Fano's inequality for random variables

Sebastien Gerchinovitz, Pierre Ménard, Gilles Stoltz

Fano's inequality for random variables*

Sébastien Gerchinovitz
Pierre Ménard

IMT — Université Paul Sabatier, Toulouse, France

sebastien.gerchinovitz@math.univ-toulouse.fr
pierre.menard@math.univ-toulouse.fr

&

Gilles Stoltz

Laboratoire de Mathématiques d'Orsay
Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France

GREGHEC — HEC Paris, CNRS

gilles.stoltz@math.u-psud.fr

September 18, 2018

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 non-stochastic sequential learning.

MSC 2000 subject classifications. Primary-62B10; secondary-62F15, 68T05.

Keywords: Multiple-hypotheses testing, Lower bounds, Information theory, Bayesian posterior concentration

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

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 $N \geqslant 2$ 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,

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

where the infimum in the right-hand side is over all probability distributions $\mathbb{Q}$ over $(\Omega, \mathcal{F})$. 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 $\mathbb{P}_i\big(\widehat{\theta} \neq i\big)$ 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 et al. [2016] further generalized Fano's inequality to continuously infinitely many distributions; see also Aeron et al. [2010]. Gushchin [2003] extended Fano's inequality in two other directions, first by considering $[0,1]$–valued random variables Z_i such that $Z_1 + \ldots + Z_N = 1$, 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 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 non-stochastic 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 et al., 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 $(p, q) \in [0, 1]^2$,

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

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

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

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 $\big(f(t) - f(x)\big)/(t - x)$ is non-decreasing as t increases. Besides, this limit does not depend on x and equals

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

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

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

Put differently,

$$\forall\, x \geqslant 0,\ y \geqslant 0, \qquad f(x+y) \leqslant f(x) + y\, M_f\,, \tag{1}$$

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

$$\mathbb{P} = \mathbb{P}_{\rm ac} + \mathbb{P}_{\rm sing}\,, \qquad \text{where} \qquad \mathbb{P}_{\rm ac} \ll \mathbb{Q} \quad \text{and} \quad \mathbb{P}_{\rm sing} \perp \mathbb{Q}\,, \tag{2}$$

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

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

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 23 (Appendix D) with $\varphi = f$ and $C = [0, +\infty)$.

Definition. *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*

$$\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) = \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{Q} + \mathbb{P}_{\rm sing}(\Omega)\, M_f\,. \tag{3}$$

Jensen's inequality of Lemma 23, together with (1), also indicates that $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) \geqslant 0$. Indeed,

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

so that by (1),

$$\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) \geqslant f\big(\mathbb{P}_{\rm ac}(\Omega)\big) + \mathbb{P}_{\rm sing}(\Omega)\, M_f \geqslant f\big(\mathbb{P}_{\rm ac}(\Omega) + \mathbb{P}_{\rm sing}(\Omega)\big) = f(1) = 0\,.$$

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 1 (Data-processing inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over 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,*

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

Corollary 2 (Data-processing inequality with expectations of random variables)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over 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,*

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

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

Corollary 3 (Joint convexity of Div_f)**.** *All f–divergences Div_f are jointly convex, i.e., for all probability distributions $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}_1, \mathbb{Q}_2$ over the same measurable space $(\Omega, \mathcal{F})$, and all $\lambda \in (0,1)$,*

$$\mathrm{Div}_f\Big((1-\lambda)\mathbb{P}_1 + \lambda\mathbb{P}_2,\, (1-\lambda)\mathbb{Q}_1 + \lambda\mathbb{Q}_2\Big) \leqslant (1-\lambda)\,\mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\,\mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)\,.$$

Lemma 1 and Corollary 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 E). The proof of Lemma 1 is extracted from Ali and Silvey [1966, Section 4.2], see also Pardo [2006, Proposition 1.2], while we derive Corollary 3 as an elementary consequence of Lemma 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 et al., 2013, in particular Exercise 4.10 therein).

On the contrary, Corollary 2 is a recent though elementary result, proved in Garivier et al. [2018] for Kullback-Leibler divergences. The proof readily extends to f–divergences.

Proof (of Corollary 2): We augment the underlying measurable space into $\Omega \times [0,1]$, where $[0,1]$ is equipped with the Borel σ–algebra $\mathcal{B}\big([0,1]\big)$ 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,

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

Now, for all events $E \in \mathcal{F} \otimes \mathcal{B}\big([0,1]\big)$, the data-processing inequality (Lemma 1) used with the indicator function $X = \mathbb{1}_E$ ensures that

$$\mathrm{Div}_f\big(\mathbb{P} \otimes \mathfrak{m},\, \mathbb{Q} \otimes \mathfrak{m}\big) \geqslant \mathrm{Div}_f\Big((\mathbb{P} \otimes \mathfrak{m})^{\mathbb{1}_E},\ (\mathbb{Q} \otimes \mathfrak{m})^{\mathbb{1}_E}\Big) = \mathrm{div}_f\big((\mathbb{P} \otimes \mathfrak{m})(E),\, (\mathbb{Q} \otimes \mathfrak{m})(E)\big)\,,$$

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 $E = \big\{(\omega, x) \in \Omega \times [0,1] :\, x \leqslant X(\omega)\big\}$, Tonelli's theorem ensures that

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

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 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*

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

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

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N}\inf\limits_{\mathbb{Q}}\displaystyle\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \ln(2)}{\ln(N)}\,.$$

Proof: Our first step is to reduce the problem to Bernoulli distributions. Using first the joint convexity of the Kullback-Leibler divergence (Corollary 3), and second the data-processing inequality with the indicator functions $X = \mathbb{1}_{A_i}$ (Lemma 1), we get

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i),\, \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{4}$$

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

$$\overline{p} = \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \qquad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) \qquad \overline{K} = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{5}$$

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

$$\mathrm{kl}\big(\overline{p}, \overline{q}\big) \geqslant \overline{p}\ln\big(1/\overline{q}\big) - \ln(2)\,, \qquad \text{thus} \qquad \overline{p} \leqslant \frac{\mathrm{kl}\big(\overline{p}, \overline{q}\big) + \ln(2)}{\ln\big(1/\overline{q}\big)} \tag{6}$$

where $\overline{q} \in (0,1)$ by assumption. Substituting the upper bound $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$ in (6) 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

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

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 $\psi(q, K) = \big(K + \ln(2)\big)/\ln(1/q)$ and

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

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 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, where $i \in \{1, 2, \ldots\}$, 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

$$\mathrm{div}_f\left(\sum_{i\geqslant 1} \alpha_i\, \mathbb{P}_i(A_i),\, \sum_{i\geqslant 1} \alpha_i\, \mathbb{Q}_i(A_i)\right) \leqslant \sum_{i\geqslant 1} \alpha_i \mathrm{div}_f\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \sum_{i\geqslant 1} \alpha_i \mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{7}$$

The second inequality of (7) follows from the data-processing inequality (Lemma 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 23 (Appendix D), by considering the convex function $\varphi = \mathrm{div}_f$ (Corollary 3) on the convex set $C = [0,1]^2$, together with the probability measure

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

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 over Θ. We assume that

$$\theta \in \Theta \longmapsto \big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big) \qquad \text{and} \qquad \theta \in \Theta \longmapsto \mathrm{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

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

$$\mathrm{div}_f\left(\int_\Theta \mathbb{P}_\theta(A_\theta)\,\mathrm{d}\nu(\theta),\, \int_\Theta \mathbb{Q}_\theta(A_\theta)\,\mathrm{d}\nu(\theta)\right) \leqslant \int_\Theta \mathrm{div}_f\big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big)\,\mathrm{d}\nu(\theta) \leqslant \int_\Theta \mathrm{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)\,\mathrm{d}\nu(\theta)\,. \tag{8}$$

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

$$\theta \in \Theta \longmapsto \Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\ \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big] \Big) \qquad \text{and} \qquad \theta \in \Theta \longmapsto \mathrm{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

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

$$\begin{aligned} \mathrm{div}_f\left(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big] \,\mathrm{d}\nu(\theta),\ \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big] \,\mathrm{d}\nu(\theta) \right) &\leqslant \int_\Theta \mathrm{div}_f\Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\ \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big] \Big) \,\mathrm{d}\nu(\theta) \\ &\leqslant \int_\Theta \mathrm{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta) \,\mathrm{d}\nu(\theta)\,, \end{aligned} \tag{9}$$

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).

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)

$$\mathrm{div}_f\big(\overline{p}, \overline{q}\big) \leqslant \overline{D}\,,$$

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 $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\ln(1/q) - \ln(2)\,, \qquad \text{thus} \qquad p \leqslant \frac{\mathrm{kl}(p,q) + \ln(2)}{\ln(1/q)}\,. \tag{10}$$

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 $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\ln(1/q) - \ln(2-q)\,, \qquad \text{thus} \qquad p \leqslant \frac{\mathrm{kl}(p,q) + \ln(2-q)}{\ln(1/q)}\,. \tag{11}$$

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

$$p \leqslant 0.21 + 0.79\,q + \frac{\mathrm{kl}(p,q)}{\ln(1/q)}\,. \tag{12}$$

The improvement (11) is a consequence of, e.g., a convexity inequality, and its proof and the one for (12) 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 (10)–(12). Namely, for all $p \in [0,1]$ and $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant \max\left\{ \ln\left(\frac{1}{q}\right),\, 2 \right\} (p-q)^2\,, \qquad \text{thus} \qquad p \leqslant q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\big\{\ln(1/q), 2\big\}}}\,. \tag{13}$$

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 $(p, q) \in [0,1]^2$,

$$\chi^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) = \frac{(p-q)^2}{q(1-q)} \geqslant \frac{(p-q)^2}{q}\,, \qquad \text{thus} \qquad p \leqslant q + \sqrt{q\,\chi^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big)}\,. \tag{14}$$

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

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

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

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

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

$$\sqrt{pq} + \sqrt{(1-q)(1-p)} \leqslant \sqrt{\big(p + (1-q)\big)\big(q + (1-p)\big)} = \sqrt{1 - (p-q)^2}\,,$$

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

$$p \leqslant q + \sqrt{1 - \big(1 - h^2(p,q)/2\big)^2} = q + \sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}\,, \tag{15}$$

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 Appendix E.

4.3. Examples of combinations

The combination of (8) and (10) yields a continuous version of Fano's inequality. (We discard again all measurability issues.)

Lemma 5. *We consider a mesurable 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 over 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

$$\int_\Theta \mathbb{P}_\theta(A_\theta)\,\mathrm{d}\nu(\theta) \leqslant \frac{\displaystyle\int_\Theta \mathrm{KL}\big(\mathbb{P}_\theta, \mathbb{Q}_\theta\big)\,\mathrm{d}\nu(\theta) + \ln(2)}{\displaystyle -\ln\left(\int_\Theta \mathbb{Q}_\theta(A_\theta)\,\mathrm{d}\nu(\theta)\right)}\,.$$

The combination of (8), used with a uniform distribution ν over N points, and (13) 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 6. *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*

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

we have

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

In particular, if $N \geqslant 2$ and $Z_1 + \ldots + Z_N = 1$ a.s., then

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

For the χ^2–divergence now, the combination of, e.g., (7) in the finite and uniform case and (14) leads to the following inequality.

Lemma 7. *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*

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

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

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

Similarly, for the Hellinger distance, the simplest reduction (7) in the finite and uniform case together with the lower bound (15) yields the following bound.

Lemma 8. *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*

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

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

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

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 E 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 E 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, i.e., $N = 1$.

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

$$\mathbb{P}_\pi(A \,|\, x_{1:n}) = \frac{\displaystyle\int_A \prod_{i=1}^n p_{\theta'}(x_i)\, \mathrm{d}\pi(\theta')}{\displaystyle\int_\Theta \prod_{i=1}^n p_{\theta'}(x_i)\, \mathrm{d}\pi(\theta')} \tag{16}$$

if the denominator lies in $(0, +\infty)$; if it is null or infinite, we set, e.g., $\mathbb{P}_\pi(A \,|\, x_{1:n}) = \pi(A)$. 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 $(\varepsilon_n)_{n \geqslant 1}$ of positive real numbers such that, for all $\theta \in \Theta$,

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

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

$$\inf_{\theta \in \Theta} \mathbb{E}_\theta \Big[\mathbb{P}_\pi \big(\theta' : \ell(\theta', \theta) \leqslant \varepsilon_n \,\big|\, X_{1:n} \big) \Big] \longrightarrow 1 \qquad \text{as } n \to +\infty\,. \tag{17}$$

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

The minimax posterior concentration rate is given by a sequence $(\varepsilon_n)_{n\geqslant 1}$ such that (17) holds for some prior π while there exists a constant $\gamma \in (0,1)$ such that for all priors π' on Θ,

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

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

Proposition 9 (A posterior concentration lower bound in the finite-dimensional Gaussian model)**.** *Let $d \geqslant 1$ be the ambient dimension, $n \geqslant 1$ the sample size, and $\sigma > 0$ the standard deviation. Assume we observe an n–sample $X_{1:n} = (X_1, \dots, X_n)$ 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* (16) *satisfies, for the Euclidean loss $\ell(\theta',\theta) = \|\theta'-\theta\|_2$ and for $\varepsilon_n = (\sigma/8)\sqrt{d/n}$,*

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

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

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, e.g., Le Cam and Yang [2000, Chapter 8].

There are at least two ways to prove the lower bound of Proposition 9. 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 et al. [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 $Z_\theta = \mathbb{P}_{\pi'}\big(\theta' : \|\theta'-\theta\|_2 \leqslant \varepsilon_n \,\big|\, X_{1:n}\big) \in [0,1]$.

Proof: 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 $B(0,\rho\varepsilon_n) = \big\{u \in \mathbb{R}^d : \|u\|_2 \leqslant \rho\varepsilon_n\big\}$ for some $\rho > 1$ to be determined by the analysis. Then, by the continuous Fano inequality in the form given by the combination of (9) and (13), 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 $Z_\theta = \mathbb{P}_{\pi'}\big(\theta' : \|\theta'-\theta\|_2 \leqslant \varepsilon_n \,\big|\, X_{1:n}\big)$, we have

$$\begin{aligned}\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\big[Z_\theta\big] &\leqslant \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_\theta\big[Z_\theta\big]\,\mathrm{d}\nu(\theta) \leqslant \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\,\mathrm{d}\nu(\theta) + \sqrt{\frac{\displaystyle\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}(\mathbb{P}_\theta,\mathbb{P}_0)\,\mathrm{d}\nu(\theta)}{\displaystyle -\ln \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\,\mathrm{d}\nu(\theta)}} \\ &\leqslant \left(\frac{1}{\rho}\right)^d + \sqrt{\frac{n\rho^2\varepsilon_n^2/(2\sigma^2)}{d\ln\rho}}\,, \end{aligned} \tag{18}$$

where the last inequality follows from (19) and (20) below. First note that, by independence, $\mathrm{KL}\big(\mathbb{P}_\theta,\mathbb{P}_0\big) = n\,\mathrm{KL}\big(\mathcal{N}(\theta,\sigma^2),\mathcal{N}(0,\sigma^2)\big) = n\,\|\theta\|_2^2/(2\sigma^2)$, so that

$$\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}\big(\mathbb{P}_\theta,\mathbb{P}_0\big)\,\mathrm{d}\nu(\theta) = \frac{n}{2\sigma^2}\int_{B(0,\rho\varepsilon_n)} \|\theta\|_2^2\,\mathrm{d}\nu(\theta) \leqslant \frac{n\rho^2\varepsilon_n^2}{2\sigma^2}\,. \tag{19}$$

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

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

we can see that

$$q \overset{\text{def}}{=} \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\,\mathrm{d}\nu(\theta) = \mathbb{E}_0\left[\int_{B(0,\rho\varepsilon_n)} \mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\big[\mathbb{1}_{\{\|\theta'-\theta\|_2\leqslant\varepsilon_n\}}\big]\,\mathrm{d}\nu(\theta)\right]$$
$$= \mathbb{E}_0\left[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\left[\int_{B(0,\rho\varepsilon_n)} \mathbb{1}_{\{\|\theta'-\theta\|_2\leqslant\varepsilon_n\}}\,\mathrm{d}\nu(\theta)\right]\right]$$
$$= \mathbb{E}_0\left[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\Big[\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big)\Big]\right] \leqslant \left(\frac{1}{\rho}\right)^d, \qquad (20)$$

where to get the last inequality we used the fact that $\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big)$ 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, e.g., $\varepsilon_n = (\sigma/8)\sqrt{d/n}$: by (18), we have

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

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

Note that, when using (13) above, we implicitly assumed that the quantity q in (20) 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 $\mathbb{P}_0\big(\mathbb{P}_{\pi'}(B(0,\varepsilon_n)\,|\,X_{1:n}) > 0\big) > 0$; indeed, for $\theta' \in B(0,\varepsilon_n)$, we have $\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big) > 0$ and thus q appears in the last equality of (20) as being lower bounded by the expectation of a positive function over a set with positive probability. If on the contrary $\mathbb{P}_0\big(\mathbb{P}_{\pi'}(B(0,\varepsilon_n)\,|\,X_{1:n}) > 0\big) = 0$, 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 9. □

Remark 1. Though the lower bound of Proposition 9 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 $t \geqslant 1$, 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 $T \geqslant 1$,

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

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 et al. [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 non-asymptotic 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 re-shuffled 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, e.g., 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 non-asymptotic. 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 10. *For all strategies of the decision-maker, for all $s \in \{0,\dots,N\}$, for all $N \geqslant 2$, for all $T \geqslant 1$, 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*

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

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 $\varepsilon \in \big(0,\, s/(2N)\big)$ 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 $1 \leqslant i \leqslant N$. Each distribution P_i over $[0,1]^N$ 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 T–th 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

$$\mathbb{E}_i\Big[\mathbb{E}\big[R_T\big]\Big] \geqslant \sum_{t=1}^{T} \mathbb{E}_i\Big[\mathbb{E}\big[\ell_{I_t,t}\big]\Big] - \min_{k=1,\dots,N} \sum_{t=1}^{T} \mathbb{E}_i\big[\ell_{k,t}\big] = \sum_{t=1}^{T} \frac{s}{2N}\left(1 - \varepsilon\,\mathbb{E}_i\Big[\mathbb{E}\big[\mathbb{1}_{\{I_t=i\}}\big]\Big]\right) - T\left(\frac{s}{2N} - \varepsilon\right)$$
$$= T\varepsilon\left(1 - \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big]\right),$$

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

$$\sup_{\ell_1,\dots,\ell_t} \mathbb{E}\big[R_T\big] \geqslant \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[R_T\big]\Big] \geqslant T\varepsilon\left(1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big]\right). \tag{21}$$

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 over $[0,1]^N$ 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 6 with $\mathbb{P}_i = \mathbb{P}\otimes P_i^T$ and $\mathbb{Q}_i = \mathbb{P}\otimes Q^T$, using that $F_1(T) + \ldots + F_N(T) = 1$ and thus $(1/N)\sum_{i=1}^{N} \mathbb{E}_Q\big[\mathbb{E}[F_i(T)]\big] = 1/N$, yields

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big] \leqslant \frac{1}{N} + \sqrt{\frac{1}{N\ln(N)}\sum_{i=1}^{N} \mathrm{KL}\big(\mathbb{P}\otimes P_i^T,\, \mathbb{P}\otimes Q^T\big)}\,. \tag{22}$$

By independence, we get, for all i,

$$\mathrm{KL}\big(\mathbb{P}\otimes P_i^T,\, \mathbb{P}\otimes Q^T\big) = \mathrm{KL}\big(P_i^T,\, Q^T\big) = T\;\mathrm{KL}\big(P_i,\, Q\big)\,. \tag{23}$$

We now show that

$$\mathrm{KL}\big(P_i,\, Q\big) \leqslant \frac{s}{N}\;\mathrm{kl}\left(\frac{1}{2} - \varepsilon\frac{N}{s},\, \frac{1}{2}\right). \tag{24}$$

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)

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

where δ_0 denotes the Dirac mass at 0, and

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

$$\bigotimes_{k=1}^{s} \mathrm{Ber}\left(\frac{1}{2}\right) \otimes \bigotimes_{k=s+1}^{N} \delta_0 \qquad \text{and} \qquad \bigotimes_{k=1}^{s} \mathrm{Ber}\left(\frac{1}{2}\right) \otimes \bigotimes_{k=s+1}^{N} \delta_0\,.$$

Only the first set of distributions contributes to the Kullback-Leibler divergence. By convexity of the Kullback-Leibler divergence (Corollary 3), we thus get the inequality

$$\begin{aligned}\mathrm{KL}\big(P_i,\, Q\big) &\leqslant \frac{\binom{N-1}{s-1}}{\binom{N}{s}}\,\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2}-\varepsilon\frac{N}{s}\right)\otimes\bigotimes_{k=2}^{s}\mathrm{Ber}\left(\frac{1}{2}\right)\otimes\bigotimes_{k=s+1}^{N}\delta_0,\ \bigotimes_{k=1}^{s}\mathrm{Ber}\left(\frac{1}{2}\right)\otimes\bigotimes_{k=s+1}^{N}\delta_0\right)\\ &= \frac{s}{N}\,\mathrm{kl}\left(\frac{1}{2}-\varepsilon\frac{N}{s},\,\frac{1}{2}\right),\end{aligned}$$

where the last equality is again by independence. Finally, the lemma stated right after this proof shows that

$$\mathrm{kl}\left(\frac{1}{2}-\varepsilon\frac{N}{s},\,\frac{1}{2}\right)\leqslant\frac{4N^2\varepsilon^2}{s^2}\,. \tag{25}$$

Combining (21)–(25), we proved so far

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

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

A standard optimization suggests the choice $\varepsilon=1/(4c)$, which is valid, i.e., 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 $T\leqslant N\ln(N)/(16s)$, we have $c\leqslant N/(2s)$ and the valid choice $\varepsilon=s/(4N)$ leads to the part of the lower bound given by $T\varepsilon(1/2-c\varepsilon)\geqslant T\varepsilon/4=sT/(16N)$. □

Lemma 11. *For all $p\in(0,1)$, for all $\varepsilon\in(0,p)$,*

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

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:

$$\mathrm{kl}(p-\varepsilon,p)=(p-\varepsilon)\ln\left(1-\frac{\varepsilon}{p}\right)+(1-p+\varepsilon)\ln\left(1+\frac{\varepsilon}{1-p}\right)\leqslant(p-\varepsilon)\frac{-\varepsilon}{p}+(1-p+\varepsilon)\frac{\varepsilon}{1-p}=\frac{\varepsilon^2}{p}+\frac{\varepsilon^2}{1-p}\,,$$

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

6. Other applications, with $N = 1$ pair of distributions

Interestingly, Proposition 4 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,

$$\mathbb{P}(A) \ln\left(\frac{1}{\mathbb{Q}(A)}\right) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \ln(2)\,.$$

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

$$\mathbb{Q}(A) \geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \ln(2)}{\mathbb{P}(A)}\right), \tag{26}$$

where the above inequality is true even if $\mathbb{P}(A) = 0$ or $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = +\infty$. We applied here a classical technique in information theory due to Haroutunian; see, for instance, Csiszár and Körner [1981, page 167].

Similarly and more generally, for all distributions $\mathbb{P}, \mathbb{Q}$ and all $[0, 1]$–valued random variables Z, we have, by Corollary 2 and the lower bound (6),

$$\mathbb{E}_{\mathbb{Q}}[Z] \geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \ln(2)}{\mathbb{E}_{\mathbb{P}}[Z]}\right), \tag{27}$$

where again the above inequality is true even if $\mathbb{E}_{\mathbb{P}}[Z] = 0$ or $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = +\infty$.

The bound (26) 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 (43). 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 (26) and (27): 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 (26), 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 12 (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 $\overline{X}_n \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} X_i$. 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 et al. 2013, pages 21–24), we have, for all $n \geqslant 1$,

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \leqslant \exp\left(-n \sup_{\lambda>0}\Big\{\lambda x - \ln \mathbb{E}_\theta\Big[e^{\lambda X_1}\Big]\Big\}\right) = \exp\Big(-n\,\mathrm{kl}(x, \theta)\Big)\,, \tag{28}$$

that is,

$$\forall n \geqslant 1, \qquad \frac{1}{n} \ln \mathbb{P}_\theta\big(\overline{X}_n > x\big) \leqslant -\operatorname{kl}(x, \theta)\,.$$

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 (26) with the distributions $\mathbb{P} = \mathbb{P}_{x+\varepsilon}$ and $\mathbb{Q} = \mathbb{P}_\theta$, and the event $A = \big\{\overline{X}_n > x\big\}$, we have

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

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

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant \exp\left(-\frac{n \operatorname{kl}(x+\varepsilon, \theta) + \ln 2}{\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big)}\right) \geqslant \exp\left(-\frac{n \operatorname{kl}(x+\varepsilon, \theta) + \ln 2}{1 - e^{-n \operatorname{kl}(x, x+\varepsilon)}}\right), \tag{29}$$

where the last bound follows from $\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big) = 1 - \mathbb{P}_{x+\varepsilon}\big(\overline{X}_n \leqslant x\big) \geqslant 1 - e^{-n \operatorname{kl}(x, x+\varepsilon)}$ by a derivation similar to (28) above. Taking the logarithms of both sides and letting $n \to +\infty$ finally yields

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

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, i.e., that transports the measure $\mathrm{Ber}(\theta)^{\otimes n}$ to $\mathrm{Ber}(x+\varepsilon)^{\otimes n}$. The bound (28), or any other large deviation inequality, is also typically used therein. However, the change of measure is usually carried out explicitly by writing

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

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

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

The empirical Kullback-Leibler divergence $\widehat{\mathrm{KL}}_n$ is then compared to its limit $\operatorname{kl}(x+\varepsilon, \theta)$ 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 explicitely, in the same spirit as for the bandit lower bounds derived by Kaufmann et al. [2016] and Garivier et al. [2018]. 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

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

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 to the minimax lower bound of Section 5.1.

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 (31) below was proved to be impossible at scale ε_n (and actually took place at a slightly larger scale ε'_n).

Theorem 13 (Distribution-dependent posterior concentration lower bound)**.** *Assume that the posterior distribution $\mathbb{P}_\pi(\,\cdot\,|\,X_{1:n})$ satisfies the uniform concentration condition*

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

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

$$\mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) > \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant 2^{-c}\,\exp\Big(-c\,n\,\psi(\varepsilon_n,\theta,\ell)\Big)\,. \tag{30}$$

The conclusion can be stated equivalently as: for all $\theta\in\Theta$,

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

The above theorem is greatly inspired from Theorem 2.1 by Hoffmann et al. [2015]. Our Fano's inequality (27) however makes the proof more direct: the change-of-measure carried out by Hoffmann et al. [2015] is now implicit, and no proof by contradiction is required. We also bypass one technical assumption (see the discussion after the proof).

Proof: We fix $\theta\in\Theta$ and $c>1$. By the uniform concentration condition, there exists $n_0\geqslant 1$ such that, for all $n\geqslant n_0$,

$$\inf_{\theta^\star\in\Theta}\ \mathbb{E}_{\theta^\star}\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta^\star) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant \frac{1}{c}\,. \tag{31}$$

We now fix $n\geqslant n_0$ and consider any $\theta^\star\in\Theta$ such that $\ell\big(\theta^\star,\theta\big)\geqslant 2\varepsilon_n$. Using Fano's inequality in the form of (27) 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 $Z_\theta=\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) > \varepsilon_n \,\big|\, X_{1:n}\big)$, we get

$$\mathbb{E}_\theta\big[Z_\theta\big] \geqslant \exp\left(-\frac{\mathrm{KL}\Big(P_{\theta^\star}^{\otimes n},P_\theta^{\otimes n}\Big)+\ln 2}{\mathbb{E}_{\theta^\star}\big[Z_\theta\big]}\right) = \exp\left(-\frac{n\,\mathrm{KL}\big(P_{\theta^\star},P_\theta\big)+\ln 2}{\mathbb{E}_{\theta^\star}\big[Z_\theta\big]}\right). \tag{32}$$

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

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

by the uniform lower bound (31). Substituting the above inequality into (32) then yields

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

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

Note that, at first sight, our result may seem a little weaker than Hoffmann et al. [2015, Theorem 2.1], because we only define $\psi\big(\varepsilon_n,\theta,\ell\big)$ in terms of KL instead of a general pre-metric d: in 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 pre-metric d by comparing d and KL after applying Theorem 13.

In the case of the pre-metric $d(\theta, \theta') = \sqrt{\mathrm{KL}(P_{\theta'}, P_\theta)}$, we bypass an additional technical assumption used for the the similar lower bound of Hoffmann et al. [2015, Theorem 2.1]; namely, that there exists a constant $C > 0$ such that

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

where the supremum is over all $\theta, \theta' \in \Theta$ satisfying $\psi\big(\varepsilon_n, \theta, \ell\big) \leqslant \mathrm{KL}\big(P_{\theta'}, P_\theta\big) \leqslant 2\psi\big(\varepsilon_n, \theta, \ell\big)$, and where $\mathcal{L}_n(\theta) = \sum_{i=1}^n \ln\big(\mathrm{d}P_\theta / \mathrm{d}\mathfrak{m}\big)(X_i)$ denotes the log-likelihood function with respect to a common dominating measure $\mathfrak{m}$. Besides, we get an improved constant in the exponential in (30), with respect to Hoffmann et al. [2015, Theorem 2.1]: by a factor of $3C/c$, which, since $C \geqslant 1$ in most cases, is $3C/c \approx 3C \geqslant 3$ 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 $2C/c \approx 2C \geqslant 2$.)

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 (9), 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 (4) 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 et al. [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 $Z_1 + \ldots + Z_N = 1$ are considered; see also Chen et al. [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 et al., 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 4 can be found, e.g., in Cover and Thomas [2006, Theorem 2.11.1], Yu [1997, Lemma 3] or Ibragimov and Has'minskii [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 (9) 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 (4) and (7) were already known and used by Han and Verdú [1994, Theorems 2, 7 and 8]. Reduction (8) is stated in spirit by Chen et al. [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 (8). Finally, as already mentioned, a reduction with random variables like (9) was stated in a special case in Gushchin [2003], for finitely many $[0,1]$–valued random variables with $Z_1 + \ldots + Z_N = 1$.

On the lower bounds on the kl **function (not really a new result).** The inequalities (10) are folklore knowledge. The first inequality in (11) can be found in Guntuboyina [2011]; the second

inequality is a new (immediate) consequence. The inequalities (13) 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 Fanos's inequality" of Chen et al. [2016] and the version of Fano's inequality by Birgé [2005], which is extremely popular among (French) statisticians.

7.1. On the "generalized Fanos's inequality" of Chen et al. [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 14.

A parameter space $(\Theta, \mathcal{G})$ is equipped with a prior probability measure ν. A family of probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$ over a measurable space $(\Omega, \mathcal{F})$, some outcome space $(\mathcal{X}, \mathcal{E})$, e.g., $\mathcal{X} = \mathbb{R}^n$, 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

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

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

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

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

Chen et al. [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 (8) with events, but we also have the reduction (9) 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 14. *In the setting described above, the Bayes risk is always larger than*

$$R_{\text{Bayes}} \geqslant 1 + \frac{\left(\inf\limits_{\mathbb{Q}} \displaystyle\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q})\, \mathrm{d}\nu(\theta)\right) + \ln\left(1 + \inf\limits_{a\in\mathcal{A}} \displaystyle\int_\Theta L(\theta, a)\, \mathrm{d}\nu(\theta)\right)}{\ln\left(1 - \inf\limits_{a\in\mathcal{A}} \displaystyle\int_\Theta L(\theta, a)\, \mathrm{d}\nu(\theta)\right)}\,,$$

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

Proof: We fix $\widehat{a}$ and an alternative $\mathbb{Q}$. The combination of (9) and (11), with $Z_\theta = 1 - L\big(\theta, \widehat{a}\big)$, yields

$$1 - \int_\Theta \mathbb{E}_\theta\Big[L\big(\theta, \widehat{a}\big)\Big]\, \mathrm{d}\nu(\theta) \leqslant \frac{\displaystyle\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q})\, \mathrm{d}\nu(\theta) + \ln\big(2 - q_{\widehat{a}}\big)}{\ln\big(1/q_{\widehat{a}}\big)}\,, \tag{33}$$

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

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

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 (33) gives

$$1 - R_{\mathrm{Bayes}} \leqslant \frac{\displaystyle\int_{\Theta} \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}) \, \mathrm{d}\nu(\theta) + \ln\big(2 - q^\star\big)}{\ln\big(1/q^\star\big)} \tag{34}$$

where

$$q^\star = \sup_{\widehat{a}} q_{\widehat{a}} = 1 - \inf_{\widehat{a}} \int_{\Theta} \mathbb{E}_{\mathbb{Q}}\Big[L\big(\theta, \widehat{a}\big)\Big] \, \mathrm{d}\nu(\theta) = 1 - \inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a) \, \mathrm{d}\nu(\theta) \, , \tag{35}$$

as is proved below. Taking the infimum of the right-hand side of (34) over $\mathbb{Q}$ and rearranging concludes the proof.

It only remains to prove the last inequality of (35) and actually, as constant elements $a \in \mathcal{A}$ are special cases of random variables $\widehat{a}$, we only need to prove that

$$\inf_{\widehat{a}} \int_{\Theta} \mathbb{E}_{\mathbb{Q}}\Big[L\big(\theta, \widehat{a}\big)\Big] \, \mathrm{d}\nu(\theta) \geqslant \inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a) \, \mathrm{d}\nu(\theta) \, . \tag{36}$$

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:

$$\int_{\Theta} \mathbb{E}_{\mathbb{Q}}\Big[L\big(\theta, \widehat{a}\big)\Big] \, \mathrm{d}\nu(\theta) = \int_{\mathcal{X}} \left(\int_{\Theta} L\big(\theta, \overline{a}(x)\big) \, \mathrm{d}\nu(\theta) \right) \mathrm{d}\mathbb{Q}(x) \geqslant \int_{\mathcal{X}} \left(\inf_{a \in \mathcal{A}} \int_{\Theta} L\big(\theta, a\big) \, \mathrm{d}\nu(\theta) \right) \mathrm{d}\mathbb{Q}(x) \, ,$$

which proves (36). □

Remark 2. As mentioned by Chen et al. [2016], one of the major results of Han and Verdú [1994], namely, their Theorem 8, is a special case of Proposition 14, 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

$$-\ln\left(1 - \inf_{\theta' \in \Theta} \int_{\Theta} L(\theta, \theta') \, \mathrm{d}\nu(\theta)\right) = -\ln\Big(\sup_{\theta \in \Theta} \nu\big(\{\theta\}\big)\Big) \stackrel{\mathrm{def}}{=} H_\infty(\nu) \, ,$$

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

This version of Fano's inequality is extremely popular among (French) statisticians. It only deal with events $A_1, \ldots, A_N$ forming a partition of the underlying measurable space. As should be clear from its proof (provided in Appendix C) this assumption is crucial.

Theorem 15 (Birgé's lemma)**.** *Given an underlying measurable space $(\Omega, \mathcal{F})$, for all $N \geqslant 2$, for all probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_N$, for all events $A_1, \dots, A_N$ forming a partition of Ω,*

$$\min_{1 \leqslant i \leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{ c_N,\ \frac{\overline{K}}{\ln(N)} \right\} \qquad \text{where} \qquad \overline{K} = \frac{1}{N-1} \sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1)$$

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

$$\frac{-\big(c \ln(c) + (1-c)\ln(1-c)\big)}{c} + \ln(1-c) = \ln\left(\frac{N-1}{N}\right). \tag{37}$$

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

The aim of this subsection is to compare this bound to the versions of Fano's inequality following from the kl lower bounds (11), (10), and (13), in this order. In the setting of the theorem above and by picking constant alternatives $\mathbb{Q}$, these lower bounds on kl respectively lead to

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N} \inf\limits_{\mathbb{Q}} \displaystyle\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \ln\left(2 - \frac{1}{N}\right)}{\ln(N)} \leqslant \frac{\dfrac{1}{N} \inf\limits_{\mathbb{Q}} \displaystyle\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \ln(2)}{\ln(N)}, \tag{38}$$

$$\text{and} \qquad \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(A_i) \leqslant \frac{1}{N} + \sqrt{\frac{\dfrac{1}{N} \inf\limits_{\mathbb{Q}} \displaystyle\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q})}{\max\big\{\ln(N),\, 2\big\}}}. \tag{39}$$

The main point of Birgé [2005] was that the most classical version of Fano's inequality, that is, the right-most side of (38), was quite unpractical for small values of N, and even useless when $N = 2$. In the latter case $N = 2$, the statistical doxa had it that one should rather resort to Pinsker's inequality, which is exactly (39) when $N = 2$ (as it then is larger than 1). One of the main motivations of Birgé [2005] was therefore to get an inequality that would be useful for all $N \geqslant 2$, so that one does not have to decide which of the classical Pinsker's inequality or the classical Fano's inequality should be applied. A drawback, however, of his bound is the $\overline{K}$ term, in which one cannot pick a convenient $\mathbb{Q}$ as in the bounds (38)–(39). Also, the result is about the minimum of the $\mathbb{P}_i(A_i)$, not about their average.

Now, we note that unlike the right-most side of (38), both the middle term in (38) and the bound (39) yield useful bounds, even for $N = 2$. The middle term in (38) was derived—with a different formulation—by Chen et al. [2016], see Proposition 14 above. Our contribution is to note that our inequality (39) provides an interpolation between Pinsker's and Fano's inequalities. More precisely, (39) implies both Pinsker's inequality and, lower bounding the maximum by $\ln(N)$, a bound as useful as Theorem 15 or Proposition 4 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.

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 (11) and (12) as well as the refined Pinsker's inequality and its consequence (13). 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 (11) and (12)

Proof: Inequality (12) follows from (11) via a function study of $q \in (0,1) \mapsto \ln(2-q)\big/\ln(1/q)$, which is dominated by $0.21 + 0.79\,\mathrm{q}$.

Now, the shortest proof of (11) notes that the duality formula for the Kullback-Leibler divergence between Bernoulli distributions—already used in (28)—ensures that, for all $p \in [0,1]$ and $q \in (0,1]$,

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

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 et al. [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

$$\forall\,(p,p_0) \in (0,1)^2, \qquad \mathrm{kl}(p,q) - \mathrm{kl}(p_0,q) \geqslant \underbrace{\ln\left(\frac{p_0(1-q)}{(1-p_0)q}\right)}_{\frac{\partial}{\partial p}\mathrm{kl}(p_0,q)}(p-p_0)\,. \tag{40}$$

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

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

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 (40) becomes

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

which proves as well the bound (11).

8.2. Proofs of the refined Pinsker's inequality and of its consequence (13)

The next theorem is a stronger version of Pinsker's inequality for Bernoulli distributions, that was proved[2] 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: $\mathrm{kl}(p,q) \geqslant 2(p-q)^2$.

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

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

where the multiplicative factor $\varphi(q) = (1-2q)^{-1}\ln\bigl((1-q)/q\bigr)$ 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: 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 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

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

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

$$\frac{\partial}{\partial p}\,\mathrm{kl}(p,q) = \ln\!\left(\frac{p(1-q)}{(1-p)q}\right) \qquad \text{and} \qquad \frac{\partial^2}{\partial^2 p}\,\mathrm{kl}(p,q) = \frac{1}{p(1-p)} \stackrel{\mathrm{def}}{=} \psi(p)\,, \tag{41}$$

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

$$f(p) = \frac{\mathrm{kl}(p,q)}{(p-q)^2} = \frac{1}{(p-q)^2}\int_q^p \frac{\psi(t)}{1!}(p-t)^1\,\mathrm{d}t = \int_0^1 \psi\bigl(q+u(p-q)\bigr)(1-u)\,\mathrm{d}u\,.$$

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 $p = 1-q$,

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

the equality to 0 follows from the fact that the function $u \mapsto \psi'\bigl(q+u(1-2q)\bigr)u(1-u)$ 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 (13). A picture is helpful, see Figure 1.

[2] 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.

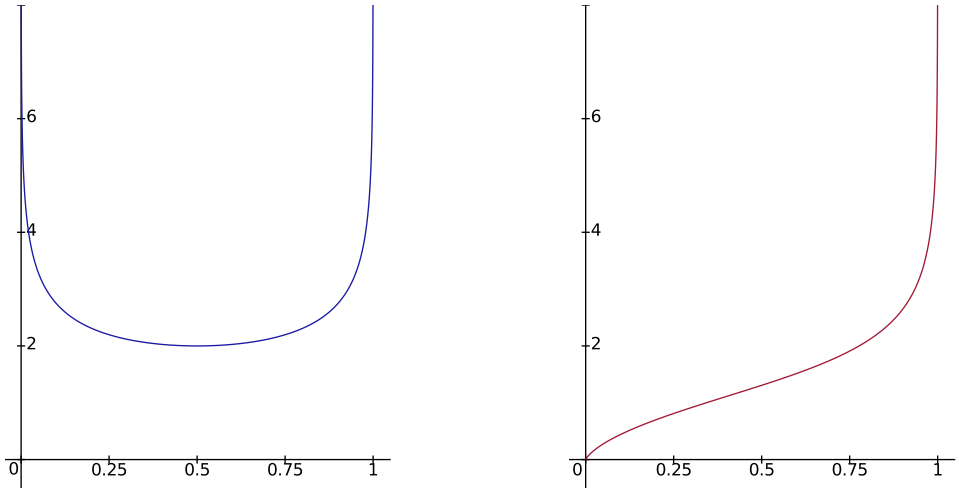

Figure 1: Plots of φ [left] and $x \in (0,1) \mapsto \varphi(x) - \ln(1/x)$ [right].

Corollary 17. *For all $q \in (0,1]$, we have $\varphi(q) \geqslant 2$ and $\varphi(q) \geqslant \ln(1/q)$. Thus, for all $p \in [0,1]$ and $q \in (0,1)$,*

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

Slightly sharper bounds are possible, like $\varphi(q) \geqslant (1+q)(1+q^2)\ln(1/q)$ or $\varphi(q) \geqslant \ln(1/q) + 2.5\,q$, 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 16 and the data-processing inequality (Lemma 1).

Theorem 18. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$. Then*

$$\forall\, A \in \mathcal{F}, \qquad \bigl|\mathbb{P}(A) - \mathbb{Q}(A)\bigr| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{\varphi\bigl(\mathbb{Q}(A)\bigr)}}\,,$$

where $\varphi \geqslant 2$ is defined in the statement of Theorem 16. In particular, the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ is bounded as

$$\sup_{A \in \mathcal{F}} \bigl|\mathbb{P}(A) - \mathbb{Q}(A)\bigr| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{\inf\limits_{A \in \mathcal{F}} \varphi\bigl(\mathbb{Q}(A)\bigr)}}\,.$$

8.3. An improved Bretagnolle-Huber inequality

The Bretagnolle-Huber inequality was introduced by Bretagnolle and Huber [1978, 1979]. The multiplicative factor $\mathrm{e}^{-1/\mathrm{e}} \geqslant 0.69$ in our statement (42) below is a slight improvement over the original $1/2$ factor. For all $p, q \in [0,1]$,

$$1 - |p-q| \geqslant \mathrm{e}^{-1/\mathrm{e}}\,\mathrm{e}^{-\,\mathrm{kl}(p,q)}\,, \qquad \text{thus} \qquad q \geqslant p - 1 + \mathrm{e}^{-1/\mathrm{e}}\,\mathrm{e}^{-\,\mathrm{kl}(p,q)}\,. \tag{42}$$

It is worth to note that Bretagnolle and Huber [1978] also proved the inequality

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

which improves as well upon the Bretagnolle-Huber inequality with the $1/2$ factor, but which is neither better nor worse than (42).

Now, via the data-processing inequality (Lemma 1), we get from (42)

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

The left-hand side can be rewritten as $\inf_{A\in\mathcal{F}}\big\{\mathbb{P}(A)+\mathbb{Q}(A^c)\big\}$, 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 (26), we can restate the general version of the Bretagnolle-Huber inequality as: for all $A \in \mathcal{F}$,

$$\mathbb{Q}(A) \geqslant \mathbb{P}(A) - 1 + \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\operatorname{KL}(\mathbb{P},\mathbb{Q})}\,. \tag{43}$$

We now provide a proof of (42); 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 $x \in (0,1) \mapsto x \ln\big(x/(1-q)\big)$ equals $1 + \ln(x) - \ln(1-q)$, so that the function achieves its minimum at $x = (1-q)/\mathrm{e}$, with value $-(1-q)/\mathrm{e} \geqslant -1/\mathrm{e}$. Therefore,

$$-\operatorname{kl}(p,q) = -p\ln\left(\frac{p}{q}\right) - (1-p)\ln\left(\frac{1-p}{1-q}\right) \leqslant -p\ln\left(\frac{p}{q}\right) + \frac{1}{\mathrm{e}} = p\left(\ln\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\frac{1}{\mathrm{e}}\,.$$

Therefore, using the convexity of the exponential,

$$\mathrm{e}^{-\operatorname{kl}(p,q)} \leqslant p\, \exp\left(\ln\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\,\mathrm{e}^{1/\mathrm{e}} = \big(q + (1-p)\big)\,\mathrm{e}^{1/\mathrm{e}}\,,$$

which shows that

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

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

$$1 - (q-p) = 1 - \big((1-p) - (1-q)\big) \geqslant \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\operatorname{kl}(1-p,1-q)} = \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\operatorname{kl}(p,q)}\,.$$

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

References

S. Aeron, V. Saligrama, and M. Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10), October 2010.

S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B. Methodological*, 28:131–142, 1966.

D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1615, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

J. Bretagnolle and C. Huber. Estimation des densités : risque minimax. *Séminaire de Probabilités de Strasbourg*, 12:342–363, 1978.

J. Bretagnolle and C. Huber. Estimation des densités : risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

R. Cerf and P. Petit. A short proof of Cramér's theorem in R. *The American Mathematical Monthly*, 118(10):925–931, 2011.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R. Schapire, and M. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label-efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2005.

X. Chen, A. Guntuboyina, and Y. Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(219):1–58, 2016.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, second edition, 2006.

H. Cramér. Sur un nouveau théorème limite de la théorie des probabilités. *Actualites Scientifiques et Industrielles*, 736:5–23, 1938.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8:85–108, 1963.

I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, Budapest, 1981.

J. Duchi and M.J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. 2013. arXiv:1311.2669.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010.

T.S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London, 1967.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: the true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018. In press.

S. Gerchinovitz, P. Ménard, and G. Stoltz. Fano's inequality for random variables, 2018. arXiv:1702.05985.

S. Ghosal, J.K. Ghosh, and A.W. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

R.M. Gray. *Entropy and Information Theory*. Springer, second edition, 2011.

A. Guntuboyina. Lower bounds for the minimax risk using-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

A.A. Gushchin. On Fano's lemma and similar inequalities for the minimax risk. *Probability Theory and Mathematical Statistics*, 67:26–37, 2003.

T.S. Han and S. Verdú. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994.

P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, 2011.

M. Hayashi. *Quantum Information Theory*. Springer, 2017.

M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. *Annals of Statistics*, 43(5):2259–2295, 2015.

I.A. Ibragimov and R.Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*, volume 16. Springer-Verlag New York, 1981.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (UAI'98)*, pages 311–319, 1998.

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(229):1–32, 2016.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

L. Le Cam and G.L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2000.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

E. Ordentlich and M.J. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, 2005.

L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman & Hall/CRC, 2006.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, second edition, 1972.

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

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G.L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. New York, NY, 1997.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

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 $\mathrm{kl}\big(\overline{p},\overline{q}\big) \leqslant \overline{K}$. 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 $y \geqslant 0$,

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

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

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

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 (10) and (11) are of the form

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

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 (10) and (11) of Section 4.2 are essentially optimal.

The bound (13) therein is of the form

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

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

$$\mathrm{kl}(p,q) \leqslant p\ln\left(\frac{1}{q}\right) \qquad \textit{thus} \qquad p \geqslant \frac{\mathrm{kl}(p,q)}{\ln(1/q)}\,.$$

Proof: We note that when $p \geqslant q$, we have $(1-p)/(1-q) \leqslant 1$, so that

$$\mathrm{kl}(p,q) = p\ln\left(\frac{1}{q}\right) + \underbrace{p\ln(p)}_{\leqslant 0} + (1-p)\underbrace{\ln\left(\frac{1-p}{1-q}\right)}_{\leqslant 0} \leqslant p\ln\left(\frac{1}{q}\right),$$

hence the first inequality. □

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 et al. [2000] (see also Chapter 12, Proposition 3 by Le Cam, 1986, as well as Section 5.1 by Hoffmann et al., 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 $(X_1, \dots, X_n)$ is the identity random variable.

Proposition 20 (From Bayesian posteriors to point estimators)**.**
Let $n \geqslant 1$, $\delta > 0$, and $\theta \in \Theta$. Let $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n)$ be any estimator satisfying, $\mathbb{P}_\theta$–almost surely,

$$\mathbb{P}_\pi\Big(\theta' : \ell(\theta', \widehat{\theta}_n) < \varepsilon_n \,\Big|\, X_{1:n}\Big) \geqslant \sup_{\widetilde{\theta} \in \Theta} \mathbb{P}_\pi\Big(\theta' : \ell\big(\theta', \widetilde{\theta}\,\big) < \varepsilon_n \,\Big|\, X_{1:n}\Big) - \delta\,. \tag{44}$$

Then,

$$\mathbb{P}_\theta\bigg(\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\bigg) \geqslant \mathbb{P}_\theta\Big(\ell\big(\widehat{\theta}_n, \theta\big) \geqslant 2\varepsilon_n\Big)\,. \tag{45}$$

This result implies that if $\widehat{\theta}_n$ is a center of a ball that almost maximizes the posterior mass—see assumption (44)—and if the posterior mass concentrates around θ at a rate $\varepsilon_n' < \varepsilon_n$—so that the left-hand side of (45) 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 (44) holds, i.e., 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 $\widetilde{\theta} \mapsto \mathbb{P}_\pi\big(\theta' : \ell(\theta', \widetilde{\theta}\,) < \varepsilon_n \,\big|\, x_{1:n}\big)$ 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 21 for more details).

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

$$\begin{aligned}
\Big\{\ell\big(\widehat{\theta}_n, \theta\big) \geqslant 2\varepsilon_n\Big\} &\subseteq \Big\{B_\ell\big(\widehat{\theta}_n, \varepsilon_n\big) \cap B_\ell(\theta, \varepsilon_n) = \varnothing\Big\} \\
&\subseteq \Big\{\mathbb{P}_\pi\big(B_\ell\big(\widehat{\theta}_n, \varepsilon_n\big) \,\big|\, X_{1:n}\big) + \mathbb{P}_\pi\big(B_\ell(\theta, \varepsilon_n) \,\big|\, X_{1:n}\big) \leqslant 1\Big\} \\
&\subseteq \bigg\{\mathbb{P}_\pi\big(B_\ell(\theta, \varepsilon_n) \,\big|\, X_{1:n}\big) \leqslant \frac{1+\delta}{2}\bigg\} \qquad (46) \\
&= \bigg\{1 - \mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) < \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\bigg\} \qquad (47) \\
&= \bigg\{\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\bigg\},
\end{aligned}$$

where (46) follows from the lower bound $\mathbb{P}_\pi\big(B_\ell\big(\widehat{\theta}_n, \varepsilon_n\big) \,\big|\, X_{1:n}\big) \geqslant \mathbb{P}_\pi\big(B_\ell(\theta, \varepsilon_n) \,\big|\, X_{1:n}\big) - \delta$, which holds by assumption (44) 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 et al. [2000] this conversion can be used the other way around: if we have a lower bound on the minimax rate of estimation, then Proposition 20 provides a lower bound on the minimax posterior concentration rate, as formalized in the following corollary. Assumption (48) below corresponds to an in-probability minimax lower bound.

Corollary 21. *Let $n \geqslant 1$. 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 $\widetilde{\theta} \mapsto \ell\big(\theta', \widetilde{\theta}\,\big)$ is continuous for all $\theta' \in \Theta$. Assume also that for some absolute constant $c < 1$, we have*

$$\inf_{\widehat{\theta}_n \text{ est.}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\Big(\ell\big(\widehat{\theta}_n, \theta\big) \geqslant 2\varepsilon_n\Big) \geqslant 1 - c\,, \tag{48}$$

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

$$\inf_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \leqslant \frac{1+c}{2} < 1\,. \tag{49}$$

Proof: Let $\delta > 0$ be a parameter that we will later take arbitrarily small. Fix any estimator $\widehat{\theta}_n$ satisfying (44) for the prior π', i.e., 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 20 used for all $\theta \in \Theta$ entails that

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta\left(\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right) \geqslant \sup_{\theta \in \Theta} \mathbb{P}_\theta\Big(\ell\big(\widehat{\theta}_n, \theta\big) \geqslant 2\varepsilon_n\Big) \geqslant 1 - c\,,$$

where the last inequality follows from the assumption (48). Now we use Markov's inequality to upper bound the left-hand side above and obtain

$$\frac{2}{1-\delta} \sup_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant \sup_{\theta \in \Theta} \mathbb{P}_\theta\left(\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right) \geqslant 1 - c\,.$$

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

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

Rearranging terms concludes the proof of (49). 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 (44) with π' in a mesurable way as soon as Θ is a separable topological space and

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

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 $(\widetilde{\theta}_i)_{i \geqslant 1}$ 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,

$$\begin{aligned}
\liminf_{i \to +\infty} \mathbb{P}_{\pi'}\Big(\theta' : \ell\big(\theta', \widetilde{\theta}_i\,\big) < \varepsilon_n \,\Big|\, x_{1:n}\Big) &= \liminf_{i \to +\infty} \mathbb{E}_{\pi'}\Big[\mathbb{1}_{\{\ell(\theta', \widetilde{\theta}_i\,) < \varepsilon_n\}} \,\Big|\, x_{1:n}\Big] \\
&\geqslant \mathbb{E}_{\pi'}\Bigg[\underbrace{\liminf_{i \to +\infty} \mathbb{1}_{\{\ell(\theta', \widetilde{\theta}_i\,) < \varepsilon_n\}}}_{= 1 \text{ if } \ell(\theta', \theta^\star) < \varepsilon_n} \,\Bigg|\, x_{1:n}\Bigg] \\
&\geqslant \mathbb{P}_{\pi'}\Big(\theta' : \ell\big(\theta', \theta^\star\,\big) < \varepsilon_n \,\Big|\, x_{1:n}\Big)\,,
\end{aligned} \tag{50}$$

where in (50) we identify that the lim inf equals 1 as soon as $\ell(\theta', \theta^\star) < \varepsilon_n$ by continuity of $\widetilde{\theta} \mapsto \ell\big(\theta', \widetilde{\theta}\big)$ at $\widetilde{\theta} = \theta^\star$. □

C. Proof of Birgé's lemma (Theorem 15)

Theorem 15 is actually a slightly simplified version of the main result by Birgé [2005] (his Corollary 1). Its proof below follows the methodology described in Section 4. In Appendix E, we also state, discuss, and prove a previous (looser) simplification by Massart [2007] and the original result [Birgé, 2005, Corollary 1].

Proof (of Theorem 15): We denote by $h : p \in [0,1] \mapsto -\big(p\ln(p) + (1-p)\ln(1-p)\big)$ the binary entropy function. The existence of c_N follows from the fact that $c \in (0,1) \mapsto h(c)/c + \ln(1-c)$ is continuous and decreasing, as the sum of two such functions; its respective limits are $+\infty$ and $-\infty$ at 0 and 1.

Reduction (4) with $\mathbb{Q}_i = \mathbb{P}_1$ for all $i \geqslant 2$ indicates that $\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) \leqslant \overline{K}$ where

$$\widetilde{p} \overset{\text{def}}{=} \frac{1}{N-1}\sum_{i=2}^{N} \mathbb{P}_i(A_i)\,, \qquad \widetilde{q} \overset{\text{def}}{=} \frac{1}{N-1}\sum_{i=2}^{N} \mathbb{P}_1(A_i) = \frac{1-\mathbb{P}_1(A_1)}{N-1}\,, \qquad \overline{K} = \frac{1}{N-1}\sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1)\,;$$

note that we used the assumption of a partition to get the alternative definition of the $\widetilde{q}$ quantity. We use the following lower bound on kl, which follows from calculations similar to the ones performed in (6), using that $c_N \geqslant 1/2$ and that the binary entropy $h : p \mapsto -\big(p\ln(p) + (1-p)\ln(1-p)\big)$ is decreasing on [1/2, 1]: for $p \geqslant c_N$,

$$\mathrm{kl}(p,q) \geqslant p\ln\left(\frac{1}{q}\right) - h(c_N) \geqslant p\ln\left(\frac{1}{q}\right) - p\,\frac{h(c_N)}{c_N}\,,$$

where $\ln(1/q) - h(c_N)/c_N > 0$ for $q < \exp\big(-h(c_N)/c_N\big)$. Hence,

$$\forall p \in [0,1],\ \ \forall\, q \in \Big(0,\, \exp\big(-h(c_N)/c_N\big)\Big), \qquad p \leqslant \max\left\{c_N,\ \frac{\mathrm{kl}(p,q)}{\ln(1/q) - h(c_N)/c_N}\right\}. \tag{51}$$

Now, we set $a = \min\limits_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i)$ and may assume $a \geqslant c_N$ (otherwise, the stated bound is obtained).

We have, by the very definition of a as a minimum and by the definition (37) of c_N,

$$a \leqslant \widetilde{p} \qquad \text{and} \qquad \widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant \frac{1-c_N}{N-1} = \frac{1}{N}\exp\left(-\frac{h(c_N)}{c_N}\right). \tag{52}$$

Note that, if $\widetilde{q} = 0$ then $\overline{K} \geqslant \mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) = +\infty$ (since $\widetilde{p} \geqslant a \geqslant c_N > 0$) so that the desired bound holds trivially. We may therefore assume that $\widetilde{q} > 0$ and combine $\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) \leqslant \overline{K}$ with (51) to get

$$a \leqslant \widetilde{p} \leqslant \max\left\{c_N,\ \frac{\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big)}{\ln\big(1/\widetilde{q}\big) - h(c_N)/c_N}\right\} \leqslant \max\left\{c_N,\ \frac{\overline{K}}{\ln(N)}\right\},$$

where, for the last inequality, we used the upper bound on $\widetilde{q}$ in (52). □

D. 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 non-open 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 $n = 1$, i.e., when C is an interval; it was dropped, e.g., 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 22 (Jensen's inequality for general convex sets; Ferguson, 1967)**.** *Let $C \subseteq \mathbb{R}^n$ be any non-empty 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 $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$ such that $\mu(C) = 1$ and $\int \|x\| \,\mathrm{d}\mu(x) < +\infty$, we have*

$$\int x \,\mathrm{d}\mu(x) \in C \qquad \text{and} \qquad \varphi\bigg(\int x \,\mathrm{d}\mu(x)\bigg) \leqslant \int_C \varphi(x) \,\mathrm{d}\mu(x)\,, \tag{53}$$

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

Our contribution is the following natural extension.

Lemma 23. *The result of Lemma 22 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 non-empty 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 $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$, such that $\mathbb{P}(X \in C) = 1$. Then

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

where $\mathbb{E}\big[\varphi(X)\big]$ 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: $\varphi(x) \geqslant a^{\mathrm{T}} x + b$ for all $x \in C$, where $(a, b) \in \mathbb{R}^2$, from which it follows that $\varphi_-(x) \leqslant \|a\|\|x\| + \|b\|$ for all $x \in C$ and thus

$$\int_C \varphi_-(x) \,\mathrm{d}\mu(x) \leqslant \int_C \Big(\|a\|\|x\| + \|b\|\Big) \,\mathrm{d}\mu(x) = \|a\| \int_C \|x\| \,\mathrm{d}\mu(x) + \|b\| < +\infty\,.$$

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 non-empty convex set, so that its relative interior R is also non-empty (see Rockafellar, 1972, Theorem 6.2); we fix $x_0 \in R$. But, by Rockafellar [1972, Theorem 23.4], the function φ admits a subgradient at x_0, that is, there exists $a \in \mathbb{R}^n$ such that $\varphi(x) \geqslant \varphi(x_0) + a^T(x - x_0)$ for all $x \in C$. This concludes the first part of this proof.

In the second part, we show the inequality (53) 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 22 with the non-empty 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 $\mu(\widetilde{C}) = 1$:

$$\int x \,\mathrm{d}\mu(x) \in \widetilde{C} \qquad \text{and} \qquad \widetilde{\varphi}\bigg(\int x \,\mathrm{d}\mu(x)\bigg) \leqslant \int_{\widetilde{C}} \widetilde{\varphi}(x) \,\mathrm{d}\mu(x)\,.$$

Using the facts that $\widetilde{\varphi}(x) = \varphi(x)$ for all $x \in \widetilde{C}$ and that $\mu\big(C \setminus \widetilde{C}\big) = 1 - 1 = 0$ entails (53). □

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

Lemma 24 (A general conditional Jensen's inequality)**.** *Let $C \subseteq \mathbb{R}^n$ be any non-empty 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 $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$, 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,*

$$\mathbb{E}[X \,|\, \mathcal{G}] \in C \qquad \textit{and} \qquad \varphi\big(\mathbb{E}[X \,|\, \mathcal{G}]\big) \leqslant \mathbb{E}\big[\varphi(X) \,|\, \mathcal{G}\big]\,,$$

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

Proof: The proof follows directly from the unconditional Jensen's inequality (Lemma 23 above) and from the existence of regular conditional distributions. More precisely, by Durrett [2010, Theorems 2.1.15 and 5.1.9] applied to the case where $(S, \mathcal{S}) = \big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$, 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 $B \in \mathcal{B}(\mathbb{R}^n)$, $\omega \in \Omega \mapsto K(\omega, B)$ is $\mathcal{G}$–measurable and $\mathbb{P}(X \in B \,\big|\, \mathcal{G}) = K(\,\cdot\,, B)$ $\mathbb{P}$–a.s.;
(P2) for $\mathbb{P}$–almost all $\omega \in \Omega$, the mapping $B \mapsto K(\omega, B)$ is a probability measure over $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$.
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,

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

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

We apply Lemma 23 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,

$$\int x\, K(\omega, \mathrm{d}x) \in C \qquad \text{and} \qquad \varphi\Big(\int x\, K(\omega, \mathrm{d}x)\Big) \leqslant \int_C \varphi(x)\, K(\omega, \mathrm{d}x)\,,$$

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 23), and $\varphi_+(X)$, which is nonnegative, we thus have proved that $\mathbb{P}$–a.s.,

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

which concludes the proof. □

E. Extended version of this article

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

Appendix F: Proofs of the data-processing inequality (Lemma 1) and of the joint convexity of Div_f (Corollary 3)

Appendix G: 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 H: Two other statements of Birgé's lemma (the original one and a simplification of it)

Supplementary material for the article

"Fano's inequality for random variables"
by Gerchinovitz, Ménard, Stoltz

F. Proofs of the data-processing inequality (Lemma 1) and of the joint convexity of Div_f (Corollary 3)

As indicated in the main body of the article, the proof of Lemma 1 is extracted from Ali and Silvey [1966, Section 4.2], see also Pardo [2006, Proposition 1.2]. Note that it can be refined: Gray [2011, Lemmas 7.5 and 7.6] establishes (54) below and then derives some (stronger) data-processing equality (not inequality).

Proof (of Lemma 1, data-processing inequality): We recall that $\mathbb{E}_{\mathbb{Q}}$ denotes the expectation with respect to a measure $\mathbb{Q}$. Let X be a random variable from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$. We write the Lebesgue decomposition (2) of $\mathbb{P}$ with respect to $\mathbb{Q}$.

We first show that $(\mathbb{P}_{\mathrm{ac}})^X \ll \mathbb{Q}^X$ and that the Radon-Nikodym derivative of $(\mathbb{P}_{\mathrm{ac}})^X$ with respect to $\mathbb{Q}^X$ equals

$$\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X} = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\,\middle|X = \cdot\right] \stackrel{\mathrm{def}}{=} \gamma\,; \tag{54}$$

i.e., γ is any measurable function such that $\mathbb{Q}$–almost surely, $\mathbb{E}_{\mathbb{Q}}\big[(\,\mathrm{d}\mathbb{P}_{\mathrm{ac}}/\,\mathrm{d}\mathbb{Q})\,\big|X\big] = \gamma(X)$. Indeed, using that $\mathbb{P}_{\mathrm{ac}} \ll \mathbb{Q}$, we have, for all $A \in \mathcal{F}'$,

$$\begin{aligned}(\mathbb{P}_{\mathrm{ac}})^X(A) = \mathbb{P}_{\mathrm{ac}}(X \in A) &= \int_\Omega \mathbb{1}_A(X)\,\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\,\mathrm{d}\mathbb{Q} = \int_\Omega \mathbb{1}_A(X)\;\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\,\middle|X\right]\mathrm{d}\mathbb{Q} \qquad (55)\\ &= \int_\Omega \mathbb{1}_A(X)\,\gamma(X)\,\mathrm{d}\mathbb{Q} = \int_{\Omega'} \mathbb{1}_A\,\gamma\,\mathrm{d}\mathbb{Q}^X\,,\end{aligned}$$

where the last equality in (55) follows by the tower rule.

Second, by unicity of the Lebesgue decomposition, the decomposition of $\mathbb{P}^X$ with respect to $\mathbb{Q}^X$ is therefore given by

$$\mathbb{P}^X = \big(\mathbb{P}^X\big)_{\mathrm{ac}} + \big(\mathbb{P}^X\big)_{\mathrm{sing}} \qquad \text{where} \qquad \big(\mathbb{P}^X\big)_{\mathrm{ac}} = (\mathbb{P}_{\mathrm{ac}})^X + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}$$
$$\text{and} \qquad \big(\mathbb{P}^X\big)_{\mathrm{sing}} = \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}\,.$$

The inner $_{\mathrm{ac}}$ and $_{\mathrm{sing}}$ symbols refer to the pair $\mathbb{P}, \mathbb{Q}$ while the outer $_{\mathrm{ac}}$ and $_{\mathrm{sing}}$ symbols refer to $\mathbb{P}^X, \mathbb{Q}^X$.

We use this decomposition for the first equality below and integrate (1) for the first inequality below:

$$\begin{aligned}\mathrm{Div}_f\big(\mathbb{P}^X, \mathbb{Q}^X\big) &= \int_{\Omega'} f\left(\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X} + \frac{\mathrm{d}\big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}^X}\right)\mathrm{d}\mathbb{Q}^X + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}(\Omega')\,M_f\\ &\leqslant \int_{\Omega'} f\left(\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X}\right)\mathrm{d}\mathbb{Q}^X + \Big(\big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}(\Omega') + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}(\Omega')\Big)\,M_f\\ &= \int_{\Omega'} f(\gamma)\,\mathrm{d}\mathbb{Q}^X + (\mathbb{P}_{\mathrm{sing}})^X(\Omega')\,M_f\\ &= \int_\Omega f\big(\gamma(X)\big)\,\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\,M_f\\ &= \int_\Omega f\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\,\middle|X\right]\right)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\,M_f\\ &\leqslant \int_\Omega \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\,\middle|X\right]\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\,M_f \qquad (56)\\ &= \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\,M_f = \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})\,,\end{aligned}$$

where the inequality in (56) is a consequence of the conditional Jensen's inequality in its general form stated in Appendix D, Lemma 24, with $\varphi = f$ and $C = [0, +\infty)$, and where the final equality follows from the tower rule. □

The joint convexity of Div_f (Corollary 3) may be proved directly, in two steps. First, the log-sum inequality is generalized into the fact that the mapping $(p, q) \in [0, +\infty)^2 \mapsto q\, f(p/q)$ is jointly convex. Second, a common dominating measure like $\mu = \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{Q}_1 + \mathbb{Q}_2$ is introduced, Radon-Nikodym derivatives p_j and q_j are introduced for the $\mathbb{P}_j$ and $\mathbb{Q}_j$ with respect to μ, and the generalized log-sum inequality is applied pointwise.

We suggest to see instead Corollary 3 as an elementary consequence of the data-processing inequality.

Proof (of Corollary 3, joint convexity of Div_f): We augment the probability space Ω into $\Omega' = \{1, 2\} \times \Omega$, which we equip with the σ–algebra $\mathcal{F}'$ generated by the events $A \times B$, where $A \in \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\}$ and $B \in \mathcal{F}$. We define the random pair (J, X) on this space by the projections

$$X : (j, \omega) \in \{1, 2\} \times \Omega \longmapsto \omega \qquad \text{and} \qquad J : (j, \omega) \in \{1, 2\} \times \Omega \longmapsto j\,,$$

and denote by $\mathbb{P}$ the joint distribution of the random pair (J, X) such that $J \sim 1 + \mathrm{Ber}(\lambda)$ and $X|J \sim \mathbb{P}_J$. More formally, $\mathbb{P}$ is the unique probability distribution on $(\Omega', \mathcal{F}')$ such that, for all $(j, B) \in \{1, 2\} \times \mathcal{F}$,

$$\mathbb{P}\big(\{j\} \times B\big) = \big((1 - \lambda)\mathbb{1}_{\{j=1\}} + \lambda \mathbb{1}_{\{j=2\}}\big)\, \mathbb{P}_j(B)\,.$$

Similarly we define the joint probability distribution $\mathbb{Q}$ on $(\Omega', \mathcal{F}')$ using the conditional distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ instead of $\mathbb{P}_1$ and $\mathbb{P}_2$.

The corollary follows directly from the data-processing inequality $\mathrm{Div}_f\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})$, as the laws of X under $\mathbb{P}$ and $\mathbb{Q}$ are respectively given by

$$\mathbb{P}^X = (1 - \lambda)\mathbb{P}_1 + \lambda \mathbb{P}_2 \qquad \text{and} \qquad \mathbb{Q}^X = (1 - \lambda)\mathbb{Q}_1 + \lambda \mathbb{Q}_2\,,$$

while elementary calculations show that $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) = (1 - \lambda)\, \mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\, \mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)$.

Indeed, for the latter point, we consider the Lebesgue decompositions of $\mathbb{P}_j$ with respect to $\mathbb{Q}_j$, where $j \in \{1, 2\}$:

$$\mathbb{P}_j = \mathbb{P}_{j,\mathrm{ac}} + \mathbb{P}_{j,\mathrm{sing}}\,, \qquad \text{where} \qquad \mathbb{P}_{j,\mathrm{ac}} \ll \mathbb{Q}_j \quad \text{and} \quad \mathbb{P}_{j,\mathrm{sing}} \perp \mathbb{Q}_j\,.$$

The (unique) Lebesgue decomposition of $\mathbb{P} = \mathbb{P}_{\mathrm{ac}} + \mathbb{P}_{\mathrm{sing}}$ with respect to $\mathbb{Q}$ is then given by

$$\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(j, \omega) = \mathbb{1}_{\{j=1\}}\, \frac{\mathrm{d}\mathbb{P}_{1,\mathrm{ac}}}{\mathrm{d}\mathbb{Q}_1}(\omega) + \mathbb{1}_{\{j=2\}}\, \frac{\mathrm{d}\mathbb{P}_{2,\mathrm{ac}}}{\mathrm{d}\mathbb{Q}_2}(\omega)$$

and for all $(j, B) \in \{1, 2\} \times \mathcal{F}$,

$$\mathbb{P}_{\mathrm{sing}}\big(\{j\} \times B\big) = \big((1 - \lambda)\mathbb{1}_{\{j=1\}} + \lambda \mathbb{1}_{\{j=2\}}\big)\, \mathbb{P}_{j,\mathrm{sing}}(B)\,.$$

This entails that

$$\begin{aligned}
\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) &= \int_{\{1,2\}\times\Omega} f\bigg(\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(j, \omega)\bigg)\, \mathrm{d}\mathbb{Q}(j, \omega) + \mathbb{P}_{\mathrm{sing}}\big(\{1, 2\} \times \Omega\big)\, M_f \\
&= (1 - \lambda) \int_\Omega f\bigg(\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(1, \omega)\bigg)\, \mathrm{d}\mathbb{Q}_1(\omega) + \lambda \int_\Omega f\bigg(\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(2, \omega)\bigg)\, \mathrm{d}\mathbb{Q}_2(\omega) \\
&\qquad + \big((1 - \lambda)\, \mathbb{P}_{1,\mathrm{sing}}(\Omega) + \lambda\, \mathbb{P}_{2,\mathrm{sing}}(\Omega)\big)\, M_f \\
&= (1 - \lambda)\, \mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\, \mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)\,.
\end{aligned}$$

□

G. Additional material on the Fano-type inequalities of Section 4

We first provide (Section G.1) a sharper bound than the bound (15) exhibited in Section 4.2 for the case of the Hellinger distance and which read

$$p \leqslant q + \sqrt{1 - \big(1 - h^2(p,q)/2\big)^2} = q + \sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}\,.$$

We then (Section G.2) study quantities of the form

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

that arise in the bounds of Section 4.3 in the case of partitions. We discuss what $\mathbb{Q}$ should be chosen and what bounds can be achieved.

G.1. A sharper lower bound on div_f for the Hellinger distance

We follow and slightly generalize Guntuboyina [2011, Example II.6]. As we prove below, we get the bound

$$p \leqslant q + (1-2q)\,h^2(p,q)\big(1 - h^2(p,q)/4\big) + 2\sqrt{q(1-q)}\,\big(1 - h^2(p,q)/2\big)\sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}\,. \quad (57)$$

It can be seen that this bound is a general expression of the bound stated by Guntuboyina [2011, Example II.6]. This bound is slightly tighter than (15), by construction (as we solve exactly an equation and perform no bounding) but it is much less readable. It anyway leads to similar conclusions in practice.

Proof: Assuming that $\underline{h}^2 = h^2(p,q)$ is given and fixed, we consider the equation, for the unknown $x \in [0,1]$,

$$\underline{h}^2 = 2\Big(1 - \Big(\sqrt{q}\sqrt{x} + \sqrt{1-q}\,\sqrt{1-x}\Big)\Big)\,;$$

this equation is satisfied for $x = p$, by definition of $h^2(p,q)$. Rearranging it, we get the equivalent equation

$$(1-x)(1-q) = \big(1 - \underline{h}^2/2 - \sqrt{q}\sqrt{x}\big)^2 = \big(1 - \underline{h}^2/2\big)^2 - 2\big(1 - \underline{h}^2/2\big)\sqrt{q}\,\sqrt{x} + qx\,,$$

or equivalently again,

$$x - 2\big(1 - \underline{h}^2/2\big)\sqrt{q}\,\sqrt{x} + \big(1 - \underline{h}^2/2\big)^2 - 1 + q = 0\,.$$

Solving this second-order equation for $\sqrt{x}$, we see that all solutions $\sqrt{x}$, including $\sqrt{p}$, are smaller than the largest root; in particular,

$$\sqrt{p} \leqslant \big(1 - \underline{h}^2/2\big)\sqrt{q} + \underbrace{\sqrt{\big(1 - \underline{h}^2/2\big)^2 q - \big(1 - \underline{h}^2/2\big)^2 + 1 - q}}_{=\sqrt{(1-q)\,\underline{h}^2(1-\underline{h}^2/4)}}\,.$$

Put differently,

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

which was the expression to obtain. □

G.2. Finding a good constant alternative $\mathbb{Q}$

The key term in the bounds of Section 4.3 in the case of a partition or of random variables summing up to 1 is given by

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

We however need a closed-form (or at least, a more concrete) expression of this quantity for these bounds to have a practical interest. This is the issue we tackle in this section.

Instead of simply studying quantities of the form indicated above, we consider

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

where $\alpha = (\alpha_1, \ldots, \alpha_N)$, with all $\alpha_i > 0$, denotes some convex combination.

Sometimes calculations are easy in practice for some specific $\mathbb{Q}$, as we illustrated, for instance, in Section 5.2. Otherwise, the lemma below indicates a good candidate, given by the weighted average $\overline{\mathbb{P}}_\alpha$ of the distributions $\mathbb{P}_i$.

To appreciate its performance, we denote by

$$B_f(\alpha) = \max_{j=1,\ldots,N} \mathrm{Div}_f(\delta_j, \alpha)$$

the maximal f–divergence between a Dirac mass δ_j at j and the convex combination α. This bound equals $\ln\bigl(1/\min\{\alpha_1, \ldots, \alpha_N\}\bigr)$ for a Kullback-Leibler divergence and $1/\min\{\alpha_1, \ldots, \alpha_N\} - 1$ for the χ^2–divergence.

Lemma 25. *Let $\mathbb{P}_1, \ldots, \mathbb{P}_N$ be N probability distributions over the same measurable space $(\Omega, \mathcal{F})$ and let $\alpha = (\alpha_1, \ldots, \alpha_N)$ be a convex combination made of positive weights. Then,*

$$\inf_{\mathbb{Q}} \sum_{i=1}^{N} \alpha_i \,\mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q}) \leqslant \sum_{i=1}^{N} \alpha_i \,\mathrm{Div}_f\bigl(\mathbb{P}_i, \overline{\mathbb{P}}_\alpha\bigr) \leqslant B_f(\alpha)\,,$$

where the infimun is over all probability distributions $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ and where $\overline{\mathbb{P}}_\alpha \stackrel{\text{def}}{=} \sum_{i=1}^{N} \alpha_i \,\mathbb{P}_i$.

The first inequality holds with equality in the case of the Kullback-Leibler divergence, as follows from the so-called compensation equality (see, e.g., Yang and Barron, 1999 or Guntuboyina, 2011, Example II.4): assuming with no loss of generality in this case (since $M_f = +\infty$) that $\mathbb{P}_j \ll \mathbb{Q}$ for all $j \in \{1, \ldots, N\}$, we have $\overline{\mathbb{P}}_\alpha \ll \mathbb{Q}$ and $\mathrm{d}\mathbb{P}_j/\mathrm{d}\mathbb{Q} = (\mathrm{d}\mathbb{P}_j/\mathrm{d}\overline{\mathbb{P}}_\alpha)(\mathrm{d}\overline{\mathbb{P}}_\alpha/\mathrm{d}\mathbb{Q})$, which entails

$$\sum_{i=1}^{N} \alpha_i \,\mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) = \sum_{i=1}^{N} \alpha_i \int \left(\ln \frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\overline{\mathbb{P}}_\alpha} + \ln \frac{\mathrm{d}\overline{\mathbb{P}}_\alpha}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P}_i = \left(\sum_{i=1}^{N} \alpha_i \,\mathrm{KL}\bigl(\mathbb{P}_i, \overline{\mathbb{P}}_\alpha\bigr)\right) + \mathrm{KL}\bigl(\overline{\mathbb{P}}_\alpha, \mathbb{Q}\bigr)\,,$$

where we used that $\sum_{i=1}^{N} \alpha_i \,\mathrm{d}\mathbb{P}_i = \mathrm{d}\overline{\mathbb{P}}_\alpha$. So, indeed, the considered infimum is achieved at $\mathbb{Q} = \overline{\mathbb{P}}$.

Proof: The first inequality follows from the choice $\mathbb{Q} = \overline{\mathbb{P}}_\alpha$. For the second inequality, we proceed as in Corollary 3 and consider the following probability distributions over $\{1, \ldots, N\} \times \Omega$: for all $j \in \{1, \ldots, N\}$ and all $B \in \mathcal{F}$,

$$\widetilde{\mathbb{P}}\bigl(\{j\} \times B\bigr) = \alpha_j \,\mathbb{P}_j(B) \qquad \text{and} \qquad \widetilde{\mathbb{Q}}\bigl(\{j\} \times B\bigr) = \alpha_j \,\overline{\mathbb{P}}_\alpha(B)\,.$$

Note that because $\alpha_i > 0$ for all i, we have $\mathbb{P}_j \ll \overline{\mathbb{P}}_\alpha$ for all j. Thus, $\widetilde{\mathbb{P}} \ll \widetilde{\mathbb{Q}}$, with Radon-Nikodym derivative given by

$$(j,\omega) \in \{1,\dots,N\} \times \Omega \;\longmapsto\; \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\widetilde{\mathbb{Q}}}(j,\omega) = \frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) \stackrel{\text{def}}{=} p_j(\omega)\,.$$

By uniqueness and linearity of the Radon-Nikodym derivatives, we thus have, for $\overline{\mathbb{P}}_\alpha$–almost all ω,

$$\sum_{j=1}^N \alpha_j\, p_j(\omega) = \sum_{j=1}^N \alpha_j\, \frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) = \frac{\mathrm{d}\overline{\mathbb{P}}_\alpha}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) = 1\,, \qquad \text{where} \qquad \forall k \in \{1,\dots,N\},\;\; \alpha_k p_k(\omega) \geqslant 0\,;$$

that is, $\alpha p(\omega) = \big(\alpha_j p_j(\omega)\big)_{1\leqslant j\leqslant N}$ is a probability distribution over $\{1,\dots,N\}$. (It corresponds to the conditional distribution of j given ω in the probabilistic model $j \sim \alpha$ and $\omega|j \sim \mathbb{P}_j$.)

We now compute $\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big)$ in two different ways. All manipulations below are valid because all integrals defining f–divergences exist (see the comments after the statement of Definition 2.2, as well as the first part of the proof of Lemma 23). Integrating over j first,

$$\begin{aligned}\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big) &= \int_{\{1,\dots,N\}\times\Omega} f\left(\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\widetilde{\mathbb{Q}}}(j,\omega)\right) \mathrm{d}\widetilde{\mathbb{Q}}(j,\omega) \\ &= \sum_{j=1}^N \alpha_j \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega)\right) \mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) = \sum_{j=1}^N \alpha_j\, \mathrm{Div}_f\big(\mathbb{P}_j,\overline{\mathbb{P}}_\alpha\big)\;.\end{aligned}$$

On the other hand, integrating over ω first,

$$\begin{aligned}\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big) &= \int_\Omega \left(\sum_{j=1}^N f\big(p_j(\omega)\big)\,\alpha_j\right) \mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) \\ &= \int_\Omega \left(\sum_{j=1}^N f\left(\frac{\alpha_j p_j(\omega)}{\alpha_j}\right)\alpha_j\right) \mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) = \int_\Omega \mathrm{Div}_f\big(\alpha p(\omega),\alpha\big)\,\mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) \leqslant B_f(\alpha)\,,\end{aligned}$$

where the last inequality follows by noting that, by joint convexity of Div_f (see Corollary 3),

$$\mathrm{Div}_f\big(\alpha p(\omega),\alpha\big) \leqslant \sum_{j=1}^n \alpha_j p_j(\omega)\,\mathrm{Div}_f\big(\delta_j,\alpha\big) \leqslant B_f(\alpha)\,.$$

Comparing the two obtained expressions for $\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big)$ concludes the proof. □

H. Two other statements of Birgé's lemma

The original result by Birgé [2005, Corollary 1] reads, with the notation of Theorem 15:

$$\min_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{ d_N,\ \frac{\overline{K}}{\ln(N)} \right\}, \tag{58}$$

where $(d_N)_{N\geqslant 2}$ is a decreasing sequence, defined as follows, based on functions $r_N : [0,1) \to \mathbb{R}$:

$$r_N(b) = \mathrm{kl}\left(b, \frac{1-b}{N-1}\right) - b\ln(N) \qquad \text{and} \qquad d_N = \max\left\{ b \in [0,1] :\ r_N(b) \leqslant 0 \right\}.$$

This original result was only stated for $N \geqslant 3$ but its proof indicates that it is also valid for $N = 2$.

On the other hand, the simplification by Massart [2007, Section 2.3.4] leads to

$$\min_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{ \frac{2\mathrm{e}-1}{2\mathrm{e}},\ \frac{\overline{K}}{\ln(N)} \right\}. \tag{59}$$

(The original constant was a larger $2\mathrm{e}/(2\mathrm{e}+1)$ in Massart, 2007, Section 2.3.4.)

Before proving these results, we compare them with Theorem 15. The values of the c_N of Theorem 1, of the d_N of (58) and of $(2\mathrm{e}-1)/(2\mathrm{e})$ are given by (values rounded upwards)

$$\frac{2\mathrm{e}-1}{2\mathrm{e}} \approx 0.8161$$

and

N	2	3	7	$+\infty$
c_N	0.7587	0.7127	< 0.67	0.63987
d_N	0.7428	0.7009	$< 2/3$	0.63987

The c_N and d_N are thus extremely close. While the c_N are slightly larger than the d_N (with, however, the same limit), they are easier to compute in practice. (See the closed-form expression for r_N below.) Also, the proof of Theorem 15 is simpler than the proof of Birgé [2005, Corollary 1]: they rely on the same proof scheme but the former involves fewer calculations than the latter. Indeed, let us now prove again Birgé [2005, Corollary 1].

Proof of (58)**.** We use the notation of the proof of Theorem 15 and its beginning. We can assume with no loss of generality that $a \geqslant d_N$, and we also have $d_N \geqslant 1/N$ as $r_N(1/N) = -\ln(N)/N \leqslant 0$. Therefore, $a \geqslant 1/N$ and using the definition of a as a minimum,

$$\widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant a \leqslant \widetilde{p}; \tag{60}$$

therefore,

$$\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) \geqslant \mathrm{kl}\big(a, \widetilde{q}\big) \geqslant \mathrm{kl}\left(a, \frac{1-a}{N-1}\right),$$

since by convexity, $p \mapsto \mathrm{kl}(p,q)$ is increasing on $[q,1]$ and $q \mapsto \mathrm{kl}(p,q)$ is decreasing on $[0,p]$. Combining this with $\overline{K} \geqslant \mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big)$, one has proved

$$\overline{K} \geqslant \mathrm{kl}\left(a, \frac{1-a}{N-1}\right) = a\ln(N) + r_N(a),$$

from which the bound (58) follows by definition of d_N. To prove that the sequence (d_N) is decreasing and to get a numerical expression via dichotomy follow from studying the variations of $r_N(b)$ in b

and N; for the latter; one should show, in particular, that $r_N(b)$ is positive before d_N and negative after d_N. This last analytical part of the proof is tedious, as

$$\begin{aligned} r_N(a) &= a\ln\left(\frac{a}{1-a}\right) + (1-a)\ln\left(\frac{1-a}{1-\dfrac{1-a}{N-1}}\right) + \big(a\ln(N-1) - a\ln(N)\big) \\ &= \big(a\ln(a) + (1-2a)\ln(1-a)\big) + a\ln\left(\frac{N-1}{N}\right) + (1-a)\ln\left(\frac{N-1}{N-2+a}\right), \end{aligned}$$

and we could overcome these heavy calculations in our proof of Theorem 15.

Proof of (59)**.** For $p \geqslant \ln(2)$ and all $q \in [0,1]$,

$$\mathrm{kl}(p,q) \geqslant p\ln\left(\frac{1}{q}\right) - \ln(2) \geqslant p\ln\left(\frac{1}{q}\right) - p = p\ln\left(\frac{1}{\mathrm{e}q}\right). \tag{61}$$

Equation (52) is adapted as

$$a \leqslant \widetilde{p} \qquad \text{and} \qquad \widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant \frac{2(1-a)}{N} \leqslant \frac{1}{\mathrm{e}N}$$

where we used respectively, for the last two inequalities, that $1/(N-1) \leqslant 2/N$ for $N \geqslant 2$ and that, with no loss of generality, $a \geqslant (2\mathrm{e}-1)/(2\mathrm{e})$. In particular, $\mathrm{e}\widetilde{q} \leqslant 1/N$. Combining this with $\overline{K} \geqslant \mathrm{kl}\big(\widetilde{p},\widetilde{q}\big)$ and (61), we have proved

$$\overline{K} \geqslant \widetilde{p}\ln\left(\frac{1}{\mathrm{e}\widetilde{q}}\right) \geqslant a\ln(N),$$

which concludes the proof.