ISSN 1066-5307, Mathematical Methods of Statistics, 2016, Vol. 25, No. 4, pp. 304–312. © Allerton Press, Inc., 2016.

The Minimum Increment of f-Divergences Given Total Variation Distances

A. A. Gushchin[1]*

[1] *Steklov Mathematical Institute of RAS, Moscow, Russia*

Received November 4, 2016

Abstract—Let $(\mathsf{P}_i, \mathsf{Q}_i)$, $i = 0, 1$, be two pairs of probability measures defined on measurable spaces $(\Omega_i, \mathscr{F}_i)$ respectively. Assume that the pair $(\mathsf{P}_1, \mathsf{Q}_1)$ is more informative than $(\mathsf{P}_0, \mathsf{Q}_0)$ for testing problems. This amounts to say that $\mathsf{I}_\mathsf{f}(\mathsf{P}_1, \mathsf{Q}_1) \geq \mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0)$, where $\mathsf{I}_\mathsf{f}(\cdot, \cdot)$ is an arbitrary f-divergence. We find a precise lower bound for the increment of f-divergences $\mathsf{I}_\mathsf{f}(\mathsf{P}_1, \mathsf{Q}_1) - \mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0)$ provided that the total variation distances $\|\mathsf{Q}_1 - \mathsf{P}_1\|$ and $\|\mathsf{Q}_0 - \mathsf{P}_0\|$ are given. This optimization problem can be reduced to the case where P_1 and Q_1 are defined on the space consisting of four points, and P_0 and Q_0 are obtained from P_1 and Q_1 respectively by merging two of these four points. The result includes the well-known lower and upper bounds for $\mathsf{I}_\mathsf{f}(\mathsf{P}, \mathsf{Q})$ given $\|\mathsf{Q} - \mathsf{P}\|$.

Keywords: comparison of experiments, f-divergence, power function, total variation distance.

2000 Mathematics Subject Classification: 94A17, 62B10, 62B15, 26D15.

DOI: 10.3103/S1066530716040049

1. INTRODUCTION AND MAIN RESULTS

A dichotomy is an (ordered) pair (P, Q) of probability measures on a common measurable space $(\Omega, \mathscr{F})$. We shall also say that $\mathbb{E} = (\Omega, \mathscr{F}, \mathsf{P}, \mathsf{Q})$ is a (binary statistical) model or experiment. A convenient way to study dichotomies is via testing problems. Namely, denote by $\Phi(\mathbb{E})$ the set of all test functions φ in $\mathbb{E}$, i.e., measurable mappings from $(\Omega, \mathscr{F})$ to $[0, 1]$. Any $\varphi(\omega)$ may be interpreted as the probability to accept the alternative 'Q' and to reject the null hypothesis 'P' if ω is observed. Let

$$\mathfrak{N}(\mathbb{E}) := \left\{ \left(\int \varphi \, d\mathsf{P}, \int \varphi \, d\mathsf{Q} \right) : \varphi \in \Phi(\mathbb{E}) \right\}$$

be the set of values of the power function. A model $\mathbb{E}_1$ is said to be *at least as informative as* $\mathbb{E}_0$, denoted $\mathbb{E}_1 \succeq \mathbb{E}_0$, if $\mathfrak{N}(\mathbb{E}_1) \supseteq \mathfrak{N}(\mathbb{E}_0)$; $\mathbb{E}_1$ and $\mathbb{E}_0$ are said to be *equivalent* ($\mathbb{E}_1 \sim \mathbb{E}_0$) if both $\mathbb{E}_1 \succeq \mathbb{E}_0$ and $\mathbb{E}_0 \succeq \mathbb{E}_1$.

Here is an instructive example. Let a model $\mathbb{E}_1 = (\Omega_1, \mathscr{F}_1, \mathsf{P}_1, \mathsf{Q}_1)$, a measurable space $(\Omega_0, \mathscr{F}_0)$, and a Markov kernel $\mathsf{K} \colon \Omega_1 \times \mathscr{F}_0 \to [0, 1]$ from $(\Omega_1, \mathscr{F}_1)$ to $(\Omega_0, \mathscr{F}_0)$ be given. Put

$$\mathsf{K}\mathsf{P}_1(B) := \int_{\Omega_1} \mathsf{K}(\omega, B) \, \mathsf{P}_1(d\omega), \quad B \in \mathscr{F}_0,$$

and define $\mathsf{K}\mathsf{Q}_1$ similarly. Put $\mathsf{P}_0 = \mathsf{K}\mathsf{P}_1$, $\mathsf{Q}_0 = \mathsf{K}\mathsf{Q}_1$, $\mathbb{E}_0 = (\Omega_0, \mathscr{F}_0, \mathsf{P}_0, \mathsf{Q}_0)$. Then, trivially, $\mathbb{E}_1 \succeq \mathbb{E}_0$. The model $\mathbb{E}_0$ is sometimes referred to as indirect observations in contrast to the model $\mathbb{E}_1$ of direct observations. A special case arises if a kernel K is defined by

$$\mathsf{K}(\omega, B) = \begin{cases} 1 & \text{if } T(\omega) \in B, \\ 0 & \text{otherwise,} \end{cases}$$

where T is a measurable mapping (statistic) from $(\Omega_1, \mathscr{F}_1)$ to $(\Omega_0, \mathscr{F}_0)$. Then P_0 and Q_0 are the images $\mathsf{P}_1 \circ T^{-1}$ and $\mathsf{Q}_1 \circ T^{-1}$ of P_1 and Q_1 respectively under T, i.e., the data are reduced by the statistic T

* E-mail: gushchin@mi.ras.ru

(which is not necessarily sufficient). An even more special case is when $\Omega_0 = \Omega_1$, $\mathscr{F}_0$ is a sub-σ-field of $\mathscr{F}_1$, and T is the identity mapping; then P_0 and Q_0 are the restrictions of P_1 and Q_1 respectively onto $\mathscr{F}_0$ (data grouping). In this situation one can interpret also the model $\mathbb{E}_1$ as obtained from $\mathbb{E}_0$ by making additional observations. It is useful to the reader to keep in mind the fact that, according to [2], if $\mathbb{E}_1$ and $\mathbb{E}_0$ are arbitrary experiments such that $\mathbb{E}_1 \succeq \mathbb{E}_0$, then one can find experiments $\mathbb{E}'_1 \sim \mathbb{E}_1$ and $\mathbb{E}'_0 \sim \mathbb{E}_0$ such that $\mathbb{E}'_0$ is obtained from $\mathbb{E}'_1$ by passing to a sub-σ-field as it has been just described.

If $\mathbb{E}_1 \succeq \mathbb{E}_0$, a natural question is how to quantify the loss of information when passing from $\mathbb{E}_1$ to $\mathbb{E}_0$, or the additional information contained in $\mathbb{E}_1$ compared to $\mathbb{E}_0$. A natural way of doing this is to use the difference $\mathsf{I}(\mathsf{P}_1, \mathsf{Q}_1) - \mathsf{I}(\mathsf{P}_0, \mathsf{Q}_0)$, where I is a functional defined on dichotomies which is increasing with respect to the partial order $\succeq$. In particular, the so-called f-*divergence* I_f is an appropriate choice for I, see the definition (1.2) and the data processing inequality (1.7) below. On the other hand, there are the notions of *deficiency* $\delta(\mathbb{E}_0, \mathbb{E}_1)$ and *insufficiency* $\eta(\mathbb{E}_0, \mathbb{E}_1)$ introduced by Le Cam [11, 12] for statistical experiments with a general parameter space. We refer to [4], [20, Chapter 3], [21], [14], and the references therein for connections between these approaches and related results.

In this paper our purpose is to find the lower bound for the increment

$$\mathsf{I}_\mathsf{f}(\mathsf{P}_1, \mathsf{Q}_1) - \mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0) \tag{1.1}$$

of f-divergences for an arbitrary f in terms of the increment

$$\|\mathsf{P}_1 - \mathsf{Q}_1\| - \|\mathsf{P}_0 - \mathsf{Q}_0\|$$

of the total variation distances (which are also f-divergences with $\mathsf{f}(x) = |x - 1|$). More precisely, we find the lower bound for the increment (1.1) provided that $\mathbb{E}_1 \succeq \mathbb{E}_0$ and the values of $\|\mathsf{P}_1 - \mathsf{Q}_1\|$ and $\|\mathsf{P}_0 - \mathsf{Q}_0\|$ are given. Our result includes the well-known lower and upper bounds for the f-divergence given the total variation distance. The special cases of these bounds corresponding to the Kullback–Leibler divergence and the Hellinger distance are widely used in probability, statistics, and information theory. Let us also mention that the problem of finding the upper bound for the increment (1.1) under the same constraints is trivial.

Before stating the main result let us give the definition of the f-divergence. We refer to [15] for the unexplained facts.

Let $\mathsf{f}\colon (0, +\infty) \to \mathbb{R}$ be a convex function. The *-conjugate f^* of f is defined by $\mathsf{f}^*(x) = x\mathsf{f}(1/x)$, $x > 0$, then f^* is also finite and convex on $(0, +\infty)$. Put $\mathsf{f}(0) := \lim_{x \downarrow 0} \mathsf{f}(x) \in (-\infty, +\infty]$; similarly, $\mathsf{f}^*(0) = \lim_{x \uparrow \infty} \mathsf{f}(x)/x \in (-\infty, +\infty]$. It is convenient to introduce the function $\mathsf{F}\colon [0, +\infty) \times [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ constructed from f by

$$\mathsf{F}(x, y) = \begin{cases} y\,\mathsf{f}(x/y) & \text{if } x > 0,\ y > 0, \\ y\,\mathsf{f}(0) & \text{if } x = 0,\ y > 0, \\ x\,\mathsf{f}^*(0) & \text{if } x > 0,\ y = 0, \\ 0 & \text{if } x = y = 0. \end{cases}$$

The function F is convex, lower semicontinuous, and positively homogeneous.

Let (P, Q) be a dichotomy. Denote by p and q the Radon–Nikodým derivatives of P and Q respectively with respect to some σ-finite measure μ dominating P and Q. The f-divergence $\mathsf{I}_\mathsf{f}(\mathsf{P}, \mathsf{Q})$ of P with respect to Q is defined by

$$\mathsf{I}_\mathsf{f}(\mathsf{P}, \mathsf{Q}) = \int \mathsf{F}(p, q)\, d\mu. \tag{1.2}$$

The definition is correct and does not depend on the choice of μ. It was introduced by Csiszár [3] and independently by Ali and Silvey [1]. Special cases of f lead to Kullback–Leibler divergence, total variation distance, squared Hellinger distance and other widely used divergences between probability measures. An equivalent definition of the f-divergence in terms of the Fenchel transform g of f was suggested in [10]. If $d\mathsf{Q}/d\mathsf{P}$ is the Radon–Nikodým derivative of the P-absolutely continuous part of Q with respect to P, then $\mathsf{P}(p = 0) = 0$, $d\mathsf{Q}/d\mathsf{P} = q/p$ P-a.s., and we can rewrite (1.2) as

$$\mathsf{I}_\mathsf{f}(\mathsf{P}, \mathsf{Q}) = \int_{\{p>0, q>0\}} q\,\mathsf{f}(p/q)\, d\mu + \mathsf{f}^*(0) \int_{\{p>0, q=0\}} p\, d\mu + f(0) \int_{\{p=0\}} q\, d\mu$$

$$= \int_{\{p>0\}} p\,\mathsf{f}^*(q/p)\,d\mu + f(0)\Big(1 - \int_{\{p>0\}} p(q/p)\,d\mu\Big)$$
$$= \int \mathsf{f}^*(d\mathsf{Q}/d\mathsf{P})\,d\mathsf{P} + f(0)\Big(1 - \int (d\mathsf{Q}/d\mathsf{P})\,d\mathsf{P}\Big). \tag{1.3}$$

The following is always true:

$$\mathsf{f}(1) \le \mathsf{I}_\mathsf{f}(\mathsf{P},\mathsf{Q}) \le \mathsf{f}(0) + \mathsf{f}^*(0),$$
$$\mathsf{I}_\mathsf{f}(\mathsf{P},\mathsf{Q}) = \mathsf{I}_{\mathsf{f}^*}(\mathsf{Q},\mathsf{P}), \tag{1.4}$$

if b and c are numbers and $\mathsf{f}_1(x) = f(x) + c + b(x-1)$, then

$$\mathsf{I}_{\mathsf{f}_1}(\mathsf{P},\mathsf{Q}) = \mathsf{I}_\mathsf{f}(\mathsf{P},\mathsf{Q}) + c. \tag{1.5}$$

In view of (1.5), we shall always assume without loss of generality that

$$\mathsf{f}(1) = 0. \tag{1.6}$$

Let $\mathbb{E}_1 = (\Omega_1, \mathscr{F}_1, \mathsf{P}_1, \mathsf{Q}_1)$ and $\mathbb{E}_0 = (\Omega_0, \mathscr{F}_0, \mathsf{P}_0, \mathsf{Q}_0)$ be two experiments. Assume that $\mathbb{E}_1$ is at least as informative as $\mathbb{E}_0$. One of the most fundamental properties of the f-divergence is the data processing inequality

$$\mathsf{I}_\mathsf{f}(\mathsf{P}_1,\mathsf{Q}_1) \ge \mathsf{I}_\mathsf{f}(\mathsf{P}_0,\mathsf{Q}_0). \tag{1.7}$$

For the sake of completeness, we give a short proof in Section 2.

Our goal is to find a quantitative version of this inequality. Namely, we find a (sharp) lower bound for the difference

$$\mathsf{I}_\mathsf{f}(\mathsf{P}_1,\mathsf{Q}_1) - \mathsf{I}_\mathsf{f}(\mathsf{P}_0,\mathsf{Q}_0) \tag{1.8}$$

provided that

$$\mathbb{E}_1 \succeq \mathbb{E}_0, \tag{1.9}$$

$$\mathsf{I}_\mathsf{f}(\mathsf{P}_0,\mathsf{Q}_0) < \infty, \tag{1.10}$$

and the total variation distances $\|\mathsf{Q}_1 - \mathsf{P}_1\|$ and $\|\mathsf{Q}_0 - \mathsf{P}_0\|$ take some fixed values $2v_1$ and $2v_0$ respectively, i.e.,

$$v_1 = \sup_{B\in\mathscr{F}_1} |\mathsf{Q}_1(B) - \mathsf{P}_1(B)|, \qquad v_0 = \sup_{B\in\mathscr{F}_0} |\mathsf{Q}_0(B) - \mathsf{P}_0(B)|, \tag{1.11}$$

$0 \le v_0 \le v_1 \le 1$.

Define, for $0 \le v_0 \le v_1 \le 1$, $0 \le a \le 1 - v_1$,

$$d_\mathsf{f}(a, v_1, v_0) := \mathsf{F}(1 - a - v_0, 1 - a - v_1) + \mathsf{F}(a, a + v_1 - v_0),$$
$$L_\mathsf{f}(v_1, v_0) := \inf_{0\le a\le 1-v_1} d_\mathsf{f}(a, v_1, v_0).$$

Proposition 1.1. *The function $L_\mathsf{f}(v_1, v_0)$, $0 \le v_0 \le v_1 \le 1$, is convex, lower semicontinuous, nonnegative, and $L_\mathsf{f}(v_1, v_0) = 0$ if $v_1 = v_0$. Moreover, $L_\mathsf{f}(v_1, v_0)$ is increasing in v_1 and decreasing in v_0, and $L_\mathsf{f}(v_1, v_0) \le L_\mathsf{f}(v_1 + h, v_0 + h)$ for $0 < h \le 1 - v_1$.*

The following theorem is the main result of the paper.

Theorem 1.1. *We have*

$$\inf\big(\mathsf{I}_\mathsf{f}(\mathsf{P}_1,\mathsf{Q}_1) - \mathsf{I}_\mathsf{f}(\mathsf{P}_0,\mathsf{Q}_0)\big) = L_\mathsf{f}(v_1, v_0), \tag{1.12}$$

where the infimum is taken over all models $\mathbb{E}_1 = (\Omega_1, \mathscr{F}_1, \mathsf{P}_1, \mathsf{Q}_1)$ and $\mathbb{E}_0 = (\Omega_0, \mathscr{F}_0, \mathsf{P}_0, \mathsf{Q}_0)$ such that (1.9)–(1.11) hold ($\inf_{\varnothing} := 0$).

Remark 1.1. Assume that $\Omega_1 = \Omega_0 = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\mathscr{F}_1$ is the σ-field of all subsets of Ω_1, $\mathscr{F}_0 = \sigma\{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}\}$, P_1 and Q_1 are defined in the following table:

	ω_1	ω_2	ω_3	ω_4
P_1	v_0	$1-a-v_0$	a	0
Q_1	0	$1-a-v_1$	$a+v_1-v_0$	v_0

and P_0 and Q_0 are the restrictions of P_1 and Q_1 respectively onto $\mathscr{F}_0$. If $\mathsf{f}(0)+\mathsf{f}^*(0)<+\infty$ or $v_0=0$, then $d_{\mathsf{f}}(a,v_1,v_0)=\mathsf{I}_{\mathsf{f}}(\mathsf{P}_1,\mathsf{Q}_1)-\mathsf{I}_{\mathsf{f}}(\mathsf{P}_0,\mathsf{Q}_0)$ and thus the infimum in (1.12) is attained on the above model with $a=a_*$, where $a_*=\arg\min d_{\mathsf{f}}(\cdot,v,v_0)$. This is not the case if $\mathsf{f}(0)+\mathsf{f}^*(0)=+\infty$ and $v_0>0$ because then $\mathsf{I}_{\mathsf{f}}(\mathsf{P}_0,\mathsf{Q}_0)=+\infty$. However, let us show that the infimum in (1.12) over *all* models can still be replaced by the infimum over models with the same spaces $(\Omega_1,\mathscr{F}_1)$ and $(\Omega_0,\mathscr{F}_0)$ as above. We consider only the case where $0<v_0<v_1<1$ and leave the remaining cases to the reader. Note also that if $f(0)=+\infty$, then $a=0$ is irrelevant for the definition of $L_{\mathsf{f}}(v_1,v_0)$ because $d_{\mathsf{f}}(0,v_1,v_0)=+\infty$; similarly, $a=1-v_1$ is irrelevant if $\mathsf{f}^*(0)=+\infty$. If $f(0)=+\infty$, then we add the same amount ε to $\mathsf{P}_1(\{\omega_4\})$ and $\mathsf{Q}_1(\{\omega_4\})$ and subtract ε from $\mathsf{P}_1(\{\omega_3\})$ and $\mathsf{Q}_1(\{\omega_3\})$. This can be done if $a>0$ and $\varepsilon>0$ is small enough, the distances $\|\mathsf{Q}_1-\mathsf{P}_1\|$ and $\|\mathsf{Q}_0-\mathsf{P}_0\|$ being unchanged. Similarly, if $f^*(0)=+\infty$, then we add ε^* to $\mathsf{P}_1(\{\omega_1\})$ and $\mathsf{Q}_1(\{\omega_1\})$ and subtract ε^* from $\mathsf{P}_1(\{\omega_2\})$ and $\mathsf{Q}_1(\{\omega_2\})$. Then the difference $\mathsf{I}_{\mathsf{f}}(\mathsf{P}_1,\mathsf{Q}_1)-\mathsf{I}_{\mathsf{f}}(\mathsf{P}_0,\mathsf{Q}_0)$ is increasing in ε and ε^*, so its infimum corresponds to the case $\varepsilon=\varepsilon^*=0$ and equals $d_{\mathsf{f}}(a,v_0,v_1)$. We leave the details to the reader, cf. also the proof of Theorem 1.1.

As a direct consequence of Theorem 1.1, we obtain the well-known lower and upper bounds for the f-divergence $\mathsf{I}_{\mathsf{f}}(\mathsf{P},\mathsf{Q})$ in terms of the total variation distance $\|\mathsf{P}-\mathsf{Q}\|$. The lower bound corresponds to the case $v_0=0$. The upper bound is trivial if $\mathsf{f}(0)+\mathsf{f}^*(0)=\infty$ and corresponds to the case $v_1=1$ otherwise.

Corollary 1.1 (Vajda [22])**.** *For every $v\in[0,1]$,*

$$L_{\mathsf{f}}(v,0)=\inf_{\|\mathsf{P}-\mathsf{Q}\|=2v}\mathsf{I}_{\mathsf{f}}(\mathsf{P},\mathsf{Q})\le\sup_{\|\mathsf{P}-\mathsf{Q}\|=2v}\mathsf{I}_{\mathsf{f}}(\mathsf{P},\mathsf{Q})=v(\mathsf{f}(0)+\mathsf{f}^*(0)).$$

It is known that the function $L_{\mathsf{f}}(v,0)$ can be represented in a simpler form in some special cases (see, e.g., [15, Proposition 8.28], [6], [7, Lemma 2.4]). The same is true for $L_{\mathsf{f}}(v_1,v_0)$.

Proposition 1.2. (i) *If $\mathsf{f}(u)=\mathsf{f}^*(u)$ for all $u\in(0,1)$, then*

$$L_{\mathsf{f}}(v_1,v_0)=\mathsf{F}(1+v_1-2v_0,1-v_1).$$

(ii) *If $\mathsf{f}(u)=\mathsf{f}(2-u)$ for all $u\in(0,1)$, then*

$$L_{\mathsf{f}}(v_1,v_0)=\begin{cases}\mathsf{F}(1+2v_1-3v_0,1-v_0) & \text{if } 1+v_0\ge 2v_1,\\ (v_1-v_0)\,\mathsf{f}(2)+\mathsf{F}(1-v_0,1-v_1) & \text{otherwise.}\end{cases}$$

(iii) *If $\mathsf{f}(u)=0$ for all $u\in(0,1)$, then*

$$L_{\mathsf{f}}(v_1,v_0)=\mathsf{F}(1-v_0,1-v_1).$$

Remark 1.2. The assumptions in (i)–(iii) can be slightly relaxed taking into account relation (1.5). Also, one may use (1.4) and to replace f by f^* in (ii) and (iii).

Remark 1.3. If a function f can be represented as $\mathsf{f}=\mathsf{f}_1+\mathsf{f}_2$, where f_1 and f_2 are convex functions, f_1 satisfies one of "symmetry" assumptions (i) or (ii), and f_2 is null either to the right or to the left of $u=1$, then one can estimate $L_{\mathsf{f}}(v_1,v_0)$ from below by $L_{\mathsf{f}_1}(v_1,v_0)+L_{\mathsf{f}_2}(v_1,v_0)$ with explicit expressions for the summands given in Proposition 1.2. This idea is due to Gilardoni [7, 8], who considers the case $v_0=0$ and assumption (i) on f_1 and shows that it works with $\mathsf{f}(u)=u\log u$, i.e., the Kullback–Leibler divergence.

Example 1.1. Let $\mathsf{f}(x) = \frac{1}{2}(\sqrt{x}-1)^2$, then $\mathsf{I}_\mathsf{f}(\mathsf{P},\mathsf{Q}) = \rho^2(\mathsf{P},\mathsf{Q})$ is the squared Hellinger distance. By Proposition 1.2(i),

$$\rho^2(\mathsf{P}_1,\mathsf{Q}_1) \ge \rho^2(\mathsf{P}_0,\mathsf{Q}_0) + 1 - \frac{1}{2}\|\mathsf{P}_0-\mathsf{Q}_0\| - \sqrt{1-\frac{1}{4}\|\mathsf{P}_1-\mathsf{Q}_1\|^2 - \|\mathsf{P}_0-\mathsf{Q}_0\|\Big(1-\frac{1}{2}\|\mathsf{P}_1-\mathsf{Q}_1\|\Big)}.$$

Inverting this inequality, we get that, if $\rho^2(\mathsf{P}_1,\mathsf{Q}_1) - \rho^2(\mathsf{P}_0,\mathsf{Q}_0) \le \kappa \le 1 - \frac{1}{2}\|\mathsf{P}_0-\mathsf{Q}_0\|$, then

$$\|\mathsf{P}_1-\mathsf{Q}_1\| \le \|\mathsf{P}_0-\mathsf{Q}_0\| + 2\sqrt{\kappa(2-\|\mathsf{P}_0-\mathsf{Q}_0\|-\kappa)}.$$

Example 1.2. Let $\mathsf{f}(x) = x\log x - x + 1$, then $\mathsf{I}_\mathsf{f}(\mathsf{P},\mathsf{Q}) = \mathsf{D}(\mathsf{P},\mathsf{Q})$ is the Kullback–Leibler divergence. According to Remark 1.3, put $\mathsf{f}_1(x) = \mathsf{f}(x)$ for $x \in (0,1)$ and $\mathsf{f}_1(x) = x\mathsf{f}_1(1/x)$ for $x > 1$; $\mathsf{f}_2(x) = \mathsf{f}(x) - \mathsf{f}_1(x)$ for $x > 0$.

It is easy to check that f_1 and f_2 are convex functions. Put $D_1 = \mathsf{D}(\mathsf{P}_1,\mathsf{Q}_1)$, $D_0 = \mathsf{D}(\mathsf{P}_0,\mathsf{Q}_0)$, $V_1 = \|\mathsf{P}_1-\mathsf{Q}_1\|$, $V_0 = \|\mathsf{P}_0-\mathsf{Q}_0\|$. Then we obtain the inequality

$$D_1 - D_0 \ge -\frac{2-V_0}{2}\log\frac{2-V_1}{2} + \frac{4-V_0-V_1}{2}\log\frac{2-V_0}{2} - \frac{2-V_1}{2}\log\frac{2+V_1-2V_0}{2}.$$

If $D_0 = V_0 = 0$, this inequality is obtained in [7, 8]. Let us emphasize that the bound is not sharp.

2. PROOFS

There are several characterizations of binary experiments, i.e., objects that determine the dichotomy up to equivalence: the power function (the power of the most powerful test of given level), the error function (the minimum Bayes risk function), the distribution of the likelihood ratio, the standard measure, etc., see [21]. In our problem, the power function has the advantage that it provides a convenient representation of experiments (see below), a simple description of the order $\succeq$ (pointwise ordering), and a simple geometric representation of the total variation distance. Feldman and Österreicher [5] use this approach to give an independent proof of Corollary 1.1. The disadvantage of using error functions is the absence of a convenient representation for experiments, but this is not so important for finite sample spaces that usually occur in solutions of optimization problems, see, e.g., [19] and [9].

First, we provide the reader with a short proof of inequality (1.7). The representation (2.1) appears in [18] for twice differentiable f and in [5] in the general case. See also [16, 17] and [14].

Given an experiment $\mathbb{E} = (\Omega, \mathscr{F}, \mathsf{P}, \mathsf{Q})$, let $l_\mathbb{E}(x,y)$, $x,y \in \mathbb{R}$, be the support function of the set $\mathfrak{N}(\mathbb{E})$. Then it satisfies

$$l_\mathbb{E}(x,y) := \sup_{(u,t)\in\mathfrak{N}(\mathbb{E})}(xu+yt) = \sup_{\varphi\in\Phi}\int(xp+yq)\varphi\,d\mu = \int(xp+yq)^+\,d\mu,$$

where μ, p, and q are as in (1.2).

Let also a convex function $\mathsf{f}\colon (0,\infty)\to\mathbb{R}$ be given. Then the right-hand derivative f'_+ is a right-continuous increasing function on $(0,\infty)$ and, hence, determines the Lebesgue–Stieltjes measure ν_f on $(0,\infty)$ satisfying $\nu_\mathsf{f}((x,y]) = \mathsf{f}'_+(y) - \mathsf{f}'_+(x)$.

Lemma 2.1. *Assume that* $\mathsf{f}(1) = 0$ *and* $\mathsf{f}'_+(1) = 0$. *Then, for every* $x,y \ge 0$,

$$\mathsf{F}(x,y) = \int_{(0,1)}(sy-x)^+\,\nu_\mathsf{f}(ds) + \int_{[1,\infty)}(x-sy)^+\,\nu_\mathsf{f}(ds).$$

The proof of the lemma is a direct consequence of the integration by parts formula. Now, by Fubini's theorem we obtain from (1.2) that

$$\mathsf{I}_{\mathsf{f}}(\mathsf{P},\mathsf{Q}) = \int_{(0,1)} l_{\mathbb{E}}(-1,s)\,\nu_{\mathsf{f}}(ds) + \int_{[1,\infty)} l_{\mathbb{E}}(1,-s)\,\nu_{\mathsf{f}}(ds). \tag{2.1}$$

Obviously, $\mathbb{E}_1 \succeq \mathbb{E}_0$ implies $l_{\mathbb{E}_1}(x,y) \geq l_{\mathbb{E}_0}(x,y)$ for all x, y, and (1.7) follows.

If $\mathbb{E}$ is a binary model, the set $\mathfrak{N}(\mathbb{E})$ is a convex and closed subset of $[0,1]\times[0,1]$, contains $(0,0)$, and is symmetric with respect to the point $(0.5, 0.5)$, see, e.g., [13, p. 62]. In Fig. 1 we present a set $\mathfrak{N}(\mathbb{E})$ of generic form.

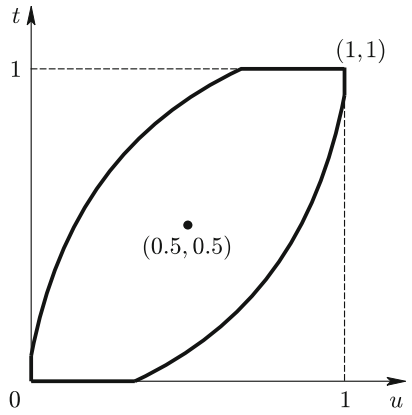

Fig. 1. The solid line is the boundary of the set $\mathfrak{N}(\mathbb{E})$. The presence of horizontal segments in this line indicates that P is not absolutely continuous with respect to Q: the length of the horizontal segments is equal to the P-measure of the singular component of P with respect to Q. Similarly, the length of the vertical segments is equal to the Q-measure of the singular component of Q with respect to P.

Let $\beta_{\mathbb{E}}(u) := \sup\{t \in [0,1]\colon (u,t) \in \mathfrak{N}(\mathbb{E})\}$, $u \in [0,1]$, be the maximal power among all tests of level u. Then, obviously, $\beta_{\mathbb{E}}(u)$ is a concave continuous function from $[0,1]$ to $[0,1]$, $\beta_{\mathbb{E}}(1) = 1$, and

$$\mathfrak{N}(\mathbb{E}) := \{(u,t) \in [0,1]\times[0,1]\colon 1 - \beta_{\mathbb{E}}(1-u) \leq t \leq \beta_{\mathbb{E}}(u)\}.$$

In particular, $\mathbb{E}_1 \succeq \mathbb{E}_0$ if and only if $\beta_{\mathbb{E}_1}(u) \geq \beta_{\mathbb{E}_0}(u)$ for all $u \in [0,1]$.

Conversely, let $\beta(u)$ be a concave continuous function from $[0,1]$ to $[0,1]$, $\beta(1) = 1$. Consider the dichotomy (λ, β) on $[0,1]$ with the Borel σ-field, where λ is the Lebesgue measure and β, by abuse of notation, is the measure on $[0,1]$ such that its distribution function coincides with $\beta(u)$ on [0,1]. Then $d\beta/d\lambda(u) = \beta'(u)\,du$-a.s., where $\beta'(u)$ denotes, say, the right-hand derivative of $\beta(u)$ at u. Since the function $\beta(u)$ is concave, the derivative is decreasing. Therefore, by the Neyman–Pearson lemma, $\mathbf{1}_{[0,u]}$ is the most powerful test of level u. Therefore, $\beta_{\mathbb{E}}(u) = \beta(u)$, $u \in [0,1]$.

Thus, to every experiment $\mathbb{E} = (\Omega, \mathscr{F}, \mathsf{P}, \mathsf{Q})$, we put into correspondence the equivalent experiment $\tilde{\mathbb{E}} := ([0,1], \mathscr{B}([0,1]), \lambda, \mu_{\mathbb{E}})$, which we consider as the representation of $\mathbb{E}$. Now (1.3) can be rewritten as

$$\mathsf{I}_{\mathsf{f}}(\mathsf{P},\mathsf{Q}) = \int_{(0,1)} \mathsf{f}^*(\beta'_{\mathbb{E}}(u))\,du + f(0)\beta_{\mathbb{E}}(0), \tag{2.2}$$

where $\beta'_{\mathbb{E}}(u)$ can be still interpreted as the right-hand derivative of the function $\beta_{\mathbb{E}}$ at u. Using this formula, we obtain, in particular,

$$\|\mathsf{Q} - \mathsf{P}\| = 2 \sup_{u\in[0,1]} (\beta_{\mathbb{E}}(u) - u).$$

Now we see that the set of dichotomies (P,Q) with $\|\mathsf{P}-\mathsf{Q}\| = 2v$ contains the greatest element with respect to the order $\succeq$, which corresponds to the function $\beta(u) = (v+u)\wedge 1$, which immediately gives the sharp upper bound in Corollary 1.1. On the other hand, the smallest element in this set does not

exist. However, every model is at least as informative as one of the models given by the probabilities on two points: $\mathsf{P} = (1-a, a)$, $\mathsf{Q} = (1-a-v, a+v)$, where a runs over $[0, 1-v]$. This proves the lower bound in Corollary 1.1. These arguments are due to [5].

Proof of Proposition 1.1. It follows from assumption (1.6) that $\mathsf{F}(x,x) \equiv 0$. Since F is convex and lower semicontinuous, the function $d_{\mathsf{f}}(a, v_1, v_0)$, $0 \le v_0 \le v_1 \le 1$, $0 \le a \le 1 - v_1$, is convex, lower semicontinuous and

$$d_{\mathsf{f}}(a, v_1, v_0) \ge 2\mathsf{F}((1-v_0)/2, (1-v_0)/2) = 0.$$

Since the infimum in the definition of $L_{\mathsf{f}}(v_1, v_0)$ is taken over a compact set, we obtain that $L_{\mathsf{f}}(v_1, v_0)$ is convex, lower semicontinuous, and nonnegative, and is equal to 0 on $\{v_0 = v_1\}$. The last fact together with convexity implies that the function is increasing in the first argument and decreasing in the second one.

Let us prove the remaining claim. Since F is convex and positively homogeneous, for all $x, y, h \ge 0$,

$$\mathsf{F}(x,y) = \frac{1}{2}\mathsf{F}(2x, 2y) + \frac{1}{2}\mathsf{F}(2h, 2h) \ge \mathsf{F}(x+h, y+h). \tag{2.3}$$

The inequality $L_{\mathsf{f}}(v_1 + h, v_0 + h) \ge L_{\mathsf{f}}(v_1, v_0)$ follows now from the definition of L_{f}. □

Proof of Theorem 1.1. Let $\mathbb{E}_1 = (\Omega_1, \mathscr{F}_1, \mathsf{P}_1, \mathsf{Q}_1)$ and $\mathbb{E}_0 = (\Omega_0, \mathscr{F}_0, \mathsf{P}_0, \mathsf{Q}_0)$ be two models satisfying (1.9) and (1.11), that is,

$$\beta_{\mathbb{E}_1}(u) \ge \beta_{\mathbb{E}_0}(u) \quad \text{for all } u \in [0,1],$$

$$v_1 = \sup_{u \in [0,1]} (\beta_{\mathbb{E}_1}(u) - u), \quad v_0 = \sup_{u \in [0,1]} (\beta_{\mathbb{E}_0}(u) - u), \tag{2.4}$$

see Fig. 2. If $v_1 = v_0$, then both sides of (1.12) are equal to 0, so we will assume that $0 \le v_0 < v_1 \le 1$. If $v_0 < v_1 = 1$ and $\mathsf{f}(0) + \mathsf{f}^*(0) = \infty$, then both sides of (1.12) are equal to $+\infty$, so this case is also excluded from further consideration. In the remaining cases $L_{\mathsf{f}}(v_1, v_0) < \infty$, so we may assume that $\mathsf{I}_{\mathsf{f}}(\mathsf{P}_1, \mathsf{Q}_1) < \infty$.

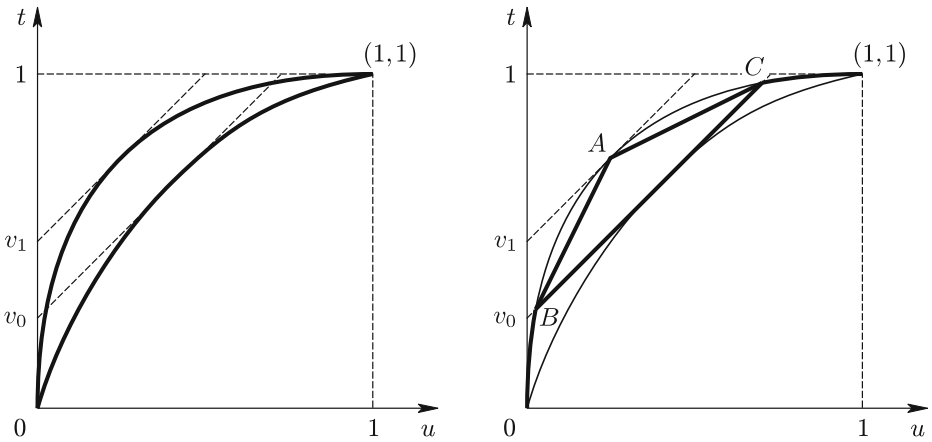

Fig. 2. In the figure on the left, the solid curves are the graphs of the functions $u \rightsquigarrow \beta_{\mathbb{E}_1}(u)$ (the upper one) and $u \rightsquigarrow \beta_{\mathbb{E}_0}(u)$ (the lower one), the dashed straight lines $u \rightsquigarrow u + v_1$ and $u \rightsquigarrow u + v_0$ are corresponding tangent lines. In the figure on the right, A is a point where the tangent line $u \rightsquigarrow u + v_1$ and the curve $u \rightsquigarrow \beta_{\mathbb{E}_1}(u)$ meet, and B and C are the points where the curve $u \rightsquigarrow \beta_{\mathbb{E}_1}(u)$ intersects the line $u \rightsquigarrow u + v_0$ (if $\beta_{\mathbb{E}_1}(0) > v_0$, then $B = (0, v_0)$). The functions $u \rightsquigarrow \beta_1(u)$ and $u \rightsquigarrow \beta_0(u)$ are defined via their graphs represented by solid curves. The graph of $u \rightsquigarrow \beta_1(u)$ contains the straight segments from B to A and from A to C, the graph of $u \rightsquigarrow \beta_0(u)$ contains the straight segment from B to C. Finally, the graphs of both functions replicate the graph of $u \rightsquigarrow \beta_{\mathbb{E}_1}(u)$ to the left of B and to the right of C.

The following construction is the core of the proof, see Fig. 2. According to (2.4), there exists a point A on the graph of $\beta_{\mathbb{E}_1}(u)$, which belongs to the line $t = u + v_1$. The graph of $\beta_{\mathbb{E}_1}(u)$ intersects

the line $t = u + v_0$ to the right of A at a (unique) point denoted by C. Similarly, if $\beta_{\mathbb{E}_1}(0) \le v_0$, the graph of $\beta_{\mathbb{E}_1}(u)$ intersects the line $t = u + v_0$ to the left of A at a (unique) point denoted by B; otherwise, $B := (0, v_0)$. Now define new functions $\beta_1(u)$ and $\beta_0(u)$ in the following way: the graph of $\beta_1(u)$ contains two straight segments BA and AC, while the graph of $\beta_0(u)$ contains the straight segment BC; for other values of u, we put $\beta_1(u) := \beta_0(u) := \beta_{\mathbb{E}_1}(u)$. It is clear that $\beta_1(u)$ and $\beta_0(u)$ are concave continuous functions equal to 1 at 1,

$$\beta_{\mathbb{E}_1}(u) \ge \beta_1(u) \ge \beta_0(u) \ge \beta_{\mathbb{E}_0}(u) \qquad \text{for all} \quad u \in [0,1],$$

$$v_1 = \sup_{u\in[0,1]} (\beta_1(u) - u), \qquad v_0 = \sup_{u\in[0,1]} (\beta_0(u) - u),$$

so

$$\mathsf{I}_\mathsf{f}(\mathsf{P}_1, \mathsf{Q}_1) - \mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0) \ge \mathsf{I}_\mathsf{f}(\lambda, \beta_1) - \mathsf{I}_\mathsf{f}(\lambda, \beta_0),$$

and the infimum in (1.12) can be replaced by the infimum over all pairs of dichotomies (λ, β_1), (λ, β_0) obtained through the above construction.

Now we note that, according to (2.2),

$$\mathsf{I}_\mathsf{f}(\lambda, \beta_1) - \mathsf{I}_\mathsf{f}(\lambda, \beta_0) = \mathsf{F}(u_A - u_B, t_A - t_B) + \mathsf{F}(u_C - u_A, t_C - t_A),$$

where $A = (u_A, t_A)$, and similarly for B and C. According to (2.3), $\mathsf{F}(x + h, y + h)$ is decreasing in h. Since the points B and C lie on the line $t = u + v_0$, for a fixed A, the infimum of $\mathsf{F}(u_A - u_B, t_A - t_B)$ over B is attained if $u_B = 0$, and the infimum of $\mathsf{F}(u_C - u_A, t_C - t_A)$ over C is attained if $t_C = 1$. Thus,

$$\begin{aligned}&\inf_{B,C}\big[\mathsf{F}(u_A - u_B, t_A - t_B) + \mathsf{F}(u_C - u_A, t_C - t_A)\big]\\ &\qquad = \mathsf{F}(u_A, t_A - v_0) + \mathsf{F}(1 - v_0 - u_A, 1 - t_A) = d(a, v_0, v_1),\end{aligned}$$

where $a = u_A$. It may happen that the set of possible positions for B does not include the extreme point $(0, v_0)$ (if $\mathsf{f}^*(0) = +\infty$) and similarly for C, but this is irrelevant for the above equality. □

Proof of Corollary 1.1. If $\|\mathsf{P}_0 - \mathsf{Q}_0\| = 0$, then $\mathsf{P}_0 = \mathsf{Q}_0$, $\mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0) = 0$ for any convex f, and $\mathbb{E}_1 \succeq \mathbb{E}_0$ for any model $\mathbb{E}_1$. Thus, if $v_0 = 0$, (1.12) reduces to the first equality in the assertion.

If $\|\mathsf{P}_1 - \mathsf{Q}_1\| = 2$, then P_1 and Q_1 are singular, $\mathsf{I}_\mathsf{f}(\mathsf{P}_1, \mathsf{Q}_1) = \mathsf{f}(0) + \mathsf{f}^*(0)$ for any convex f, and $\mathbb{E}_1 \succeq \mathbb{E}_0$ for any model $\mathbb{E}_0$. This means that if $v_1 = 1$ and $\mathsf{f}(0) + \mathsf{f}^*(0) < +\infty$, (1.12) reduces to

$$\sup_{\|\mathsf{P}-\mathsf{Q}\|=2v} \mathsf{I}_\mathsf{f}(\mathsf{P}, \mathsf{Q}) = \mathsf{f}(0) + \mathsf{f}^*(0) - L_\mathsf{f}(1, v) = v(\mathsf{f}(0) + \mathsf{f}^*(0)).$$

If $\mathsf{f}(0) + \mathsf{f}^*(0) = +\infty$ and $v > 0$, it is enough to note that, in the model in Remark 1.1 with $v_0 = v$, $v_1 = 1$, $a = 0$, we have $\|\mathsf{P}_0 - \mathsf{Q}_0\| = 2v$ and $\mathsf{I}_\mathsf{f}(\mathsf{P}_0, \mathsf{Q}_0) = +\infty$. □

Proof of Proposition 1.2. (i) Our assumption on f implies $\mathsf{F}(a, a + v_1 - v_0) = \mathsf{F}(a + v_1 - v_0, a)$. Hence, by convexity and positive homogeneity of F,

$$d_\mathsf{f}(a, v_1, v_0) = \mathsf{F}(1 - a - v_0, 1 - a - v_1) + \mathsf{F}(a + v_1 - v_0, a) \ge \mathsf{F}(1 + v_1 - 2v_0, 1 - v_1),$$

and equality holds for $a = (1 - v_1)/2$.

(ii) Our assumption on f implies $\mathsf{F}(a, a + v_1 - v_0) = \mathsf{F}(a + 2v_1 - 2v_0, a + v_1 - v_0)$, hence

$$d_\mathsf{f}(a, v_1, v_0) = \mathsf{F}(1 - a - v_0, 1 - a - v_1) + \mathsf{F}(a + 2v_1 - 2v_0, a + v_1 - v_0).$$

The expression on the right is well defined for all real $a \le 1 - v_1$, is a convex function of a and, similarly to the previous case, attains the minimum at $a_* = (1 + v_0 - 2v_1)/2$. This implies that the minimum of $d_\mathsf{f}(a, v_1, v_0)$ over $a \in [0, 1 - v_1]$ is attained at $a = a_*$ if $a_* \ge 0$, and at $a = 0$ otherwise. The claim follows.

(iii) Our assumption on f implies $\mathsf{F}(a, a + v_1 - v_0) = 0$, so the claim follows from (2.3). □

REFERENCES

1. S. M. Ali and S. D. Silvey, "A General Class of Coefficients of Divergence of One Distribution from Another", J. Roy. Statist. Soc., Ser. B (Methodological) **28**, 131–142 (1966).
2. D. Blackwell, "Comparison of Experiments", in *Proc. Second Berkeley Symp. on Math. Statist. and Probab.* (Univ. of Calif. Press, 1951), pp. 93–102.
3. I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten", Magyar Tud. Akad. Mat. Kutató Int. Közl. **8**, 85–108 (1963).
4. I. Csiszár, "Information-Type Measures of Difference of Probability Distributions and Indirect Observations", Stud. Sci. Math. Hung. **2**, 299–318 (1967).
5. D. Feldman and F. Österreicher, "A Note on f-Divergences", Stud. Sci. Math. Hung. **24**, 191–200 (1989).
6. G. L. Gilardoni, "On the Minimum f-Divergence for Given Total Variation", C. R. Math. Acad. Sci. Paris **343**, 763–766 (2006).
7. G. L. Gilardoni, "An Improvement on Vajda's Inequality", in *Progress in Probability* (Birkhäuser, Basel, 2008), Vol. 60, pp. 299–304.
8. G. L. Gilardoni, "On Pinsker's and Vajda's Type Inequalities for Csiszár's f-Divergences", IEEE Trans. Inform. Theory **56**, 5377–5386 (2010).
9. A. Guntuboyina, S. Saha, and G. Schiebinger, "Sharp Inequalities for f-divergences", IEEE Trans. Inform. Theory **60**, 104–121 (2014).
10. A. A. Gushchin, "On an Extension of the Notion of f-Divergence", Theory Probab. Appl. **52**, 439–455 (2008).
11. L. Le Cam, "Sufficiency and Approximate Sufficiency", Ann. Math. Statist. **35**, 1419–1455 (1964).
12. L. Le Cam, "On the Information Contained in Additional Observations", Ann. Statist. **2**, 630–649 (1974).
13. E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses* (Springer Science & Business Media, New York, 2006).
14. F. Liese, "ϕ-Divergences, Sufficiency, Bayes Sufficiency, and Deficiency", Kybernetika **48**, 690–713 (2012).
15. F. Liese and I. Vajda, *Convex Statistical Distances* (Teubner, Leipzig, 1987).
16. F. Liese and I. Vajda, "On Divergences and Information in Statistics and Information Theory", IEEE Trans. Inform. Theory **52**, 4394–4412 (2006).
17. F. Liese and I. Vajda, "f-Divergences: Sufficiency, Deficiency and Testing of Hypotheses", in *Advances in Inequalities from Probability Theory and Statistics*, Ed. by N. S. Barnett and S. S. Dragomir (Nova Science, New York, 2008), pp. 113–158.
18. F. Österreicher and D. Feldman, "Divergenzen von Wahrscheinlichkeitsverteilungen Integralgeometrisch Betrachtet", Acta Math. Sci. Hungar. **37**, 329–337 (1981).
19. M. D. Reid and R. C. Williamson, "Generalized Pinsker Inequalities", in *The 22nd Ann. Conf. on Learning Theory (COLT 2009)* (2009), pp. 1–10.
20. H. Strasser, *Mathematical Theory of Statistics* (de Gruyter, Berlin, New York, 1985).
21. E. Torgersen, *Comparison of Statistical Experiments* (Cambridge Univ. Press, Cambridge, 1991).
22. I. Vajda, "On the f-Divergence and Singularity of Probability Measures", Period. Math. Hung. **2**, 223–234 (1972).