Tight Bounds for Symmetric Divergence Measures and a New Inequality Relating f-Divergences

Igal Sason
Department of Electrical Engineering
Technion, Haifa 32000, Israel
E-mail: sason@ee.technion.ac.il

***Abstract*—Tight bounds for several symmetric divergence measures are introduced, given in terms of the total variation distance. Each of these bounds is attained by a pair of 2 or 3-element probability distributions. An application of these bounds for lossless source coding is provided, refining and improving a certain bound by Csiszár. A new inequality relating f-divergences is derived, and its use is exemplified. The last section of this conference paper is not included in the recent journal paper [16], as well as some new remarks that are linked to new references.**

I. Introduction and Preliminaries

Divergence measures are widely used in information theory, machine learning, statistics, and other theoretical and applied branches of mathematics (see, e.g., [3], [5], [15]). The class of f-divergences forms an important class of divergence measures. Their properties, including relations to statistical tests and estimators, were studied, e.g., in [5] and [13].

In [9], Gilardoni studied the problem of minimizing an arbitrary *symmetric* f-divergence for a given total variation distance (these terms are defined later in this section), providing a closed-form solution of this optimization problem. In a follow-up paper by the same author [10], 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 fixed total variation distance was further studied.

One of the main results in [10] was a further derivation of a simple closed-form lower bound on the relative entropy in terms of the total variation distance. The relative entropy is an asymmetric f-divergence, as it is clarified in the continuation to this section. The lower bound on the relative entropy suggests an improvement over Pinsker's and Vajda's inequalities. A derivation of a simple and reasonably tight closed-form upper bound on the infimum of the relative entropy has been also provided in [10] in terms of the total variation distance. An exact characterization of the minimum of the relative entropy subject to a fixed total variation distance has been derived in [8] and [9].

Sharp inequalities for f-divergences were recently studied in [11] as a general 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 [11] is that such infinite-dimensional optimization problems are equivalent to optimization problems over finite-dimensional spaces where the latter are numerically solvable.

The total variation distance has been further studied from an information-theoretic perspective by Verdú [19], providing upper and lower bounds on the total variation distance between two probability measures P and Q in terms of the distribution of the relative information $\log \frac{dP}{dQ}(X)$ and $\log \frac{dP}{dQ}(Y)$ where X and Y are distributed according to P and Q, respectively.

Following previous work, *tight* bounds on symmetric f-divergences and related distances are introduced in this paper. An application of these bounds for lossless source coding is provided, refining and improving a certain bound by Csiszár [4]. The material in this conference paper appears in the recently published journal paper by the same author [16]. However, we also provide in this conference paper a new inequality relating f-divergences, and its use is exemplified; this material is not included in the journal paper [16] since it does not necessarily refer to symmetric f-divergences.

The paper is organized as follows: tight bounds for several symmetric divergence measures, which are either symmetric f-divergences or related symmetric distances, are introduced without proofs in Section II; these bounds are expressed in terms of the total variation distance. An application for the derivation of an improved and refined bound in the context of lossless source coding is provided in Section III. The full version of this work, including proofs of the tight bounds in Section III, appears in [16]. Section IV provides a new inequality that relates between f-divergences; this inequality is proved since it is not included in the journal paper [16].

We end this section by introducing some preliminaries.

Definition 1: Let P and Q be two probability distributions with a common σ-algebra $\mathcal{F}$. The *total variation distance* between P and Q is $d_{\text{TV}}(P,Q) \triangleq \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$. If P and Q are defined on a countable set, it is simplified to

$$d_{\text{TV}}(P,Q) = \frac{1}{2} \sum_x \bigl|P(x) - Q(x)\bigr| = \frac{||P-Q||_1}{2}. \quad (1)$$

Definition 2: 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. The *f-divergence* from P to Q is defined by

$$D_f(P||Q) \triangleq \sum_x Q(x)\, f\left(\frac{P(x)}{Q(x)}\right) \quad (2)$$

with the convention that

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

Definition 3: An f-divergence is said to be *symmetric* if $D_f(P||Q)=D_f(Q||P)$ for every P and Q.

Symmetric f-divergences include (among others) the squared Hellinger distance where

$$f(t)=(\sqrt{t}-1)^2, \quad D_f(P||Q)=\sum_x\left(\sqrt{P(x)}-\sqrt{Q(x)}\right)^2,$$

and the total variation distance in (1) where $f(t)=\frac{1}{2}\,|t-1|$.

An f-divergence is symmetric if and only if the function f satisfies the equality (see [9, p. 765])

$$f(u)=u\,f\left(\frac{1}{u}\right)+a(u-1), \quad \forall\, u\in(0,\infty) \qquad (3)$$

for some constant a. If f is differentiable at $u=1$ then a differentiation of both sides of equality (3) at $u=1$ gives that $a=2f'(1)$.

Note that the relative entropy (a.k.a. the Kullback-Leibler divergence) $D(P||Q) \triangleq \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right)$ is an f-divergence with $f(t)=t\log(t),\ t>0$; its dual, $D(Q||P)$, is an f-divergence with $f(t)=-\log(t),\ t>0$; clearly, it is an asymmetric f-divergence since $D(P||Q)\neq D(Q||P)$.

The following result, which was derived by Gilardoni (see [9], [10]), refers to the infimum of a symmetric f-divergence for a fixed value of the total variation distance:

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

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

be the infimum of the f-divergence for a given total variation distance. If D_f is a symmetric f-divergence, and f is differentiable at $u=1$, then

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

II. Tight Bounds on Symmetric Divergence Measures

The following section introduces tight bounds for several symmetric divergence measures (where part of them are not f-divergences) for a fixed value of the total variation distance.

A. Tight Bounds on the Bhattacharyya Coefficient

Definition 4: Let P and Q be two probability distributions that are defined on the same set. The *Bhattacharyya coefficient* between P and Q is given by $Z(P,Q)\triangleq\sum_x\sqrt{P(x)\,Q(x)}$.

Proposition 1: Let P and Q be two probability distributions. Then, for a fixed value $\varepsilon\in[0,1]$ of the total variation distance (i.e., if $d_{\text{TV}}(P,Q)=\varepsilon$), the respective Bhattacharyya coefficient satisfies the inequality $1-\varepsilon\le Z(P,Q)\le\sqrt{1-\varepsilon^2}$. Both upper and lower bounds are tight: the upper bound is attained by the pair of 2-element probability distributions $P=\left(\frac{1-\varepsilon}{2},\,\frac{1+\varepsilon}{2}\right)$, and $Q=\left(\frac{1+\varepsilon}{2},\,\frac{1-\varepsilon}{2}\right)$, and the lower bound is attained by the pair of 3-element probability distributions $P=(\varepsilon,1-\varepsilon,0)$, and $Q=(0,1-\varepsilon,\varepsilon)$.

Remark 1: Although derived independently in this work, Proposition 1 is a known result in quantum information theory (on the relation between the trace distance and fidelity [21]).

B. A Tight Bound on the Chernoff Information

Definition 5: The *Chernoff information* between two probability distributions P and Q, defined on the same set, is

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

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

Proposition 2: Let

$$C(\varepsilon)\triangleq\min_{P,Q\colon\, d_{\text{TV}}(P,Q)=\varepsilon} C(P,Q), \quad \forall\,\varepsilon\in[0,1] \qquad (4)$$

be the minimum of the Chernoff information for a fixed value $\varepsilon\in[0,1]$ of the total variation distance. This minimum indeed exists, and it is equal to

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

For $\varepsilon\in[0,1)$, it is achieved by the pair of 2-element probability distributions $P=\left(\frac{1-\varepsilon}{2},\,\frac{1+\varepsilon}{2}\right)$, and $Q=\left(\frac{1+\varepsilon}{2},\,\frac{1-\varepsilon}{2}\right)$.

Outline of the proof: Definition 5 with a possibly sub-optimal value of $\lambda=\frac{1}{2}$, and Proposition 1 yield that

$$\begin{aligned} C(P,Q) &\ge -\log\left(\sum_x\sqrt{P(x)\,Q(x)}\right)\\ &= -\log Z(P,Q)\\ &\ge -\frac{1}{2}\left(1-\left(d_{\text{TV}}(P,Q)\right)^2\right). \end{aligned} \qquad (5)$$

Consequently, from (4), $C(\varepsilon)\ge-\frac{1}{2}\log(1-\varepsilon^2)$ for $\varepsilon\in[0,1)$. It can be verified that the lower bound on $C(P,Q)$ is achieved for $P=\left(\frac{1-\varepsilon}{2},\,\frac{1+\varepsilon}{2}\right)$, and $Q=\left(\frac{1+\varepsilon}{2},\,\frac{1-\varepsilon}{2}\right)$.

Remark 2: A geometric interpretation of the minimum of the Chernoff information subject to a minimal total variation distance has been recently provided in [17, Section 3].

Remark 3 (An Application): From (5), a lower bound on the total variation distance implies 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. This approach has been recently used in [20] to obtain a lower bound on the Chernoff information for studying a communication problem that is related to channel-code detection via the likelihood ratio test.

C. A Tight Bound on the Capacitory Discrimination

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

Definition 6: Let P and Q be two probability distributions. The capacitory discrimination between P and Q is given by

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

This divergence measure was studied, e.g., in [14] and [18].

Proposition 3: For every $\varepsilon \in [0,1)$,

$$\min_{P,Q:\, d_{\text{TV}}(P,Q)=\varepsilon} \overline{C}(P,Q) = 2\,d\left(\frac{1-\varepsilon}{2} \,||\, \frac{1}{2}\right) \tag{6}$$

and it is achieved by the 2-element probability distributions $P = \left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right)$, and $Q = \left(\frac{1+\varepsilon}{2}, \frac{1-\varepsilon}{2}\right)$. In (6),

$$d(p||q) \triangleq p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right), \quad p,q \in [0,1],$$

with the convention that $0\log 0 = 0$.

Outline of the proof: In [11, p. 119], $\overline{C}(P,Q) = D_f(P\|Q)$ with $f(t) = t\log t - (t+1)\log(1+t) + 2\log 2$ for $t > 0$. This is a symmetric f-divergence where f is convex with $f(1) = 0$, $f'(1) = -\log 2$. Eq. (6) follows from Theorem 1.

D. A Tight Bound on Jeffreys' divergence

Definition 7: Let P and Q be two probability distributions. Jeffreys' divergence [12] is a symmetrized version of the relative entropy, which is defined as

$$J(P,Q) \triangleq \frac{D(P||Q) + D(Q||P)}{2}. \tag{7}$$

Proposition 4: For every $\varepsilon \in [0,1)$,

$$\min_{P,Q:\, d_{\text{TV}}(P,Q)=\varepsilon} J(P,Q) = \varepsilon\,\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right). \tag{8}$$

The minimum in (8) is achieved by the pair of 2-element distributions $P = \left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right)$ and $Q = \left(\frac{1+\varepsilon}{2}, \frac{1-\varepsilon}{2}\right)$.

Outline of the proof: Jeffreys' divergence can be expressed as a symmetric f-divergence where $f(t) = \frac{1}{2}(t-1)\log t$ for $t > 0$. Note that f is convex, and $f(1) = f'(1) = 0$. Eq. (8) follows from Theorem 1.

III. A Bound for Lossless Source Coding

We illustrate in the following a use of Proposition 4 for lossless source coding. This tightens, and also refines under a certain condition, a bound by Csiszár [4].

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 $l(u) = -\log_d P(u)$ for every $u \in \mathcal{U}$. 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, the inequality $c_{d,l} \leq 1$ holds in general for UD codes, and the equality $c_{d,l} = 1$ holds if the code achieves the entropy of the source (i.e., $\overline{L} = H_d(U)$).

Define the probability distribution $Q_{d,l}(u) \triangleq \left(\frac{1}{c_{d,l}}\right) d^{-l(u)}$ for every $u \in \mathcal{U}$, and let $\Delta_d \triangleq \overline{L} - H_d(U)$ designate the redundancy of the code. Note that for a UD code that achieves the entropy of the source, its probability distribution P is equal to $Q_{d,l}$ (since $c_{d,l} = 1$, and $P(u) = d^{-l(u)}$ for every $u \in \mathcal{U}$).

In [4], 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

$$D(P||Q_{d,l}) = \Delta_d\,\log d + \log\bigl(c_{d,l}\bigr). \tag{9}$$

The interest in [4] 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 [4] that

$$D(P||Q_{d,l}) \leq \Delta_d\,\log d. \tag{10}$$

Consequently, it follows from Pinsker's inequality that

$$\sum_{u\in\mathcal{U}} \bigl|P(u) - Q_{d,l}(u)\bigr| \leq \min\bigl\{\sqrt{2\Delta_d \log d},\, 2\bigr\} \tag{11}$$

where it is also taken into account that, from the triangle inequality, the sum on the left-hand side of (11) 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 [4]. One possible improvement of the bound in (11) follows by replacing Pinsker's inequality with the result in [8], 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:

$$\sum_{u\in\mathcal{U}} \bigl|P(u) - Q_{d,l}(u)\bigr| \leq 2\;L^{-1}(\Delta_d \log d) \tag{12}$$

where L^{-1} is the inverse function of L, given as follows [15]:

$$\begin{aligned}L(\varepsilon) &\triangleq \inf_{P,Q:\, d_{\text{TV}}(P,Q)=\varepsilon} D(P||Q)\\ &= \min_{\beta\in[\varepsilon-1,\,1-\varepsilon]} \left\{\left(\frac{\varepsilon+1-\beta}{2}\right)\log\left(\frac{\beta-1-\varepsilon}{\beta-1+\varepsilon}\right)\right.\\ &\qquad\left. + \left(\frac{\beta+1-\varepsilon}{2}\right)\log\left(\frac{\beta+1-\varepsilon}{\beta+1+\varepsilon}\right)\right\}.\end{aligned} \tag{13}$$

It can be verified that the numerical minimization w.r.t. β in (13) can be restricted to the interval $[\varepsilon - 1,\, 0]$ (it is calculated numerically).

In the following, the utility of Proposition 4 is shown by refining the bound in (12). Let $\delta(u) \triangleq l(u) + \log_d P(u)$ for every $u \in \mathcal{U}$. Calculation of the dual divergence gives

$$D(Q_{d,l}||P) = -\log\bigl(c_{d,l}\bigr) - \left(\frac{\log d}{c_{d,l}}\right) \mathbb{E}\bigl[\delta(U)\, d^{-\delta(U)}\bigr] \quad (14)$$

and the combination of (7), (9) and (14) yields that

$$J(P, Q_{d,l}) = \frac{1}{2}\left[\Delta_d \log d - \left(\frac{\log d}{c_{d,l}}\right) \mathbb{E}\bigl[\delta(U)\, d^{-\delta(U)}\bigr]\right]. \quad (15)$$

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

$$l(u) \geq \left\lceil \log_d \frac{1}{P(u)} \right\rceil, \quad \forall\, u \in \mathcal{U}. \quad (16)$$

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. Since (16) is equivalent to the condition that δ is non-negative on $\mathcal{U}$, it follows from (15) that

$$J(P, Q_{d,l}) \leq \frac{\Delta_d \log d}{2} \quad (17)$$

so, the upper bound on Jeffreys' divergence in (17) is twice smaller than the upper bound on the relative entropy in (10). It is partially because the term $\log c_{d,l}$ is canceled out along the derivation of the bound in (17), in contrast to the derivation of the bound in (10) 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 4, for an arbitrary $x \geq 0$, let $\varepsilon \triangleq \varepsilon(x)$ be the solution in the interval $[0, 1)$ of the equation

$$\varepsilon \, \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) = x. \quad (18)$$

The combination of (8) and (17) implies that

$$\sum_{u \in \mathcal{U}} \bigl|P(u) - Q_{d,l}(u)\bigr| \leq 2\, \varepsilon\left(\frac{\Delta_d \log d}{2}\right). \quad (19)$$

In the following, the bounds in (12) and (19) are compared analytically for the case where the average redundancy is small (i.e., $\Delta_d \approx 0$). Under this approximation, the bound in (11) (i.e., the original bound from [4]) coincides with its tightened version in (12). On the other hand, since for $\varepsilon \approx 0$, the left-hand side of (18) is approximately $2\varepsilon^2$, it follows from (18) that, for $x \approx 0$, we have $\varepsilon(x) \approx \sqrt{\frac{x}{2}}$. It follows that, if $\Delta_d \approx 0$, inequality (19) gets approximately the form

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

Hence, even for a small redundancy, the bound in (19) improves (11) by a factor of $\sqrt{2}$.

A numerical comparison of the bounds in (11), (12) and (19) is provided in the journal paper, see [16, Figure 2].

Remark 4: Another application of Jeffreys' divergence has been recently studied in [1, Section 5] where the mutual information $I(X;Y) = D(P_{X,Y} \| P_X P_Y)$ has been upper bounded by the symmetrized divergence

$$\begin{aligned} D_{\text{sym}}(P_{X,Y} \| P_X P_Y) &= D(P_{X,Y} \| P_X P_Y) + D(P_X P_Y \| P_{X,Y}) \\ &= 2J(P_{X,Y}, P_X P_Y). \end{aligned}$$

Consequently, the channel capacity satisfies the upper bound $C = \max_{P_X} I(X;Y) \leq 2 \max_{P_X} J(P_{X,Y}, P_X P_Y)$. This provides a good bound on the channel capacity in the low SNR regime (see [1, Section 5]). It has been applied in [1, Section 6] to obtain a bound on the capacity of a linear-time invariant Poisson channel; this bound is improved by increasing the parameter of the background noise (λ_0) [1].

IV. A New Inequality Relating f-Divergences

We introduce in the following an inequality which relates f-divergences, and its use is exemplified. This inequality is proved here since it is not included in the journal paper [16].

Recall the following definition of the χ^2-divergence.

Definition 8: The *chi-squared divergence* between two probability distributions P and Q on a set $\mathcal{A}$ is given by

$$\chi^2(P, Q) \triangleq \sum_{x \in \mathcal{A}} \frac{\bigl(P(x) - Q(x)\bigr)^2}{Q(x)} = \sum_{x \in \mathcal{A}} \frac{P(x)^2}{Q(x)} - 1\,. \quad (20)$$

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

Proposition 5: 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 $g(t) = -t f(t)$ 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:

$$\begin{aligned} &\min_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D_f(P||Q) \\ &\leq -D_g(P||Q) - f\bigl(1 + \chi^2(P, Q)\bigr) \\ &\leq \max_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D_f(P||Q). \end{aligned} \quad (21)$$

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

$$J_n(f, \underline{u}, P) \triangleq \sum_{i=1}^{n} P(x_i)\, f(u_i) - f\left(\sum_{i=1}^{n} P(x_i) u_i\right). \quad (22)$$

The following refinement of Jensen's inequality appears in [7, Theorem 1] for a convex function $f\colon (0, \infty) \to \mathbb{R}$, and it has been extended in [2, Theorem 1] to hold for a convex f over an arbitrary interval $[a, b]$:

$$\begin{aligned} &\min_{i \in \{1, \ldots, n\}} \left\{\frac{P(x_i)}{Q(x_i)}\right\} J_n(f, \underline{u}, Q) \leq J_n(f, \underline{u}, P) \\ &\leq \max_{i \in \{1, \ldots, n\}} \left\{\frac{P(x_i)}{Q(x_i)}\right\} J_n(f, \underline{u}, Q). \end{aligned} \quad (23)$$

The refined version of Jensen's inequality in (23) is applied in the following to prove (21). Let $u_i \triangleq \frac{P(x_i)}{Q(x_i)}$ for $i \in \{1, \ldots, n\}$. Calculation of (22) gives that

$$\begin{aligned} J_n(f, \underline{u}, Q) &= \sum_{i=1}^{n} Q(x_i)\, f\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} Q(x_i) \cdot \frac{P(x_i)}{Q(x_i)}\right) \\ &= \sum_{x \in \mathcal{A}} Q(x)\, f\left(\frac{P(x)}{Q(x)}\right) - f(1) = D_f(P||Q), \quad (24) \end{aligned}$$

$$\begin{aligned} J_n(f, \underline{u}, P) &= \sum_{i=1}^{n} P(x_i)\, f\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(x_i)^2}{Q(x_i)}\right) \\ &\overset{\text{(a)}}{=} -\sum_{i=1}^{n} Q(x_i)\, g\left(\frac{P(x_i)}{Q(x_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(x_i)^2}{Q(x_i)}\right) \\ &\overset{\text{(b)}}{=} -D_g(P||Q) - f\big(1 + \chi^2(P,Q)\big) \quad (25) \end{aligned}$$

where equality (a) holds by the definition of g, and equality (b) follows from equalities (2) and (20). The substitution of (24) and (25) in (23) completes the proof.

As a consequence of Proposition 5, we introduce the following inequality which relates between the relative entropy, its dual and the chi-squared divergence.

Corollary 1: 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:

$$\begin{aligned} &\min_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D(Q||P) \\ &\leq \log\big(1 + \chi^2(P,Q)\big) - D(P||Q) \\ &\leq \max_{x \in \mathcal{A}} \frac{P(x)}{Q(x)} \cdot D(Q||P). \quad (26) \end{aligned}$$

Proof Let $f(t) = -\log(t)$ for $t > 0$. The function $f\colon (0,\infty) \to \mathrm{I\!R}$ is convex with $f(1) = 0$, and $g(t) = -tf(t) = t\log(t)$ for $t > 0$ defines a convex function with $g(1) = 0$. Inequality (26) follows by substituting f, g in (21) where $D_f(P||Q) = D(Q||P)$ and $D_g(P||Q) = D(P||Q)$.

Remark 5: Inequality (26) strengthens the inequality

$$\chi^2(P,Q) \geq e^{D(P||Q)} - 1 \quad (27)$$

which is derived by using Jensen's inequality as follows [6]:

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

The following inequality is another consequence of Proposition 5, relating the chi-squared divergence and its dual:

Corollary 2: Under the same conditions of Corollary 1, the following inequality holds:

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

Proof This follows from Proposition 5 where $f(t) = \frac{1}{t} - 1$, and $g(t) = -tf(t) = t - 1$ for $t > 0$. Consequently, we have $D_g(P||Q) = 0$, $D_f(P||Q) = \chi^2(Q,P)$.

Acknowledgment

The author thanks the anonymous reviewers of the journal paper in the *IEEE Trans. on Information Theory* [16], and of this conference paper at the proceedings of the 2015 IEEE Information Theory Workshop in Israel for their helpful comments. This research work has been supported by the Israeli Science Foundation (ISF), grant number 12/12.

References

[1] H. Arjmandi, G. Aminian, A. Gohari, M. N. Kenari and U. Mitra, "Capacity of diffusion based molecular communication networks over LTI-Poisson channels," October 2014. [Online]. Available at http://arxiv.org/abs/1410.3988.

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society A*, vol. 186, no. 1007, pp. 453–461, September 1946.

[13] 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.

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

[15] 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.

[16] I. Sason, "Tight bounds on symmetric divergence measures and a refined bound for lossless source coding," *IEEE Trans. on Information Theory*, vol. 61, no. 2, pp. 701–707, February 2015.

[17] I. Sason, "On the Rényi divergence and the joint range of relative entropies," 2015. [Online]. Available at http://arxiv.org/abs/1501.03616.

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

[19] S. Verdú, "Total variation distance and the distribution of relative information," presented at the 2014 *Information Theory and Applications Workshop*, February 2014. [Online]. Available: https://www.princeton.edu/∼verdu/reprints/VERDU-ITA2014.pdf.

[20] A. D. Yardi. A. Kumar, and S. Vijayakumaran, "Channel-code detection by a third-party receiver via the likelihood ratio test," *Proceedings of the 2014 IEEE International Symposium on Information Theory*, pp. 1051–1055, Honolulu, Hawaii, USA, July 2014.

[21] Fidelity of quantum states and relationship to trace distance. [Online]. Available at http://en.wikipedia.org/wiki/Fidelity_of_quantum_states.