depending on the operational meaning and corresponding probability bound one needs. In this subsection, we will consider the following conditional version of I_α:

$$I_\alpha^{Y|Z}(X,Y|Z) = \min_{\mathcal{Q}_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ}\|\mathcal{P}_{X|Z}\mathcal{Q}_{Y|Z}\mathcal{P}_Z). \tag{174}$$

The choice of this specific definition is necessary to provide a conditional version of Theorem 8 and Equation (26) similar to (Xu and Raginsky, 2017, Theorem 1, Eq. (5)). Leveraging said definition and the fact that:

$$I_\alpha^{Y|Z}(X,Y|Z) \xrightarrow{\alpha\to\infty} \mathcal{L}\left(X\to Y|Z\right)$$

one can thus give a conditional version of Theorem 8 and Equation (26), introducing the following notion of conditional small-ball probability, $L_{W|U}(U,\rho) = \sup_{\hat{w}\in\hat{\mathcal{W}}}\mathcal{P}_{W|U}(\ell(W,\hat{w}) < \rho)$:

Theorem 23. *Consider the Bayesian framework described in Section 1.3,*

$$R_B \geq \sup_{\mathcal{P}_{U|W,X}} \sup_{\rho>0,\alpha\geq 1} \rho\left(1-\exp\left(\frac{\alpha-1}{\alpha}(I_\alpha(W,X|U)+\log(\mathcal{P}_U(L_{W|U}(U,\rho)))\right)\right), \tag{175}$$

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

$$R_B \geq \sup_{\mathcal{P}_{U|W,X}} \sup_{\rho>0} \rho\left(1-\exp\left(\mathcal{L}\left(W\to X|U\right)+\log(\mathcal{P}_U(L_{W|U}(U,\rho)))\right)\right). \tag{176}$$

The proof will follow at the end of this section. The main idea behind using conditional Mutual Information, as presented in Xu and Raginsky (2017), is that by choosing an appropriate U it is possible to control the growth of $I(W;X|U)$ and obtain tighter bounds in some cases. In particular, consider the sequence of 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 $\mathbb{R}^d$, choosing 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, (Xu and Raginsky, 2017)). This property seems to be specific to Shannon's Mutual Information. In the examples addressed in this manuscript, there does not appear to be a suitable U that tightens the bounds further for the divergences considered. Nonetheless, we stated the result as it may be of interest in other settings.

Proof. For the selected choice of conditional Sibson Mutual Information (see Equation (174)) one has that

$$I_\alpha(W,\hat{W}|U) = \frac{\alpha}{\alpha-1}\log\left\|\left\|\left\|\frac{d\mathcal{P}_{W\hat{W}U}}{d\mathcal{P}_U\mathcal{P}_{\hat{W}|U}\mathcal{P}_{W|U}}\right\|_{L^\alpha(\mathcal{P}_{W|U})}\right\|_{L^1(\mathcal{P}_{\hat{W}|U})}\right\|_{L^\alpha(\mathcal{P}_U)}. \tag{177}$$

Consequently, one can prove via Hölder's inequality a result analogous to Theorem 5 (cf. (Esposito, 2022, Theorem 17)) which implies then, selecting $f = \mathbb{1}_{\{\ell(W,\hat{W})\leq\rho\}}$, the following for every $\rho > 0, \alpha \geq 1$ and every U

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}) < \rho) = \mathcal{P}_{W\hat{W}U}(\ell(W,\hat{W}) < \rho) \tag{178}$$

$$\leq \mathcal{P}_U^{\frac{\alpha-1}{\alpha}}\left(\operatorname*{ess\,sup}_{\mathcal{P}_{\hat{W}|U}} \mathcal{P}_{W|U}(\ell(W,\hat{W}) \leq \rho)\right) \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,\hat{W}|U)\right). \tag{179}$$

$$\leq \mathcal{P}_U^{\frac{\alpha-1}{\alpha}}\left(L_{W|U}(U,\rho)\right) \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W,X|U)\right) \tag{180}$$

$$= \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(W,X|U) + \log(\mathcal{P}_U(L_{W|U}(U,\rho)))\right)\right) \tag{181}$$

The statement of the theorem then follows from the same sequence of steps (involving Lemma 7) that led to Theorem 8. Moreover, starting from Equation (177) and taking the limit of $\alpha \to +\infty$ one recovers the following:

$$I_\infty(W,X|U) = \log \operatorname*{ess\,sup}_{\mathcal{P}_U} \left\| \operatorname*{ess\,sup}_{\mathcal{P}_{W|U}} \frac{d\mathcal{P}_{WXU}}{d\mathcal{P}_U\mathcal{P}_{X|U}\mathcal{P}_{W|U}} \right\|_{L^1(\mathcal{P}_{X|U})}, \tag{182}$$

which can be seen as being equal to $\mathcal{L}\left(W \to X|U\right)$ (see Issa et al. (2020, Section III.E)). □

E.2 Inverting the roles

The Sibson's α-Mutual Information is an asymmetric quantity. A natural question is: can one provide a result similar to Theorem 8 involving $I_\alpha(X,W)$ instead? Indeed, by inverting the roles of W and $\hat{W}$, such a bound can be given but it will involve the small ball probability for $\hat{W}$ *i.e.*,

$$L_{\hat{W}}(\rho) = \sup_w \mathcal{P}_{\hat{W}}(\ell(w,\hat{W}) \geq \rho). \tag{183}$$

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

Lemma 24. *Consider the Bayesian framework described in Section 1.3. The following holds for every $\alpha > 1$ and $\rho > 0$:*

$$R_B \geq \rho\left(1 - \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(X,W) + \log(L_{\hat{W}}(\rho))\right)\right)\right). \tag{184}$$

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

$$R_B \geq \rho \left(1 - \exp\left(\mathcal{L}\left(X \to W\right) + \log(L_{\hat{W}}(\rho))\right)\right). \tag{185}$$

To apply this lemma in concrete cases, one needs to compute or upper bound the small ball probability $L_{\hat{W}}(\rho)$. Leveraging the 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 $\rho \geq 0$ (Bobkov and Chistyakov, 2015, Theorem 1.1),

$$L_{\sum_i^m Y_i}(\rho) \leq \frac{2\rho}{\sqrt{\mathrm{Var}(\sum_{i=1}^m Y_i) + \rho^2/3}} = \frac{2\rho}{\sqrt{m\mathrm{Var}(Y_1) + \rho^2/3}}. \tag{186}$$

More general statements can be made, assuming $\phi(Y^m) = \sum_{i=1}^m a_i Y_i$ under different constraints over a_i (Nguyen and Vu, 2013). To appreciate the promise of this approach, let us also discuss the behaviors of $I_\alpha(W, X)$ and $I_\alpha(X, W)$. More specifically, let us consider again the "Hide-and-Seek" problem. Assuming, as in Xu and Raginsky (2017, Example 12), that $\mathcal{P}_W$ is uniform over $[d]$, one has that

$$\mathcal{L}\left(X^{n\times m} \to W\right) = \log \frac{d(1/2 + \rho)}{(d-1)(1/2 - \rho) + (1/2 + \rho)} = \log \kappa(d, \rho) < \log d. \tag{187}$$

In case ρ and d are constant and the estimator ϕ is a linear combination of the observations, using Equation (186) in Lemma 24 one gets:

$$R_B \geq \rho \left(1 - \frac{\kappa(d,\rho) 2\rho}{\sqrt{m\mathrm{Var}(Y_1) + \rho^2/3}}\right). \tag{188}$$

This lower bound approaches ρ as m grows, rather than providing the trivial lower bound of 0, as it happens in Equation (102).
The assumptions required, along with the need to specify a prior over W, clearly restrict the domain of applicability of Lemma 24 with respect to Theorem 8 and Equation (26). However, this approach can provide results in settings where Theorem 8 and Equation (26) become vacuous.

E.3 Lower bounding the risk directly

An alternative route can be undertaken that does not use Markov's inequality as a first step and can possibly lead to tighter bounds. Since our purpose is to provide *lower bounds* on the risk (essentially, an inner-product between the joint measure of the parameter and the estimation and the loss function, $\langle \mathcal{P}_{W\hat{W}}, \ell \rangle$) one can also consider the application of reverse Hölder's inequality in order to directly lower bound the risk. Consider $\alpha < 1$, the following result can be easily proven:

Corollary 25. *Consider the Bayesian framework described in Section 1.3. The following holds for every $\alpha, \alpha' < 1$*

$$\mathcal{P}_{W\hat{W}}(\ell) \geq \mathcal{P}_{\hat{W}}^{\frac{1}{\beta'}} \left(\mathcal{P}_W^{\frac{\beta'}{\beta}}\left(\ell^\beta\right)\right) \cdot \mathcal{P}_{\hat{W}}^{\frac{1}{\alpha'}} \left(\mathcal{P}_W^{\frac{\alpha'}{\alpha}} \left(\left(\frac{d\mathcal{P}_{W\hat{W}}}{d\mathcal{P}_W \mathcal{P}_{\hat{W}}}\right)^\alpha\right)\right), \tag{189}$$

where $\frac{1}{\alpha}+\frac{1}{\beta}=1=\frac{1}{\alpha'}+\frac{1}{\beta'}$ and $\alpha,\alpha'<1$. Moreover, if one takes the limit of $\alpha'\to 1^-$, which implies $\beta'\to-\infty$, then one recovers the following with $0<\alpha<1$:

$$R_B \geq \operatorname*{ess\,inf}_{\mathcal{P}_{\hat{W}}} \left(\mathcal{P}_W^{\frac{1}{\beta}}\left(\ell(W,\hat{W})^\beta\right)\right)\cdot\exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W,X)\right). \tag{190}$$

Proof. The proof of Equation (189) follows from the same proof of Theorem 5 (cf. (Esposito, 2022, Theorem 15)) with $f=\ell$ but using reverse Hölder's inequality rather than regular Hölder's inequality. Considering the limit of $\alpha'\to 1^-$ in Equation (189) one recovers the following:

$$\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) \geq \operatorname*{ess\,inf}_{\mathcal{P}_{\hat{W}}} \left(\mathcal{P}_W^{\frac{1}{\beta}}\left(\ell(W,\hat{W})^\beta\right)\right)\cdot\exp\left(\operatorname{sign}(\alpha)\cdot\frac{\alpha-1}{\alpha}I_\alpha(W,\hat{W})\right) \tag{191}$$

Now, if $0<\alpha<1$ then $\frac{1}{\beta}<0$. By the Data-Processing Inequality for I_α with $0<\alpha<1$ (along with the negativity of $\frac{1}{\beta}$) one has that

$$\exp\left(\operatorname{sign}(\alpha)\cdot\frac{\alpha-1}{\alpha}I_\alpha(W,\hat{W})\right) = \exp\left(\frac{1}{\beta}I_\alpha(W,\hat{W})\right) \geq \exp\left(\frac{1}{\beta}I_\alpha(W,X)\right). \tag{192}$$

The lower bound on the Risk follows by noticing that the right-hand side Equation (191) can be rendered independent of $\hat{W}$ for every $\alpha<1$ (*i.e.*, it will only depend on the support of $\hat{W}$ through the ess inf) via Equation (192). □

Remark 26 (Extending to $\alpha<0$)**.** *One could also extend the result to $\alpha<0$ (which implies $0<\beta<1$), however, this would lead to a notion of I_α for $\alpha<0$ (see (Esposito et al., 2022)) which is outside the scope of this work. However, in that case, one would have the following interesting limiting behavior when $\alpha\to-\infty$:*

$$\begin{aligned}\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W})) &\geq \left(\operatorname*{ess\,inf}_{\mathcal{P}_{\hat{W}}}\mathcal{P}_W\left(\ell(W,\hat{w})\right)\right)\left(\int_{\hat{\mathcal{W}}}\operatorname*{ess\,inf}_{\mathcal{P}_W}\mathcal{P}_{\hat{W}|W}\right) &(193)\\ &= \left(\operatorname*{ess\,inf}_{\mathcal{P}_{\hat{W}}}\mathcal{P}_W\left(\ell(W,\hat{w})\right)\right)\exp\left(-\mathcal{L}^c(W\to\hat{W})\right), &(194)\end{aligned}$$

where $\mathcal{L}^c(W\to\hat{W})$ represents maximal cost-leakage (Issa et al., 2020, Definition 11).

Corollary 25 is different from the results presented in the previous section. While in Section 4 the only dependence on ℓ was through the small-ball probability, in Corollary 25 one is required to have access to the expected value of the β-th moments of ℓ with respect to $\mathcal{P}_X$. Such an object may not be as easy to bound as the small-ball probability.

Remark 27. *If $W=\hat{W}$ then $\ell(W,\hat{W})=0$ and $I_\alpha(W,\hat{W})=0$. If $0<\alpha<1$, given that $\beta<0$, one recovers the following lower bound on the risk, which matches with our intuition: $\mathcal{P}_{W\hat{W}}(\ell(W,\hat{W}))\geq 0$.*

References

Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, pages 925–939, 1976.

Venkat Anantharam. A variational characterization of Rényi divergences. *IEEE Transactions on Information Theory*, 64(11):6979–6989, 2018. doi: 10.1109/TIT.2018.2861013.

Shahab Asoodeh, Maryam Aliakbarpour, and Flavio P. Calmon. Local differential privacy is equivalent to contraction of an f-divergence. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 545–550, 2021. doi: 10.1109/ISIT45174.2021.9517999.

T. Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of Rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021. doi: 10.1137/20M1368926.

Sergey G. Bobkov and Gennadiy P. Chistyakov. On concentration functions of random variables. *Journal of Theoretical Probability volume*, 28, 2015.

Michel Broniatowski and Amor Keziou. Minimization of divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006. doi: 10.1556/SScMath.43.2006.4.2.

Lawrence D. Brown and Leslaw Gajek. Information inequalities for the Bayes risk. *The Annals of Statistics*, 18(4):1578–1594, 1990. ISSN 00905364.

L.D. Brown and R.C. Liu. Bounds on the Bayes and minimax risk for signal parameter estimation. *IEEE Transactions on Information Theory*, 39(4):1386–1394, 1993. doi: 10.1109/18.243453.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *J. Mach. Learn. Res.*, 17(1):7687–7744, jan 2016. ISSN 1532-4435.

Joel E. Cohen, Yoh Iwasa, Gh. Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993. ISSN 0024-3795.

Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972.

Pierre Del Moral, Michel Ledoux, and Laurent Miclo. On contraction properties of markov kernels. *Probability Theory and Related Fields*, 126:395–420, 2003.

John C. Duchi and Martin J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *CoRR*, abs/1311.2669, 2013.

Amedeo Roberto Esposito. *A Functional Perspective on Information Measures*. PhD thesis, EPFL, Lausanne, 2022.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021a. doi: 10.1109/TIT.2021.3085190.

Amedeo Roberto Esposito, Diyuan Wu, and Michael Gastpar. On conditional Sibson's α-mutual information. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1796–1801, 2021b. doi: 10.1109/ISIT45174.2021.9517944.

Amedeo Roberto Esposito, Adrien Vandenbroucque, and Michael Gastpar. On Sibson's α-mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, June 2022. doi: 10.1109/isit50566.2022.9834428.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968. ISBN 0471257087.

Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, 1989.

I. Issa, A. B. Wagner, and S. Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2020. doi: 10.1109/TIT.2019.2962804.

Wenbo Li and Qi-Man Shao. Gaussian processes: Inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19, 12 2001. doi: 10.1016/S0169-7161(01)19019-X.

F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, Leipzig, 1987.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.881731.

A. Makur and L Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56:103–156, 2020. doi: https://doi.org/10.1134/S0032946020020015.

Hoi H. Nguyen and Van H. Vu. *Small Ball Probability, Inverse Theorems, and Applications*, pages 409–463. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-39286-3. doi: 10.1007/978-3-642-39286-3_16.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010. doi: 10.1109/TIT.2010.2043769.

Maxim Raginsky. Strong data processing inequalities and ϕ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.

Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures.* 05 2015. ISBN 978-0-8218-7578-0.

R. Tyrrell Rockafellar. *Convex analysis.* Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151.

Michikazu Sato and Masafumi Akahira. An information inequality for the Bayes risk. *The Annals of Statistics*, 24(5):2288–2295, 1996. ISSN 00905364.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 163–171. Curran Associates, Inc., 2014.

R. Sibson. Information radius. *Z. Wahrscheinlichkeitstheorie verw Gebiete 14*, pages 149–160, 1969.

Te Sun Han and S. Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, 1998. doi: 10.1109/18.720540.

T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014.

Harry L. Van Trees. *Detection, Estimation, and Modulation Theory.* John Wiley & Sons, Ltd, 2001. ISBN 9780471221081.

Harry L. Van Trees and Kristine L. Bell. *Bounds on the Bayes and Minimax Risk for Signal Parameter Estimation*, pages 329–337. 2007. doi: 10.1109/9780470544198.ch28.

S.R.S. Varadhan. *Large Deviations and Applications.* 1984.

Sergio Verdú. α-mutual information. In *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, pages 1–6, 2015.

A. Xu and M. Raginsky. Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.