IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Bounds on f-Divergences and Related Distances

Igal Sason

CCIT Report #859
May 2014

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #859 May 2014

Bounds on f-Divergences and Related Distances

Igal Sason

Abstract

Derivation of tight bounds on f-divergences and related distances is of interest in information theory and statistics. This paper improves some existing bounds on f-divergences. In some cases, an alternative approach leads to a simplified proof of an existing bound. Following bounds on the chi-squared divergence, an improved version of a reversed Pinsker's inequality is derived for an arbitrary pair of probability distributions on a finite set. Following bounds on the relative entropy and Jeffreys' divergence, a tightened inequality for lossless source coding is derived and considered. Finally, a new inequality relating f-divergences is derived and studied.

Index Terms – Bhattacharyya distance, Chernoff information, chi-squared divergence, f-divergence, Hellinger distance, Jeffreys' divergence, lossless source coding, relative entropy (Kullback-Leibler divergence), total variation distance.

I. Introduction

Divergence measures are widely used in information theory, machine learning, statistics, and other theoretical and applied branches of mathematics (see, e.g., [3], [12], [15], [34], [40] and [44]). The class of f-divergences, introduced independently in [1], [8] and [32], forms an important class of divergence measures which includes the relative entropy (a.k.a. information divergence or the Kullback-Leibler divergence), its dual and symmetrized divergences, the total variation distance, squared Hellinger distance, chi-squared divergence, etc. Properties of f-divergences, including relations to statistical tests and estimators, were extensively studied in [30].

In the following, some related papers that are most relevant to the scope of this work are briefly reviewed. Pinsker's inequality (a.k.a the Csiszár-Kemperman-Kullback-Pinsker inequality) and Vajda's inequality [44] have been derived during the sixties to provide lower bounds on the relative entropy in terms of the total variation distance. Following these bounds, Fedotov *et al.* [21] derived an exact parametrization of the infimum of the relative entropy with respect to all possible pairs of probability distributions with a given total variation distance. The derivation of the parametrization in [21] relies on the data processing theorem for the relative entropy, leading to a maximization problem of the convex conjugate function of the relative entropy between two-element probability distributions; this approach leads to closed-form solutions that are used to identify a possible form for the required parametrization. As an extension to this problem, Harremoës and Vajda studied in [26] the joint range of pairs of f-divergences, characterizing all the possible points in $[0,\infty]^2$ that are achievable by a given pair of f-divergences. It was shown that this region is convex where each point is a convex combination of two achievable points that are obtained by a pair of probability distributions over two elements; hence, every such an achievable point is obtained by a pair of probability distributions over at most 4 elements.

In [23], Gilardoni studied the problem of minimizing an arbitrary *symmetric* f-divergence for a given total variation distance, and provided a closed-form solution of this optimization problem. Furthermore, an alternative parametrization of the infimum of the relative entropy for a given total variation distance was derived in [23]. In a follow-up paper by the same author [24], Pinsker's and Vajda's type inequalities were studied for symmetric f-divergences, and the issue of obtaining lower bounds on f-divergences for a given total variation distance was further studied. One of the main results in [24] was a derivation of an improved and simple closed-form lower bound on the relative entropy in terms of the total variation distance, as well as a simple and reasonably tight closed-form upper bound on the infimum of the relative entropy for a given total variation distance.

Sharp inequalities for f-divergences were recently studied in [25] as a problem of maximizing or minimizing an arbitrary f-divergence between two probability measures subject to a finite number of inequality constraints on other f-divergences. The main result stated in [25] is that such infinite-dimensional optimization problems are equivalent to optimization problems over finite-dimensional spaces where the latter are numerically solvable.

I. Sason is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sason@ee.technion.ac.il).

Following previous work, some new bounds on f-divergences and related distances are derived in this paper. This work improves some existing bounds on f-divergences; in some cases, an alternative approach leads to a simplified proof of an existing bound. An improved version of a reversed Pinsker's inequality is derived in this paper for an arbitrary pair of probability distributions on a finite set. Following [9], two tightened inequalities for lossless source coding are derived via bounds on the relative entropy and Jeffreys' divergence. Finally, a new general inequality relating f-divergences is derived and studied.

The paper is organized as follows: preliminary material is introduced in Section II, and the bounds and their proofs are introduced in Section III, followed by discussions and remarks that link the new bounds to the literature. Section III is separated into seven subsections that refer to various f-divergences and related distances. The paper is concluded in Section IV where the contributions of this work are outlined, and an additional proof that provides some insight is relegated to an appendix.

II. Preliminaries

We introduce, in the following, some preliminaries and notation that are essential to the analysis in this paper.

Definition 1: Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let P and Q be two probability distributions on a set $\mathcal{A}$. The *f-divergence* from P to Q is defined by

$$D_f(P||Q) \triangleq \sum_{x\in\mathcal{A}} Q(x)\, f\left(\frac{P(x)}{Q(x)}\right) \tag{1}$$

with the convention that

$$0f\Big(\frac{0}{0}\Big) = 0, \quad f(0) = \lim_{t\to 0^+} f(t),$$
$$0f\Big(\frac{a}{0}\Big) = \lim_{t\to 0^+} tf\Big(\frac{a}{t}\Big) = a\lim_{u\to\infty}\frac{f(u)}{u},\ \forall\, a>0.$$

The relative entropy $D(P||Q) \triangleq \sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$ is an f-divergence where $f(t) = t\log(t)$ for $t>0$.

An f-divergence $D_f(P||Q)$ is in general jointly convex in the two probability distributions P and Q, and it is non-negative [8]; these basic properties, which hold for the relative entropy, are preserved for f-divergences in general.

Definition 2: The *dual of an f-divergence* is given by $D_f^\star(P||Q) \triangleq D_f(Q||P)$.

It is easy to verify that $D_f^\star(P||Q) = D_{f^\star}(P||Q)$ with $f^\star$ that is of the form

$$f^\star(t) = t\, f\left(\frac{1}{t}\right) + a\,(t-1), \quad \forall\, t>0$$

for an arbitrary $a\in\mathbb{R}$. Since the perspective of a convex function is also convex (see [5, p. 89]), the convexity of $f^\star$ follows and $f^\star(1) = 0$.

Definition 3: An f-divergence is said to be *symmetric* if it is equal to its dual.

An f-divergence is symmetric if and only if there exists a constant $a\in\mathbb{R}$ such that the function f in Definition 1 satisfies the equality

$$f(t) = t\, f\Big(\frac{1}{t}\Big) + a(t-1), \quad \forall\, t>0.$$

The sufficiency of this condition for ensuring the symmetry of the f-divergence is easy to verify from (1), and the necessity of this condition was proved in [44, Theorem 9.6].

Definition 4: Let P and Q be two probability distributions on a set $\mathcal{A}$. The *total variation distance* between P and Q is defined by

$$d_{\mathrm{TV}}(P,Q) \triangleq \sup_{A\subseteq\mathcal{A}} |P(A) - Q(A)| \tag{2}$$

where the supremum is taken over all the subsets A of $\mathcal{A}$ for which $P(A)$ and $Q(A)$ are well defined.

If $\mathcal{A}$ is a countable set, (2) is simplified to

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{A}}\big|P(x) - Q(x)\big| = \frac{||P-Q||_1}{2} \tag{3}$$

so, the total variation distance forms a symmetric f-divergence where $f(t) = \frac{1}{2}|t-1|$ for $t \in \mathbb{R}$.

The following result refers to the infimum of a symmetric f-divergence for a given total variation distance [23]:

Theorem 1: Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and assume that f is twice differentiable. Let

$$L_{D_f}(\varepsilon) \triangleq \inf_{P,Q\colon\, d_{\mathrm{TV}}(P,Q)=\varepsilon} D_f(P||Q), \quad \forall\, \varepsilon \in [0,1] \tag{4}$$

be the infimum of the f-divergence for a given total variation distance. If D_f is a symmetric f-divergence, then

$$L_{D_f}(\varepsilon) = (1-\varepsilon)\, f\left(\frac{1+\varepsilon}{1-\varepsilon}\right) - 2f'(1)\,\varepsilon, \quad \forall\, \varepsilon \in [0,1]. \tag{5}$$

Consequently, for a symmetric f-divergence, and for every pair of probability distributions P and Q, we have

$$D_f(P||Q) \geq \big(1 - d_{\mathrm{TV}}(P,Q)\big)\, f\left(\frac{1+d_{\mathrm{TV}}(P,Q)}{1-d_{\mathrm{TV}}(P,Q)}\right) - 2f'(1)\, d_{\mathrm{TV}}(P,Q) \tag{6}$$

and this bound, expressed in terms of the total variation distance, is tight.

The following corollary from [23] and [25, Corollary 5.4] is useful for the analysis in this paper:

Corollary 1: If $f\colon (0,\infty) \to \mathbb{R}$ satisfies the equality

$$f(t) = t\, f\Big(\frac{1}{t}\Big), \quad \forall\, t > 0 \tag{7}$$

and it is convex with $f(1) = 0$, then D_f is symmetric, and

$$L_{D_f}(\varepsilon) = (1-\varepsilon)\, f\left(\frac{1+\varepsilon}{1-\varepsilon}\right), \quad \forall\, \varepsilon \in [0,1]. \tag{8}$$

Proof: It is easy to verify from (7) that D_f is symmetric. For simplicity, assume that f is twice differentiable. Equality (7) with $f(1) = 0$ implies that $f'(1) = 0$, and the result follows from Theorem 1. ■

Definition 5: Let P and Q be two probability distributions on a set $\mathcal{A}$. The *Hellinger* and *Bhattacharyya distances* between P and Q are, respectively,

$$d_{\mathrm{H}}(P,Q) \triangleq \left(\frac{1}{2}\sum_{x\in\mathcal{A}}\Big(\sqrt{P(x)} - \sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}} \tag{9}$$

$$Z(P,Q) \triangleq \sum_{x\in\mathcal{A}} \sqrt{P(x)\,Q(x)}\,. \tag{10}$$

The divergence measures in (4), (9) and (10) are bounded between 0 and 1. Furthermore, it is easy to verify that

$$d_{\mathrm{H}}(P,Q) = \sqrt{1 - Z(P,Q)}. \tag{11}$$

The squared Hellinger distance is a symmetric f-divergence since

$$\big(d_{\mathrm{H}}(P,Q)\big)^2 = D_f(P||Q) \tag{12}$$

where

$$f(x) = \frac{1}{2}\,(1-\sqrt{x})^2, \quad x \geq 0 \tag{13}$$

is a convex function with $f(1) = 0$.

Definition 6: The *Chernoff information* between two probability distributions P and Q on a set $\mathcal{A}$ is given by

$$C(P,Q) \triangleq -\min_{\lambda\in[0,1]} \log\left(\sum_{x\in\mathcal{A}} P(x)^\lambda\, Q(x)^{1-\lambda}\right) \tag{14}$$

where throughout this paper, the logarithms are on base e. Note that

$$C(P,Q) = \max_{\lambda\in[0,1]} \left\{ -\log\left(\sum_{x\in\mathcal{A}} P(x)^{\lambda}\, Q(x)^{1-\lambda}\right)\right\}$$
$$= \max_{\lambda\in(0,1)} \Big\{(1-\lambda)\, D_{\lambda}(P,Q)\Big\} \tag{15}$$

where $D_{\lambda}(P,Q)$ designates the Rényi divergence of order λ [19]. The endpoints of the interval $[0,1]$ are excluded in the second line of (15) since the Chernoff information is non-negative, and the logarithmic function in the first line of (15) is equal to zero at both endpoints.

Proposition 1: For two probability distributions P and Q on a set $\mathcal{A}$

$$d_{\mathrm{TV}}(P,Q) \leq \sqrt{2}\, d_{\mathrm{H}}(P,Q) \leq \sqrt{D(P||Q)}. \tag{16}$$

Proof: The left-hand side of (16) is proved in [35, p. 99], and the right-hand side is proved in [35, p. 328]. ■

Definition 7: The *chi-squared divergence* between two probability distributions P and Q on a set $\mathcal{A}$ is given by

$$\chi^2(P,Q) \triangleq \sum_{x\in\mathcal{A}} \frac{\big(P(x)-Q(x)\big)^2}{Q(x)} = \sum_{x\in\mathcal{A}} \frac{P(x)^2}{Q(x)} - 1\,. \tag{17}$$

The chi-squared divergence is an asymmetric f-divergence where $f(t) = (t-1)^2$ for $t \geq 0$.

For further study of f-divergences and related distances between probability distributions, the reader is referred to [12, Chapter 4], [15], [44] and references therein.

III. Bounds on f-Divergences and Related Distances

The following section is separated into six subsections that correspond to the derivation of bounds for various f-divergences and related distances.

A. Bounds on the Hellinger and Bhattacharyya Distances

The following proposition introduces a sharpened version of Proposition 1.

Proposition 2: Let P and Q be two probability distributions on a set $\mathcal{A}$. Then, the following inequality suggests a tightened version of inequality (16)

$$1 - \sqrt{1 - \big(d_{\mathrm{TV}}(P,Q)\big)^2} \leq \big(d_{\mathrm{H}}(P,Q)\big)^2 \leq \min\left\{1 - \exp\left(-\frac{D(P||Q)}{2}\right),\, d_{\mathrm{TV}}(P,Q)\right\} \tag{18}$$

and

$$\max\left\{\exp\left(-\frac{D(P||Q)}{2}\right),\, 1 - d_{\mathrm{TV}}(P,Q)\right\} \leq Z(P,Q) \leq \sqrt{1 - \big(d_{\mathrm{TV}}(P,Q)\big)^2}. \tag{19}$$

Proof: We start with the proof of the left-hand side of (18). From (3)– (11), and the Cauchy-Schwartz inequality

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{A}} |P(x)-Q(x)|$$
$$= \frac{1}{2}\sum_{x\in\mathcal{A}} \Big|\sqrt{P(x)} - \sqrt{Q(x)}\Big|\Big(\sqrt{P(x)} + \sqrt{Q(x)}\Big)$$
$$\leq \frac{1}{2}\left(\sum_{x\in\mathcal{A}} \Big(\sqrt{P(x)} - \sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}} \left(\sum_{x\in\mathcal{A}} \Big(\sqrt{P(x)} + \sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}}$$
$$= d_{\mathrm{H}}(P,Q)\, \left(1 + \sum_{x\in\mathcal{A}} \sqrt{P(x)\,Q(x)}\right)^{\frac{1}{2}}$$
$$= d_{\mathrm{H}}(P,Q)\, \Big(2 - \big(d_{\mathrm{H}}(P,Q)\big)^2\Big)^{\frac{1}{2}}. \tag{20}$$

Let $c \triangleq \big(d_{\mathrm{TV}}(P,Q)\big)^2$ and $x \triangleq \big(d_{\mathrm{H}}(P,Q)\big)^2$. Squaring both sides of (20), gives $x(2-x) \geq c$, which implies that

$$1 - \sqrt{1-c} \leq x \leq 1 + \sqrt{1-c}\,. \tag{21}$$

The right-hand side of (21) is satisfied automatically since $0 \leq d_{\mathrm{H}}(P,Q) \leq 1$ implies that $x \leq 1$. The left-hand side of (21) gives the lower bound on the left-hand side of (18). Next, we prove the upper bound on the right-hand side of (18). Similarly to [27, p. 711] (or the proof of [28, Eq. (50)]), combining equations (10), (11) and Jensen's inequality give

$$\begin{aligned}
\big(d_{\mathrm{H}}(P,Q)\big)^2 &= 1 - Z(P,Q) \\
&= 1 - \sum_{x \in \mathcal{A}} P(x)\, e^{\frac{1}{2}\log\left(\frac{Q(x)}{P(x)}\right)} \\
&\leq 1 - e^{\frac{1}{2}\sum_{x\in\mathcal{A}} P(x)\log\left(\frac{Q(x)}{P(x)}\right)} \\
&= 1 - e^{-\frac{1}{2}D(P||Q)}
\end{aligned} \tag{22}$$

and the inequality $\big(d_{\mathrm{H}}(P,Q)\big)^2 \leq d_{\mathrm{TV}}(P,Q)$ is due to [29, Lemma 1]; its (somewhat simplified) proof is

$$\begin{aligned}
\big(d_{\mathrm{H}}(P,Q)\big)^2 &= \frac{1}{2}\sum_{x\in\mathcal{A}}\Big(\sqrt{P(x)} - \sqrt{Q(x)}\Big)^2 \\
&= \frac{1}{2}\sum_{x\in\mathcal{A}} |P(x)-Q(x)| \left(\frac{|\sqrt{P(x)} - \sqrt{Q(x)}|}{\sqrt{P(x)} + \sqrt{Q(x)}}\right) \\
&\leq \frac{1}{2}\sum_{x\in\mathcal{A}} |P(x)-Q(x)| = d_{\mathrm{TV}}(P,Q).
\end{aligned} \tag{23}$$

The two upper bounds on the squared Hellinger distance in (22) and (23) lead to the upper bound on the right-hand side of (18). The other bound on the Bhattacharyya distance in (19) follows from (11) and (18). ■

Discussion 1: The proof of Proposition 2 is elementary. It is interesting to realize that the sharpened lower bound on the Hellinger distance in (18), expressed in terms of the total variation distance, also follows from the more advanced result in (8) (see [25, Corollary 5.4]). To verify this, a combination of Corollary 1, (12) and (13) yields that

$$\begin{aligned}
\big(d_{\mathrm{H}}(P,Q)\big)^2 &\geq \frac{1 - d_{\mathrm{TV}}(P,Q)}{2}\left(1 - \sqrt{\frac{1+d_{\mathrm{TV}}(P,Q)}{1-d_{\mathrm{TV}}(P,Q)}}\right)^2 \\
&= 1 - \sqrt{1 - \big(d_{\mathrm{TV}}(P,Q)\big)^2}
\end{aligned}$$

which coincides with the lower bound on the left-hand side of (18). Similarly, the right-hand side of (19) follows from the equality in (11) and the left-hand side of (18).

Remark 1: Since the total variation distance $d_{\mathrm{TV}}(P,Q)$ and the Hellinger distance $d_{\mathrm{H}}(P,Q)$ are symmetric in P and Q, in contrast to the relative entropy $D(P||Q)$, one can improve the upper bound on the Hellinger distance in (18) to

$$d_{\mathrm{H}}(P,Q) \leq \sqrt{\min\left\{1 - \exp\left(-\frac{d}{2}\right),\, d_{\mathrm{TV}}(P,Q)\right\}} \tag{24}$$

where $d \triangleq \min\big\{D(P||Q),\, D(Q||P)\big\}$. From (11), the lower bound on the Bhattacharyya distance in (19) is improved to

$$Z(P,Q) \geq \max\left\{\exp\left(-\frac{d}{2}\right),\, 1 - d_{\mathrm{TV}}(P,Q)\right\}. \tag{25}$$

Remark 2: The bounds in (16) follow from a loosening of the bounds in (18) by a use of the inequalities $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for $x \in [0,1]$, and $e^{-x} \geq 1 - x$ for $x \geq 0$.

B. Bounds on the Chi-Squared Divergence

Proposition 3: Let P and Q be two probability distributions on a set $\mathcal{A}$. Then, the chi-squared divergence between P and Q is lower bounded in terms of the relative entropy as follows:

$$\chi^2(P,Q) \geq e^{D(P||Q)} - 1 \tag{26}$$

and, it is also lower bounded in terms of the total variation distance as follows:

$$\chi^2(P,Q) \geq \max\left\{ \frac{\big(1+d_{\mathrm{TV}}(P,Q)\big)^{d_{\mathrm{TV}}(P,Q)}}{1-\big(d_{\mathrm{TV}}(P,Q)\big)^2} - 1, \big(2d_{\mathrm{TV}}(P,Q)\big)^2 \right\}. \tag{27}$$

Furthermore, if $\mathcal{A}$ is a finite set, the following inequality holds:

$$\chi^2(P,Q) \leq \frac{2\big(d_{\mathrm{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}} Q(x)}. \tag{28}$$

Proof: From (17), it follows from Jensen's inequality that

$$\begin{aligned}
\chi^2(P,Q) &= \sum_{x\in\mathcal{A}} \Big\{ P(x) e^{\log\left(\frac{P(x)}{Q(x)}\right)} \Big\} - 1 \\
&\geq e^{\sum_{x\in\mathcal{A}} P(x)\,\log\left(\frac{P(x)}{Q(x)}\right)} - 1 \\
&= e^{D(P||Q)} - 1.
\end{aligned}$$

This proves the inequality in (26).

The second lower bound on the chi-squared divergence in (27), expressed in terms of the total variation distance, follows from a combination of the first lower bound in (26) with the improvement in [24] of Vajda's inequality:

$$D(P||Q) \geq \log\left(\frac{1}{1-d_{\mathrm{TV}}(P,Q)}\right) - \big(1-d_{\mathrm{TV}}(P,Q)\big)\,\log\big(1+d_{\mathrm{TV}}(P,Q)\big). \tag{29}$$

Furthermore, the inequality $\chi^2(P,Q) \geq \big(2d_{\mathrm{TV}}(P,Q)\big)^2$ is derived in [22, p. 429] using the Cauchy-Schwartz inequality:

$$\big(2d_{\mathrm{TV}}(P,Q)\big)^2 = \left(\sum_{x\in\mathcal{A}} \frac{|P(x)-Q(x)|}{\sqrt{Q(x)}} \cdot \sqrt{Q(x)}\right)^2 \leq \sum_{x\in\mathcal{A}} \frac{\big(P(x)-Q(x)\big)^2}{Q(x)} \cdot \sum_{x\in\mathcal{A}} Q(x) = \chi^2(P,Q).$$

It is verified numerically that the first term on the right-hand side of (27) improves the second term, $\big(2d_{\mathrm{TV}}(P,Q)\big)^2$, if the total variation distance satisfies $0.721 \leq d_{\mathrm{TV}}(P,Q) \leq 1$. This improvement is especially significant in the limit where the total variation distance tends to 1; in this limiting case, the first term of (27) tends to infinity, whereas the second term tends to 4. Hence, the new lower bound on the chi-squared divergence improves the existing bound for values of the total variation distance that lie between 0.721 and 1, where the improvement is especially significant as the value of the total variation distance gets closer to 1.

For the derivation of (28), note that

$$\begin{aligned}
\chi^2(P,Q) &= \sum_{x\in\mathcal{A}} \frac{\big(P(x)-Q(x)\big)^2}{Q(x)} \\
&\leq \frac{\sum_{x\in\mathcal{A}}\big(P(x)-Q(x)\big)^2}{\min_{x\in\mathcal{A}} Q(x)} \qquad (30) \\
&\leq \frac{\Big(\sum_{x\in\mathcal{A}} |P(x)-Q(x)|\Big)^2}{\min_{x\in\mathcal{A}} Q(x)} \\
&= \frac{4\big(d_{\mathrm{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}} Q(x)} \qquad (31)
\end{aligned}$$

where the last equality follows from (3). However, the upper bound in (28) is twice smaller than (31). In order to prove the tightened upper bound on the chi-squared divergence in (28), we rely on (30), and the following lemma:

Lemma 1: Let

$$d_{\text{loc}}(P,Q) \triangleq ||P-Q||_\infty = \sup_{x\in\mathcal{A}} |P(x)-Q(x)| \tag{32}$$

be the *local distance* between a pair of probability distributions P and Q on a set $\mathcal{A}$. Then, the inequality $d_{\text{loc}}(P,Q) \leq d_{\text{TV}}(P,Q)$ holds, which means that the l_∞-norm of $P-Q$ does not exceed *one-half* of its l_1-norm.

Proof: This known inequality follows from (2) and (3). ■

As a continuation to the proof of (28), it follows from (30) and Lemma 1 that

$$\begin{aligned}
\chi^2(P,Q) &\leq \frac{\sum_{x\in\mathcal{A}}\big(P(x)-Q(x)\big)^2}{\min_{x\in\mathcal{A}} Q(x)} \\
&\leq \frac{\max_{x\in\mathcal{A}}|P(x)-Q(x)| \ \sum_{x\in\mathcal{A}}|P(x)-Q(x)|}{\min_{x\in\mathcal{A}} Q(x)} \\
&\overset{\text{(a)}}{=} \frac{2\, d_{\text{loc}}(P,Q)\, d_{\text{TV}}(P,Q)}{\min_{x\in\mathcal{A}} Q(x)} \\
&\overset{\text{(b)}}{\leq} \frac{2\big(d_{\text{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}} Q(x)}
\end{aligned}$$

where equality (a) follows from (3) and (32) (note that $\mathcal{A}$ is a finite set), and inequality (b) follows from Lemma 1. To conclude, the upper bound on the chi-squared divergence in (28) is improved by a factor of 2, as compared to (31); this improvement is obtained by the use of Lemma 1, instead of the transition from (30) to (31). ■

Remark 3: Inequality (26) dates back to Dragomir and Gluščević (see [16, Theorem 4]).[1] The lower bound on the chi-squared divergence in (26) significantly improves the Csiszár-Györfi-Talata bound[2] in [13, Lemma 6.3] which states that $\chi^2(P,Q) \geq D(P||Q)$. Inequality (26) is refined in the continuation of this paper (see Corollary 5).

Remark 4: The transition from (a) to (b) in the derivation of the new upper bound in (28) implies that the improvement by a factor of 2 that is obtained there, as compared to (31), can be further enhanced under a mild condition. Specifically, a further improvement is obtained if the ratio $\frac{d_{\text{loc}}(P,Q)}{d_{\text{TV}}(P,Q)}$, which according to Lemma 1 is no more than 1, is strictly below 1 (for such possible examples, the reader is referred to [38, Section 4]); in this case, the improvement over the upper bound on the chi-squared divergence in (31) is by a factor of $\frac{2\, d_{\text{TV}}(P,Q)}{d_{\text{loc}}(P,Q)}$.

Remark 5: Inequalities (26) and (27) both imply that when the total variation distance tends to 1, the chi-squared divergence should necessarily tend to infinity (according to Vajda's-type inequalities in [24] and [43], if the total variation distance tends to 1 then the relative entropy tends to infinity). It is noted that the claimed achievability of the points on the parabolic curve in [26, Example 4.B] (see [26, Eq. (12)]) is partially inconsistent with the lower bound in (27) (the quadratic term that appears as a second term on the right-hand side of (27) is below the lower bound in (27) when the total variation distance lies between 0.721 and 1). Inequality (27) provides a lower bound on a non-symmetric f-divergence in terms of a symmetric f-divergence; for other such inequalities, see [41].

C. A New Reversed Pinsker's Inequality for Probability Distributions on a Finite Set

As a consequence of Proposition 3, a sort of a reversed Pinsker's inequality is obtained in the following.

Corollary 2: Let P and Q be two probability distributions on a finite set $\mathcal{A}$. Then, the following inequality holds:

$$D(P||Q) \leq \log\left(1 + \frac{2\big(d_{\text{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}} Q(x)}\right). \tag{33}$$

Proof: This result follows from the bounds on the chi-squared divergence in (26) and (28). ■

[1] Inequality (26) is missing a proof in [16]; it was recently proved in [39, Theorem 3.1], and it was derived independently in this work (before being aware of [16] and [39]).

[2] As a historical note, Györfi was acknowledged for pointing out the inequality $\chi^2(P,Q) \geq D(P||Q)$ in [13, Lemma 6.3]; this inequality was earlier stated in [10, Lemma 4] under a redundant requirement (see also [14, Lemma A.7], stated with a variant of this requirement).

Remark 6: The bound in (33) improves the bound that follows by combining Csiszár-Györfi-Talata bound in [13, Lemma 6.3] (see Remark 3) and the bound in (31). This combination gives the Csiszár-Györfi-Talata bound

$$\min_{x\in\mathcal{A}} Q(x)\, D(P||Q) \leq 4\big(d_{\mathrm{TV}}(P,Q)\big)^2. \tag{34}$$

The improvement that is suggested in (33) over (34) is twofold: the logarithm on the right-hand side of (33) follows from the lower bound on the chi-squared divergence in (26) (as compared to the inequality $\chi^2(P,Q) \geq D(P||Q)$ in [13, Lemma 6.3]); another improvement, obtained by a replacement of the factor 4 on the right-hand side of (34) by a factor 2 inside the logarithm on the right-hand side of (33), follows from the improvement of the upper bound on the chi-squared divergence in (28) over the bound in (31).

Note that when the distributions P and Q are close enough in total variation, the upper bounds on the relative entropy in (33) and (34) scale like the squared total variation distance (although the former bound improves the latter bound by a factor of 2).

Remark 7: In the context of Corollary 2, another inequality has been recently introduced by Verdú [45]:

$$d_{\mathrm{TV}}(P,Q) \geq \left(\frac{1-\beta}{\log\frac{1}{\beta}}\right) D(P||Q)$$

where $\beta^{-1} \triangleq \sup_{x\in\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$. The reader is also referred to [20, Lemma 3.10] where a related inequality is provided.

Remark 8: The lower bound on the chi-squared divergence in (27) is looser than the bound in (26) (due to the additional use of the inequality in (29) for the derivation of (27)); nevertheless, the bound in (27) is expressed in terms of the total variation distance, whereas the bound in (26) is expressed in terms of the relative entropy.

Remark 9: As an addition to Proposition 3, a parameterized upper bound on the chi-squared divergence is introduced in [25, Corollary 5.6] where this bound is expressed in terms of some power divergences.

Remark 10: A related problem to the result in Corollary 2 has been recently studied in [4]. Consider an arbitrary distribution Q, and an arbitrary $\varepsilon > 0$. The problem studied in [4] is the characterization of $D^*(\varepsilon, Q)$, defined as the infimum of $D(P||Q)$ over all distributions P that are at least ε-far away from Q in total variation; from Sanov's theorem, $D^*(\varepsilon, Q)$ is equal to the asymptotic exponential decay of the probability that the L_1 distance between the empirical distribution of a sequence of i.i.d. random variables and the true distribution (Q) is more than a specified value (2ε, according to (3)) [33, Section 3]. It is demonstrated in [4, Theorem 1] that $D^*(\varepsilon, Q)$ scales like $C\varepsilon^2 + O(\varepsilon^3)$ for a certain constant C (with explicit upper and lower bounds on C that match when Q is a 'balanced' distribution [4]). If the support of the distribution Q is a finite set $\mathcal{A}$, the linear scaling of $D^*(\varepsilon, Q)$ in ε^2 (for $\varepsilon \ll 1$) also follows from a combination of Corollary 2 and Pinsker's inequality. Corollary 2 further implies that, for such an atomic distribution Q,

$$D^*(\varepsilon, Q) \triangleq \inf_{P:\, d_{\mathrm{TV}}(P,Q)\geq\varepsilon} D(P||Q) \leq \log\left(1 + \frac{2\varepsilon^2}{\min_{x\in\mathcal{A}} Q(x)}\right).$$

It is noted that, for a certain class of distributions Q which includes all the non-atomic distributions on $\mathbb{R}$, a full characterization of $D^*(\varepsilon, Q)$ is provided in [4, Theorem 2]; for this class of distributions Q, it satisfies the equality

$$D^*(\varepsilon, Q) = L(\varepsilon) \triangleq \inf_{P,R:\ d_{\mathrm{TV}}(P,R)=\varepsilon} D(P||R)$$

where, due to the data processing theorem for the relative entropy, it is sufficient to restrict attention to (P, Q) defined on a binary alphabet (see [21, p. 1492]); also, whenever the infimum for $L(\varepsilon)$ is finite, it is also a minimum (see the proof of [21, Theorem 1]). The exact parametric equation of the curve $(\varepsilon, L(\varepsilon))_{0<\varepsilon<1}$ is introduced in [21, Eq. (3)]; due to (3), the total variation distance contains an extra factor of one-half (as compared to [21]), and consequently this parametric equation gets the form

$$\begin{aligned} \varepsilon(t) &= \frac{t}{2}\left[1 - \left(\coth(t) - \frac{1}{t}\right)^2\right], \qquad t > 0 \\ L\big(\varepsilon(t)\big) &= \log\left(\frac{t}{\sinh(t)}\right) + t\coth(t) - \left(\frac{t}{\sinh(t)}\right)^2. \end{aligned} \tag{35}$$

D. Bounds on the Capacitory Discrimination

The capacitory discrimination (a.k.a. the Jensen-Shannon divergence) is defined as follows:

Definition 8: Let P and Q be two probability distributions. The capacitory discrimination between P and Q is given by

$$\begin{aligned}\overline{C}(P,Q) &\triangleq D\left(P\,||\,\frac{P+Q}{2}\right)+D\left(Q\,||\,\frac{P+Q}{2}\right)\\ &= 2\left[H\left(\frac{P+Q}{2}\right)-\frac{H(P)+H(Q)}{2}\right].\end{aligned} \tag{36}$$

This information measure is studied in [18], [25], [31] and [42]. Due to the parallelogram identity for relative entropy (see, e.g., [11, Problem 3.20]), it follows that

$$\overline{C}(P,Q)=\min_{R}\big\{D(P||R)+D(Q||R)\big\}$$

where the minimization is w.r.t. all probability distributions R.

Proposition 4: The capacitory discrimination satisfies the following equality, for every $\varepsilon\in[0,1]$,

$$\inf_{P,Q:\ d_{\mathrm{TV}}(P,Q)=\varepsilon}\overline{C}(P,Q)=2\,D\left(\frac{1-\varepsilon}{2}\,||\,\frac{1}{2}\right) \tag{37}$$

where this infimum is achievable (i.e., it is a minimum), and $D(p||q)\triangleq p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)$ for $p,q\in[0,1]$ (with the convention that $0\log 0=0$).

If $\mathcal{A}$ is a finite set, the following inequality holds:

$$\overline{C}(P,Q)\le 2\ \min\left\{d_{\mathrm{TV}}(P,Q)\,\log 2,\,\log\left(1+\frac{\big(d_{\mathrm{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}}\big(P(x)+Q(x)\big)}\right)\right\}. \tag{38}$$

Proof: In [25, p. 119], the capacitory discrimination is expressed as an f-divergence where

$$f(x)=x\log x-(x+1)\log(1+x)+2\log 2,\quad x>0 \tag{39}$$

is a convex function with $f(1)=0$. The combination of (5) and (39) implies that

$$\begin{aligned}&\inf_{P,Q:\ d_{\mathrm{TV}}(P,Q)=\varepsilon}\overline{C}(P,Q)\\ &=(1-\varepsilon)\,f\left(\frac{1+\varepsilon}{1-\varepsilon}\right)-2\varepsilon f'(1)\\ &=(1+\varepsilon)\,\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)-2\,\log\left(\frac{2}{1-\varepsilon}\right)+2\log 2\\ &=(1+\varepsilon)\,\log(1+\varepsilon)+(1-\varepsilon)\,\log(1-\varepsilon)\\ &=2\left[\log 2-h\left(\frac{1-\varepsilon}{2}\right)\right]\\ &=2\,D\left(\frac{1-\varepsilon}{2}\,||\,\frac{1}{2}\right).\end{aligned}$$

The last equality holds since $D(p||\frac{1}{2})=\log 2-h(p)$ for $p\in[0,1]$ where h denotes the binary entropy function. This proves the lower bound in (37). The appendix provides an alternative proof of (37) which only relies on basics of information theory and the use of maximal coupling; this proof is of interest by itself, and is elementary.

The derivation of the upper bound in (38) relies on a combination of (33) (see Corollary 2), and the equality $d_{\mathrm{TV}}\left(P,\frac{P+Q}{2}\right)=d_{\mathrm{TV}}\left(Q,\frac{P+Q}{2}\right)=\frac{d_{\mathrm{TV}}(P,Q)}{2}$. This gives the inequality

$$\overline{C}(P,Q)\le 2\,\log\left(1+\frac{\big(d_{\mathrm{TV}}(P,Q)\big)^2}{\min_{x\in\mathcal{A}}\big(P(x)+Q(x)\big)}\right).$$

Since also $0 \leq \overline{C}(P,Q) \leq 2 \log 2 \, d_{\mathrm{TV}}(P,Q)$ (see [31, Theorem 3]), inequality (38) follows. ■

Discussion 2: The lower bound on the capacitory discrimination in (37), expressed in terms of the total variation distance, forms a closed-form expression of the bound by Topsøe in [42, Theorem 5]. This bound is given by

$$\overline{C}(P,Q) \geq \sum_{\nu=1}^{\infty} \frac{\big(d_{\mathrm{TV}}(P,Q)\big)^{2\nu}}{\nu(2\nu-1)}. \tag{40}$$

The equivalence of (37) and (40) follows from the power series expansion of the binary entropy function (to the natural base) around one-half:

$$h(x) = \log 2 - \sum_{\nu=1}^{\infty} \frac{(1-2x)^{2\nu}}{2\nu(2\nu-1)}, \quad \forall x \in [0,1]$$

which yields that

$$\begin{aligned} \sum_{\nu=1}^{\infty} \frac{\big(d_{\mathrm{TV}}(P,Q)\big)^{2\nu}}{\nu(2\nu-1)} &= 2\left[\log 2 - h\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2}\right)\right] \\ &= 2D\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2} \,||\, \frac{1}{2}\right). \end{aligned}$$

Note, however, that the proof here is much shorter than the proof of [42, Theorem 5] (which relies on properties of the triangular discrimination in [42] and previous theorems of this paper), and it also leads directly to a closed-form expression of this bound. Consequently, one concludes that the lower bound in [42, Theorem 5] is a special case of (8) (see [23] and [25, Corollary 5.4]), which provides a lower bound on a symmetric f-divergence in terms of the total variation distance. The lower bound on the capacitory discrimination was obtained independently of the work by Briët and Harremoës (see [6, Eq. (18)] for $\alpha = 1$) whose derivation was done in a different approach.

The upper bound on the capacitory discrimination in (38) is new, and it is based on Corollary 2 which provides an improvement of the Csiszár-Györfi-Talata bound (see Remarks 3, 6). The upper bound $\overline{C}(P,Q) \leq 2 \log 2 \, d_{\mathrm{TV}}(P,Q)$ is looser, and it was derived independently in [6, Theorem 9] (as a special case where $\alpha \to 1$) and in [31, Theorem 3].

The following result provides a measure of the concavity of the entropy function:

Corollary 3: For arbitrary probability distributions P and Q, the following inequality holds:

$$H\left(\frac{P+Q}{2}\right) - \frac{H(P)+H(Q)}{2} \geq D\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2} \,||\, \frac{1}{2}\right)$$

and this inequality is tight for a given total variation distance.

Proof: This result follows from (36) and Proposition 4. ■

E. An Exact Characterization of the Minimum of the Chernoff Information for a Given Total Variation Distance

Proposition 5: Let $C(P,Q)$ denote the Chernoff information between two probability distributions P and Q (see (14)), and let

$$C(\varepsilon) \triangleq \inf_{P,Q\colon \, d_{\mathrm{TV}}(P,Q)=\varepsilon} C(P,Q), \quad \forall\, \varepsilon \in [0,1] \tag{41}$$

be the infimum of the Chernoff information for a given total variation distance (ε). Then, the following equality holds:

$$C(\varepsilon) = \begin{cases} -\frac{1}{2}\log(1-\varepsilon^2) & \text{if } \varepsilon \in [0,1) \\ +\infty & \text{if } \varepsilon = 1. \end{cases} \tag{42}$$

For $\varepsilon \in [0,1)$, the infimum in (41) is achievable by the pair of probability distributions

$$P = \left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right), \quad Q = \left(\frac{1+\varepsilon}{2}, \frac{1-\varepsilon}{2}\right) \tag{43}$$

so, the infimum in (41) is a minimum that is obtained by the pair of 2-element probability distributions in (43).

Proof:

$$\begin{aligned} C(P,Q) &\overset{\text{(a)}}{\geq} -\log\left(\sum_{x\in\mathcal{A}}\sqrt{P(x)\,Q(x)}\right) \\ &\overset{\text{(b)}}{=} -\log\, Z(P,Q) \\ &\overset{\text{(c)}}{\geq} -\frac{1}{2}\log\Big(1-\big(d_{\mathrm{TV}}(P,Q)\big)^2\Big) \end{aligned}$$

where inequality (a) follows by selecting the possibly sub-optimal choice $\lambda = \frac{1}{2}$ in (14), equality (b) holds by definition (see (10)), and inequality (c) follows from the right-hand side of (19). By the definition in (41), it follows that $C(\varepsilon)$ satisfies the inequality

$$C(\varepsilon) \geq -\frac{1}{2}\,\log(1-\varepsilon^2). \tag{44}$$

In order to show that (44) provides a tight lower bound for a given total variation distance (ε), it is sufficient to show the existence of a pair of probability distributions P and Q where $d_{\mathrm{TV}}(P,Q)=\varepsilon$ and $C(P,Q) = -\frac{1}{2}\,\log(1-\varepsilon^2)$. For the pair of 2-element probability distributions P and Q in (43), the Chernoff information in (14) satisfies the equality

$$C(P,Q) = -\min_{\lambda\in[0,1]}\log\left(\frac{1-\varepsilon}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{\lambda}+\frac{1+\varepsilon}{2}\left(\frac{1-\varepsilon}{1+\varepsilon}\right)^{\lambda}\right). \tag{45}$$

Minimization of the logarithmic function in (45), by setting its derivative to zero, gives $\lambda = \frac{1}{2}$. For the pair of 2-element probability distributions P and Q in (43) with $\lambda = \frac{1}{2}$, the Chernoff information is equal to

$$C(P,Q) = -\frac{1}{2}\,\log(1-\varepsilon^2)$$

so, the lower bound on $C(\varepsilon)$ in (44) is tight. This concludes the proof of Proposition 5. ■

Corollary 4: For any pair of probability distributions P and Q, the Chernoff information between P and Q satisfies

$$C(P,Q) \geq -\frac{1}{2}\log\Big(1-\big(d_{\mathrm{TV}}(P,Q)\big)^2\Big). \tag{46}$$

and (46) is obtained with equality for the pair of 2-element probability distributions in (43) where $d_{\mathrm{TV}}(P,Q)=\varepsilon$.

Proof: Inequality (46) follows directly from the equality in (42), and it turns to hold with equality for the 2-element probability distributions in (43) where $d_{\mathrm{TV}}(P,Q)=\varepsilon$. ■

Remark 11: The fact that, subject to a given total variation distance, the Chernoff information achieves its minimum by a pair of 2-element probability distributions can be also justified by the same reasoning as in [21, first paragraph of Section 2]. The reasoning in [21] refers to a minimization of the relative entropy, subject to the same equality constraint on the total variation distance, and it is a simple consequence of the data processing theorem for the relative entropy. The same concept of proof also applies to the minimization of the Chernoff information, for a given total variation distance, since the Chernoff information also satisfies a data processing theorem. The satisfiability of a data processing theorem by the Chernoff information can be justified by combining the data processing theorem for the Rényi divergence (see [19, Theorem 1]) with equation (15) that relates the Chernoff information to the Rényi divergence.

Remark 12: Following Corollary 4, a lower bound on the total variation distance gives a lower bound on the Chernoff information; consequently, it provides an upper bound on the best achievable Bayesian probability of error for binary hypothesis testing (see, e.g., [7, Theorem 11.9.1]). This approach was recently used in [46] to obtain a lower bound on the Chernoff information for studying a communication problem.

Discussion 3: Let

$$L(\varepsilon) \triangleq \inf_{P,Q:\ d_{\mathrm{TV}}(P,Q)=\varepsilon} D(P||Q). \tag{47}$$

The exact parametric equation of the curve $(\varepsilon, L(\varepsilon))_{0<\varepsilon<1}$ is introduced in [21, Eq. (3)] (see (35) in Remark 10). From the satisfiability of the inequality (see [7, Section 11.9])

$$C(P,Q) \leq \min\big\{D(P||Q), D(Q||P)\big\} \tag{48}$$

it follows from (41), (47) and (48) that

$$C(\varepsilon) \leq L(\varepsilon), \quad \forall\, \varepsilon \in [0, 1) \tag{49}$$

where the left and right-hand sides of (49) correspond to the minima of the Chernoff information and relative entropy, respectively, given the value of the total variation distance (ε). Figure 1 plots these minima as a function of the total variation distance, supporting inequality (49). For small values of ε, $C(\varepsilon)$ and $L(\varepsilon)$, respectively, are approximately equal to $\frac{\varepsilon^2}{2}$ and $2\varepsilon^2$ (note that Pinsker's inequality is tight for $\varepsilon \ll 1$), so $\lim_{\varepsilon \to 0} \frac{L(\varepsilon)}{C(\varepsilon)} = 4$.

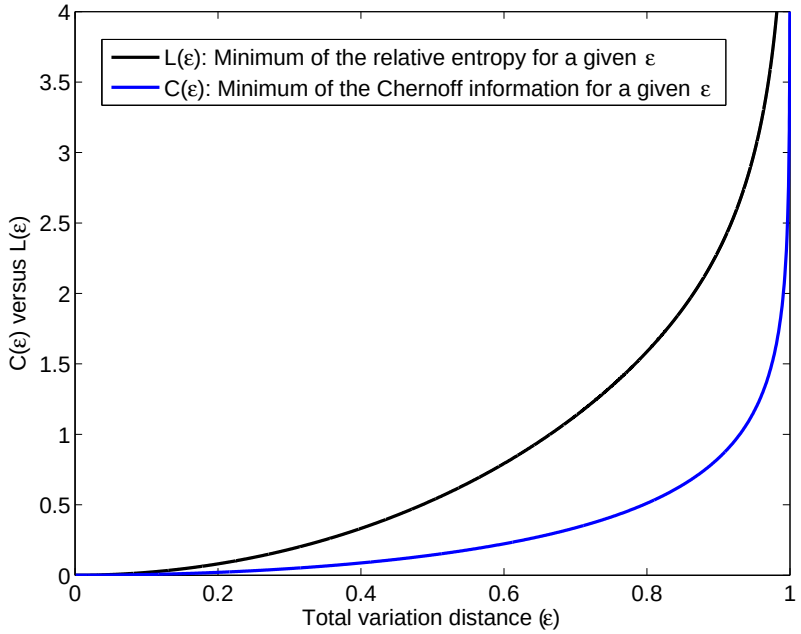

Fig. 1. A plot of the minima of the Chernoff information and the relative entropy for a given total variation distance $\varepsilon \in [0, 1]$, denoted by $C(\varepsilon)$ and $L(\varepsilon)$, respectively; C and L are given in Proposition 5 and [21, Theorem 2] (see (35)).

F. On Jeffreys' Divergence and Lossless Source Coding

Definition 9: Let P and Q be two probability distributions. Jeffreys' divergence is a symmetrized version of the relative entropy, which is defined as

$$J(P, Q) \triangleq \frac{D(P||Q) + D(Q||P)}{2}. \tag{50}$$

It is easy to verify that it is a symmetric f-divergence where

$$f(t) = \frac{(t-1)\log(t)}{2}, \quad t > 0 \tag{51}$$

is a convex function on $(0, \infty)$ with $f(1) = 0$. Relying on [23], [24], the following equalities hold:

Proposition 6:

$$\inf_{P,Q:\, d_{\mathrm{TV}}(P,Q)=\varepsilon} J(P, Q) = \varepsilon \, \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right), \quad \forall\, \varepsilon \in [0, 1), \tag{52}$$

$$\inf_{P,Q:\, D(P||Q)=\varepsilon} J(P, Q) = \frac{\varepsilon}{2}, \quad \forall\, \varepsilon > 0, \tag{53}$$

and the two respective suprema are equal to $+\infty$.

Proof: Jeffreys' divergence is a symmetric f-divergence where the convex function f in (51) satisfies the equality $f(t) = tf(\frac{1}{t})$ for every $t > 0$ with $f(1) = 0$. The equality in (52) follows from Corollary 1. Eq. (53) follows from (50) and the fact that, given the value of the relative entropy $D(P||Q)$, its dual $(D(Q||P))$ can be

made arbitrarily small. The two corresponding suprema are equal to infinity because given the value of the total variation distance or the relative entropy, the dual of the relative entropy can be made arbitrarily large. ■

We exemplify in the following a use of Proposition 6 for lossless source coding. This tightens, and also refines under a certain condition, a bound by Csiszár [9].

Consider a memoryless and stationary source with alphabet $\mathcal{U}$ that emits symbols according to a probability distribution P, and assume a uniquely decodable (UD) code with an alphabet of size d. It is well known that such a UD code achieves the entropy of the source if and only if the length $l(u)$ of the codeword that is assigned to each symbol $u \in \mathcal{U}$ satisfies the equality

$$l(u) = -\log_d P(u), \quad \forall\, u \in \mathcal{U}.$$

This corresponds to a dyadic source where, for every $u \in \mathcal{U}$, we have $P(u) = d^{-n_u}$ with a natural number n_u; in this case, $l(u) = n_u$ for every symbol $u \in \mathcal{U}$. Let $\overline{L} \triangleq \mathbb{E}[L]$ designate the average length of the codewords, and $H_d(U) \triangleq -\sum_{u\in\mathcal{U}} P(u)\, \log_d P(u)$ be the entropy of the source (to the base d). Furthermore, let $c_{d,l} \triangleq \sum_{u\in\mathcal{U}} d^{-l(u)}$. According to the Kraft-McMillian inequality (see [7, Theorem 5.5.1]), the inequality $c_{d,l} \leq 1$ holds in general for UD codes, and the equality $c_{d,l} = 1$ holds if and only if the code achieves the entropy of the source (i.e., $\overline{L} = H_d(U)$). Hence, for a UD code that achieves the entropy of the source, the probability distribution P satisfies the equality

$$P(u) = \left(\frac{1}{c_{d,l}}\right) d^{-l(u)}, \quad \forall\, u \in \mathcal{U}. \tag{54}$$

Note that the right-hand side of (54) is in general a probability distribution. Let's designate it by $Q_{d,l}$, i.e.,

$$Q_{d,l}(u) \triangleq \left(\frac{1}{c_{d,l}}\right) d^{-l(u)}, \quad \forall\, u \in \mathcal{U} \tag{55}$$

and let $\Delta_d \triangleq \overline{L} - H_d(U)$ designate the redundancy of the code.

In [9], a generalization for UD source codes has been studied by a derivation of an upper bound on the L_1 norm between the two probability distributions P and $Q_{d,l}$ as a function of the redundancy Δ_d of the code. To this end, straightforward calculation shows that the relative entropy from P to $Q_{d,l}$ is given by

$$D(P||Q_{d,l}) = \Delta_d\, \log d + \log\bigl(c_{d,l}\bigr). \tag{56}$$

The interest in [9] is in getting an upper bound that only depends on the (average) redundancy Δ_d of the code, but is independent of the specific distribution of the length of each codeword. Hence, since the Kraft-McMillian inequality states that $c_{d,l} \leq 1$ for general UD codes, it is concluded in [9] that

$$D(P||Q_{d,l}) \leq \Delta_d\, \log d. \tag{57}$$

Consequently, it follows from Pinsker's inequality that

$$\sum_{u\in\mathcal{U}} \bigl|P(u) - Q_{d,l}(u)\bigr| \leq \min\bigl\{\sqrt{2\Delta_d \log d},\, 2\bigr\} \tag{58}$$

where it is also taken into account that, from the triangle inequality, the sum on the left-hand side of (58) cannot exceed 2. This inequality is indeed consistent with the fact that the probability distributions P and $Q_{d,l}$ coincide when $\Delta_d = 0$ (i.e., for a UD code which achieves the entropy of the source).

At this point we deviate from the analysis in [9]. One possible improvement of the bound in (58) follows by replacing Pinsker's inequality with the result in [21], i.e., by taking into account the exact parametrization of the infimum of the relative entropy for a given total variation distance. This gives the following tightened bound:

$$\sum_{u\in\mathcal{U}} \bigl|P(u) - Q_{d,l}(u)\bigr| \leq 2\, \varepsilon\bigl(L^{-1}(\Delta_d \log d)\bigr) \tag{59}$$

where the parametric functions ε and L are introduced in (35), and L^{-1} is the inverse function of L (calculated numerically).

In the following, the use of Proposition 6 is exemplified in refining the latter bound in (59). Let

$$\delta(u) \triangleq l(u) + \log_d P(u), \quad \forall\, u \in \mathcal{U}.$$

Calculation of the dual divergence gives

$$\begin{aligned}
&D(Q_{d,l}||P)\\
&= \log d \sum_{u\in\mathcal{U}} Q_{d,l}(u) \log_d\left(\frac{Q_{d,l}(u)}{P(u)}\right)\\
&= \log d \left[-\frac{\log_d(c_{d,l})}{c_{d,l}} \sum_{u\in\mathcal{U}} d^{-l(u)} - \frac{1}{c_{d,l}} \sum_{u\in\mathcal{U}} l(u) d^{-l(u)} - \frac{1}{c_{d,l}} \sum_{u\in\mathcal{U}} \log_d P(u)\, d^{-l(u)}\right]\\
&= -\log(c_{d,l}) - \frac{\log d}{c_{d,l}} \sum_{u\in\mathcal{U}} \delta(u)\, d^{-l(u)}\\
&= -\log\big(c_{d,l}\big) - \frac{\log d}{c_{d,l}} \sum_{u\in\mathcal{U}} P(u)\,\delta(u)\, d^{-\delta(u)}\\
&= -\log\big(c_{d,l}\big) - \left(\frac{\log d}{c_{d,l}}\right) \mathbb{E}\big[\delta(U)\, d^{-\delta(U)}\big] \qquad (60)
\end{aligned}$$

and the combination of (50), (56) and (60) yields that

$$J(P, Q_{d,l}) = \frac{1}{2}\left[\Delta_d \log d - \left(\frac{\log d}{c_{d,l}}\right) \mathbb{E}\big[\delta(U)\, d^{-\delta(U)}\big]\right]. \qquad (61)$$

For the simplicity of the continuation of the analysis, we restrict our attention to UD codes that satisfy the condition

$$l(u) \geq \left\lceil \log_d \frac{1}{P(u)} \right\rceil, \quad \forall\, u \in \mathcal{U}. \qquad (62)$$

In general, it excludes Huffman codes; nevertheless, it is satisfied by some other important UD codes such as the Shannon code, Shannon-Fano-Elias code, and arithmetic coding (see, e.g., [7, Chapter 5]). Since (62) is equivalent to the condition that δ is non-negative on $\mathcal{U}$, it follows from (61) that

$$J(P, Q_{d,l}) \leq \frac{\Delta_d \log d}{2} \qquad (63)$$

so, the upper bound on Jeffreys' divergence in (63) is twice smaller than the upper bound on the relative entropy in (57). It is partially because the term $\log c_{d,l}$ is canceled out along the derivation of the bound in (63), in contrast to the derivation of the bound in (57) where this term was removed from the bound in order to avoid its dependence on the length of the codeword for each individual symbol.

Following Proposition 6, for an arbitrary $x \geq 0$, let $\varepsilon \triangleq \varepsilon(x)$ be the solution in the interval $[0, 1)$ of the equation

$$\varepsilon\, \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) = x. \qquad (64)$$

The combination of (52) and (63) implies that

$$\sum_{u\in\mathcal{U}} \big|P(u) - Q_{d,l}(u)\big| \leq 2\, \varepsilon\left(\frac{\Delta_d \log d}{2}\right). \qquad (65)$$

The bounds in (58), (59) and (65) are depicted in Figure 2 for UD codes where the size of their alphabet is $d = 10$.

In the following, the bounds in (59) and (65) are compared analytically for the case where the average redundancy is small (i.e., $\Delta_d \approx 0$). Under this approximation, the bound in (58) (i.e., the original bound from [9]) coincides with its tightened version in (65). On the other hand, since for $\varepsilon \approx 0$, the left-hand side of (64) is approximately $2\varepsilon^2$, it follows from (64) that, for $x \approx 0$, we have $\varepsilon(x) \approx \sqrt{\frac{x}{2}}$. It follows that, if $\Delta_d \approx 0$, inequality (65) gets approximately the form

$$\sum_{u\in\mathcal{U}} \big|P(u) - Q_{d,l}(u)\big| \leq \sqrt{\Delta_d \log d}.$$

Hence, even for a small redundancy, the bound in (65) improves (58) by a factor of $\sqrt{2}$. This conclusion is consistent with the plot in Figure 2.

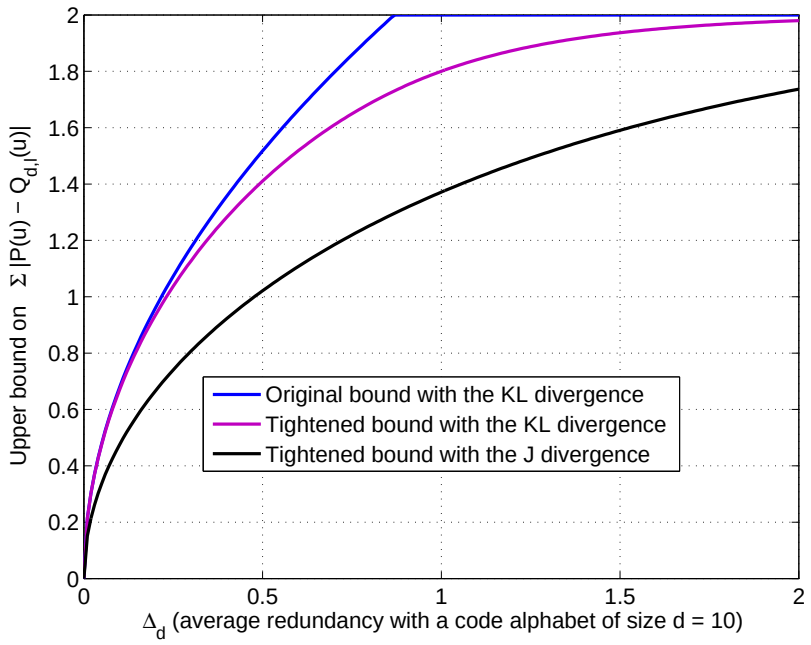

Fig. 2. Upper bounds on $\sum |P(u) - Q_{d,l}(u)|$ as a function of the (average) redundancy $\Delta_d \triangleq \mathbb{E}[L] - H_d$ for a UD code with an alphabet of size $d = 10$. The original bound in (58) appears in [9], and the tightened bound that relies on the Kullback-Leibler (KL) divergence is given in (59). The further tightening of this bound is restricted in this plot to UD codes whose codewords satisfy the condition in (62). The latter bound relies on Proposition 6 for Jeffreys' (J) divergence, and it is given in (65).

G. A New Inequality Relating f-Divergences

We introduce in the following a new inequality which relates f-divergences, and study some of its consequences.

Proposition 7: Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and further assume that the function $g\colon (0,\infty) \to \mathbb{R}$, defined by $g(t) = -tf(t)$ for every $t > 0$, is also convex. Let P and Q be two probability distributions on a finite set $\mathcal{A}$, and assume that P, Q are positive on this set. Then, the following inequality holds:

$$\min_{x\in\mathcal{A}} \frac{P(x)}{Q(x)} \cdot D_f(P||Q) \leq -D_g(P||Q) - f\big(1+\chi^2(P,Q)\big) \leq \max_{x\in\mathcal{A}} \frac{P(x)}{Q(x)} \cdot D_f(P||Q). \tag{66}$$

Proof: Let $|\mathcal{A}| = n$ be the size of the finite set $\mathcal{A}$, and let $\mathcal{A} = \{x_1, \ldots, x_n\}$. Let $\underline{u} = (u_1, \ldots, u_n) \in \mathbb{R}_+^n$ be an arbitrary n-tuple with positive entries. Define

$$\begin{aligned} J_n(f, \underline{u}, P) &\triangleq \sum_{i=1}^n P(x_i)\, f(u_i) - f\left(\sum_{i=1}^n P(x_i) u_i\right), \\ J_n(Q, \underline{u}, P) &\triangleq \sum_{i=1}^n Q(x_i)\, f(u_i) - f\left(\sum_{i=1}^n Q(x_i) u_i\right). \end{aligned} \tag{67}$$

The following refinement of Jensen's inequality has been proved in [17, Theorem 1] for a convex function $f\colon (0,\infty) \to \mathbb{R}$ (and it was extended in [2, Theorem 1] to hold for a convex f over an arbitrary interval $[a,b]$):

$$\min_{i\in\{1,\ldots,n\}} \left\{\frac{P(x_i)}{Q(x_i)}\right\} J_n(f, \underline{u}, Q) \leq J_n(f, \underline{u}, P) \leq \max_{i\in\{1,\ldots,n\}} \left\{\frac{P(x_i)}{Q(x_i)}\right\} J_n(f, \underline{u}, Q). \tag{68}$$

The refined version of Jensen's inequality in (68) is applied in the following to prove (66). Let

$$u_i \triangleq \frac{P(x_i)}{Q(x_i)}, \quad \forall\, i \in \{1, \ldots, n\}.$$

Calculation of (67) gives that

$$\begin{aligned} &J_n(f, \underline{u}, Q) \\ &= \sum_{i=1}^{n} Q(x_i)\, f\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} Q(x_i) \cdot \frac{P(x_i)}{Q(x_i)}\right) \\ &= \sum_{x \in \mathcal{A}} Q(x)\, f\left(\frac{P(x)}{Q(x)}\right) - f(1) \\ &= D_f(P||Q) \end{aligned} \tag{69}$$

and

$$\begin{aligned} &J_n(f, \underline{u}, P) \\ &= \sum_{i=1}^{n} P(x_i)\, f\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(x_i)^2}{Q(x_i)}\right) \\ &\overset{\text{(a)}}{=} -\sum_{i=1}^{n} Q(x_i)\, g\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(x_i)^2}{Q(x_i)}\right) \\ &\overset{\text{(b)}}{=} -D_g(P||Q) - f\big(1 + \chi^2(P, Q)\big) \end{aligned} \tag{70}$$

where equality (a) holds by the definition of g, and equality (b) follows from equalities (1) and (17). The substitution of (69) and (70) in (68) completes the proof. ■

As a consequence of Proposition 7, we introduce the following inequality which relates between the relative entropy, its dual and the chi-squared divergence.

Corollary 5: Let P and Q be two probability distributions on a finite set $\mathcal{A}$, and assume that P, Q are positive on $\mathcal{A}$. Then, the following inequality holds:

$$\min_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D(Q||P) \leq \log\big(1 + \chi^2(P, Q)\big) - D(P||Q) \leq \max_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D(Q||P). \tag{71}$$

Proof: Let $f(t) = -\log(t)$ for $t > 0$. The function $f\colon (0, \infty) \to \mathbb{R}$ is convex with $f(1) = 0$. Furthermore, $g(t) = -tf(t) = t\log(t)$ for $t > 0$ defines a convex function with $g(1) = 0$. The inequality in (71) follows by substituting f and g in (66) where $D_f(P||Q) = D(Q||P)$ and $D_g(P||Q) = D(P||Q)$. ■

Remark 13: Inequality (71) forms a refinement of (26). Combining it with (28) also refines the inequality in (33), giving

$$D(P||Q) + \min_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D(Q||P) \leq \log\left(1 + \frac{2\big(d_{\text{TV}}(P, Q)\big)^2}{\min_{x \in \mathcal{A}} Q(x)}\right). \tag{72}$$

The following inequality is another consequence of Proposition 7, relating the chi-squared divergence and its dual:

Corollary 6: Under the same conditions of Corollary 5, the following inequality holds:

$$\min_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot \chi^2(Q, P) \leq \frac{\chi^2(P, Q)}{1 + \chi^2(P, Q)} \leq \max_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot \chi^2(Q, P). \tag{73}$$

Proof: The parametric function $f(t) = t^\alpha - 1$ satisfies the conditions in Proposition 7 for $\alpha \in [-1, 0]$. For $\alpha = -1$, the inequality in (73) follows from (66) where

$$D_f(P||Q) = \chi^2(Q, P), \quad D_g(P||Q) = 0.$$

■

IV. SUMMARY

Derivation of bounds on f-divergences and related distances is considered in this paper. In some cases, existing recent bounds are reproduced by elementary proofs, and in some other cases, significant improvements are obtained. The contributions of this work are outlined in the following:

- Upper and lower bounds on both the Hellinger distance and the Bhattacharyya distance are expressed in terms of the total variation distance and relative entropy (see Proposition 2). This tightens the bounds introduced in Proposition 1 (see, e.g., [35]). The proof of these bounds is elementary, replacing an alternative proof that is based on an advanced result (see Discussion 1). The simple proof of the other two bounds dates back to Hoeffding and Wolfowitz [27], and Kraft [29].
- Three bounds on the chi-squared divergence are introduced in Proposition 3. The first lower bound in (26) dates back to Dragomir and Gluščević [16] (see Remark 3). A second lower bound on the chi-squared divergence is expressed in terms of the total variation distance (see (27)); this new bound improves the bound in [22, p. 429] when the total variation distance lies between 0.721 and 1, and the improvement is especially significant when the total variation distance tends to 1 (where the existing lower bound tends to 4, whereas the new bound tends to infinity). The upper bound on the chi-squared divergence in (28) is new as well, and it suggests an improvement over the bound in (31) by a factor of 2 (according to Remark 4, this gain can be further improved under a mild condition).
- The improvements of the bounds on the chi-squared divergence in Proposition 3 lead to a new improved upper bound on the relative entropy in terms of the total variation distance for an arbitrary pair of probability distributions on a finite set (see Corollary 2, followed by Remarks 6–10). This forms a new sort of a reversed Pinsker's inequality which improves the Csiszár-Györfi-Talata bound in (34).
- Bounds on the capacitory discrimination are provided in terms of the total variation distance (see Proposition 4). The lower bound on the capacitory discrimination forms a closed-form expression of the bound by Topsøe in [42, Theorem 5]; it has two proofs in this paper: the first proof relies on an advanced result (see [23] and [25, Corollary 5.4]), and the second proof relies on basics of information theory and coupling between random variables (see the appendix). Both proofs do not involve properties of the triangular discrimination that are used in the original proof in [42]. The lower bound on the capacitory discrimination was obtained independently by Briët and Harremoës (see [6, Eq. (18)] for $\alpha = 1$) with a different approach. Furthermore, the upper bound on the capacitory discrimination in (38) is new (see Discussion 2), and it sharpens a bound that was derived in [6, Theorem 9] and [31, Theorem 3].
- Proposition 5 provides an exact characterization of the minimum of the Chernoff information for a given total variation distance, which is obtained by a pair of 2-element probability distributions. The minima of the Chernoff information and the relative entropy for a given total variation distance are plotted in Figure 1 where the former is less than or equal to the latter (see (49)), and their ratio is approximately 4 for small values of the total variation distance. The lower bound on the Chernoff information for a given total variation distance (see Corollary 4) is therefore tight, and it is achieved with equality for a pair of 2-element probability distributions. This lower bound has been recently applied in [46] in the context of a channel codebook detection in a binary hypothesis testing problem where the receiver needs to detect the channel code upon observing noise-affected codewords through a noisy channel (the authors referred to the bound in [37, Proposition 5] (un-published), which was stated there as a lower bound without proving its tightness for a given total variation distance).
- A lower bound on Jeffreys' divergence in terms of the total variation distance was readily obtained from the analysis by Gilardoni ([23], [24]). This bound was used in Section III-F to tighten a bound by Csiszár in the context of lossless source coding [9] (the original and tightened bounds are plotted in Figure 2).
- A new inequality which relates f-divergences was derived in Proposition 7, based on a refinement of Jensen's inequality in [2] and [17]. Corollaries of Proposition 7 include an inequality relating the relative entropy, its dual and the chi-squared divergence, and another inequality which relates the chi-squared divergence and its dual (see Corollaries 5 and 6, respectively).

ACKNOWLEDGMENT

This research work was supported by the Israeli Science Foundation (ISF), grant number 12/12.

APPENDIX: AN ELEMENTARY PROOF OF (37)

The following proof of (37) relies on basics of information theory, and coupling between random variables. By the definition of the capacitory discrimination in (36), it follows that for any pair of probability distributions (P and Q) on an arbitrary set $\mathcal{A}$

$$\begin{aligned}
&\overline{C}(P,Q) \\
&= D\left(P \,||\, \frac{P+Q}{2}\right) + D\left(Q \,||\, \frac{P+Q}{2}\right) \\
&= 2\log 2 + \sum_{x\in\mathcal{A}} P(x)\,\log\left(\frac{P(x)}{P(x)+Q(x)}\right) + \sum_{x\in\mathcal{A}} Q(x)\,\log\left(\frac{Q(x)}{P(x)+Q(x)}\right) \qquad (74)\\
&= 2\left[\log 2 - \sum_{x\in\mathcal{A}} \left(\frac{P(x)+Q(x)}{2}\right)\, h\left(\frac{P(x)}{P(x)+Q(x)}\right)\right]
\end{aligned}$$

where h denotes the binary entropy function.

Let Θ, X_1 and X_2 be random variables where $X_1 \sim P$, $X_2 \sim Q$, and $\Theta \sim \mathrm{Ber}(\frac{1}{2})$ is a Bernoulli random variable that gets the values 1 or 2 with equal probability ($\frac{1}{2}$). Further assume that (X_1, X_2) is independent of Θ. A basic result on a coupling $(\hat{X}_1, \hat{X}_2)$ of (X_1, X_2) (see, e.g., [36, Proposition 2.7] or [38, Theorem 2]) states that, since $\hat{X}_1 \sim P$ and $\hat{X}_2 \sim Q$,

$$\Pr(\hat{X}_1 = \hat{X}_2) \leq 1 - d_{\mathrm{TV}}(P,Q) \qquad (75)$$

and equality in (75) holds in the case of maximal coupling (for such a construction of maximal coupling, the reader is referred, e.g., to [36, p. 58] or [38, p. 7119]).

Let $\hat{X}_\Theta$ be equal to $\hat{X}_1$ or $\hat{X}_2$ when $\Theta = 1$ or $\Theta = 2$, respectively. Then, for every $x \in \mathcal{A}$,

$$\Pr(\hat{X}_\Theta = x) = \frac{P(x)+Q(x)}{2}\,, \qquad (76)$$

$$\Pr(\Theta = 1 \,|\, \hat{X}_\Theta = x) = \frac{\Pr(\Theta=1,\ \hat{X}_1 = x)}{\Pr(\hat{X}_\Theta = x)} = \frac{\Pr(\Theta=1)\ \Pr(\hat{X}_1 = x)}{\Pr(\hat{X}_\Theta = x)} = \frac{P(x)}{P(x)+Q(x)}\,, \qquad (77)$$

$$\Pr(\Theta = 2 \,|\, \hat{X}_\Theta = x) = 1 - \Pr(\Theta = 1 \,|\, \hat{X}_\Theta = x) = \frac{Q(x)}{P(x)+Q(x)}\,. \qquad (78)$$

The combination of (74)–(78) gives

$$\begin{aligned}
\overline{C}(P,Q) \ &= 2\left[\log 2 - \sum_{x\in\mathcal{A}} \Pr(\hat{X}_\Theta = x)\, H(\Theta \,|\, \hat{X}_\Theta = x)\right] \\
&= 2\Big[\log 2 - H(\Theta \,|\, \hat{X}_\Theta)\Big]. \qquad (79)
\end{aligned}$$

Let $\widetilde{\Theta}\colon \mathcal{A} \to \{1,2\}$ be an arbitrary estimator of Θ, based on the value of $\hat{X}_\Theta$, and let $P_{\mathrm{e}} = \mathbb{E}\big[\Pr(\Theta \neq \widetilde{\Theta} \,|\, \hat{X}_\Theta)\big]$ denote the average probability of error given $\hat{X}_\Theta$. Since Θ is a Bernoulli random variable, it follows that

$$H(\Theta \,|\, \hat{X}_\Theta) = h(P_{\mathrm{e}}). \qquad (80)$$

Let E be a Bernoulli random variable that gets the value 1 in case of an error event in the estimation of Θ (i.e., $E = 1$ if $\Theta \neq \widetilde{\Theta}$), and it is zero otherwise. The average probability of error is equal to

$$P_{\mathrm{e}} = \mathbb{E}\big[\Pr(E = 1 \,|\, \hat{X}_\Theta)\big] \qquad (81)$$

and

$$\Pr(E = 1 \,|\, \hat{X}_\Theta) = \Pr(E = 1 \,|\, \hat{X}_\Theta,\ \hat{X}_1 = \hat{X}_2)\ \Pr(\hat{X}_1 = \hat{X}_2) + \Pr(E = 1,\ \hat{X}_1 \neq \hat{X}_2 \,|\, \hat{X}_\Theta). \qquad (82)$$

If $\hat{X}_1 = \hat{X}_2$, the knowledge of the value of $\hat{X}_\Theta$ does not help in estimating Θ, and

$$\Pr(E = 1 \,|\, \hat{X}_\Theta,\ \hat{X}_1 = \hat{X}_2) = \frac{1}{2} \qquad (83)$$

since Θ is equally likely to be either 1 or 2. Consider the second term on the right-hand side of (82):

$$\Pr(E=1,\,\hat{X}_1\neq\hat{X}_2\,|\,\hat{X}_\Theta)=\sum_{\hat{x}_1,\hat{x}_2\,:\,\hat{x}_1\neq\hat{x}_2}\Big\{\Pr(E=1\,|\,\hat{X}_\Theta,\,\hat{X}_1=\hat{x}_1,\,\hat{X}_2=\hat{x}_2)\cdot\Pr(\hat{X}_1=\hat{x}_1,\,\hat{X}_2=\hat{x}_2)\Big\}.$$

If the values of $\hat{X}_1,\hat{X}_2,\hat{X}_\Theta$ are known, $\hat{X}_1=\hat{x}_1$ and $\hat{X}_2=\hat{x}_2$ where $\hat{x}_1\neq\hat{x}_2$, one can determine the value of Θ without any ambiguity (i.e., $\Theta=1$ if and only if $\hat{X}_\Theta=\hat{x}_1$, and $\Theta=2$ if and only if $\hat{X}_\Theta=\hat{x}_2$), and

$$\Pr(E=1,\,\hat{X}_1\neq\hat{X}_2\,|\,\hat{X}_\Theta)=0. \tag{84}$$

Combining (75) and (80)–(84) gives

$$H(\Theta\,|\,\hat{X}_\Theta)\leq h\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2}\right) \tag{85}$$

and (85) is obtained with equality when (75) holds with equality (i.e., for a maximal coupling). The combination of (79) and (85) finally gives

$$\begin{aligned}\overline{C}(P,Q)\;&\geq 2\left[\log 2-h\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2}\right)\right]\\&=2D\left(\frac{1-d_{\mathrm{TV}}(P,Q)}{2}\,||\,\frac{1}{2}\right)\end{aligned}$$

where, for a given total variation distance, equality is obtained for a maximal coupling that indeed implies that (75) holds with equality. This completes the proof of (37) where it is shown that the infimum in this equality is in fact a minimum.

REFERENCES

[1] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, series B, vol. 28, no. 1, pp. 131–142, 1966.

[2] J. Barić and A. Matković, "Bounds for the normalized Jensen-Mercer functional," *Journal of Mathematical Inequalities*, vol. 3, no. 4, pp. 529–541, 2009.

[3] M. Basseville, "Divergence measures for statistical data processing - an annotated bibliography," *Signal Processing*, vol. 93, no. 4, pp. 621–633, 2013.

[4] D. Berend, P. Harremoës and A. Kontorovich, "Minimum KL-divergence on complements of L_1 balls," to appear in the *IEEE Trans. on Information Theory*, vol. 60, 2014.

[5] S. Boyd and L. Vanderberghe, *Convex Optimization*, Cambridge University Press, 2004. [Online]. Available: http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

[6] J. Briët P. Harremoës, "Properties of classical and quantum Jensen-Shannon divergence," *Physical Review A*, vol. 79, 052311, May 2009.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, second edition, 2006.

[8] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations,"*Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[9] I. Csiszár, "Two remarks to noiseless coding," *Information and Control*, vol. 11, no. 3, pp. 317–322, September 1967.

[10] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators," *IEEE Trans. on Information Theory*, vol. 48, no. 6, pp. 1616–1628, June 2002.

[11] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, 2011.

[12] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, Foundations and Trends in Communications and Information Theory, vol. 1, no. 4, pp. 417–528, 2004.

[13] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[14] I. Csiszár and Z. Talata, "Consisitent estimation of the basic neighborhood of Markov random fields," *Annals of Statistics*, vol. 34, no. 1, pp. 123–145, May 2006.

[15] S. S. Dragomir, *Inequalities for Csiszár f-Divergences in Information Theory*, RGMIA Monographs, Victoria University, 2000. [Online]. Available: http://rgmia.org/monographs/csiszar.htm.

[16] S. S. Dragomir and V. Gluščević, "Some inequalities for the Kullback-Leibler and χ^2-distances in information theory and applications," *Tamsui Oxford Journal of Mathematical Sciences*, vol. 17, no. 2, pp. 97–111, 2001.

[17] S. S. Dragomir, "Bounds for the normalized Jensen functional," *Bulletin of the Australian Mathematical Society*, vol. 74, no. 3, pp. 471–478, 2006.

[18] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. on Information Theory*, vol. 49, no. 7, pp. 1858–1860, July 2003.

[19] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," June 2012. [Online]. Available: http://arxiv.org/pdf/1206.2459v1.pdf.

[20] E. Even-Dar, S. M. Kakade and Y. Mansour, "The value of observation for monitoring dynamical systems," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2474–2479, Hyderabad, India, January 2007.

[21] A. A. Fedotov, P. Harremoës and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. on Information Theory*, vol. 49, no. 6, pp. 1491–1498, June 2003.

[22] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[23] G. L. Gilardoni, "On the minimum f-divergence for given total variation," *Comptes Rendus Mathematique*, vol. 343, no. 11–12, pp. 763–766, 2006.

[24] G. L. Gilardoni, "On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences," *IEEE Trans. on Information Theory*, vol. 56, no. 11, pp. 5377–5386, November 2010.

[25] A. Guntuboyina, S. Saha, and G. Schiebinger, "Sharp inequalities for f-divergences," *IEEE Trans. on Information Theory*, vol. 60, no. 1, pp. 104–121, January 2014.

[26] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range,' *IEEE Trans. on Information Theory*, vol. 57, no. 6, pp. 3230–3235, June 2011.

[27] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions," *Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 700–718, September 1958.

[28] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. on Communication Technology*, vol. 15, no. 1, pp. 52–60, February 1967.

[29] C. Kraft, "Some conditions for consistency and uniform consistency of statistical procedures," *University of California Publications in Statistics*, vol. 1, pp. 125–142, 1955.

[30] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[31] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[32] T. Morimoto, "Markov processes and the H-theorem," *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, 1963.

[33] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[34] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *Journal of Machine Learning Research*, vol. 12, pp. 731–817, March 2011.

[35] R. D. Reiss, *Approximate Distributions of Order Statistics with Applications to Non-Parametric Statistics*, Springer Series in Statistics, Springer-Verlag, 1989.

[36] S. M. Ross and E. A. Peköz, *A Second Course in Probability*, Probability Bookstore, 2007.

[37] I. Sason, "An information-theoretic perspective of the Poisson approximation via the Chen-Stein method," un-published manuscript, *arXiv:1206.6811v4*, 2012.

[38] I. Sason, "Entropy bounds for discrete random variables via maximal coupling," *IEEE Trans. on Information Theory*, vol. 59, no. 11, pp. 7118–7131, November 2013.

[39] A. Sayyareh, "A new upper bound for Kullback-Leibler divergence," *Applied Mathematical Sciences*, vol. 5, no. 67, pp. 3303–3317, 2011.

[40] I. J. Taneja, *Generalized Information Measures and Their Applications*, online book, 2001. [Online]. Available: http://mtm.ufsc.br/~taneja/book/book.html.

[41] I. J. Taneja, "Bounds on non-symmetric divergence measures in terms of symmetric divergence measures," *Journal of Combinatorics, Information, and System Sciences*, vol. 29, pp. 115–134, 2005.

[42] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. on Information Theory*, vol. 46, pp. 1602–1609, July 2000.

[43] I. Vajda, "Note on discrimination information and variation," *IEEE Trans. on Information Theory*, vol. 16, no. 6, pp. 771–773, Nov. 1970.

[44] I. Vajda, *Theory of Statistical Inference and Information*, London, U.K, Kluwer Academic Press, 1989.

[45] S. Verdú, "Total variation distance and the distribution of the relative information," presented at the *9th Workshop on Information Theory and Applications (ITA 2014)*, San-Diego, California, USA, February 2014.

[46] A. D. Yardi. A. Kumar, and S. Vijayakumaran, "Channel-code detection by a third-party receiver via the likelihood ratio test," accepted to the *2014 IEEE International Symposium on Information Theory*, Honolulu, Hawaii, USA, July 2014.