Statistical Science
2020, Vol. 35, No. 2, 178–201
https://doi.org/10.1214/19-STS716

Fano's Inequality for Random Variables

Sébastien Gerchinovitz, Pierre Ménard and Gilles Stoltz

Abstract. We extend Fano's inequality, which controls the average probability of events in terms of the average of some f-divergences, to work with arbitrary events (not necessarily forming a partition) and even with arbitrary $[0, 1]$-valued random variables, possibly in continuously infinite number. We provide two applications of these extensions, in which the consideration of random variables is particularly handy: we offer new and elegant proofs for existing lower bounds, on Bayesian posterior concentration (minimax or distribution-dependent) rates and on the regret in nonstochastic sequential learning.

Key words and phrases: Multiple-hypotheses testing, lower bounds, information theory, Bayesian posterior concentration.

1. INTRODUCTION

Fano's inequality is a popular information-theoretical result that provides a lower bound on worst-case error probabilities in multiple-hypotheses testing problems. It has important consequences in information theory (Cover and Thomas, 2006) and related fields. In mathematical statistics, it has become a key tool to derive lower bounds on minimax (worst-case) rates of convergence for various statistical problems such as nonparametric density estimation, regression and classification (see, e.g., Tsybakov, 2009, Massart, 2007).

Multiple variants of Fano's inequality have been derived in the literature. They can handle a finite, countable or even continuously infinite number of hypotheses. Depending on the community, it has been stated in various ways. In this article, we focus on statistical versions of Fano's inequality. For instance, its most classical version states that for all sequences of $N \geq 2$ probability distributions $\mathbb{P}_1, \ldots, \mathbb{P}_N$ on the same measurable space $(\Omega, \mathcal{F})$, and all events $A_1, \ldots, A_N$ forming a partition of Ω,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leq \frac{\frac{1}{N}\inf_{\mathbb{Q}} \sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \ln(2)}{\ln(N)},$$

where the infimum in the right-hand side is over all probability distributions $\mathbb{Q}$ on $(\Omega, \mathcal{F})$. The link to multiple-hypotheses testing is by considering events of the form $A_i = \{\widehat{\theta} = i\}$, where $\widehat{\theta}$ is an estimator of θ. Lower bounds on the average of the $\mathbb{P}_i(\widehat{\theta} \neq i)$ are then obtained.

Several extensions to more complex settings were derived in the past. For example, Han and Verdú (1994) addressed the case of countably infinitely many probability distributions, while Duchi and Wainwright (2013) and Chen, Guntuboyina and Zhang (2016) further generalized Fano's inequality to continuously infinitely many distributions; see also Aeron, Saligrama and Zhao (2010). Gushchin (2003) extended Fano's inequality in two other directions, first by considering $[0, 1]$-valued random variables Z_i such that $Z_1 + \cdots + Z_N = 1$, instead of the special case $Z_i = \mathbb{1}_{A_i}$, and second, by considering f-divergences. All these extensions, as well as others recalled in Section 7, provide a variety of tools that adapt nicely to the variety of statistical problems.

Content and Outline of This Article

In this article, we first revisit and extend Fano's inequality and then provide new applications. More precisely, Section 2 recalls the definition of f-divergences and states our main ingredient for our extended Fano's inequality, namely, a data-processing inequality with expectations of random variables. The short Section 3 is a pedagogical version of the longer Section 4, where we explain and illustrate our two-step methodology to establish new versions of Fano's inequality: a Bernoulli reduction is followed by careful lower bounds on the f-divergences between two Bernoulli distributions. In particular, we are able to extend Fano's inequality to both continuously many distributions $\mathbb{P}_\theta$ and arbitrary events

Sébastien Gerchinovitz is Assistant Professor, Institut de Mathématiques de Toulouse—Université Paul Sabatier, Toulouse, France (e-mail: sebastien.gerchinovitz@math.univ-toulouse.fr). Pierre Ménard is Ph.D. Student, Institut de Mathématiques de Toulouse—Université Paul Sabatier, Toulouse, France (e-mail: pierre.menard@math.univ-toulouse.fr). Gilles Stoltz is Senior research fellow, Laboratoire de Mathématiques d'Orsay—Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France & Affiliate Professor, GREGHEC—HEC Paris, CNRS, Jouy-en-Josas, France (e-mail: gilles.stoltz@math.u-psud.fr).

A_θ that do not necessarily form a partition or to arbitrary $[0,1]$-valued random variables Z_θ that are not required to sum up (or integrate) to 1. We also point out that the alternative distribution $\mathbb{Q}$ could vary with θ. We then move on in Section 5 to our main new statistical applications, illustrating in particular that it is handy to be able to consider random variables not necessarily summing up to 1. The two main such applications deal with Bayesian posterior concentration lower bounds and a regret lower bound in nonstochastic sequential learning. (The latter application, however, could be obtained by the extension by Gushchin, 2003.) Section 6 presents two other applications which—perhaps surprisingly—follow from the special case $N = 1$ in Fano's inequality. One of these applications is about distribution-dependent lower bounds on Bayesian posterior concentration (elaborating on results by Hoffmann, Rousseau and Schmidt-Hieber, 2015). The end of the article provides a review of the literature in Section 7; it explains, in particular, that the Bernoulli reduction lying at the heart of our analysis was already present, at various levels of clarity, in earlier works. Finally, Section 8 provides new and simpler proofs of some important lower bounds on the Kullback–Leibler divergence, the main contributions being a short and enlightening proof of the refined Pinsker's inequality by Ordentlich and Weinberger (2005), and a sharper Bretagnolle and Huber (1978, 1979) inequality.

2. DATA-PROCESSING INEQUALITY WITH EXPECTATIONS OF RANDOM VARIABLES

This section collects the definition of and some well-known results about f-divergences, a special case of which is given by the Kullback–Leibler divergence. It also states a recent and less known result, called the data-processing inequality with expectations of random variables; it will be at the heart of the derivation of our new Fano's inequality for random variables.

2.1 Kullback–Leibler Divergence

Let $\mathbb{P}, \mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$. We write $\mathbb{P} \ll \mathbb{Q}$ to indicate that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. The Kullback–Leibler divergence $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ is defined by

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \begin{cases} \displaystyle\int_\Omega \ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} & \text{if } \mathbb{P} \ll \mathbb{Q}; \\ +\infty & \text{otherwise.} \end{cases}$$

We write $\mathrm{Ber}(p)$ for the Bernoulli distribution with parameter p. We also use the usual measure-theoretic conventions in $\mathbb{R} \cup \{+\infty\}$; in particular $0 \times (+\infty) = 0$ and $1/0 = +\infty$, as well as $0/0 = 0$. We also set $\ln(0) = -\infty$ and $0 \ln(0) = 0$.

The Kullback–Leibler divergence function kl between Bernoulli distributions equals, for all $(p, q) \in [0,1]^2$,

$$\begin{aligned} \mathrm{kl}(p, q) &\stackrel{\mathrm{def}}{=} \mathrm{KL}\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) \\ &= p \ln\left(\frac{p}{q}\right) + (1-p) \ln\left(\frac{1-p}{1-q}\right). \end{aligned}$$

Kullback–Leibler divergences are actually a special case of f-divergences with $f(x) = x \ln x$; see Csiszár (1963), Ali and Silvey (1966) and Gushchin (2003) for further details.

2.2 f-Divergences

Let $f : (0, +\infty) \to \mathbb{R}$ be any convex function satisfying $f(1) = 0$. By convexity, we can define

$$f(0) \stackrel{\mathrm{def}}{=} \lim_{t \downarrow 0} f(t) \in \mathbb{R} \cup \{+\infty\};$$

the extended function $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ is still convex.

Before we may actually state the definition of f-divergences, we recall the definition of the maximal slope M_f of a convex function f and provide notation for the Lebesgue decomposition of measures.

Maximal slope. For any $x > 0$, the limit

$$\lim_{t \to +\infty} \frac{f(t) - f(x)}{t - x} = \sup_{t > 0} \frac{f(t) - f(x)}{t - x} \in [0, +\infty]$$

exists since (by convexity) the slope $(f(t) - f(x))/(t - x)$ is nondecreasing as t increases. Besides, this limit does not depend on x and equals

$$M_f \stackrel{\mathrm{def}}{=} \lim_{t \to +\infty} \frac{f(t)}{t} \in (-\infty, +\infty],$$

which thus represents the maximal slope of f. A useful inequality following from the two equations above with $t = x + y$ is

$$\forall x > 0,\, y > 0, \quad \frac{f(x+y) - f(x)}{y} \le M_f.$$

Put differently,

$$\forall x \ge 0,\, y \ge 0, \quad f(x+y) \le f(x) + y M_f, \tag{2.1}$$

where the extension to $y = 0$ is immediate and the one to $x = 0$ follows by continuity of f on $(0, +\infty)$, which itself follows from its convexity.

Lebesgue decomposition of measures. We recall that $\ll$ denotes the absolute continuity between measures and we let $\perp$ denote the fact that two measures are singular. For distributions $\mathbb{P}$ and $\mathbb{Q}$ defined on the same measurable space $(\Omega, \mathcal{F})$, the Lebesgue decomposition of $\mathbb{P}$ with respect to $\mathbb{Q}$ is denoted by

$$\begin{aligned} &\mathbb{P} = \mathbb{P}_{\mathrm{ac}} + \mathbb{P}_{\mathrm{sing}} \\ &\text{where } \mathbb{P}_{\mathrm{ac}} \ll \mathbb{Q} \text{ and } \mathbb{P}_{\mathrm{sing}} \perp \mathbb{Q}, \end{aligned} \tag{2.2}$$

so that $\mathbb{P}_{\mathrm{ac}}$ and $\mathbb{P}_{\mathrm{sing}}$ are both sub-probabilities (positive measures with total mass smaller than or equal to 1) and, by definition,

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}.$$

Definition of f-divergences. The existence of the integral in the right-hand side of the definition below follows from the general form of Jensen's inequality stated in Lemma C.2 (Appendix C) with $\varphi = f$ and $C = [0, +\infty)$.

DEFINITION 1. Given a convex function $f : (0, +\infty) \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q})$ between two probability distributions on the same measurable space $(\Omega, \mathcal{F})$ is defined as

$$\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) = \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega) M_f. \tag{2.3}$$

Jensen's inequality of Lemma C.2, together with (2.1), also indicates that $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) \geq 0$. Indeed,

$$\int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{Q} \geq f\left(\int_\Omega \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \mathrm{d}\mathbb{Q}\right) = f\big(\mathbb{P}_{\mathrm{ac}}(\Omega)\big),$$

so that by (2.1),

$$\begin{aligned}\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) &\geq f\big(\mathbb{P}_{\mathrm{ac}}(\Omega)\big) + \mathbb{P}_{\mathrm{sing}}(\Omega) M_f \\ &\geq f\big(\mathbb{P}_{\mathrm{ac}}(\Omega) + \mathbb{P}_{\mathrm{sing}}(\Omega)\big) = f(1) = 0.\end{aligned}$$

Concrete and important examples of f-divergences, such as the Hellinger distance and the χ^2-divergence, are discussed in details in Section 4. The Kullback–Leibler divergence corresponds to Div_f with the function $f : x \mapsto x \ln(x)$. We have $M_f = +\infty$ for the Kullback–Leibler and χ^2-divergences, while $M_f = 1$ for the Hellinger distance.

2.3 The Data-Processing Inequality and Two Major Consequences

The data-processing inequality (also called contraction of relative entropy in the case of the Kullback–Leibler divergence) indicates that transforming the data at hand can only reduce the ability to distinguish between two probability distributions.

LEMMA 2.1 (Data-processing inequality). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}^X$ and $\mathbb{Q}^X$ the associated pushforward measures (the laws of X under $\mathbb{P}$ and $\mathbb{Q}$). Then,*

$$\mathrm{Div}_f(\mathbb{P}^X, \mathbb{Q}^X) \leq \mathrm{Div}_f(\mathbb{P}, \mathbb{Q}).$$

COROLLARY 2.2 (Data-processing inequality with expectations of random variables). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$ taking values in $[0, 1]$. Denote by $\mathbb{E}_{\mathbb{P}}[X]$ and $\mathbb{E}_{\mathbb{Q}}[X]$ the expectations of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{div}_f\big(\mathbb{E}_{\mathbb{P}}[X], \mathbb{E}_{\mathbb{Q}}[X]\big) \leq \mathrm{Div}_f(\mathbb{P}, \mathbb{Q}),$$

where $\mathrm{div}_f(p, q) = \mathrm{Div}_f(\mathrm{Ber}(p), \mathrm{Ber}(q))$ denotes the f-divergence between Bernoulli distributions with respective parameters p and q.

COROLLARY 2.3 (Joint convexity of Div_f). *All f-divergences Div_f are jointly convex, that is, for all probability distributions $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}_1, \mathbb{Q}_2$ on the same measurable space $(\Omega, \mathcal{F})$, and all $\lambda \in (0, 1)$,*

$$\begin{aligned}&\mathrm{Div}_f\big((1-\lambda)\mathbb{P}_1 + \lambda\mathbb{P}_2, (1-\lambda)\mathbb{Q}_1 + \lambda\mathbb{Q}_2\big) \\ &\quad\leq (1-\lambda)\mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2).\end{aligned}$$

Lemma 2.1 and Corollary 2.3 are folklore knowledge. However, for the sake of self-completeness, we provide complete and elementary proofs thereof in the extended version of this article (see Appendix D). The proof of Lemma 2.1 is extracted from Ali and Silvey (1966), Section 4.2 (see also Pardo (2006), Proposition 1.2), while we derive Corollary 2.3 as an elementary consequence of Lemma 2.1 applied to an augmented probability space. These proof techniques do not seem to be well known; indeed, in the literature many proofs of the elementary properties above for the Kullback–Leibler divergence focus on the discrete case (Cover and Thomas, 2006) or use the duality formula for the Kullback–Leibler divergence (Massart, 2007 or Boucheron, Lugosi and Massart, 2013, in particular Exercise 4.10 therein).

On the contrary, Corollary 2.2 is a recent though elementary result, proved in Garivier, Ménard and Stoltz (2019) for Kullback–Leibler divergences. The proof readily extends to f-divergences.

PROOF OF COROLLARY 2.2. We augment the underlying measurable space into $\Omega \times [0, 1]$, where $[0, 1]$ is equipped with the Borel σ-algebra $\mathcal{B}([0, 1])$ and the Lebesgue measure $\mathfrak{m}$. We denote by $\mathbb{P} \otimes \mathfrak{m}$ and $\mathbb{Q} \otimes \mathfrak{m}$ the product distributions of $\mathbb{P}$ and $\mathfrak{m}$, $\mathbb{Q}$ and $\mathfrak{m}$. We write the Lebesgue decomposition $\mathbb{P} = \mathbb{P}_{\mathrm{ac}} + \mathbb{P}_{\mathrm{sing}}$ of $\mathbb{P}$ with respect to $\mathbb{Q}$, and deduce from it the Lebesgue decomposition of $\mathbb{P} \otimes \mathfrak{m}$ with respect to $\mathbb{Q} \otimes \mathfrak{m}$: the absolutely continuous part is given by $\mathbb{P}_{\mathrm{ac}} \otimes \mathfrak{m}$, with density

$$(\omega, x) \in \Omega \times [0, 1] \longmapsto \frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}} \otimes \mathfrak{m})}{\mathrm{d}(\mathbb{Q} \otimes \mathfrak{m})}(\omega, x) = \frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(\omega),$$

while the singular part is given by $\mathbb{P}_{\mathrm{sing}} \otimes \mathfrak{m}$, a subprobability with total mass $\mathbb{P}_{\mathrm{sing}}(\Omega)$. In particular,

$$\mathrm{Div}_f(\mathbb{P} \otimes \mathfrak{m}, \mathbb{Q} \otimes \mathfrak{m}) = \mathrm{Div}_f(\mathbb{P}, \mathbb{Q}).$$

Now, for all events $E \in \mathcal{F} \otimes \mathcal{B}([0, 1])$, the data-processing inequality (Lemma 2.1) used with the indicator function

$X = \mathbb{1}_E$ ensures that

$$\begin{aligned}&\mathrm{Div}_f(\mathbb{P}\otimes\mathfrak{m}, \mathbb{Q}\otimes\mathfrak{m})\\ &\quad\geq \mathrm{Div}_f\big((\mathbb{P}\otimes\mathfrak{m})^{\mathbb{1}_E}, (\mathbb{Q}\otimes\mathfrak{m})^{\mathbb{1}_E}\big)\\ &\quad= \mathrm{div}_f\big((\mathbb{P}\otimes\mathfrak{m})(E), (\mathbb{Q}\otimes\mathfrak{m})(E)\big),\end{aligned}$$

where the final equality is by mere definition of div_f as the f-divergence between Bernoulli distributions. The proof is concluded by noting that for the choice of $E = \{(\omega, x) \in \Omega \times [0,1] : x \leq X(\omega)\}$, Tonelli's theorem ensures that

$$\begin{aligned}(\mathbb{P}\otimes\mathfrak{m})(E) &= \int_\Omega \left(\int_{[0,1]} \mathbb{1}_{\{x\leq X(\omega)\}}\,\mathrm{d}\mathfrak{m}(x)\right)\mathrm{d}\mathbb{P}(\omega)\\ &= \mathbb{E}_{\mathbb{P}}[X]\end{aligned}$$

and, similarly, $(\mathbb{Q}\otimes\mathfrak{m})(E) = \mathbb{E}_{\mathbb{Q}}[X]$. □

3. HOW TO DERIVE A FANO-TYPE INEQUALITY: AN EXAMPLE

In this section, we explain on an example the methodology to derive Fano-type inequalities. We will present the generalization of the approach and the resulting bounds in Section 4, but the proof below already contains the two key arguments: a reduction to Bernoulli distributions, and a lower bound on the f-divergence between Bernoulli distributions. For the sake of concreteness, we focus on the Kullback–Leibler divergence in this section. We recall that we will discuss how novel (or not novel) our results and approaches are in Section 7.

PROPOSITION 3.1. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i, \mathbb{Q}_i$ and all events A_i (not necessarily disjoint), where $i \in \{1,\dots,N\}$, with $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, we have*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leq \frac{\frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}_i) + \ln(2)}{-\ln(\frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i))}.$$

In particular, if $N \geq 2$ and the A_i form a partition,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leq \frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}) + \ln(2)}{\ln(N)}.$$

PROOF. Our first step is to reduce the problem to Bernoulli distributions. Using first the joint convexity of the Kullback–Leibler divergence (Corollary 2.3), and second the data-processing inequality with the indicator functions $X = \mathbb{1}_{A_i}$ (Lemma 2.1), we get

$$\begin{aligned}&\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i), \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\right)\\ &\quad\leq \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \qquad (3.1)\\ &\quad\leq \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}_i).\end{aligned}$$

Therefore, we have $\mathrm{kl}(\overline{p},\overline{q}) \leq \overline{K}$ with

$$\begin{aligned}\overline{p} &= \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i), \quad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i),\\ \overline{K} &= \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}_i).\end{aligned} \qquad (3.2)$$

Our second and last step is to lower bound $\mathrm{kl}(\overline{p},\overline{q})$ to extract an upper bound on $\overline{p}$. Noting that $\overline{p}\ln(\overline{p}) + (1-\overline{p})\ln(1-\overline{p}) \geq -\ln(2)$, we have, by definition of $\mathrm{kl}(\overline{p},\overline{q})$,

$$\begin{aligned}&\mathrm{kl}(\overline{p},\overline{q}) \geq \overline{p}\ln(1/\overline{q}) - \ln(2)\\ &\text{thus } \overline{p} \leq \frac{\mathrm{kl}(\overline{p},\overline{q}) + \ln(2)}{\ln(1/\overline{q})},\end{aligned} \qquad (3.3)$$

where $\overline{q} \in (0,1)$ by assumption. Substituting the upper bound $\mathrm{kl}(\overline{p},\overline{q}) \leq \overline{K}$ in (3.3) concludes the proof. □

4. VARIOUS FANO-TYPE INEQUALITIES, WITH THE SAME TWO INGREDIENTS

We extend the approach of Section 3 and derive a broad family of Fano-type inequalities, which will be of the form

$$\overline{p} \leq \psi(\overline{q}, \overline{K}),$$

where the average quantities $\overline{p}$, $\overline{q}$ and $\overline{K}$ are described in Section 4.1 (first ingredient) and where the functions ψ are described in Section 4.2 (second ingredient). The simplest example that we considered in Section 3 corresponds to $\psi(q,K) = (K + \ln(2))/\ln(1/q)$ and

$$\begin{aligned}\overline{p} &= \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i), \quad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i),\\ \overline{K} &= \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}_i).\end{aligned}$$

We address here the more general cases where the finite averages are replaced with integrals over any measurable space Θ and where the indicator functions $\mathbb{1}_{A_i}$ are replaced with arbitrary $[0,1]$-valued random variables Z_θ, where $\theta \in \Theta$.

We recall that the novelty (or lack of novelty) of our results will be discussed in detail in Section 7; of particular interest therein is the discussion of the (lack of) novelty of our first ingredient, namely the reduction to Bernoulli distributions.

4.1 Reduction to Bernoulli Distributions

As in Section 3, we can resort to the data-processing inequality (Lemma 2.1) to lower bound any f-divergence by that of suitably chosen Bernoulli distributions. We present three such reductions, in increasing degree of generality. We only indicate how to prove the first one, since they are all similar.

Countably many distributions. We consider some underlying measurable space, countably many pairs of probability distributions $\mathbb{P}_i, \mathbb{Q}_i$ on this space, not necessarily disjoint events A_i, all indexed by $i \in \{1, 2, \ldots\}$, as well as a convex combination $\alpha = (\alpha_1, \alpha_2, \ldots)$. The latter can be thought of as a prior distribution. The inequality reads

$$\begin{aligned} &\operatorname{div}_f\Big(\sum_{i\geq 1}\alpha_i\mathbb{P}_i(A_i), \sum_{i\geq 1}\alpha_i\mathbb{Q}_i(A_i)\Big) \\ &\leq \sum_{i\geq 1}\alpha_i \operatorname{div}_f\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \\ &\leq \sum_{i\geq 1}\alpha_i \operatorname{Div}_f(\mathbb{P}_i, \mathbb{Q}_i). \end{aligned} \tag{4.1}$$

The second inequality of (4.1) follows from the data-processing inequality (Lemma 2.1) by considering the indicator functions $X = \mathbb{1}_{A_i}$. For the first inequality, we resort to a general version of Jensen's inequality stated in Lemma C.2 (Appendix C), by considering the convex function $\varphi = \operatorname{div}_f$ (Corollary 2.3) on the convex set $C = [0, 1]^2$, together with the probability measure

$$\mu = \sum_i \alpha_i \delta_{(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i))},$$

where $\delta_{(x,y)}$ denotes the Dirac mass at $(x, y) \in \mathbb{R}^2$.

Distributions indexed by a possibly continuous set. Up to measurability issues (that are absent in the countable case), the reduction above immediately extends to the case of statistical models $\mathbb{P}_\theta, \mathbb{Q}_\theta$ and not necessarily disjoint events A_θ indexed by a measurable parameter space $(\Theta, \mathcal{G})$, equipped with a prior probability distribution ν on Θ. We assume that

$$\begin{aligned} &\theta \in \Theta \longmapsto \big(\mathbb{P}_\theta(A_\theta), \mathbb{Q}_\theta(A_\theta)\big) \quad \text{and} \\ &\theta \in \Theta \longmapsto \operatorname{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta) \end{aligned}$$

are $\mathcal{G}$-measurable and get the reduction

$$\begin{aligned} &\operatorname{div}_f\Big(\int_\Theta \mathbb{P}_\theta(A_\theta)\,\mathrm{d}\nu(\theta), \int_\Theta \mathbb{Q}_\theta(A_\theta)\,\mathrm{d}\nu(\theta)\Big) \\ &\leq \int_\Theta \operatorname{div}_f\big(\mathbb{P}_\theta(A_\theta), \mathbb{Q}_\theta(A_\theta)\big)\,\mathrm{d}\nu(\theta) \\ &\leq \int_\Theta \operatorname{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)\,\mathrm{d}\nu(\theta). \end{aligned} \tag{4.2}$$

Random variables. In the reduction above, it was unnecessary that the sets A_θ form a partition or even be disjoint. It is therefore not surprising that it can be generalized by replacing the indicator functions $\mathbb{1}_{A_\theta}$ with arbitrary [0, 1]-valued random variables Z_θ. We denote the expectations of the latter with respect to $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta$ by $\mathbb{E}_{\mathbb{P}_\theta}$ and $\mathbb{E}_{\mathbb{Q}_\theta}$ and assume that

$$\begin{aligned} &\theta \in \Theta \longmapsto \big(\mathbb{E}_{\mathbb{P}_\theta}[Z_\theta], \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]\big) \quad \text{and} \\ &\theta \in \Theta \longmapsto \operatorname{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta) \end{aligned}$$

are $\mathcal{G}$-measurable. The reduction reads in this case

$$\begin{aligned} &\operatorname{div}_f\Big(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}[Z_\theta]\,\mathrm{d}\nu(\theta), \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]\,\mathrm{d}\nu(\theta)\Big) \\ &\leq \int_\Theta \operatorname{div}_f\big(\mathbb{E}_{\mathbb{P}_\theta}[Z_\theta], \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]\big)\,\mathrm{d}\nu(\theta) \\ &\leq \int_\Theta \operatorname{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)\,\mathrm{d}\nu(\theta), \end{aligned} \tag{4.3}$$

where the first inequality relies on convexity of div_f and on Jensen's inequality, and the second inequality follows from the data-processing inequality with expectations of random variables (Lemma 2.2).

4.2 Any Lower Bound on div_f Leads to a Fano-Type Inequality

The section above indicates that after the reduction to the Bernoulli case, we get inequations of the form ($\overline{p}$ is usually the unknown)

$$\operatorname{div}_f(\overline{p}, \overline{q}) \leq \overline{D},$$

where $\overline{D}$ is an average of f-divergences, and $\overline{p}$ and $\overline{q}$ are averages of probabilities of events or expectations of [0, 1]-valued random variables. We thus proceed by lower bounding the div_f function. The lower bounds are idiosyncratic to each f-divergence and we start with the most important one, namely, the Kullback–Leibler divergence.

Lower bounds on kl. The most classical bound was already used in Section 3: for all $p \in [0, 1]$ and $q \in (0, 1)$,

$$\begin{aligned} &\operatorname{kl}(p, q) \geq p\ln(1/q) - \ln(2), \\ &\text{thus } p \leq \frac{\operatorname{kl}(p, q) + \ln(2)}{\ln(1/q)}. \end{aligned} \tag{4.4}$$

It is well known that this bound can be improved by replacing the term $\ln(2)$ with $\ln(2 - q)$: for all $p \in [0, 1]$ and $q \in (0, 1)$,

$$\begin{aligned} &\operatorname{kl}(p, q) \geq p\ln(1/q) - \ln(2 - q), \\ &\text{thus } p \leq \frac{\operatorname{kl}(p, q) + \ln(2 - q)}{\ln(1/q)}. \end{aligned} \tag{4.5}$$

This leads to a nontrivial bound even if $q = 1/2$ (as is the case in some applications). A (novel) consequence of this bound is that

$$p \leq 0.21 + 0.79q + \frac{\operatorname{kl}(p, q)}{\ln(1/q)}. \tag{4.6}$$

The improvement (4.5) is a consequence of, for example, a convexity inequality, and its proof and the one for (4.6) can be found in Section 8.1.

The next and final bound makes a connection between Pinsker's and Fano's inequalities: on the one hand, it is a refined Pinsker's inequality and on the other hand, it leads

to a bound on p of the same flavor as (4.4)–(4.6). Namely, for all $p \in [0,1]$ and $q \in (0,1)$,

$$\begin{aligned} &\mathrm{kl}(p,q) \geq \max\left\{\ln\left(\frac{1}{q}\right), 2\right\}(p-q)^2, \\ &\text{thus } p \leq q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\{\ln(1/q), 2\}}}. \end{aligned} \tag{4.7}$$

The first inequality was stated and proved by Ordentlich and Weinberger (2005), the second is a novel but straightforward consequence of it. We provide their proofs and additional references in Section 8.2.

Lower bound on div_f *for the* χ^2*-divergence.* This case corresponds to $f(x) = x^2 - 1$. The associated divergence equals $+\infty$ when $\mathbb{P} \not\ll \mathbb{Q}$, and when $\mathbb{P} \ll \mathbb{Q}$,

$$\chi^2(\mathbb{P}, \mathbb{Q}) = \int_\Omega \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)^2 \mathrm{d}\mathbb{Q} - 1.$$

A direct calculation and the usual measure-theoretic conventions entail the following simple lower bound: for all $(p,q) \in [0,1]^2$,

$$\begin{aligned} &\chi^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) \\ &\quad = \frac{(p-q)^2}{q(1-q)} \geq \frac{(p-q)^2}{q}, \\ &\qquad \text{thus } p \leq q + \sqrt{q\chi^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big)}. \end{aligned} \tag{4.8}$$

Lower bound on div_f *for the Hellinger distance.* This case corresponds to $f(x) = (\sqrt{x} - 1)^2$, for which $M_f = 1$. The associated divergence equals, when $\mathbb{P} \ll \mathbb{Q}$,

$$\begin{aligned} H^2(\mathbb{P}, \mathbb{Q}) &= \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}} - 1\right)^2 \mathrm{d}\mathbb{Q} \\ &= 2\left(1 - \int_\Omega \sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}}\,\mathrm{d}\mathbb{Q}\right) \end{aligned}$$

and always lies in $[0,2]$. A direct calculation indicates that for all $p \in [0,1]$ and $q \in (0,1)$,

$$\begin{aligned} h^2(p,q) &\stackrel{\mathrm{def}}{=} H^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) \\ &= 2\big(1 - \big(\sqrt{pq} + \sqrt{(1-p)(1-q)}\big)\big), \end{aligned}$$

and further direct calculations in the cases $q = 0$ and $q = 1$ show that this formula remains valid in these cases. To get a lower bound on $h^2(p,q)$, we proceed as follows. The Cauchy–Schwarz inequality indicates that

$$\begin{aligned} &\sqrt{pq} + \sqrt{(1-q)(1-p)} \\ &\quad \leq \sqrt{\big(p + (1-q)\big)\big(q + (1-p)\big)} \\ &\quad = \sqrt{1 - (p-q)^2}, \end{aligned}$$

or put differently, that $h^2(p,q) \geq 2(1 - \sqrt{1-(p-q)^2})$, thus

$$\begin{aligned} p &\leq q + \sqrt{1 - \big(1 - h^2(p,q)/2\big)^2} \\ &= q + \sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}, \end{aligned} \tag{4.9}$$

which is one of Le Cam's inequalities. A slightly sharper but less readable bound was exhibited by Guntuboyina (2011), Example II.6, and is provided, for the sake of completeness, in an extended version of this article; see Appendix D.

4.3 Examples of Combinations

The combination of (4.2) and (4.4) yields a continuous version of Fano's inequality. (We discard again all measurability issues.)

LEMMA 4.1. *We consider a measurable space* $(\Theta, \mathcal{E})$ *equipped with a probability distribution* ν*. Given an underlying measurable space* $(\Omega, \mathcal{F})$*, for all two collections* $\mathbb{P}_\theta, \mathbb{Q}_\theta$*, of probability distributions on this space and all collections of events* A_θ *of* $(\Omega, \mathcal{F})$*, where* $\theta \in \Theta$*, with*

$$0 < \int_\Theta \mathbb{Q}_\theta(A_\theta)\,\mathrm{d}\nu(\theta) < 1,$$

we have

$$\begin{aligned} &\int_\Theta \mathbb{P}_\theta(A_\theta)\,\mathrm{d}\nu(\theta) \\ &\quad \leq \frac{\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)\,\mathrm{d}\nu(\theta) + \ln(2)}{-\ln\big(\int_\Theta \mathbb{Q}_\theta(A_\theta)\,\mathrm{d}\nu(\theta)\big)}. \end{aligned}$$

The combination of (4.2), used with a uniform distribution ν on N points, and (4.7) ensures the following Fano-type inequality for finitely many random variables, whose sum does not need to be 1. It will be used in our second application, in Section 5.2.

LEMMA 4.2. *Given an underlying measurable space, for all probability pairs* $\mathbb{P}_i, \mathbb{Q}_i$ *and for all* $[0,1]$*-valued random variables* Z_i *defined on this measurable space, where* $i \in \{1, \dots, N\}$*, with*

$$0 < \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[Z_i] < 1,$$

we have

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{P}_i}[Z_i] \\ &\quad \leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[Z_i] + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)}{-\ln\big(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[Z_i]\big)}}. \end{aligned}$$

In particular, if $N \geq 2$ *and* $Z_1 + \cdots + Z_N = 1$ *a.s., then*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{P}_i}[Z_i] \leq \frac{1}{N} + \sqrt{\frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q})}{\ln(N)}}.$$

For the χ^2-divergence now, the combination of, for example, (4.1) in the finite and uniform case and (4.8) leads to the following inequality.

LEMMA 4.3. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i, \mathbb{Q}_i$ and all events A_i (not necessarily disjoint), where $i \in \{1, \dots, N\}$, with $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, we have*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \le \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) + \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)}\sqrt{\frac{1}{N}\sum_{i=1}^N \chi^2(\mathbb{P}_i, \mathbb{Q}_i)}.$$

In particular, if $N \ge 2$ and the A_i form a partition,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \le \frac{1}{N} + \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \chi^2(\mathbb{P}_i, \mathbb{Q})}.$$

Similarly, for the Hellinger distance, the simplest reduction (4.1) in the finite and uniform case together with the lower bound (4.9) yields the following bound.

LEMMA 4.4. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i, \mathbb{Q}_i$ and all events A_i (not necessarily disjoint), where $i \in \{1, \dots, N\}$, with $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, we have*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \le \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) + \sqrt{\frac{1}{N}\sum_{i=1}^N H^2(\mathbb{P}_i, \mathbb{Q}_i)}\sqrt{1 - \frac{1}{4N}\sum_{i=1}^N H^2(\mathbb{P}_i, \mathbb{Q}_i)}.$$

In particular, if $N \ge 2$ and the A_i form a partition,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \le \frac{1}{N} + \inf_{\mathbb{Q}}\sqrt{\frac{1}{N}\sum_{i=1}^N H^2(\mathbb{P}_i, \mathbb{Q})} \times \sqrt{1 - \frac{1}{4N}\sum_{i=1}^N H^2(\mathbb{P}_i, \mathbb{Q})} \le \frac{1}{N} + \inf_{\mathbb{Q}}\sqrt{\frac{1}{N}\sum_{i=1}^N H^2(\mathbb{P}_i, \mathbb{Q})}.$$

4.4 Comments on These Bounds

Section A in Appendix discusses the sharpness of the bounds obtained above, for the case of the Kullback–Leibler divergence.

Section D provides a pointer to an extended version of this article where the choice of a good constant alternative distribution $\mathbb{Q}$ is studied. The examples of bounds derived in Section 4.3 show indeed that when the A_i form a partition, the upper bounds feature an average f-divergence of the form

$$\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q})$$

and one may indeed wonder what $\mathbb{Q}$ should be chosen and what bound can be achieved. Section D points to a discussion of these matters.

5. MAIN APPLICATIONS

We present two new applications of Fano's inequality, with $[0, 1]$-valued random variables Z_i or Z_θ. The topics covered are:

- Bayesian posterior concentration rates;
- robust sequential learning (prediction of individual sequences) in the case of sparse losses.

As can be seen below, the fact that we are now able to consider arbitrary $[0, 1]$-valued random variables Z_θ on a continuous parameter space Θ makes the proof of the Bayesian posterior concentration lower bound quite simple.

Two more applications will also be presented in Section 6; they have a different technical flavor, as they rely on only one pair of distributions, that is, $N = 1$.

5.1 Lower Bounds on Bayesian Posterior Concentration Rates

In the next paragraphs we show how our continuous Fano's inequality can be used in a simple fashion to derive lower bounds for posterior concentration rates.

Setting and Bayesian terminology. We consider the following density estimation setting: we observe a sample of independent and identically distributed random variables $X_{1:n} = (X_1, \dots, X_n)$ drawn from a probability distribution P_θ on $(\mathcal{X}, \mathcal{F})$, with a fixed but unknown $\theta \in \Theta$. We assume that the measurable parameter space $(\Theta, \mathcal{G})$ is equipped with a prior distribution π and that all $P_{\theta'}$ have a density $p_{\theta'}$ with respect to some reference measure $\mathfrak{m}$ on $(\mathcal{X}, \mathcal{F})$. We also assume that $(x, \theta') \mapsto p_{\theta'}(x)$ is $\mathcal{F} \otimes \mathcal{G}$-measurable. We can thus consider the transition kernel $(x_{1:n}, A) \mapsto \mathbb{P}_\pi(A|x_{1:n})$ defined for all $x_{1:n} \in \mathcal{X}^n$ and all sets $A \in \mathcal{G}$ by

$$\mathbb{P}_\pi(A|x_{1:n}) = \frac{\int_A \prod_{i=1}^n p_{\theta'}(x_i)\,\mathrm{d}\pi(\theta')}{\int_\Theta \prod_{i=1}^n p_{\theta'}(x_i)\,\mathrm{d}\pi(\theta')} \tag{5.1}$$

if the denominator lies in $(0, +\infty)$; if it is null or infinite, we set, for example, $\mathbb{P}_\pi(A|x_{1:n}) = \pi(A)$. The resulting random measure $\mathbb{P}_\pi(\cdot|X_{1:n})$ is known as the *posterior* distribution.

Let $\ell : \Theta \times \Theta \to \mathbb{R}_+$ be a measurable loss function that we assume to be a pseudo-metric.[1] A posterior concentration rate with respect to ℓ is a sequence $(\varepsilon_n)_{n\geq 1}$ of positive real numbers such that, for all $\theta \in \Theta$,

$$\mathbb{E}_\theta\big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) \leq \varepsilon_n | X_{1:n}\big)\big] \longrightarrow 1 \quad \text{as } n \to +\infty,$$

where $\mathbb{E}_\theta$ denotes the expectation with respect to $X_{1:n}$ where each X_j has the P_θ law. The above convergence guarantee means that, as the size n of the sample increases, the posterior mass concentrates in expectation on an ε_n-neighborhood of the true parameter θ. Several variants of this definition exist (e.g., convergence in probability or almost surely; or ε_n that may depend on θ). Though most of these definitions can be handled with the techniques provided below, we only consider this one for the sake of conciseness.

Minimax posterior concentration rate. As our sequence $(\varepsilon_n)_{n\geq 1}$ does not depend on the specific $\theta \in \Theta$ at hand, we may study uniform posterior concentration rates: sequences $(\varepsilon_n)_{n\geq 1}$ such that

$$\inf_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) \leq \varepsilon_n | X_{1:n}\big)\big] \longrightarrow 1 \quad \text{as } n \to +\infty. \tag{5.2}$$

The minimax posterior concentration rate is given by a sequence $(\varepsilon_n)_{n\geq 1}$ such that (5.2) holds for some prior π while there exists a constant $\gamma \in (0, 1)$ such that for all priors π' on Θ,

$$\limsup_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \leq \gamma\varepsilon_n | X_{1:n}\big)\big] < 1.$$

We focus on proving the latter statement and provide a general technique to do so.

PROPOSITION 5.1 (A posterior concentration lower bound in the finite-dimensional Gaussian model). *Let $d \geq 1$ be the ambient dimension, $n \geq 1$ the sample size, and $\sigma > 0$ the standard deviation. Assume we observe an n-sample $X_{1:n} = (X_1, \ldots, X_n)$ distributed according to $\mathcal{N}(\theta, \sigma^2 I_d)$ for some unknown $\theta \in \mathbb{R}^d$. Let π' be any prior distribution on $\mathbb{R}^d$. Then the posterior distribution $\mathbb{P}_{\pi'}(\cdot|X_{1:n})$ defined in (5.1) satisfies, for the Euclidean loss $\ell(\theta', \theta) = \|\theta' - \theta\|_2$ and for $\varepsilon_n = (\sigma/8)\sqrt{d/n}$,*

$$\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\big[\mathbb{P}_{\pi'}\big(\theta' : \|\theta' - \theta\|_2 \leq \varepsilon_n | X_{1:n}\big)\big] \leq c_d,$$

where $(c_d)_{d\geq 1}$ is a decreasing sequence such that $c_1 \leq 0.55$, $c_2 \leq 0.37$, and $c_d \to 0.21$ as $d \to +\infty$.

[1] The only difference with a metric is that we allow $\ell(\theta, \theta') = 0$ for $\theta \neq \theta'$.

This proposition indicates that the best possible posterior concentration rate is at best $\sigma\sqrt{d/n}$ up to a multiplicative constant; actually, this order of magnitude is the best achievable posterior concentration rate, see, for example, Le Cam and Yang (2000), Chapter 8.

There are at least two ways to prove the lower bound of Proposition 5.1. A first one is to use a well-known conversion of "good" Bayesian posteriors into "good" point estimators, which indicates that lower bounds for point estimation can be turned into lower bounds for posterior concentration. For the sake of completeness, we recall this conversion in Appendix B and provide a nonasymptotic variant of Theorem 2.5 by Ghosal, Ghosh and van der Vaart (2000).

The second method—followed in the proof below—is, however, more direct. We use our most general continuous Fano's inequality with the random variables $Z_\theta = \mathbb{P}_{\pi'}(\theta' : \|\theta' - \theta\|_2 \leq \varepsilon_n | X_{1:n}) \in [0, 1]$.

PROOF OF PROPOSITION 5.1. We may assume, with no loss of generality, that the probability space on which $X_{1:n}$ is defined is $(\mathbb{R}^d)^n$ endowed with its Borel σ-field and the probability measure $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^{\otimes n}$. Let ν denote the uniform distribution on the Euclidean ball $B(0, \rho\varepsilon_n) = \{u \in \mathbb{R}^d : \|u\|_2 \leq \rho\varepsilon_n\}$ for some $\rho > 1$ to be determined by the analysis. Then, by the continuous Fano inequality in the form given by the combination of (4.3) and (4.7), with $\mathbb{Q}_\theta = \mathbb{P}_0 = \mathcal{N}(0, \sigma^2)^{\otimes n}$, where 0 denotes the null vector of $\mathbb{R}^d$, and with the $[0, 1]$-valued random variables $Z_\theta = \mathbb{P}_{\pi'}(\theta' : \|\theta' - \theta\|_2 \leq \varepsilon_n | X_{1:n})$, we have

$$\begin{aligned}
\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta[Z_\theta] &\leq \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_\theta[Z_\theta]\,\mathrm{d}\nu(\theta) \\
&\leq \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0[Z_\theta]\,\mathrm{d}\nu(\theta) \\
&\quad + \sqrt{\frac{\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0)\,\mathrm{d}\nu(\theta)}{-\ln \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0[Z_\theta]\,\mathrm{d}\nu(\theta)}} \\
&\leq \left(\frac{1}{\rho}\right)^d + \sqrt{\frac{n\rho^2\varepsilon_n^2/(2\sigma^2)}{d\ln\rho}},
\end{aligned} \tag{5.3}$$

where the last inequality follows from (5.4) and (5.5) below. First note that, by independence, $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0) = n\,\mathrm{KL}(\mathcal{N}(\theta, \sigma^2), \mathcal{N}(0, \sigma^2)) = n\|\theta\|_2^2/(2\sigma^2)$, so that

$$\begin{aligned}
\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0)\,\mathrm{d}\nu(\theta) &= \frac{n}{2\sigma^2} \int_{B(0,\rho\varepsilon_n)} \|\theta\|_2^2\,\mathrm{d}\nu(\theta) \\
&\leq \frac{n\rho^2\varepsilon_n^2}{2\sigma^2}.
\end{aligned} \tag{5.4}$$

Second, using the Fubini–Tonelli theorem (twice) and the definition of

$$\begin{aligned}
Z_\theta &= \mathbb{P}_{\pi'}\big(\theta' : \|\theta' - \theta\|_2 \leq \varepsilon_n | X_{1:n}\big) \\
&= \mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot|X_{1:n})}\big[\mathbb{1}_{\{\|\theta'-\theta\|_2\leq\varepsilon_n\}}\big],
\end{aligned}$$

we can see that

$$
\begin{aligned}
q &\overset{\text{def}}{=} \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0[Z_\theta]\,\mathrm{d}\nu(\theta) \\
&= \mathbb{E}_0\Big[\int_{B(0,\rho\varepsilon_n)} \mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot|X_{1:n})}\big[\mathbb{1}_{\{\|\theta'-\theta\|_2\le\varepsilon_n\}}\big]\,\mathrm{d}\nu(\theta)\Big] \\
&= \mathbb{E}_0\Big[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot|X_{1:n})}\Big[\int_{B(0,\rho\varepsilon_n)} \mathbb{1}_{\{\|\theta'-\theta\|_2\le\varepsilon_n\}}\,\mathrm{d}\nu(\theta)\Big]\Big] \qquad (5.5)\\
&= \mathbb{E}_0\big[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot|X_{1:n})}\big[\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big)\big]\big] \\
&\le \Big(\frac{1}{\rho}\Big)^d,
\end{aligned}
$$

where to get the last inequality we used the fact that $\nu(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n))$ is the ratio of the volume of the (possibly truncated) Euclidean ball $B(\theta',\varepsilon_n)$ of radius ε_n and center θ' with the volume of the support of ν, namely, the larger Euclidean ball $B(0,\rho\varepsilon_n)$, in dimension d.

The proof is then concluded by recalling that $\rho>1$ was a parameter of the analysis and by picking, for example, $\varepsilon_n=(\sigma/8)\sqrt{d/n}$: by (5.4), we have

$$
\begin{aligned}
&\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\big[\mathbb{P}_\pi\big(\theta' : \|\theta'-\theta\|_2 \le \varepsilon_n | X_{1:n}\big)\big] \\
&\quad= \inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta[Z_\theta] \\
&\quad\le \inf_{\rho>1}\left\{\Big(\frac{1}{\rho}\Big)^d + \frac{\rho}{8\sqrt{2\ln\rho}}\right\} \overset{\text{def}}{=} c_d.
\end{aligned}
$$

We can see that $c_1\le 0.55$ and $c_2\le 0.37$ via the respective choices $\rho=5$ and $\rho=3$, while the fact that the limit is smaller than (and actually equal to) $\sqrt{e}/8\le 0.21$ follows from the choice $\rho=\sqrt{e}$.

Note that, when using (4.7) above, we implicitly assumed that the quantity q in (5.5) lies in $(0,1)$. The fact that $q<1$ follows directly from the upper bound $(1/\rho)^d$ and from $\rho>1$. Besides, the condition $q>0$ is met as soon as $\mathbb{P}_0(\mathbb{P}_{\pi'}(B(0,\varepsilon_n)|X_{1:n})>0)>0$; indeed, for $\theta'\in B(0,\varepsilon_n)$, we have $\nu(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n))>0$ and thus q appears in the last equality of (5.5) as being lower bounded by the expectation of a positive function over a set with positive probability. If on the contrary $\mathbb{P}_0(\mathbb{P}_{\pi'}(B(0,\varepsilon_n)|X_{1:n})>0)=0$, then $\mathbb{P}_0(Z_0>0)=0$, so that $\inf_\theta \mathbb{E}_\theta[Z_\theta]=\mathbb{E}_0[Z_0]=0$, which immediately implies the bound of Proposition 5.1. □

REMARK 1. Though the lower bound of Proposition 5.1 is only stated for the posterior distributions $\mathbb{P}_{\pi'}(\cdot|X_{1:n})$, it is actually valid for any transition kernel $Q(\cdot|X_{1:n})$. This is because the proof above relies on general information-theoretic arguments and does not use the particular form of $\mathbb{P}_{\pi'}(\cdot|X_{1:n})$. This is in the same spirit as for minimax lower bounds for point estimation.

In Section 6.2, we derive another type of posterior concentration lower bound that is no longer uniform. More precisely, we prove a distribution-dependent lower bound that specifies how the posterior mass fails to concentrate on ε_n-neighborhoods of θ for every $\theta\in\Theta$.

5.2 Lower Bounds in Robust Sequential Learning with Sparse Losses

We consider a framework of robust sequential learning called prediction of individual sequences. Its origins and core results are described in the monography by Cesa-Bianchi and Lugosi (2006). In its simplest version, a decision-maker and an environment play repeatedly as follows: at each round $t\ge 1$, and simultaneously, the environment chooses a vector of losses $\ell_t=(\ell_{1,t},\dots,\ell_{N,t})\in[0,1]^N$ while the decision-maker picks an index $I_t\in\{1,\dots,N\}$, possibly at random. Both players then observe ℓ_t and I_t. The decision-maker wants to minimize her cumulative regret, the difference between her cumulative loss and the cumulative loss associated with the best constant choice of an index: for $T\ge 1$,

$$
R_T=\sum_{t=1}^T \ell_{I_t,t} - \min_{k=1,\dots,N}\sum_{t=1}^T \ell_{k,t}.
$$

In this setting, the optimal regret in the worst-case is of the order of $\sqrt{T\ln(N)}$. Cesa-Bianchi et al. (1997) exhibited an asymptotic lower bound of $\sqrt{T\ln(N)/2}$, based on the central limit theorem and on the fact that the expectation of the maximum of N independent standard Gaussian random variables is of the order of $\sqrt{\ln(N)}$. To do so, they considered stochastic environments drawing independently the loss vectors ℓ_t according to a well-chosen distribution.

Cesa-Bianchi, Lugosi and Stoltz (2005) extended this result to a variant called label-efficient prediction, in which loss vectors are observed upon choosing and with a budget constraint: no more than m observations within T rounds. They prove an optimal and nonasymptotic lower bound on the regret of the order of $T\sqrt{\ln(N)/m}$, based on several applications of Fano's inequality to deterministic strategies of the decision-maker, and then, an application of Fubini's theorem to handle general, randomized, strategies. Our reshuffled proof technique below shows that a single application of Fano's inequality to general strategies would be sufficient there (details omitted).

Recently, Kwon and Perchet (2016) considered a setting of sparse loss vectors, in which at each round at most s of the N components of the loss vectors ℓ_t are different from zero. They prove an optimal and asymptotic lower bound on the regret of the order of $\sqrt{Ts\ln(N)/N}$, which generalizes the result for the basic framework, in which $s=N$. Their proof is an extension of the proof of Cesa-Bianchi et al. (1997) and is based on the central limit theorem together with additional technicalities, for example, the use of Slepian's lemma to deal with some dependencies arising from the sparsity assumption.

The aim of this section is to provide a short and elementary proof of this optimal asymptotic $\sqrt{Ts\ln(N)/N}$ bound. As a side result, our bound will even be nonasymptotic. However, for small values of T, given that s/N is

small, picking components I_t uniformly at random ensures an expected cumulative loss thus an expected cumulative regret less than sT/N. The latter is smaller than $\sqrt{Ts\ln(N)/N}$ for values of T of the order of $N\ln(N)/s$. This is why the bound below involves a minimum between quantities of the order of $\sqrt{Ts\ln(N)/N}$ and sT/N; it matches the upper bounds on the regret that can be guaranteed and is therefore optimal.

The expectation in the statement below is with respect to the internal randomization used by the decision-maker's strategy.

THEOREM 5.2. *For all strategies of the decision-maker, for all $s \in \{0,\dots,N\}$, for all $N \geq 2$, for all $T \geq 1$, there exists a fixed-in-advance sequence of loss vectors $\ell_1,\dots,\ell_T$ in $[0,1]^N$ that are each s-sparse such that*

$$\begin{aligned}\mathbb{E}[R_T] &= \sum_{t=1}^{T}\mathbb{E}[\ell_{I_t,t}] - \min_{k=1,\dots,N}\sum_{t=1}^{T}\ell_{k,t}\\ &\geq \min\left\{\frac{s}{16N}T, \frac{1}{32}\sqrt{T\frac{s}{N}\ln N}\right\}.\end{aligned}$$

PROOF. The case $s=0$ corresponds to instantaneous losses $\ell_{j,t}$ that are all null, so that the regret is null as well. Our lower bound holds in this case, but is uninteresting. We therefore focus in the rest of this proof on the case $s \in \{1,\dots,N\}$.

We fix $\varepsilon \in (0, s/(2N))$ and consider, as Kwon and Perchet (2016) did, independent and identically distributed loss vectors $\ell_t \in [0,1]^N$, drawn according to one distribution among P_i, where $1 \leq i \leq N$. Each distribution P_i on $[0,1]^N$ is defined as the law of a random vector L drawn in two steps as follows. We pick s components uniformly at random among $\{1,\dots,N\}$. Then, the components k not picked are associated with zero losses, $L_k = 0$. The losses L_k for picked components $k \neq i$ are drawn according to a Bernoulli distribution with parameter $1/2$. If component i is picked, its loss L_i is drawn according to a Bernoulli distribution with parameter $1/2 - \varepsilon N/s$. The loss vector $L \in [0,1]^N$ thus generated is indeed s-sparse. We denote by P_i^T the Tth product distribution $P_i \otimes \cdots \otimes P_i$. We will actually identify the underlying probability and the law P_i^T. Finally, we denote the expectation under P_i^T by $\mathbb{E}_i$.

Now, under P_i^T, the components $\ell_{k,t}$ of the loss vectors are all distributed according to Bernoulli distributions, with parameters $s/(2N)$ if $k \neq i$ and $s/(2N) - \varepsilon$ if $k = i$. The expected regret, where the expectation $\mathbb{E}$ is with respect to the strategy's internal randomization and the expectation $\mathbb{E}_i$ is with respect to the random choice of the loss vectors, is thus larger than

$$\begin{aligned}&\mathbb{E}_i\big[\mathbb{E}[R_T]\big]\\ &\quad\geq \sum_{t=1}^{T}\mathbb{E}_i\big[\mathbb{E}[\ell_{I_t,t}]\big] - \min_{k=1,\dots,N}\sum_{t=1}^{T}\mathbb{E}_i[\ell_{k,t}]\\ &\quad= \sum_{t=1}^{T}\frac{s}{2N}\big(1-\varepsilon\mathbb{E}_i\big[\mathbb{E}\big[\mathbb{1}_{\{I_t=i\}}\big]\big]\big) - T\left(\frac{s}{2N}-\varepsilon\right)\\ &\quad= T\varepsilon\big(1-\mathbb{E}_i\big[\mathbb{E}\big[F_i(T)\big]\big]\big),\end{aligned} \tag{5.6}$$

where

$$F_i(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}.$$

All in all, we copied almost word for word the (standard) beginning of the proof by Kwon and Perchet (2016), whose first lower bound is exactly

$$\begin{aligned}&\sup_{\ell_1,\dots,\ell_t}\mathbb{E}[R_T]\\ &\quad\geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\big[\mathbb{E}[R_T]\big]\\ &\quad\geq T\varepsilon\left(1-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\big[\mathbb{E}\big[F_i(T)\big]\big]\right).\end{aligned} \tag{5.7}$$

The main differences arise now: we replace a long asymptotic argument (based on the central limit theorem and the study of the limit via Slepian's lemma) by a single application of Fano's inequality.

We introduce the distribution Q on $[0,1]^N$ corresponding to the same randomization scheme as for the P_i, except that no picked component is favored and that all their corresponding losses are drawn according to the Bernoulli distribution with parameter $1/2$. We also denote by $\mathbb{P}$ the probability distribution that underlies the internal randomization of the strategy. An application of Lemma 4.2 with $\mathbb{P}_i = \mathbb{P}\otimes P_i^T$ and $\mathbb{Q}_i = \mathbb{P}\otimes Q^T$, using that $F_1(T) + \cdots + F_N(T) = 1$ and thus $(1/N)\sum_{i=1}^{N}\mathbb{E}_Q[\mathbb{E}[F_i(T)]] = 1/N$, yields

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\big[\mathbb{E}\big[F_i(T)\big]\big]\\ &\quad\leq \frac{1}{N} + \sqrt{\frac{1}{N\ln(N)}\sum_{i=1}^{N}\mathrm{KL}\big(\mathbb{P}\otimes P_i^T, \mathbb{P}\otimes Q^T\big)}.\end{aligned} \tag{5.8}$$

By independence, we get, for all i,

$$\begin{aligned}\mathrm{KL}\big(\mathbb{P}\otimes P_i^T, \mathbb{P}\otimes Q^T\big) &= \mathrm{KL}\big(P_i^T, Q^T\big)\\ &= T\,\mathrm{KL}(P_i, Q).\end{aligned} \tag{5.9}$$

We now show that

$$\mathrm{KL}(P_i, Q) \leq \frac{s}{N}\,\mathrm{kl}\left(\frac{1}{2}-\varepsilon\frac{N}{s}, \frac{1}{2}\right). \tag{5.10}$$

Indeed, both P_i and Q can be seen as uniform convex combinations of probability distributions of the following form, indexed by the subsets of $\{1,\dots,N\}$ with s elements and up to permutations of the Bernoulli distributions in the products below (which does not change the value of the Kullback–Leibler divergences between them):

$\binom{N-1}{s-1}$ distributions of the form (when i is picked)

$$\mathrm{Ber}\Big(\frac{1}{2}-\varepsilon\frac{N}{s}\Big)\otimes\bigotimes_{k=2}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0 \quad\text{and}$$

$$\bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0,$$

where δ_0 denotes the Dirac mass at 0, and

$\binom{N-1}{s}$ distributions of the form (when i is not picked)

$$\bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0 \quad\text{and}$$

$$\bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0.$$

Only the first set of distributions contributes to the Kullback–Leibler divergence. By convexity of the Kullback–Leibler divergence (Corollary 2.3), we thus get the inequality

$$\mathrm{KL}(P_i,Q)\le\frac{\binom{N-1}{s-1}}{\binom{N}{s}}\widetilde{K}_{\varepsilon,N,s}=\frac{s}{N}\,\mathrm{kl}\Big(\frac{1}{2}-\varepsilon\frac{N}{s},\frac{1}{2}\Big),$$

where $\widetilde{K}_{\varepsilon,N,s}$ denotes

$$\mathrm{KL}\bigg(\mathrm{Ber}\Big(\frac{1}{2}-\varepsilon\frac{N}{s}\Big)\otimes\bigotimes_{k=2}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0,\ \bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0\bigg)$$

and where the equality $\widetilde{K}_{\varepsilon,N,s}=\mathrm{kl}(1/2-\varepsilon N/s,1/2)$ is again by independence. Finally, the lemma stated right after this proof shows that

$$\mathrm{kl}\Big(\frac{1}{2}-\varepsilon\frac{N}{s},\frac{1}{2}\Big)\le\frac{4N^2\varepsilon^2}{s^2}. \tag{5.11}$$

Combining (5.7)–(5.11), we proved so far

$$\begin{aligned}&\forall\varepsilon\in\big(0,s/(2N)\big),\\ &\sup_{\ell_1,\dots,\ell_t}\mathbb{E}[R_T]\ge T\varepsilon\Big(1-\frac{1}{N}-\sqrt{\frac{4NT\varepsilon^2}{s\ln(N)}}\Big)\\ &\qquad\ge T\varepsilon\Big(\frac{1}{2}-c\varepsilon\Big),\end{aligned}$$

where we used $1/N\le 1/2$ and denoted $c=2\sqrt{NT}/\sqrt{s\ln(N)}$.

A standard optimization suggests the choice $\varepsilon=1/(4c)$, which is valid, that is, is indeed $<s/(2N)$ as required, as soon as $T>N\ln(N)/(16s)$. In that case, we get a lower bound $T\varepsilon/4$, which corresponds to the $\sqrt{Ts\ln(N)/N}/32$ part of the lower bound.

In case $T\le N\ln(N)/(16s)$, we have $c\le N/(2s)$ and the valid choice $\varepsilon=s/(4N)$ leads to the part of the lower bound given by $T\varepsilon(1/2-c\varepsilon)\ge T\varepsilon/4=sT/(16N)$. □

LEMMA 5.3. *For all $p\in(0,1)$, for all $\varepsilon\in(0,p)$,*

$$\mathrm{kl}(p-\varepsilon,p)\le\frac{\varepsilon^2}{p(1-p)}.$$

PROOF. This result is a special case of the fact that the KL divergence is upper bounded by the χ^2-divergence. We recall, in our particular case, how this is seen:

$$\begin{aligned}&\mathrm{kl}(p-\varepsilon,p)\\ &\quad=(p-\varepsilon)\ln\Big(1-\frac{\varepsilon}{p}\Big)+(1-p+\varepsilon)\ln\Big(1+\frac{\varepsilon}{1-p}\Big)\\ &\quad\le(p-\varepsilon)\frac{-\varepsilon}{p}+(1-p+\varepsilon)\frac{\varepsilon}{1-p}\\ &\quad=\frac{\varepsilon^2}{p}+\frac{\varepsilon^2}{1-p},\end{aligned}$$

where we used $\ln(1+u)\le u$ for all $u>-1$ to get the stated inequality. □

6. OTHER APPLICATIONS, WITH $N=1$ PAIR OF DISTRIBUTIONS

Interestingly, Proposition 3.1 can be useful even for $N=1$ pair of distributions. Rewriting it slightly differently, we indeed have, for all distributions $\mathbb{P},\mathbb{Q}$ and all events A with $\mathbb{Q}(A)\in(0,1)$,

$$\mathbb{P}(A)\ln\Big(\frac{1}{\mathbb{Q}(A)}\Big)\le\mathrm{KL}(\mathbb{P},\mathbb{Q})+\ln(2).$$

Solving for $\mathbb{Q}(A)$—and not for $\mathbb{P}(A)$ as was previously the case—we get

$$\mathbb{Q}(A)\ge\exp\Big(-\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})+\ln(2)}{\mathbb{P}(A)}\Big). \tag{6.1}$$

We applied here a classical technique in information theory due to Haroutunian; see, for instance, Csiszár and

Körner (1981), page 167. The inequality above also holds in the case $\mathbb{Q}(A) = 1$, as the right-hand side is the exponential of a nonpositive quantity, and in the case $\mathbb{Q}(A) = 0$. Indeed, we either have $\mathbb{P}(A) > 0$, which entails, by the data-processing inequality (Lemma 2.1),

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) \geq \mathrm{kl}\big(\mathbb{P}(A), \mathbb{Q}(A)\big) = +\infty,$$

or $\mathbb{P}(A) = 0$; that is, when $\mathbb{Q}(A) = 0$, no matter the value of $\mathbb{P}(A)$, the inequality features the exponential of $-\infty$ in its right-hand side.

Similarly and more generally, for all distributions $\mathbb{P}, \mathbb{Q}$ and all $[0, 1]$-valued random variables Z, we have, by Corollary 2.2 and the lower bound (3.3),

$$\mathbb{E}_{\mathbb{Q}}[Z] \geq \exp\left(-\frac{\mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \ln(2)}{\mathbb{E}_{\mathbb{P}}[Z]}\right). \tag{6.2}$$

The bound (6.1) is similar in spirit to (a consequence of) the Bretagnolle–Huber inequality, recalled and actually improved in Section 8.3; see details therein, and in particular its consequence (8.4). Both bounds can indeed be useful when $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ is larger than a constant and $\mathbb{P}(A)$ is close to 1.

Next, we show two applications of (6.1) and (6.2): a simple proof of a large deviation lower bound for Bernoulli distributions, and a distribution-dependent posterior concentration lower bound.

6.1 A Simple Proof of Cramér's Theorem for Bernoulli Distributions

The next proposition is a well-known large deviation result on the sample mean of independent and identically distributed Bernoulli random variables. It is a particular case of Cramér's theorem that dates back to Cramér (1938), Chernoff (1952); see also Cerf and Petit (2011) for further references and a proof in a very general context. Thanks to Fano's inequality (6.1), the proof of the lower bound that we provide below avoids any explicit change of measure (see the remark after the proof). We are grateful to Aurélien Garivier for suggesting this proof technique to us; see also strong connections with an approach followed by Hayashi (2017), Section 2.4.2.

PROPOSITION 6.1 (Cramér's theorem for Bernoulli distributions). *Let $\theta \in (0, 1)$. Assume that $X_1, \ldots, X_n$ are independent and identically distributed random variables drawn from* $\mathrm{Ber}(\theta)$. *Denoting by $\mathbb{P}_\theta$ the underlying probability measure, we have, for all $x \in (\theta, 1)$,*

$$\lim_{n\to+\infty} \frac{1}{n} \ln \mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) = -\mathrm{kl}(x, \theta).$$

PROOF. We set $\overline{X}_n \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^n X_i$. For the convenience of the reader, we first briefly recall how to prove the upper bound, and then proceed with a new proof for the lower bound.

Upper bound: By the Cramér–Chernoff method and the duality formula for the Kullback–Leibler divergence between Bernoulli distributions (see, e.g., Boucheron, Lugosi and Massart, 2013, pp. 21–24), we have, for all $n \geq 1$,

$$\begin{aligned}\mathbb{P}_\theta(\overline{X}_n > x) &\leq \exp\Big(-n \sup_{\lambda>0}\big\{\lambda x - \ln \mathbb{E}_\theta\big[e^{\lambda X_1}\big]\big\}\Big) \\ &= \exp\big(-n\,\mathrm{kl}(x, \theta)\big),\end{aligned} \tag{6.3}$$

that is,

$$\forall n \geq 1, \quad \frac{1}{n}\ln \mathbb{P}_\theta(\overline{X}_n > x) \leq -\mathrm{kl}(x, \theta).$$

Lower bound: Choose $\varepsilon > 0$ small enough such that $x + \varepsilon < 1$. We may assume with no loss of generality that the underlying distribution is $\mathbb{P}_\theta = \mathrm{Ber}(\theta)^{\otimes n}$. By Fano's inequality in the form (6.1) with the distributions $\mathbb{P} = \mathbb{P}_{x+\varepsilon}$ and $\mathbb{Q} = \mathbb{P}_\theta$, and the event $A = \{\overline{X}_n > x\}$, we have

$$\mathbb{P}_\theta(\overline{X}_n > x) \geq \exp\left(-\frac{\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) + \ln(2)}{\mathbb{P}_{x+\varepsilon}(\overline{X}_n > x)}\right).$$

Noting that $\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) = n\,\mathrm{kl}(x + \varepsilon, \theta)$ we get

$$\begin{aligned}&\mathbb{P}_\theta(\overline{X}_n > x) \\ &\quad\geq \exp\left(-\frac{n\,\mathrm{kl}(x+\varepsilon, \theta) + \ln 2}{\mathbb{P}_{x+\varepsilon}(\overline{X}_n > x)}\right) \\ &\quad\geq \exp\left(-\frac{n\,\mathrm{kl}(x+\varepsilon, \theta) + \ln 2}{1 - e^{-n\,\mathrm{kl}(x, x+\varepsilon)}}\right),\end{aligned} \tag{6.4}$$

where the last bound follows from $\mathbb{P}_{x+\varepsilon}(\overline{X}_n > x) = 1 - \mathbb{P}_{x+\varepsilon}(\overline{X}_n \leq x) \geq 1 - e^{-n\,\mathrm{kl}(x, x+\varepsilon)}$ by a derivation similar to (6.3) above. Taking the logarithms of both sides and letting $n \to +\infty$ finally yields

$$\liminf_{n\to+\infty} \frac{1}{n}\ln \mathbb{P}_\theta(\overline{X}_n > x) \geq -\mathrm{kl}(x + \varepsilon, \theta).$$

We conclude the proof by letting $\varepsilon \to 0$, and by combining the upper and lower bounds. □

Comparison with an historical proof. A classical proof for the lower bound relies on the same change of measure as the one used above, that is, that transports the measure $\mathrm{Ber}(\theta)^{\otimes n}$ to $\mathrm{Ber}(x + \varepsilon)^{\otimes n}$. The bound (6.3), or any other large deviation inequality, is also typically used therein. However, the change of measure is usually carried out explicitly by writing

$$\begin{aligned}\mathbb{P}_\theta(\overline{X}_n > x) &= \mathbb{E}_\theta\big[\mathbb{1}_{\{\overline{X}_n > x\}}\big] \\ &= \mathbb{E}_{x+\varepsilon}\left[\mathbb{1}_{\{\overline{X}_n > x\}} \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_{x+\varepsilon}}(X_1, \ldots, X_n)\right] \\ &= \mathbb{E}_{x+\varepsilon}\big[\mathbb{1}_{\{\overline{X}_n > x\}} e^{-n\widehat{\mathrm{KL}}_n}\big],\end{aligned}$$

where the empirical Kullback–Leibler divergence $\widehat{\mathrm{KL}}_n$ is defined by

$$\widehat{\mathrm{KL}}_n \stackrel{\text{def}}{=} \frac{1}{n}\ln\left(\frac{\mathrm{d}\mathbb{P}_{x+\varepsilon}}{\mathrm{d}\mathbb{P}_\theta}(X_1,\dots,X_n)\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{X_i=1\}}\ln\left(\frac{x+\varepsilon}{\theta}\right) + \mathbb{1}_{\{X_i=0\}}\ln\left(\frac{1-(x+\varepsilon)}{1-\theta}\right)\right).$$

The empirical Kullback–Leibler divergence $\widehat{\mathrm{KL}}_n$ is then compared to its limit $\mathrm{kl}(x+\varepsilon,\theta)$ via the law of large numbers. On the contrary, our short proof above bypasses any call to the law of large numbers and does not perform the change of measure explicitly, in the same spirit as for the bandit lower bounds derived by Kaufmann, Cappé and Garivier (2016) and Garivier, Ménard and Stoltz (2019). Note that the different and more general proof of Cerf and Petit (2011) also bypassed any call to the law of large numbers thanks to other convex duality arguments.

6.2 Distribution-Dependent Posterior Concentration Lower Bounds

In this section, we consider the same Bayesian setting as the one described at the beginning of Section 5.1. In addition, we define the global modulus of continuity between KL and ℓ around $\theta\in\Theta$ and at scale $\varepsilon_n>0$ by

$$\psi(\varepsilon_n,\theta,\ell) \stackrel{\text{def}}{=} \inf\{\mathrm{KL}(P_{\theta'},P_\theta):\ell(\theta',\theta)\geq 2\varepsilon_n,\theta'\in\Theta\};$$

the infimum is set to $+\infty$ if the set is empty.

Next, we provide a distribution-dependent lower bound for posterior concentration rates, that is, a lower bound that holds true for every $\theta\in\Theta$, as opposed[2] to the minimax lower bound of Section 5.1. Theorem 6.2 below indicates that, if the ℓ-ball around θ with radius ε_n has an expected posterior mass close to 1 uniformly over all $\theta\in\Theta$, then this posterior mass cannot be too close to 1 either. Indeed, inequality (6.5) provides a lower bound on the expected posterior mass outside of this ball. The term $n\psi(\varepsilon_n,\theta,\ell)$ within the exponential is a way to quantify how difficult it can be to distinguish between the two product measures $P_{\theta'}^{\otimes n}$ and $P_\theta^{\otimes n}$ when $\ell(\theta',\theta)\geq 2\varepsilon_n$.

THEOREM 6.2 (Distribution-dependent posterior concentration lower bound). *Assume that the posterior distribution $\mathbb{P}_\pi(\cdot|X_{1:n})$ satisfies the uniform concentration condition*

$$\inf_{\theta\in\Theta}\mathbb{E}_\theta\big[\mathbb{P}_\pi\big(\theta':\ell(\theta',\theta)<\varepsilon_n|X_{1:n}\big)\big]\longrightarrow 1 \quad \text{as } n\to+\infty.$$

Then, for all $c>1$, for all n large enough, for all $\theta\in\Theta$,

$$\mathbb{E}_\theta\big[\mathbb{P}_\pi\big(\theta':\ell(\theta',\theta)>\varepsilon_n|X_{1:n}\big)\big] \geq 2^{-c}\exp\big(-cn\psi(\varepsilon_n,\theta,\ell)\big). \tag{6.5}$$

The conclusion can be stated equivalently as

$$\liminf_{n\to+\infty}\inf_{\theta\in\Theta}\frac{\ln(\mathbb{E}_\theta[\mathbb{P}_\pi(\theta':\ell(\theta',\theta)>\varepsilon_n|X_{1:n})])}{\ln(2)+n\psi(\varepsilon_n,\theta,\ell)}\geq -1.$$

The above theorem is greatly inspired from Theorem 2.1 by Hoffmann, Rousseau and Schmidt-Hieber (2015). Our Fano's inequality (6.2), however, makes the proof more direct: the change-of-measure carried out by Hoffmann, Rousseau and Schmidt-Hieber (2015) is now implicit, and no proof by contradiction is required. We also bypass one technical assumption (see the discussion after the proof).

PROOF OF THEOREM 6.2. We fix $c>1$. By the uniform concentration condition, there exists $n_0\geq 1$ such that, for all $n\geq n_0$,

$$\inf_{\theta^\star\in\Theta}\mathbb{E}_{\theta^\star}\big[\mathbb{P}_\pi\big(\theta':\ell(\theta',\theta^\star)<\varepsilon_n|X_{1:n}\big)\big]\geq\frac{1}{c}. \tag{6.6}$$

We now fix $n\geq n_0$ and $\theta\in\Theta$. We consider any $\theta^\star\in\Theta$ such that $\ell(\theta^\star,\theta)\geq 2\varepsilon_n$. Using Fano's inequality in the form of (6.2) with the distributions $\mathbb{P}=P_{\theta^\star}^{\otimes n}$ and $\mathbb{Q}=P_\theta^{\otimes n}$, together with the $[0,1]$-valued random variable $Z_\theta=\mathbb{P}_\pi(\theta':\ell(\theta',\theta)>\varepsilon_n|X_{1:n})$, we get

$$\mathbb{E}_\theta[Z_\theta]\geq\exp\left(-\frac{\mathrm{KL}(P_{\theta^\star}^{\otimes n},P_\theta^{\otimes n})+\ln 2}{\mathbb{E}_{\theta^\star}[Z_\theta]}\right) = \exp\left(-\frac{n\,\mathrm{KL}(P_{\theta^\star},P_\theta)+\ln 2}{\mathbb{E}_{\theta^\star}[Z_\theta]}\right). \tag{6.7}$$

By the triangle inequality and the assumption $\ell(\theta^\star,\theta)\geq 2\varepsilon_n$, we can see that $\{\theta':\ell(\theta',\theta)>\varepsilon_n\}\supseteq\{\theta':\ell(\theta',\theta^\star)<\varepsilon_n\}$, so that

$$\mathbb{E}_{\theta^\star}[Z_\theta]\geq\mathbb{E}_{\theta^\star}\big[\mathbb{P}_\pi\big(\theta':\ell(\theta',\theta^\star)<\varepsilon_n|X_{1:n}\big)\big]\geq\frac{1}{c}$$

by the uniform lower bound (6.6). Substituting the above inequality into (6.7) then yields

$$\mathbb{E}_\theta[Z_\theta]\geq\exp\big(-c\big(n\,\mathrm{KL}(P_{\theta^\star},P_\theta)+\ln 2\big)\big).$$

To conclude the proof, it suffices to take the supremum of the right-hand side over all $\theta^\star\in\Theta$ such that $\ell(\theta^\star,\theta)\geq 2\varepsilon_n$, and to identify the definition of $\psi(\varepsilon_n,\theta,\ell)$. □

Note that, at first sight, our result may seem a little weaker than Hoffmann, Rousseau and Schmidt-Hieber (2015), Theorem 2.1, because we only define $\psi(\varepsilon_n,\theta,\ell)$ in terms of KL instead of a general premetric d: in

[2] Note, however, that we are here in a slightly different regime than in Section 5.1, where we addressed cases for which the uniform posterior concentration condition (6.6) was proved to be impossible at scale ε_n (and actually took place at a slightly larger scale ε_n').

other words, we only consider the case $d(\theta,\theta') = \sqrt{\mathrm{KL}(P_{\theta'}, P_\theta)}$. However, it is still possible to derive a bound in terms of an arbitrary premetric d by comparing d and KL after applying Theorem 6.2.

In the case of the premetric $d(\theta,\theta') = \sqrt{\mathrm{KL}(P_{\theta'}, P_\theta)}$, we bypass an additional technical assumption used for the the similar lower bound of Hoffmann, Rousseau and Schmidt-Hieber (2015), Theorem 2.1, namely, that there exists a constant $C > 0$ such that

$$\sup_{\theta,\theta'} P_{\theta'}^{\otimes n}\big(\mathcal{L}_n(\theta') - \mathcal{L}_n(\theta) \geq Cn\,\mathrm{KL}(P_{\theta'}, P_\theta)\big) \longrightarrow 0$$
$$\text{as } n \to +\infty,$$

where the supremum is over all $\theta,\theta' \in \Theta$ satisfying $\psi(\varepsilon_n,\theta,\ell) \leq \mathrm{KL}(P_{\theta'}, P_\theta) \leq 2\psi(\varepsilon_n,\theta,\ell)$, and where $\mathcal{L}_n(\theta) = \sum_{i=1}^n \ln(\mathrm{d}P_\theta/\mathrm{d}\mathfrak{m})(X_i)$ denotes the log-likelihood function with respect to a common dominating measure $\mathfrak{m}$. Besides, we get an improved constant in the exponential in (6.5), with respect to Hoffmann, Rousseau and Schmidt-Hieber (2015), Theorem 2.1: by a factor of $3C/c$, which, since $C \geq 1$ in most cases, is $3C/c \approx 3C \geq 3$ when $c \approx 1$. (A closer look at their proof can yield a constant arbitrarily close to $2C$, which is still larger than our c by a factor of $2C/c \approx 2C \geq 2$.)

7. REFERENCES AND COMPARISON TO THE LITERATURE

We discuss in this section how novel (or not novel) our results and approaches are. We first state where our main innovation lie in our eyes, and then discuss the novelty or lack of novelty through a series of specific points.

Main innovations in a nutshell. We could find no reference indicating that the alternative distributions $\mathbb{Q}_i$ and $\mathbb{Q}_\theta$ could vary and do not need to be set to a fixed alternative $\mathbb{Q}_0$, nor that arbitrary $[0,1]$-valued random variables Z_i or Z_θ (i.e., not summing up to 1) could be considered. These two elements are encompassed in the reduction (4.3), which is to be considered our main new result. The first application in Section 5 relies on such arbitrary $[0,1]$-valued random variables Z_θ (but in the second application the finitely many Z_i sum up to 1).

That the sets A_i considered in the reduction (3.1) form a partition of the underlying measurable space or that the finitely many random variables Z_i sum up to 1 (see Gushchin, 2003) were typical requirements in the literature until recently, with one exception. Indeed, Chen, Guntuboyina and Zhang (2016) noted in spirit that the requirement of forming a partition was unnecessary, which we too had been aware of as early as Stoltz (2007), where we also already mentioned the fact that in particular the alternative distribution $\mathbb{Q}$ had not to be fixed and could depend on i or θ.

Generalization to f-divergences (not a new result). Gushchin (2003) generalized Fano-type inequalities with the Kullback–Leibler divergence to arbitrary f-divergences, in the case where finitely many $[0,1]$-valued random variables $Z_1 + \cdots + Z_N = 1$ are considered; see also Chen, Guntuboyina and Zhang (2016). Most of the literature focuses, however, on Fano-type inequalities with the Kullback–Leibler divergence, like all references discussed below.

On the two-step methodology used (not a new result). The two-step methodology of Section 4, which simply notes that Bernoulli distributions are the main case to study when establishing Fano-type inequalities, was well known in the cases of disjoint events or $[0,1]$-valued random variables summing up to 1. This follows at various levels of clarity from references that will be discussed in details in this section for other matters (Han and Verdú, 1994, Gushchin, 2003, and Chen, Guntuboyina and Zhang, 2016) and other references (Zhang, 2006, Section D, and Harremoës and Vajda, 2011, which is further discussed at the beginning of Section 8). In particular, the conjunction of a Bernoulli reduction and the use of a lower bound on the kl function was already present in Han and Verdú (1994).

Other, more information-theoretic statements and proof techniques of Fano's inequalities for finitely many hypotheses as in Proposition 3.1 can be found, for example, in Cover and Thomas (2006), Theorem 2.11.1, Yu (1997), Lemma 3, or Ibragimov and Has'minskiĭ (1981), Chapter VII, Lemma 1.1; they resort to classical formulas on the Shannon entropy, the conditional entropy and the mutual information.

On the reductions to Bernoulli distributions. Reduction (4.3) is new at this level of generality, as we indicated, but all other reductions were known, though sometimes proved in a more involved way. Reduction (3.1) and (4.1) were already known and used by Han and Verdú (1994), Theorems 2, 7 and 8. Reduction (4.2) is stated in spirit by Chen, Guntuboyina and Zhang (2016) with a constant alternative $\mathbb{Q}_\theta \equiv \mathbb{Q}$; see also a detailed discussion and comparison below between their approach and the general approach we took in Section 4. We should also mention that Duchi and Wainwright (2013) provided preliminary (though more involved) results towards the continuous reduction (4.2). Finally, as already mentioned, a reduction with random variables like (4.3) was stated in a special case in Gushchin (2003), for finitely many $[0,1]$-valued random variables with $Z_1 + \cdots + Z_N = 1$.

On the lower bounds on the kl *function (not really a new result)*. The inequalities (4.4) are folklore knowledge. The first inequality in (4.5) can be found in Guntuboyina (2011); the second inequality is a new (immediate) consequence. The inequalities (4.7) are a consequence, which we derived on our own, of a refined Pinsker's inequality stated by Ordentlich and Weinberger (2005).

In-depth discussion of two articles. We now discuss two earlier contributions and indicate how our results encompass them: the "generalized Fano's inequality" of Chen, Guntuboyina and Zhang (2016) and the version of Fano's inequality by Birgé (2005), which was designed to also cover the case where $N = 2$.

7.1 On the "Generalized Fano's Inequality" of Chen, Guntuboyina and Zhang (2016)

The Bayesian setting considered therein is the following; it generalizes the setting of Han and Verdú (1994), whose results we discuss in a remark after the proof of Proposition 7.1.

A parameter space $(\Theta, \mathcal{G})$ is equipped with a prior probability measure ν. A family of probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$ on a measurable space $(\Omega, \mathcal{F})$, some outcome space $(\mathcal{X}, \mathcal{E})$, for example, $\mathcal{X} = \mathbb{R}^n$, and a random variable $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{E})$ are considered. We denote by $\mathbb{E}_\theta$ the expectation under $\mathbb{P}_\theta$. Of course we may have $(\Omega, \mathcal{F}) = (\mathcal{X}, \mathcal{E})$ and X be the identity, in which case $\mathbb{P}_\theta$ will be the law of X under $\mathbb{P}_\theta$.

The goal is either to estimate θ or to take good actions: we consider a measurable target space $(\mathcal{A}, \mathcal{H})$, that may or may not be equal to Θ. The quality of a prediction or of an action is measured by a measurable loss function $L : \Theta \times \mathcal{A} \to [0, 1]$. The random variable X is our observation, based on which we construct a $\sigma(X)$-measurable random variable $\widehat{a}$ with values in $\mathcal{A}$. Putting aside all measurability issues (here and in the rest of this subsection), the risk of $\widehat{a}$ in this model equals

$$R(\widehat{a}) = \int_\Theta \mathbb{E}_\theta\big[L(\theta, \widehat{a})\big]\,\mathrm{d}\nu(\theta)$$

and the Bayes risk in this model is the smallest such possible risk,

$$R_{\mathrm{Bayes}} = \inf_{\widehat{a}} R(\widehat{a}),$$

where the infimum is over all $\sigma(X)$-measurable random variables with values in $\mathcal{A}$.

Chen, Guntuboyina and Zhang (2016) call their main result (Corollary 5) a "generalized Fano's inequality"; we state it and prove it below not only for $\{0, 1\}$-valued loss functions L as in the original article, but for any $[0, 1]$-valued loss function. The reason behind this extension is that we not only have the reduction (4.2) with events, but we also have the reduction (4.3) with $[0, 1]$-valued random variables. We also feel that our proof technique is more direct and more natural.

We only deal with Kullback–Leibler divergences, but the result and proof below readily extend to f-divergences.

PROPOSITION 7.1. *In the setting described above, the Bayes risk is always larger than*

$$\begin{aligned} &R_{\mathrm{Bayes}} \\ &\quad\ge 1 + \Bigg(\bigg(\inf_{\mathbb{Q}} \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q})\,\mathrm{d}\nu(\theta)\bigg) \\ &\qquad + \ln\bigg(1 + \inf_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\,\mathrm{d}\nu(\theta)\bigg)\Bigg) \\ &\qquad \Big/\bigg(\ln\bigg(1 - \inf_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\,\mathrm{d}\nu(\theta)\bigg)\bigg), \end{aligned}$$

where the infimum in the numerator is over all probability measures $\mathbb{Q}$ on $(\Omega, \mathcal{F})$.

PROOF. We fix $\widehat{a}$ and an alternative $\mathbb{Q}$. The combination of (4.3) and (4.5), with $Z_\theta = 1 - L(\theta, \widehat{a})$, yields

$$\begin{aligned} &1 - \int_\Theta \mathbb{E}_\theta\big[L(\theta, \widehat{a})\big]\,\mathrm{d}\nu(\theta) \\ &\quad\le \frac{\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q})\,\mathrm{d}\nu(\theta) + \ln(2 - q_{\widehat{a}})}{\ln(1/q_{\widehat{a}})}, \end{aligned} \tag{7.1}$$

where $\mathbb{E}_\mathbb{Q}$ denotes the expectation with respect to $\mathbb{Q}$ and

$$q_{\widehat{a}} = 1 - \int_\Theta \mathbb{E}_\mathbb{Q}\big[L(\theta, \widehat{a})\big]\,\mathrm{d}\nu(\theta).$$

As $q \mapsto 1/\ln(1/q)$ and $q \mapsto \ln(2 - q)/\ln(1/q)$ are both increasing, taking the supremum over the $\sigma(X)$-measurable random variables $\widehat{a}$ in both sides of (7.1) gives

$$\begin{aligned} &1 - R_{\mathrm{Bayes}} \\ &\quad\le \frac{\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q})\,\mathrm{d}\nu(\theta) + \ln(2 - q^\star)}{\ln(1/q^\star)}, \end{aligned} \tag{7.2}$$

where

$$\begin{aligned} q^\star &= \sup_{\widehat{a}} q_{\widehat{a}} = 1 - \inf_{\widehat{a}} \int_\Theta \mathbb{E}_\mathbb{Q}\big[L(\theta, \widehat{a})\big]\,\mathrm{d}\nu(\theta) \\ &= 1 - \inf_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\,\mathrm{d}\nu(\theta), \end{aligned} \tag{7.3}$$

as is proved below. Taking the infimum of the right-hand side of (7.2) over $\mathbb{Q}$ and rearranging concludes the proof.

It only remains to prove the last inequality of (7.3) and actually, as constant elements $a \in \mathcal{A}$ are special cases of random variables $\widehat{a}$, we only need to prove that

$$\begin{aligned} &\inf_{\widehat{a}} \int_\Theta \mathbb{E}_\mathbb{Q}\big[L(\theta, \widehat{a})\big]\,\mathrm{d}\nu(\theta) \\ &\quad\ge \inf_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\,\mathrm{d}\nu(\theta). \end{aligned} \tag{7.4}$$

Now, each $\widehat{a}$ that is $\sigma(X)$-measurable can be rewritten $\widehat{a} = \overline{a}(X)$ for some measurable function $\overline{a} : \mathcal{X} \to \mathcal{A}$; then,

by the Fubini–Tonelli theorem,

$$\begin{aligned}\int_\Theta \mathbb{E}_{\mathbb{Q}}\big[L(\theta,\widehat{a})\big]\,\mathrm{d}\nu(\theta) &= \int_{\mathcal{X}}\bigg(\int_\Theta L\big(\theta,\overline{a}(x)\big)\,\mathrm{d}\nu(\theta)\bigg)\,\mathrm{d}\mathbb{Q}(x)\\ &\geq \int_{\mathcal{X}}\bigg(\inf_{a\in\mathcal{A}}\int_\Theta L(\theta,a)\,\mathrm{d}\nu(\theta)\bigg)\,\mathrm{d}\mathbb{Q}(x),\end{aligned}$$

which proves (7.4). □

REMARK 2. As mentioned by Chen, Guntuboyina and Zhang (2016), one of the major results of Han and Verdú (1994), namely, their Theorem 8, is a special case of Proposition 7.1, with $\Theta = \mathcal{A}$ and the loss function $L(\theta,\theta') = \mathbb{1}_{\{\theta\neq\theta'\}}$. The (opposite of the) denominator in the lower bound on the Bayes risk then takes the simple form

$$\begin{aligned}&-\ln\bigg(1-\inf_{\theta'\in\Theta}\int_\Theta L(\theta,\theta')\,\mathrm{d}\nu(\theta)\bigg)\\ &\quad= -\ln\Big(\sup_{\theta\in\Theta}\nu\big(\{\theta\}\big)\Big)\stackrel{\text{def}}{=} H_\infty(\nu),\end{aligned}$$

which is called the infinite-order Rényi entropy of the probability distribution ν. Han and Verdú (1994) only dealt with the case of discrete sets Θ but the extension to continuous Θ is immediate, as we showed in Section 4.

7.2 Comparison to Birgé (2005): An Interpolation Between Pinsker's and Fano's Inequalities

The most classical version of Fano's inequality, that is, the right-most side of (7.6) below, is quite impractical for small values of N (cf. Birgé, 2005), and even useless when $N = 2$, the latter case being straightforward to deal with by several well-known tools, for example, by Pinsker's inequality. One of the main motivations of Birgé (2005) was therefore to get an inequality that would be useful for all $N \geq 2$. His inequality is stated next; it only deals with events $A_1,\dots,A_N$ forming a partition of the underlying measurable space. As should be clear from its proof this assumption is crucial. (See Appendix D for a pointer to an extended version of this article where a proof following the methodology described in Section 4 is provided.)

THEOREM 7.2 (Birgé's lemma). *Given an underlying measurable space $(\Omega,\mathcal{F})$, for all $N \geq 2$, for all probability distributions $\mathbb{P}_1,\dots,\mathbb{P}_N$, for all events $A_1,\dots,A_N$ forming a partition of Ω,*

$$\min_{1\leq i\leq N}\mathbb{P}_i(A_i) \leq \max\bigg\{c_N,\frac{\overline{K}}{\ln(N)}\bigg\}$$

$$\text{where } \overline{K} = \frac{1}{N-1}\sum_{i=2}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{P}_1)$$

and where $(c_N)_{N\geq 2}$ is a decreasing sequence, where each term c_N is defined as the unique $c\in(0,1)$ such that

$$\begin{aligned}&\frac{-(c\ln(c)+(1-c)\ln(1-c))}{c}+\ln(1-c)\\ &\quad=\ln\bigg(\frac{N-1}{N}\bigg).\end{aligned}\tag{7.5}$$

We have, for instance, $c_2 \approx 0.7587$ and $c_3 \approx 0.7127$, while $\lim c_N = 0.63987$.

However, a first drawback of the bound above lies in the $\overline{K}$ term: one cannot pick a convenient $\mathbb{Q}$ as in the bounds (7.6)–(7.7) below. A second drawback is that the result is about the minimum of the $\mathbb{P}_i(A_i)$, not about their average. In contrast, the versions of Fano's inequality based the kl lower bounds (4.5), (4.4) and (4.7) respectively lead to the following inequalities, stated in the setting of Theorem 7.2 and by picking constant alternatives $\mathbb{Q}$:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i)\\ &\quad\leq \frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q})+\ln(2-\frac{1}{N})}{\ln(N)}\\ &\quad\leq \frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q})+\ln(2)}{\ln(N)},\end{aligned}\tag{7.6}$$

and

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i)\\ &\quad\leq \frac{1}{N}+\sqrt{\frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i,\mathbb{Q})}{\max\{\ln(N),2\}}}.\end{aligned}\tag{7.7}$$

The middle term in (7.6) was derived—with a different formulation—by Chen, Guntuboyina and Zhang (2016), see Proposition 7.1 above.

Discussion. We note that unlike the right-most side of (7.6), both the middle term in (7.6) and the bound (7.7) yield useful bounds for all $N \geq 2$, and in particular, for $N = 2$. Even better, (7.7) implies both Pinsker's inequality and, lower bounding the maximum by $\ln(N)$, a bound as useful as Theorem 7.2 or Proposition 3.1 in case of a partition. Indeed, in practice, the additional additive $1/N$ term and the additional square root do not prevent from obtaining the desired lower bounds, as illustrated in Section 5.2.

Therefore, our inequality (7.7) provides some interpolation between Pinsker's and Fano's inequalities: it simultaneously deals with all values $N \geq 2$.

8. PROOFS OF THE LOWER BOUNDS ON kl STATED IN SECTION 4.2 (AND PROOF OF AN IMPROVED BRETAGNOLLE–HUBER INEQUALITY)

We prove in this section the convexity inequalities (4.5) and (4.6) as well as the refined Pinsker's inequality and its consequence (4.7). Using the same techniques and methodology as for establishing these bounds, we also improve in passing the Bretagnolle–Huber inequality.

The main advantage of the Bernoulli reductions of Section 4.1 is that we could then capitalize in Section 4.3 (and also in Section 6) on any lower bound on the Kullback–Leibler divergence $\mathrm{kl}(p,q)$ between Bernoulli distributions. In the same spirit, our key argument below to prove the refined Pinsker's inequality and the Bretagnolle–Huber inequality (which hold for arbitrary probability distributions) is in both cases an inequality between the Kullback–Leibler divergence and the total variation distance between Bernoulli distributions. This simple but deep observation was made in great generality by Harremoës and Vajda (2011).

8.1 Proofs of the Convexity Inequalities (4.5) and (4.6)

PROOF. Inequality (4.6) follows from (4.5) by noting that the function $q \in (0,1) \mapsto \ln(2-q)/\ln(1/q)$ is dominated by $q \in (0,1) \mapsto 0.21 + 0.79q$.

Now, the shortest proof of (4.5) notes that the duality formula for the Kullback–Leibler divergence between Bernoulli distributions—already used in (6.3)—ensures that, for all $p \in [0,1]$ and $q \in (0,1]$,

$$\mathrm{kl}(p,q) = \sup_{\lambda \in \mathbb{R}} \{\lambda p - \ln(q(\mathrm{e}^{\lambda} - 1) + 1)\}$$
$$\geq p \ln\left(\frac{1}{q}\right) - \ln(2-q)$$

for the choice $\lambda = \ln(1/q)$. □

An alternative, longer but more elementary proof uses a direct convexity argument, as in Guntuboyina (2011), Example II.4, which already included the inequality of interest in the special case when $q = 1/N$; see also Chen, Guntuboyina and Zhang (2016). We deal separately with $p = 0$ and $p = 1$, and thus restrict our attention to $p \in (0,1)$ in the sequel. For $q \in (0,1)$, as $p \mapsto \mathrm{kl}(p,q)$ is convex and differentiable on $(0,1)$, we have

$$\forall (p, p_0) \in (0,1)^2,$$
$$\mathrm{kl}(p,q) - \mathrm{kl}(p_0,q) \geq \underbrace{\ln\left(\frac{p_0(1-q)}{(1-p_0)q}\right)}_{\frac{\partial}{\partial p}\mathrm{kl}(p_0,q)}(p - p_0). \tag{8.1}$$

The choice $p_0 = 1/(2-q)$ is such that

$$\frac{p_0}{1-p_0} = \frac{1}{1-q} \quad \text{thus} \quad \ln\left(\frac{p_0(1-q)}{(1-p_0)q}\right) = \ln\left(\frac{1}{q}\right)$$

and

$$\mathrm{kl}(p_0,q) = \frac{1}{2-q}\ln\left(\frac{1/(2-q)}{q}\right) + \frac{1-q}{2-q}\ln\left(\frac{(1-q)/(2-q)}{1-q}\right)$$
$$= \frac{1}{2-q}\ln\left(\frac{1}{q}\right) + \ln\left(\frac{1}{2-q}\right).$$

Inequality (8.1) becomes

$$\forall p \in (0,1),$$
$$\mathrm{kl}(p,q) - \frac{1}{2-q}\ln\left(\frac{1}{q}\right) + \ln(2-q) \geq \left(p - \frac{1}{2-q}\right)\ln\left(\frac{1}{q}\right),$$

which proves as well the bound (4.5).

8.2 Proofs of the Refined Pinsker's Inequality and of Its Consequence (4.7)

The next theorem is a stronger version of Pinsker's inequality for Bernoulli distributions, that was proved[3] by Ordentlich and Weinberger (2005). Indeed, note that the function φ defined below satisfies $\min \varphi = 2$, so that the next theorem always yields an improvement over the most classical version of Pinsker's inequality: $\mathrm{kl}(p,q) \geq 2(p-q)^2$.

We provide below an alternative elementary proof for Bernoulli distributions of this refined Pinsker's inequality. The extension to the case of general distributions, via the contraction-of-entropy property, is stated at the end of this section.

THEOREM 8.1 (A refined Pinsker's inequality by Ordentlich and Weinberger, 2005). *For all $p, q \in [0,1]$,*

$$\mathrm{kl}(p,q) \geq \frac{\ln((1-q)/q)}{1-2q}(p-q)^2 \stackrel{\text{def}}{=} \varphi(q)(p-q)^2,$$

where the multiplicative factor $\varphi(q) = (1-2q)^{-1}\ln((1-q)/q)$ is defined for all $q \in [0,1]$ by extending it by continuity as $\varphi(1/2) = 2$ and $\varphi(0) = \varphi(1) = +\infty$.

The proof shows that $\varphi(q)$ is the optimal multiplicative factor in front of $(p-q)^2$ when the bounds needs to hold for all $p \in [0,1]$; the proof also provides a natural explanation for the value of φ.

PROOF OF THEOREM 8.1. The stated inequality is satisfied for $q \in \{0,1\}$ as $\mathrm{kl}(p,q) = +\infty$ in these cases unless $p = q$. The special case $q = 1/2$ is addressed at

[3] We also refer the reader to Kearns and Saul (1998), Lemma 1, and Berend and Kontorovich (2013), Theorem 3.2, for dual inequalities upper bounding the moment-generating function of the Bernoulli distributions.

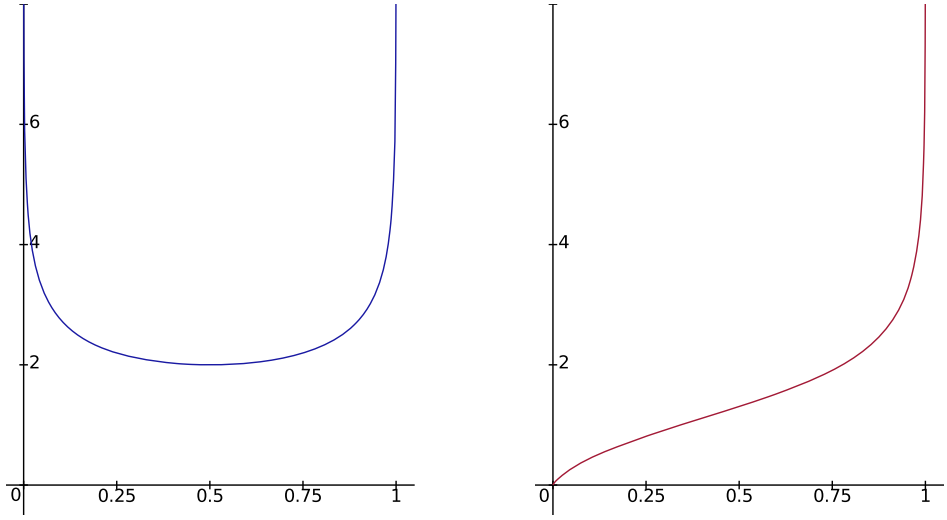

FIG. 1. *Plots of* φ [*left*] *and* $x \in (0,1) \mapsto \varphi(x) - \ln(1/x)$ [*right*].

the end of the proof. We thus fix $q \in (0,1) \setminus \{1/2\}$ and set $f(p) = \mathrm{kl}(p,q)/(p-q)^2$ for $p \neq q$, with a continuity extension at $p = q$. We exactly show that f attains its minimum at $p = 1-q$, from which the result (and its optimality) follow by noting that

$$f(1-q) = \frac{\mathrm{kl}(1-q,q)}{(1-2q)^2} = \frac{\ln((1-q)/q)}{1-2q} = \varphi(q).$$

Given the form of f, it is natural to perform a second-order Taylor expansion of $\mathrm{kl}(p,q)$ around q. We have

$$\begin{aligned} &\frac{\partial}{\partial p}\mathrm{kl}(p,q) = \ln\left(\frac{p(1-q)}{(1-p)q}\right) \quad \text{and} \\ &\frac{\partial^2}{\partial^2 p}\mathrm{kl}(p,q) = \frac{1}{p(1-p)} \stackrel{\text{def}}{=} \psi(p), \end{aligned} \tag{8.2}$$

so that Taylor's formula with integral remainder reveals that for $p \neq q$,

$$\begin{aligned} f(p) = \frac{\mathrm{kl}(p,q)}{(p-q)^2} &= \frac{1}{(p-q)^2}\int_q^p \frac{\psi(t)}{1!}(p-t)^1\,\mathrm{d}t \\ &= \int_0^1 \psi\big(q+u(p-q)\big)(1-u)\,\mathrm{d}u. \end{aligned}$$

This rewriting of f shows that f is strictly convex (as ψ is so). Its global minimum is achieved at the unique point where its derivative vanishes. But by differentiating under the integral sign, we have, at $p = 1-q$,

$$f'(1-q) = \int_0^1 \psi'\big(q+u(1-2q)\big)u(1-u)\,\mathrm{d}u = 0;$$

the equality to 0 follows from the fact that the function $u \mapsto \psi'(q+u(1-2q))u(1-u)$ is antisymmetric around $u = 1/2$ (essentially because ψ' is antisymmetric itself around $1/2$). As a consequence, the convex function f attains its global minimum at $1-q$, which concludes the proof for the case where $q \in (0,1) \setminus \{1/2\}$.

It only remains to deal with $q = 1/2$: we use the continuity of $\mathrm{kl}(p,\cdot)$ and φ to extend the obtained inequality from $q \in [0,1] \setminus \{1/2\}$ to $q = 1/2$. □

We now prove the second inequality of (4.7). A picture is helpful; see Figure 1.

COROLLARY 8.2. *For all* $q \in (0,1]$, *we have* $\varphi(q) \geq 2$ *and* $\varphi(q) \geq \ln(1/q)$. *Thus, for all* $p \in [0,1]$ *and* $q \in (0,1)$,

$$p \leq q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\{\ln(1/q), 2\}}}.$$

Slightly sharper bounds are possible, like $\varphi(q) \geq (1+q)(1+q^2)\ln(1/q)$ or $\varphi(q) \geq \ln(1/q) + 2.5q$, but we were unable to exploit these refinements in our applications.

General refined Pinsker's inequality. The following result, which improves on Pinsker's inequality, is due to Ordentlich and Weinberger (2005). Our approach through Bernoulli distributions enables to derive it in an elementary (and enlightening) way: by combining Theorem 8.1 and the data-processing inequality (Lemma 2.1).

THEOREM 8.3. *Let* $\mathbb{P}$ *and* $\mathbb{Q}$ *be two probability distributions on the same measurable space* $(\Omega, \mathcal{F})$. *Then*

$$\forall A \in \mathcal{F}, \quad \big|\mathbb{P}(A) - \mathbb{Q}(A)\big| \leq \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{\varphi(\mathbb{Q}(A))}},$$

where $\varphi \geq 2$ *is defined in the statement of Theorem* 8.1. *In particular, the total variation distance between* $\mathbb{P}$ *and* $\mathbb{Q}$ *is bounded as*

$$\sup_{A\in\mathcal{F}} \big|\mathbb{P}(A) - \mathbb{Q}(A)\big| \leq \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{\inf_{A\in\mathcal{F}} \varphi(\mathbb{Q}(A))}}.$$

8.3 An Improved Bretagnolle–Huber Inequality

The Bretagnolle–Huber inequality was introduced by Bretagnolle and Huber (1978, 1979). The multiplicative factor $\mathrm{e}^{-1/\mathrm{e}} \geq 0.69$ in our statement (8.3) below is a slight improvement over the original $1/2$ factor. For all $p, q \in [0,1]$,

$$\begin{aligned} 1 - |p-q| &\geq \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{kl}(p,q)} \\ \text{thus } q &\geq p - 1 + \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{kl}(p,q)}. \end{aligned} \tag{8.3}$$

It is worth to note that Bretagnolle and Huber (1978) also proved the inequality

$$|p-q| \leq \sqrt{1 - \exp\bigl(-\mathrm{kl}(p,q)\bigr)},$$

which improves as well upon the Bretagnolle–Huber inequality with the $1/2$ factor, but which is neither better nor worse than (8.3).

Now, via the data-processing inequality (Lemma 2.1), we get from (8.3)

$$1 - \sup_{A \in \mathcal{F}} \bigl|\mathbb{P}(A) - \mathbb{Q}(A)\bigr| \geq \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{KL}(\mathbb{P},\mathbb{Q})}.$$

The left-hand side can be rewritten as $\inf_{A \in \mathcal{F}} \{\mathbb{P}(A) + \mathbb{Q}(A^c)\}$, where A^c denotes the complement of A. Therefore, the above inequality is a lower bound on the test affinity between $\mathbb{P}$ and $\mathbb{Q}$. For the sake of comparison to (6.1), we can restate the general version of the Bretagnolle–Huber inequality as: for all $A \in \mathcal{F}$,

$$\mathbb{Q}(A) \geq \mathbb{P}(A) - 1 + \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{KL}(\mathbb{P},\mathbb{Q})}. \tag{8.4}$$

We now provide a proof of (8.3); note that our improvement was made possible because we reduced the proof to very elementary arguments in the case of Bernoulli distributions.

PROOF. The case where $p \in \{0,1\}$ or $q \in \{0,1\}$ can be handled separately; we consider $(p,q) \in (0,1)^2$ in the sequel. The derivative of the function $x \in (0,1) \mapsto x \ln(x/(1-q))$ equals $1 + \ln(x) - \ln(1-q)$, so that the function achieves its minimum at $x = (1-q)/\mathrm{e}$, with value $-(1-q)/\mathrm{e} \geq -1/\mathrm{e}$. Therefore,

$$\begin{aligned} -\mathrm{kl}(p,q) &= -p \ln\left(\frac{p}{q}\right) - (1-p) \ln\left(\frac{1-p}{1-q}\right) \\ &\leq -p \ln\left(\frac{p}{q}\right) + \frac{1}{\mathrm{e}} \\ &= p\left(\ln\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\frac{1}{\mathrm{e}}. \end{aligned}$$

Therefore, using the convexity of the exponential,

$$\begin{aligned} \mathrm{e}^{-\mathrm{kl}(p,q)} &\leq p \exp\left(\ln\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\mathrm{e}^{1/\mathrm{e}} \\ &= \bigl(q + (1-p)\bigr)\mathrm{e}^{1/\mathrm{e}}, \end{aligned}$$

which shows that

$$1 - (p-q) \geq \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{kl}(p,q)}.$$

By replacing q by $1-q$ and p by $1-p$, we also get

$$\begin{aligned} 1 - (q-p) &= 1 - \bigl((1-p) - (1-q)\bigr) \\ &\geq \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{kl}(1-p,1-q)} \\ &= \mathrm{e}^{-1/\mathrm{e}} \mathrm{e}^{-\mathrm{kl}(p,q)}. \end{aligned}$$

This concludes the proof, as $1 - |p-q|$ is equal to the smallest value between $1-(p-q)$ and $1-(q-p)$. □

APPENDIX A: ON THE SHARPNESS OF FANO-TYPE INEQUALITIES OF SECTION 4

The reductions of Section 4.1 are sharp in the sense that they can hold with equality (they cannot be improved at this level of generality).

For the Kullback–Leibler divergence, they lead to inequalities of the form $\mathrm{kl}(\overline{p}, \overline{q}) \leq \overline{K}$. We are interested in upper bounds on $\overline{p}$. We introduce the generalized inverse of kl in its second argument: for all $q \in [0,1]$ and all $y \geq 0$,

$$\mathrm{kl}(\cdot, q)^{(-1)}(y) \stackrel{\mathrm{def}}{=} \sup\bigl\{p \in [0,1] : \mathrm{kl}(p,q) \leq y\bigr\};$$

when $q \in (0,1)$, it is thus equal to the largest root q of the equation $\mathrm{kl}(p,q) = y$ if $y \leq \ln(1/q)$ or to 1 otherwise. From $\mathrm{kl}(\overline{p}, \overline{q}) \leq \overline{K}$ the best general upper bound on $\overline{p}$ is

$$\overline{p} \leq \mathrm{kl}(\cdot, \overline{q})^{(-1)}(\overline{K}).$$

This formulation should be reminiscent of Birgé (2005), Theorem 2, but has one major practical drawback: it is unreadable, and this is why we considered the lower bounds of Section 4.2.

Question is now how sharp these lower bounds on kl are. Bounds (4.4) and (4.5) are of the form

$$p \leq \frac{\mathrm{kl}(p,q)}{\ln(1/q)} + \varepsilon(q),$$

where the $\varepsilon(q)$ quantity vanishes when $q \to 0$. Now, in the applications, q is typically small and the main term $\mathrm{kl}(p,q)/\ln(1/q)$ is of the order of a constant. Therefore, the lemma below explains that up to the ε quantity, the bounds (4.4) and (4.5) of Section 4.2 are essentially optimal.

The bound (4.7) therein is of the form

$$p \leq \sqrt{\frac{\mathrm{kl}(p,q)}{\ln(1/q)}} + \varepsilon(q),$$

but given the discussion above, it can also be considered optimal in spirit, as in the applications q is typically small and the main term $\mathrm{kl}(p,q)/\ln(1/q)$ is of the order of a constant.

LEMMA A.1. *For all $q \in (0,1)$ and $p \in [0,1]$, whenever $p \geq q$, we have*

$$\mathrm{kl}(p,q) \leq p \ln\biggl(\frac{1}{q}\biggr), \quad \textit{thus } p \geq \frac{\mathrm{kl}(p,q)}{\ln(1/q)}.$$

PROOF. We note that when $p \geq q$, we have $(1-p)/(1-q) \leq 1$, so that

$$\begin{aligned}\mathrm{kl}(p,q) &= p \ln\biggl(\frac{1}{q}\biggr) + \underbrace{p \ln(p)}_{\leq 0} \\ &\quad + (1-p) \underbrace{\ln\biggl(\frac{1-p}{1-q}\biggr)}_{\leq 0} \leq p \ln\biggl(\frac{1}{q}\biggr),\end{aligned}$$

hence the first inequality. □

APPENDIX B: FROM BAYESIAN POSTERIORS TO POINT ESTIMATORS

We recall below a well-known result that indicates how to construct good point estimators from good Bayesian posteriors (Section B.1 below). One theoretical benefit is that this result can be used to convert known minimax lower bounds for point estimation into minimax lower bounds for posterior concentration rates (Section B.2 below). This technique is thus a—less direct—alternative to the method we presented in Section 5.1.

B.1 The Conversion

The following statement is a nonasymptotic variant of Theorem 2.5 by Ghosal, Ghosh and van der Vaart (2000); see also Proposition 3 in Chapter 12 by Le Cam (1986), as well as Section 5 by Hoffmann, Rousseau and Schmidt-Hieber (2015).

We consider the same setting as in Section 5.1 and assume in particular that the underlying probability measure is given by $\mathbb{P}_\theta = P_\theta^{\otimes n}$, that is, that $(X_1, \ldots, X_n)$ is the identity random variable.

PROPOSITION B.1 (From Bayesian posteriors to point estimators). *Let $n \geq 1$, $\delta > 0$ and $\theta \in \Theta$. Let $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ be any estimator satisfying, $\mathbb{P}_\theta$-almost surely,*

$$\begin{aligned}&\mathbb{P}_\pi\bigl(\theta' : \ell\bigl(\theta', \widehat{\theta}_n\bigr) < \varepsilon_n | X_{1:n}\bigr) \\ &\quad \geq \sup_{\widetilde{\theta} \in \Theta} \mathbb{P}_\pi\bigl(\theta' : \ell\bigl(\theta', \widetilde{\theta}\bigr) < \varepsilon_n | X_{1:n}\bigr) - \delta.\end{aligned} \tag{B.1}$$

Then,

$$\begin{aligned}&\mathbb{P}_\theta\biggl(\mathbb{P}_\pi\bigl(\theta' : \ell\bigl(\theta', \theta\bigr) \geq \varepsilon_n | X_{1:n}\bigr) \geq \frac{1-\delta}{2}\biggr) \\ &\quad \geq \mathbb{P}_\theta\bigl(\ell\bigl(\widehat{\theta}_n, \theta\bigr) \geq 2\varepsilon_n\bigr).\end{aligned} \tag{B.2}$$

This result implies that if $\widehat{\theta}_n$ is a center of a ball that almost maximizes the posterior mass—see assumption (B.1)—and if the posterior mass concentrates around θ at a rate $\varepsilon_n' < \varepsilon_n$—so that the left-hand side of (B.2) vanishes by Markov's inequality—then $\widehat{\theta}_n$ is $(2\varepsilon_n)$-close to θ with high probability. Therefore, at least from a theoretical viewpoint, a good posterior distribution can be converted into a good point estimator, by defining $\widehat{\theta}_n$ based on $\mathbb{P}_\pi(\cdot | X_{1:n})$ such that (B.1) holds, that is, by taking an approximate argument of the supremum. A measurable such $\widehat{\theta}_n$ exists as soon as Θ is a separable topological space and the function $\widetilde{\theta} \mapsto \mathbb{P}_\pi(\theta' : \ell(\theta', \widetilde{\theta}) < \varepsilon_n | x_{1:n})$ is lower-semicontinuous for $\mathfrak{m}^{\otimes n}$-almost every $x_{1:n} \in \mathcal{X}^n$ (see the end of the proof of Corollary B.2 for more details).

PROOF. Denote by $B_\ell(\theta, \varepsilon) \stackrel{\mathrm{def}}{=} \{\theta' \in \Theta : \ell(\theta', \theta) < \varepsilon\}$ the open ℓ-ball of center θ and radius ε. By the triangle inequality we have the following inclusions of events:

$$\begin{aligned}&\bigl\{\ell\bigl(\widehat{\theta}_n, \theta\bigr) \geq 2\varepsilon_n\bigr\} \\ &\quad \subseteq \bigl\{B_\ell\bigl(\widehat{\theta}_n, \varepsilon_n\bigr) \cap B_\ell(\theta, \varepsilon_n) = \varnothing\bigr\} \\ &\quad \subseteq \bigl\{\mathbb{P}_\pi\bigl(B_\ell\bigl(\widehat{\theta}_n, \varepsilon_n\bigr) | X_{1:n}\bigr) \\ &\qquad + \mathbb{P}_\pi\bigl(B_\ell(\theta, \varepsilon_n) | X_{1:n}\bigr) \leq 1\bigr\} \qquad \text{(B.3)} \\ &\quad \subseteq \biggl\{\mathbb{P}_\pi\bigl(B_\ell(\theta, \varepsilon_n) | X_{1:n}\bigr) \leq \frac{1+\delta}{2}\biggr\} \\ &\quad = \biggl\{1 - \mathbb{P}_\pi\bigl(\theta' : \ell\bigl(\theta', \theta\bigr) < \varepsilon_n | X_{1:n}\bigr) \geq \frac{1-\delta}{2}\biggr\} \\ &\quad = \biggl\{\mathbb{P}_\pi\bigl(\theta' : \ell\bigl(\theta', \theta\bigr) \geq \varepsilon_n | X_{1:n}\bigr) \geq \frac{1-\delta}{2}\biggr\},\end{aligned}$$

where (B.3) follows from the lower bound $\mathbb{P}_\pi(B_\ell(\widehat{\theta}_n, \varepsilon_n) | X_{1:n}) \geq \mathbb{P}_\pi(B_\ell(\theta, \varepsilon_n) | X_{1:n}) - \delta$, which holds by assumption (B.1) on $\widehat{\theta}_n$. This concludes the proof. □

B.2 Application to Posterior Concentration Lower Bounds

We explained above that a good posterior distribution can be converted into a good point estimator. As noted by Ghosal, Ghosh and van der Vaart (2000) this conversion can be used the other way around: if we have a lower bound on the minimax rate of estimation, then Proposition B.1 provides a lower bound on the minimax posterior concentration rate, as formalized in the following corollary. Assumption (B.4) below corresponds to an in-probability minimax lower bound.

COROLLARY B.2. *Let $n \geq 1$. Consider the setting of Section 5.1, with underlying probability measure $\mathbb{P}_\theta = P_\theta^{\otimes n}$ when the unknown parameter is θ. Assume that Θ is a separable topological space and that $\widetilde{\theta} \mapsto \ell(\theta', \widetilde{\theta})$ is continuous for all $\theta' \in \Theta$. Assume also that for some absolute constant $c < 1$, we have*

$$\inf_{\widehat{\theta}_n \text{est.}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\bigl(\ell\bigl(\widehat{\theta}_n, \theta\bigr) \geq 2\varepsilon_n\bigr) \geq 1 - c, \tag{B.4}$$

where the infimum is taken over all estimators $\widehat{\theta}_n$. Then, for all priors π' on Θ,

$$\inf_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta\big) < \varepsilon_n | X_{1:n}\big)\big] \le \frac{1+c}{2} < 1. \tag{B.5}$$

PROOF. Let $\delta > 0$ be a parameter that we will later take arbitrarily small. Fix any estimator $\widehat{\theta}_n$ satisfying (B.1) for the prior π', that is, that almost maximizes the posterior mass on an open ball of radius ε_n. (See the end of the proof for details on why such a measurable $\widehat{\theta}_n$ exists.) Then, Proposition B.1 used for all $\theta \in \Theta$ entails that

$$\begin{aligned}
&\sup_{\theta\in\Theta} \mathbb{P}_\theta\bigg(\mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta\big) \ge \varepsilon_n | X_{1:n}\big) \ge \frac{1-\delta}{2}\bigg) \\
&\quad\ge \sup_{\theta\in\Theta} \mathbb{P}_\theta\big(\ell\big(\widehat{\theta}_n, \theta\big) \ge 2\varepsilon_n\big) \\
&\quad\ge 1 - c,
\end{aligned}$$

where the last inequality follows from the assumption (B.4). Now we use Markov's inequality to upper bound the left-hand side above and obtain

$$\begin{aligned}
&\frac{2}{1-\delta} \sup_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta\big) \ge \varepsilon_n | X_{1:n}\big)\big] \\
&\quad\ge \sup_{\theta\in\Theta} \mathbb{P}_\theta\bigg(\mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta\big) \ge \varepsilon_n | X_{1:n}\big) \ge \frac{1-\delta}{2}\bigg) \\
&\quad\ge 1 - c.
\end{aligned}$$

Letting $\delta \to 0$ and dividing both sides by 2 yields

$$1 - \inf_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta\big) < \varepsilon_n | X_{1:n}\big)\big] \ge \frac{1-c}{2}.$$

Rearranging terms concludes the proof of (B.5). We now address the technical issue mentioned at the beginning of the proof.

Why a measurable $\widehat{\theta}_n$ exists. Note that it is possible to choose $\widehat{\theta}_n$ satisfying (B.1) with π' in a measurable way as soon as Θ is a separable topological space and

$$\psi : \widetilde{\theta} \in \Theta \longmapsto \mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \widetilde{\theta}\big) < \varepsilon_n | x_{1:n}\big)$$

is lower-semicontinuous for $\mathfrak{m}^{\otimes n}$-almost every $x_{1:n} \in \mathcal{X}^n$, and thus $\mathbb{P}_\theta$-almost surely for all $\theta \in \Theta$. The reason is that, in that case, it is possible to equate the supremum of ψ over Θ to a supremum on a countable subset of Θ. Next, and thanks to the continuity assumption on ℓ, we prove that the desired lower-semicontinuity holds true for all $x_{1:n} \in \mathcal{X}^n$ (not just almost all of them).

To that end, we show the lower-semicontinuity at any fixed $\theta^\star \in \Theta$. Consider any sequence $(\widetilde{\theta}_i)_{i\ge1}$ in Θ converging to $\theta^\star$. For all $x_{1:n} \in \mathcal{X}^n$, by Fatou's lemma applied to the well-defined probability distribution $\mathbb{P}_{\pi'}(\cdot|x_{1:n})$, we have

$$\begin{aligned}
&\liminf_{i\to+\infty} \mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \widetilde{\theta}_i\big) < \varepsilon_n | x_{1:n}\big) \\
&\quad= \liminf_{i\to+\infty} \mathbb{E}_{\pi'}\big[\mathbb{1}_{\{\ell(\theta', \widetilde{\theta}_i) < \varepsilon_n\}} | x_{1:n}\big] \\
&\quad\ge \mathbb{E}_{\pi'}\Big[\underbrace{\liminf_{i\to+\infty} \mathbb{1}_{\{\ell(\theta', \widetilde{\theta}_i) < \varepsilon_n\}}}_{=1 \text{ if } \ell(\theta', \theta^\star) < \varepsilon_n} | x_{1:n}\Big] \\
&\quad\ge \mathbb{P}_{\pi'}\big(\theta' : \ell\big(\theta', \theta^\star\big) < \varepsilon_n | x_{1:n}\big),
\end{aligned} \tag{B.6}$$

where in (B.7) we identify that the lim inf equals 1 as soon as $\ell(\theta', \theta^\star) < \varepsilon_n$ by continuity of $\widetilde{\theta} \mapsto \ell(\theta', \widetilde{\theta})$ at $\widetilde{\theta} = \theta^\star$. □

APPENDIX C: ON JENSEN'S INEQUALITY

Classical statements of Jensen's inequality for convex functions φ on $C \subseteq \mathbb{R}^n$ either assume that the underlying probability measure is supported on a finite number of points or that the convex subset C is open. In the first case, the proof follows directly from the definition of convexity, while in the second case, it is a consequence of the existence of subgradients. In both cases, it is assumed that the function φ under consideration only takes finite values. In this article, Jensen's inequality is applied several times to nonopen convex sets C, like $C = [0, 1]^2$ or $C = [0, +\infty)$ and/or convex functions φ that can possibly be equal to $+\infty$ at some points.

The restriction of C being open is easy to drop when the dimension equals $n = 1$, that is, when C is an interval; it was dropped, for example, by Ferguson (1967), pages 74–76, in higher dimensions, thanks to a proof by induction to address possible boundary effects with respect to the arbitrary convex set C. Let $\mathcal{B}(\mathbb{R}^n)$ denote the Borel σ-field of $\mathbb{R}^n$.

LEMMA C.1 (Jensen's inequality for general convex sets; Ferguson, 1967). *Let $C \subseteq \mathbb{R}^n$ be any nonempty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R}$ be any convex Borel function. Then, for all probability measures μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\mu(C) = 1$ and $\int \|x\| \,\mathrm{d}\mu(x) < +\infty$, we have*

$$\begin{aligned}
&\int x \,\mathrm{d}\mu(x) \in C \quad \text{and} \\
&\varphi\bigg(\int x \,\mathrm{d}\mu(x)\bigg) \le \int_C \varphi(x) \,\mathrm{d}\mu(x),
\end{aligned} \tag{C.1}$$

where the integral of φ against μ is well defined in $\mathbb{R} \cup \{+\infty\}$.

Our contribution is the following natural extension.

LEMMA C.2. *The result of Lemma C.1 also holds for any convex Borel function $\varphi : C \to \mathbb{R} \cup \{+\infty\}$.*

We rephrase this extension in terms of random variables. Let $C \subseteq \mathbb{R}^n$ be any nonempty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R} \cup \{+\infty\}$ be any convex Borel function. Let X be an integrable random variable from any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, such that $\mathbb{P}(X \in C) = 1$. Then

$$\mathbb{E}[X] \in C \quad \text{and} \quad \varphi\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[\varphi(X)\big],$$

where $\mathbb{E}[\varphi(X)]$ is well defined in $\mathbb{R} \cup \{+\infty\}$.

PROOF. We first check that $\varphi_- = \max\{-\varphi, 0\}$ is μ-integrable on C, so that the integral of φ against μ is well defined in $\mathbb{R} \cup \{+\infty\}$. To that end, we will prove that φ is lower bounded on C by an affine function: $\varphi(x) \geq a^{\mathrm{T}}x + b$ for all $x \in C$, where $(a, b) \in \mathbb{R}^2$, from which it follows that $\varphi_-(x) \leq \|a\|\|x\| + \|b\|$ for all $x \in C$ and thus

$$\begin{aligned}\int_C \varphi_-(x)\,\mathrm{d}\mu(x) &\leq \int_C \big(\|a\|\|x\| + \|b\|\big)\,\mathrm{d}\mu(x) \\ &= \|a\| \int_C \|x\|\,\mathrm{d}\mu(x) + \|b\| < +\infty.\end{aligned}$$

So, it only remains to prove the affine lower bound. If the domain $\{\varphi < +\infty\}$ is empty, any affine function is suitable. Otherwise, $\{\varphi < +\infty\}$ is a nonempty convex set, so that its relative interior R is also nonempty (see Rockafellar, 1970, Theorem 6.2); we fix $x_0 \in R$. But, by Rockafellar (1970), Theorem 23.4, the function φ admits a subgradient at x_0, that is, there exists $a \in \mathbb{R}^n$ such that $\varphi(x) \geq \varphi(x_0) + a^T(x - x_0)$ for all $x \in C$. This concludes the first part of this proof.

In the second part, we show the inequality (C.1) via a reduction to the case of real-valued functions. Indeed, note that if $\mu(\varphi = +\infty) > 0$ then the desired inequality is immediate. We can thus assume that $\mu(\varphi < +\infty) = 1$. But, using Lemma C.1 with the nonempty convex Borel subset $\widetilde{C} = \{\varphi < +\infty\}$ and the real-valued convex Borel function $\widetilde{\varphi} : \widetilde{C} \to \mathbb{R}$ defined by $\widetilde{\varphi}(x) = \varphi(x)$, we get, since $\mu(\widetilde{C}) = 1$,

$$\begin{aligned}&\int x\,\mathrm{d}\mu(x) \in \widetilde{C} \quad \text{and} \\ &\widetilde{\varphi}\bigg(\int x\,\mathrm{d}\mu(x)\bigg) \leq \int_{\widetilde{C}} \widetilde{\varphi}(x)\,\mathrm{d}\mu(x).\end{aligned}$$

Using the facts that $\widetilde{\varphi}(x) = \varphi(x)$ for all $x \in \widetilde{C}$ and that $\mu(C \setminus \widetilde{C}) = 1 - 1 = 0$ entails (C.1). □

We now complete our extension by tackling the conditional form of Jensen's inequality.

LEMMA C.3 (A general conditional Jensen's inequality). *Let $C \subseteq \mathbb{R}^n$ be any nonempty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R} \cup \{+\infty\}$ be any convex Borel function. Let X be an integrable random variable from any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, such that $\mathbb{P}(X \in C) = 1$. Then, for every sub-σ-field $\mathcal{G}$ of $\mathcal{F}$, we have, $\mathbb{P}$-almost surely,*

$$\mathbb{E}[X|\mathcal{G}] \in C \quad \text{and} \quad \varphi\big(\mathbb{E}[X|\mathcal{G}]\big) \leq \mathbb{E}\big[\varphi(X)|\mathcal{G}\big],$$

where $\mathbb{E}[\varphi(X)|\mathcal{G}]$ is $\mathbb{P}$-almost-surely well defined in $\mathbb{R} \cup \{+\infty\}$.

PROOF. The proof follows directly from the unconditional Jensen's inequality (Lemma C.2 above) and from the existence of regular conditional distributions. More precisely, by Theorems 2.1.15 and 5.1.9 of Durrett (2010), applied to the case where $(S, \mathcal{S}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, there exists a regular conditional distribution of X given $\mathcal{G}$. That is, there exists a function $K : \Omega \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ such that:

(P1) for every $B \in \mathcal{B}(\mathbb{R}^n)$, the mapping $\omega \in \Omega \mapsto K(\omega, B)$ is $\mathcal{G}$-measurable and $\mathbb{P}(X \in B|\mathcal{G}) = K(\cdot, B)$ $\mathbb{P}$-a.s.;

(P2) for $\mathbb{P}$-almost all $\omega \in \Omega$, the mapping $B \mapsto K(\omega, B)$ is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Moreover, as a consequence of (P1):

(P1') for every Borel function $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(X)$ is $\mathbb{P}$-integrable or such that g is nonnegative,

$$\int g(x)K(\cdot, \mathrm{d}x) = \mathbb{E}\big[g(X)|\mathcal{G}\big] \quad \mathbb{P}\text{-a.s.}$$

Now, given our assumptions and thanks to (P1) and (P1'):

(P3) by $\mathbb{P}(X \in C) = 1$ we also have $K(\cdot, C) = \mathbb{P}(X \in C|\mathcal{G}) = 1$ $\mathbb{P}$-a.s.;

(P4) since X is $\mathbb{P}$-integrable, so is $\int \|x\| K(\cdot, \mathrm{d}x) = \mathbb{E}[\|X\||\mathcal{G}]$, which is therefore $\mathbb{P}$-a.s. finite.

We apply Lemma C.2 with the probability measures $\mu_\omega = K(\omega, \cdot)$, for those ω for which the properties stated in (P2), (P3) and (P4) actually hold; these ω are $\mathbb{P}$-almost all elements of Ω. We get, for these ω,

$$\begin{aligned}&\int xK(\omega, \mathrm{d}x) \in C \quad \text{and} \\ &\varphi\bigg(\int xK(\omega, \mathrm{d}x)\bigg) \leq \int_C \varphi(x)K(\omega, \mathrm{d}x),\end{aligned}$$

where the integral in the right-hand side is well defined in $\mathbb{R} \cup \{+\infty\}$. Thanks to (P1'), and by decomposing $\varphi(X)$ into $\varphi_-(X)$, which is integrable (see the beginning of the proof of Lemma C.2), and $\varphi_+(X)$, which is nonnegative, we thus have proved that $\mathbb{P}$-a.s.,

$$\mathbb{E}[X|\mathcal{G}] \in C \quad \text{and} \quad \varphi\big(\mathbb{E}[X|\mathcal{G}]\big) \leq \mathbb{E}\big[\varphi(X)|\mathcal{G}\big],$$

which concludes the proof. □

APPENDIX D: EXTENDED VERSION OF THIS ARTICLE

An extended version of this article is available on ArXiv (Gerchinovitz, Ménard and Stoltz, 2018, arXiv:1702.05985) and features the following additional appendices.

APPENDIX E

Proofs of the data-processing inequality (Lemma 2.1) and of the joint convexity of Div_f (Corollary 2.3).

APPENDIX F

Additional material on the Fano-type inequalities of Section 4, namely:

- A sharper lower bound on div_f for the Hellinger distance.
- Finding a good constant alternative $\mathbb{Q}$.

APPENDIX G

On Birgé's lemma, namely:

- A proof of Theorem 7.2.
- Two other statements of this lemma (the original one and a simplification of it).

ACKNOWLEDGMENTS

This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA). Gilles Stoltz would like to thank Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) for financial support.

The authors would like to thank Aurélien Garivier, Jean-Baptiste Hiriart-Urruty and Vincent Tan Yan Fu for their insightful comments and suggestions.

REFERENCES

AERON, S., SALIGRAMA, V. and ZHAO, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inform. Theory* **56** 5111–5130. MR2808668 https://doi.org/10.1109/TIT.2010.2059891

ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777

BEREND, D. and KONTOROVICH, A. (2013). On the concentration of the missing mass. *Electron. Commun. Probab.* **18** no. 3, 7. MR3011530 https://doi.org/10.1214/ECP.v18-2359

BIRGÉ, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory* **51** 1611–1615. MR2241522 https://doi.org/10.1109/TIT.2005.844101

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BRETAGNOLLE, J. and HUBER, C. (1978). Estimation des densités: Risque minimax. In *Séminaire de Probabilités*, *XII* (*Univ. Strasbourg*, *Strasbourg*, 1976/1977). *Lecture Notes in Math.* **649** 342–363. Springer, Berlin. MR0520011

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137. MR0523165 https://doi.org/10.1007/BF00535278

CERF, R. and PETIT, P. (2011). A short proof of Cramér's theorem in $\mathbb{R}$. *Amer. Math. Monthly* **118** 925–931. MR2869520 https://doi.org/10.4169/amer.math.monthly.118.10.925

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction*, *Learning*, *and Games*. Cambridge Univ. Press, Cambridge. MR2409394 https://doi.org/10.1017/CBO9780511546921

CESA-BIANCHI, N., LUGOSI, G. and STOLTZ, G. (2005). Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory* **51** 2152–2162. MR2235288 https://doi.org/10.1109/TIT.2005.847729

CESA-BIANCHI, N., FREUND, Y., HAUSSLER, D., HELMBOLD, D. P., SCHAPIRE, R. E. and WARMUTH, M. K. (1997). How to use expert advice. *J. ACM* **44** 427–485. MR1470152 https://doi.org/10.1145/258128.258179

CHEN, X., GUNTUBOYINA, A. and ZHANG, Y. (2016). On Bayes risk lower bounds. *J. Mach. Learn. Res.* **17** Paper No. 218, 58. MR3595153

CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23** 493–507. MR0057518 https://doi.org/10.1214/aoms/1177729330

COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

CRAMÉR, H. (1938). Sur un nouveau théorème limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* **736** 5–23.

CSISZÁR, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **8** 85–108. MR0164374

CSISZÁR, I. and KÖRNER, J. (1981). *Information Theory*: *Coding Theorems for Discrete Memoryless Systems*, 3rd ed. Akadémiai Kiadó, Budapest. MR1492180

DUCHI, J. and WAINWRIGHT, M. J. (2013). Distance-based and continuum Fano inequalities with applications to statistical estimation. Available at arXiv:1311.2669.

DURRETT, R. (2010). *Probability*: *Theory and Examples*, 4th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **31**. Cambridge Univ. Press, Cambridge. MR2722836 https://doi.org/10.1017/CBO9780511779398

FERGUSON, T. S. (1967). *Mathematical Statistics*: *A Decision Theoretic Approach*. *Probability and Mathematical Statistics*, *Vol.* 1. Academic Press, New York. MR0215390

GARIVIER, A., MÉNARD, P. and STOLTZ, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.* **44** 377–399. MR3959077 https://doi.org/10.1287/moor.2017.0928

GERCHINOVITZ, S., MÉNARD, P. and STOLTZ, G. (2018). Fano's inequality for random variables. Available at arXiv:1702.05985.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

GUNTUBOYINA, A. (2011). Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Trans. Inform. Theory* **57** 2386–2399. MR2809097 https://doi.org/10.1109/TIT.2011.2110791

GUSHCHIN, A. A. (2003). Fano's lemma and analogous inequalities for minimax risk. *Theory Probab. Math. Statist.* **67** 26–37.

HAN, T. S. and VERDÚ, S. (1994). Generalizing the Fano inequality. *IEEE Trans. Inform. Theory* **40** 1247–1251. MR1301430 https://doi.org/10.1109/18.335943

HARREMOËS, P. and VAJDA, I. (2011). On pairs of f-divergences and their joint range. *IEEE Trans. Inform. Theory* **57** 3230–3235. MR2817015 https://doi.org/10.1109/TIT.2011.2137353

HAYASHI, M. (2017). *Quantum Information Theory*, 2nd ed. *Graduate Texts in Physics*. Springer, Berlin. MR3558531 https://doi.org/10.1007/978-3-662-49725-8

HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, New York. MR0620321

KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.* **17** Paper No. 1, 42. MR3482921

KEARNS, M. and SAUL, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (*UAI*'98) 311–319.

KWON, J. and PERCHET, V. (2016). Gains and losses are fundamentally different in regret minimization: The sparse case. *J. Mach. Learn. Res.* **17** Paper No. 227, 32. MR3595163

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7

LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1784901 https://doi.org/10.1007/978-1-4612-1166-2

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

ORDENTLICH, E. and WEINBERGER, M. J. (2005). A distribution dependent refinement of Pinsker's inequality. *IEEE Trans. Inform. Theory* **51** 1836–1840. MR2235683 https://doi.org/10.1109/TIT.2005.846407

PARDO, L. (2006). *Statistical Inference Based on Divergence Measures*. *Statistics*: *Textbooks and Monographs* **185**. CRC Press/CRC, Boca Raton, FL. MR2183173

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. *Princeton Mathematical Series*, *No*. 28. Princeton Univ. Press, Princeton, NJ. MR0274683

STOLTZ, G. (2007). An introduction to the prediction of individual sequences: (1) oracle inequalities; (2) prediction with partial monitoring. Statistics seminar of Univ. Paris VI and Paris VII, Chevaleret, November 12 and 26, 2007; written version of the pair of seminar talks available upon request.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York. MR1462963

ZHANG, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory* **52** 1307–1321. MR2241190 https://doi.org/10.1109/TIT.2005.864439