entropy

MDPI

Article

On f-Divergences: Integral Representations, Local Behavior, and Inequalities

Igal Sason

Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 3200003, Israel; sason@ee.technion.ac.il

Received: 15 April 2018; Accepted: 15 May 2018; Published: 19 May 2018

Abstract: This paper is focused on f-divergences, consisting of three main contributions. The first one introduces integral representations of a general f-divergence by means of the relative information spectrum. The second part provides a new approach for the derivation of f-divergence inequalities, and it exemplifies their utility in the setup of Bayesian binary hypothesis testing. The last part of this paper further studies the local behavior of f-divergences.

Keywords: DeGroot statistical information; f-divergences; local behavior; relative information spectrum; Rényi divergence

1. Introduction

Probability theory, information theory, learning theory, statistical signal processing and other related disciplines, greatly benefit from non-negative measures of dissimilarity (a.k.a. divergence measures) between pairs of probability measures defined on the same measurable space (see, e.g., [1–7]). An axiomatic characterization of information measures, including divergence measures, was provided by Csiszár [8]. Many useful divergence measures belong to the set of f-divergences, independently introduced by Ali and Silvey [9], Csiszár [10–13], and Morimoto [14] in the early sixties. The family of f-divergences generalizes the relative entropy (a.k.a. the Kullback- Leibler divergence) while also satisfying the data processing inequality among other pleasing properties (see, e.g., [3] and references therein).

Integral representations of f-divergences serve to study properties of these information measures, and they are also used to establish relations among these divergences. An integral representation of f-divergences, expressed by means of the DeGroot statistical information, was provided in [3] with a simplified proof in [15]. The importance of this integral representation stems from the operational meaning of the DeGroot statistical information [16], which is strongly linked to Bayesian binary hypothesis testing. Some earlier specialized versions of this integral representation were introduced in [17–21], and a variation of it also appears in [22] Section 5.B. Implications of the integral representation of f-divergences, by means of the DeGroot statistical information, include an alternative proof of the data processing inequality, and a study of conditions for the sufficiency or ε-deficiency of observation channels [3,15].

Since many distance measures of interest fall under the paradigm of an f-divergence [23], bounds among f-divergences are very useful in many instances such as the analysis of rates of convergence and concentration of measure bounds, hypothesis testing, testing goodness of fit, minimax risk in estimation and modeling, strong data processing inequalities and contraction coefficients, etc. Earlier studies developed systematic approaches to obtain f-divergence inequalities while dealing with pairs of probability measures defined on arbitrary alphabets. A list of some notable existing f-divergence inequalities is provided, e.g., in [22] Section 1 and [23] Section 3. State-of-the-art techniques which serve to derive bounds among f-divergences include:

(1) Moment inequalities which rely on log-convexity arguments ([22] Section 5.D, [24–28]);
(2) Inequalities which rely on a characterization of the exact locus of the joint range of f-divergences [29];
(3) f-divergence inequalities via functional domination ([22] Section 3, [30–32]);
(4) Sharp f-divergence inequalities by using numerical tools for maximizing or minimizing an f-divergence subject to a finite number of constraints on other f-divergences [33];
(5) Inequalities which rely on powers of f-divergences defining a distance [34–37];
(6) Vajda and Pinsker-type inequalities for f-divergences ([4,10,13], [22] Sections 6–7, [38,39]);
(7) Bounds among f-divergences when the relative information is bounded ([22] Sections 4–5, [40–47]), and reverse Pinsker inequalities ([22] Section 6, [40,48]);
(8) Inequalities which rely on the minimum of an f-divergence for a given total variation distance and related bounds [4,33,37,38,49–53];
(9) Bounds among f-divergences (or functions of f-divergences such as the Rényi divergence) via integral representations of these divergence measures [22] Section 8;
(10) Inequalities which rely on variational representations of f-divergences (e.g., [54] Section 2).

Following earlier studies of the local behavior of f-divergences and their asymptotic properties (see related results by Csiszár and Shields [55] Theorem 4.1, Pardo and Vajda [56] Section 3, and Sason and Vérdu [22] Section 3.F), it is known that the local behavior of f-divergences scales, such as the chi-square divergence (up to a scaling factor which depends on f) provided that the first distribution approaches the reference measure in a certain strong sense. The study of the local behavior of f-divergences is an important aspect of their properties, and we further study it in this work.

This paper considers properties of f-divergences, while first introducing in Section 2 the basic definitions and notation needed, and in particular the various measures of dissimilarity between probability measures used throughout this paper. The presentation of our new results is then structured as follows:

Section 3 is focused on the derivation of new integral representations of f-divergences, expressed as a function of the relative information spectrum of the pair of probability measures, and the convex function f. The novelty of Section 3 is in the unified approach which leads to integral representations of f-divergences by means of the relative information spectrum, where the latter cumulative distribution function plays an important role in information theory and statistical decision theory (see, e.g., [7,54]). Particular integral representations of the type of results introduced in Section 3 have been recently derived by Sason and Verdú in a case-by-case basis for some f-divergences (see [22] Theorems 13 and 32), while lacking the approach which is developed in Section 3 for general f-divergences. In essence, an f-divergence $D_f(P\|Q)$ is expressed in Section 3 as an inner product of a simple function of the relative information spectrum (depending only on the probability measures P and Q), and a non-negative weight function $\omega_f\colon (0,\infty) \mapsto [0,\infty)$ which only depends on f. This kind of representation, followed by a generalized result, serves to provide new integral representations of various useful f-divergences. This also enables in Section 3 to characterize the interplay between the DeGroot statistical information (or between another useful family of f-divergence, named the E_γ divergence with $\gamma \geq 1$) and the relative information spectrum.

Section 4 provides a new approach for the derivation of f-divergence inequalities, where an arbitrary f-divergence is lower bounded by means of the E_γ divergence [57] or the DeGroot statistical information [16]. The approach used in Section 4 yields several generalizations of the Bretagnole-Huber inequality [58], which provides a closed-form and simple upper bound on the total variation distance as a function of the relative entropy; the Bretagnole-Huber inequality has been proved to be useful, e.g., in the context of lower bounding the minimax risk in non-parametric estimation (see, e.g., [5] pp. 89–90, 94), and in the problem of density estimation (see, e.g., [6] Section 1.6). Although Vajda's tight lower bound in [59] is slightly tighter everywhere than the Bretagnole-Huber inequality, our motivation for the generalization of the latter bound is justified later in this paper. The utility of the new inequalities is exemplified in the setup of Bayesian binary hypothesis testing.

Section 5 finally derives new results on the local behavior of f-divergences, i.e., the characterization of their scaling when the pair of probability measures are sufficiently close to each other. The starting point

of our analysis in Section 5 relies on the analysis in [56] Section 3, regarding the asymptotic properties of f-divergences.

The reading of Sections 3–5 can be done in any order since the analysis in these sections is independent.

2. Preliminaries and Notation

We assume throughout that the probability measures P and Q are defined on a common measurable space $(\mathcal{A}, \mathscr{F})$, and $P \ll Q$ denotes that P is absolutely continuous with respect to Q, namely there is no event $\mathcal{F} \in \mathscr{F}$ such that $P(\mathcal{F}) > 0 = Q(\mathcal{F})$.

Definition 1. *The relative information provided by $a \in \mathcal{A}$ according to (P, Q), where $P \ll Q$, is given by*

$$\imath_{P\|Q}(a) := \log \frac{dP}{dQ}(a). \tag{1}$$

More generally, even if $P \not\ll Q$, let R be an arbitrary dominating probability measure such that $P, Q \ll R$ (e.g., $R = \frac{1}{2}(P+Q)$); irrespectively of the choice of R, the relative information is defined to be

$$\imath_{P\|Q}(a) := \imath_{P\|R}(a) - \imath_{Q\|R}(a), \quad a \in \mathcal{A}. \tag{2}$$

The following asymmetry property follows from (2):

$$\imath_{P\|Q} = -\imath_{Q\|P}. \tag{3}$$

Definition 2. *The relative information spectrum is the cumulative distribution function*

$$\mathbb{F}_{P\|Q}(x) = \mathbb{P}\big[\imath_{P\|Q}(X) \le x\big], \quad x \in \mathbb{R},\ X \sim P. \tag{4}$$

The relative entropy is the expected valued of the relative information when it is distributed according to P:

$$D(P\|Q) := \mathbb{E}\big[\imath_{P\|Q}(X)\big], \quad X \sim P. \tag{5}$$

Throughout this paper, $\mathcal{C}$ denotes the set of convex functions $f\colon (0,\infty) \mapsto \mathbb{R}$ with $f(1) = 0$. Hence, the function $f \equiv 0$ is in $\mathcal{C}$; if $f \in \mathcal{C}$, then $af \in \mathcal{C}$ for all $a > 0$; and if $f, g \in \mathcal{C}$, then $f + g \in \mathcal{C}$. We next provide a general definition for the family of f-divergences (see [3] p. 4398).

Definition 3 (f-divergence [9,10,12])**.** *Let P and Q be probability measures, let μ be a dominating measure of P and Q (i.e., $P, Q \ll \mu$; e.g., $\mu = P + Q$), and let $p := \frac{dP}{d\mu}$ and $q := \frac{dQ}{d\mu}$. The f-divergence from P to Q is given, independently of μ, by*

$$D_f(P\|Q) := \int q\, f\Big(\frac{p}{q}\Big)\, d\mu, \tag{6}$$

where

$$f(0) := \lim_{t \downarrow 0} f(t), \tag{7}$$

$$0 f\Big(\frac{0}{0}\Big) := 0, \tag{8}$$

$$0 f\Big(\frac{a}{0}\Big) := \lim_{t \downarrow 0} t f\Big(\frac{a}{t}\Big) = a \lim_{u \to \infty} \frac{f(u)}{u}, \quad a > 0. \tag{9}$$

We rely in this paper on the following properties of f-divergences:

Proposition 1. *Let $f, g \in \mathcal{C}$. The following conditions are equivalent:*

(1)

$$D_f(P\|Q) = D_g(P\|Q), \quad \forall\, P, Q; \tag{10}$$

(2) there exists a constant $c \in \mathbb{R}$ such that

$$f(t) - g(t) = c\ (t-1), \quad \forall\, t \in (0,\infty). \tag{11}$$

Proposition 2. *Let $f \in \mathcal{C}$, and let $f^*\colon (0,\infty) \mapsto \mathbb{R}$ be the conjugate function, given by*

$$f^*(t) = t\, f\left(\tfrac{1}{t}\right) \tag{12}$$

for $t > 0$. Then, $f^* \in \mathcal{C}$; $f^{**} = f$, and for every pair of probability measures (P, Q),

$$D_f(P\|Q) = D_{f^*}(Q\|P). \tag{13}$$

By an analytic extension of f^* in (12) at $t = 0$, let

$$f^*(0) := \lim_{t \downarrow 0} f^*(t) = \lim_{u \to \infty} \frac{f(u)}{u}. \tag{14}$$

Note that the convexity of f^* implies that $f^*(0) \in (-\infty, \infty]$. In continuation to Definition 3, we get

$$D_f(P\|Q) = \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu \tag{15}$$

$$= \int_{\{pq>0\}} q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu + Q(p=0)\, f(0) + P(q=0)\, f^*(0) \tag{16}$$

with the convention in (16) that $0 \cdot \infty = 0$, We refer in this paper to the following f-divergences:

(1) *Relative entropy*:

$$D(P\|Q) = D_f(P\|Q). \tag{17}$$

with

$$f(t) = t \log t, \quad t > 0. \tag{18}$$

(2) *Jeffrey's divergence [60]*:

$$J(P\|Q) := D(P\|Q) + D(Q\|P) \tag{19}$$

$$= D_f(P\|Q) \tag{20}$$

with

$$f(t) = (t-1) \log t, \quad t > 0. \tag{21}$$

(3) *Hellinger divergence of order $\alpha \in (0,1) \cup (1,\infty)$* [2] Definition 2.10:

$$\mathscr{H}_\alpha(P\|Q) = D_{f_\alpha}(P\|Q) \tag{22}$$

with

$$f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}, \quad t > 0. \tag{23}$$

Some of the significance of the Hellinger divergence stems from the following facts:

– The analytic extension of $\mathscr{H}_\alpha(P\|Q)$ at $\alpha = 1$ yields

$$D(P\|Q) = H_1(P\|Q) \log e. \tag{24}$$

– The *chi-squared divergence* [61] is the second order Hellinger divergence (see, e.g., [62] p. 48), i.e.,

$$\chi^2(P\|Q) = \mathscr{H}_2(P\|Q). \tag{25}$$

Note that, due to Proposition 1,

$$\chi^2(P\|Q) = D_f(P\|Q), \tag{26}$$

where $f\colon (0,\infty) \mapsto \mathbb{R}$ can be defined as

$$f(t) = (t-1)^2, \quad t > 0. \tag{27}$$

– The *squared Hellinger distance* (see, e.g., [62] p. 47), denoted by $\mathscr{H}^2(P\|Q)$, satisfies the identity

$$\mathscr{H}^2(P\|Q) = \tfrac{1}{2}\, \mathscr{H}_{\frac{1}{2}}(P\|Q). \tag{28}$$

– The *Bhattacharyya distance* [63], denoted by $B(P\|Q)$, satisfies

$$B(P\|Q) = \log \frac{1}{1 - \mathscr{H}^2(P\|Q)}. \tag{29}$$

– The *Rényi divergence* of order $\alpha \in (0,1) \cup (1,\infty)$ is a one-to-one transformation of the Hellinger divergence of the same order [11] (14):

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\bigl(1 + (\alpha - 1)\, \mathscr{H}_\alpha(P\|Q)\bigr). \tag{30}$$

– The *Alpha-divergence* of order α, as it is defined in [64] and ([65] (4)), is a generalized relative entropy which (up to a scaling factor) is equal to the Hellinger divergence of the same order α. More explicitly,

$$D_{\mathrm{A}}^{(\alpha)}(P\|Q) = \frac{1}{\alpha}\, \mathscr{H}_\alpha(P\|Q), \tag{31}$$

where $D_{\mathrm{A}}^{(\alpha)}(\cdot\|\cdot)$ denotes the Alpha-divergence of order α. Note, however, that the Beta and Gamma-divergences in [65], as well as the generalized divergences in [66,67], are not f-divergences in general.

(4) χ^s *divergence* for $s \geq 1$ [2] (2.31), and the *total variation distance*: The function

$$f_s(t) = |t-1|^s, \quad t > 0 \tag{32}$$

results in

$$\chi^s(P\|Q) = D_{f_s}(P\|Q). \tag{33}$$

Specifically, for $s = 1$, let

$$f(t) := f_1(t) = |t-1|, \quad t > 0, \tag{34}$$

and the total variation distance is expressed as an f-divergence:

$$|P - Q| = D_f(P\|Q). \tag{35}$$

(5) *Triangular Discrimination [39] (a.k.a. Vincze-Le Cam distance)*:

$$\Delta(P\|Q) = D_f(P\|Q) \tag{36}$$

with

$$f(t) = \frac{(t-1)^2}{t+1}, \quad t > 0. \tag{37}$$

Note that

$$\tfrac{1}{2}\,\Delta(P\|Q) = \chi^2(P \,\|\, \tfrac{1}{2}P + \tfrac{1}{2}Q) = \chi^2(Q \,\|\, \tfrac{1}{2}P + \tfrac{1}{2}Q). \tag{38}$$

(6) *Lin's measure* [68] (4.1):

$$L_\theta(P\|Q) := H\big(\theta P + (1-\theta)Q\big) - \theta H(P) - (1-\theta)H(Q) \tag{39}$$

$$= \theta\, D\big(P \,\|\, \theta P + (1-\theta)Q\big) + (1-\theta)\, D\big(Q \,\|\, \theta P + (1-\theta)Q\big), \tag{40}$$

for $\theta \in [0,1]$. This measure can be expressed by the following f-divergence:

$$L_\theta(P\|Q) = D_{f_\theta}(P\|Q), \tag{41}$$

with

$$f_\theta(t) := \theta t \log t - \big(\theta t + 1 - \theta\big)\, \log\big(\theta t + 1 - \theta\big), \quad t > 0. \tag{42}$$

The special case of (41) with $\theta = \frac{1}{2}$ gives the *Jensen-Shannon divergence* (a.k.a. capacitory discrimination):

$$\mathrm{JS}(P\|Q) := L_{\frac{1}{2}}(P\|Q) \tag{43}$$

$$= \tfrac{1}{2} D\big(P \,\|\, \tfrac{1}{2}P + \tfrac{1}{2}Q\big) + \tfrac{1}{2} D\big(Q \,\|\, \tfrac{1}{2}P + \tfrac{1}{2}Q\big). \tag{44}$$

(7) E_γ *divergence* [57] p. 2314: For $\gamma \geq 1$,

$$E_\gamma(P\|Q) := \max_{\mathcal{U} \in \mathscr{F}} \big(P(\mathcal{U}) - \gamma\, Q(\mathcal{U})\big) \tag{45}$$

$$= \mathbb{P}[\imath_{P\|Q}(X) > \log\gamma] - \gamma\, \mathbb{P}[\imath_{P\|Q}(Y) > \log\gamma] \tag{46}$$

with $X \sim P$ and $Y \sim Q$, and where (46) follows from the Neyman-Pearson lemma. The E_γ divergence can be identified as an f-divergence:

$$E_\gamma(P\|Q) = D_{f_\gamma}(P\|Q) \tag{47}$$

with

$$f_\gamma(t) := (t-\gamma)^+, \quad t > 0 \tag{48}$$

where $(x)^+ := \max\{x, 0\}$. The following relation to the total variation distance holds:

$$E_1(P\|Q) = \tfrac{1}{2}\,|P - Q|. \tag{49}$$

(8) *DeGroot statistical information* [3,16]: For $\omega \in (0,1)$,

$$\mathcal{I}_\omega(P\|Q) = D_{\phi_\omega}(P\|Q) \tag{50}$$

with

$$\phi_\omega(t) = \min\{\omega, 1-\omega\} - \min\{\omega t, 1-\omega\}, \quad t > 0. \tag{51}$$

The following relation to the total variation distance holds:

$$\mathcal{I}_{\frac{1}{2}}(P\|Q) = \tfrac{1}{4}\,|P - Q|, \tag{52}$$

and the DeGroot statistical information and the E_γ divergence are related as follows [22] (384):

$$\mathcal{I}_\omega(P\|Q) = \begin{cases} \omega\, E_{\frac{1-\omega}{\omega}}(P\|Q), & \omega \in \left(0, \frac{1}{2}\right], \\ (1-\omega)\, E_{\frac{\omega}{1-\omega}}(Q\|P), & \omega \in \left(\frac{1}{2}, 1\right). \end{cases} \tag{53}$$

3. New Integral Representations of f-Divergences

The main result in this section provides new integral representations of f-divergences as a function of the relative information spectrum (see Definition 2). The reader is referred to other integral representations (see [15] Section 2, [4] Section 5, [22] Section 5.B, and references therein), expressing a general f-divergence by means of the DeGroot statistical information or the E_γ divergence.

Lemma 1. *Let $f \in \mathcal{C}$ be a strictly convex function at 1. Let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be defined as*

$$g(x) := \exp(-x)\, f\big(\exp(x)\big) - f'_+(1)\,\big(1 - \exp(-x)\big), \qquad x \in \mathbb{R} \tag{54}$$

where $f'_+(1)$ denotes the right-hand derivative of f at 1 (due to the convexity of f on $(0,\infty)$, it exists and it is finite). Then, the function g is non-negative, it is strictly monotonically decreasing on $(-\infty, 0]$, and it is strictly monotonically increasing on $[0,\infty)$ with $g(0) = 0$.

Proof. For any function $u \in \mathcal{C}$, let $\widetilde{u} \in \mathcal{C}$ be given by

$$\widetilde{u}(t) = u(t) - u'_+(1)(t-1), \quad t \in (0,\infty), \tag{55}$$

and let $u^* \in \mathcal{C}$ be the conjugate function, as given in (12). The function g in (54) can be expressed in the form

$$g(x) = (\widetilde{f})^*\big(\exp(-x)\big), \quad x \in \mathbb{R}, \tag{56}$$

as it is next verified. For $t > 0$, we get from (12) and (55),

$$(\widetilde{f})^*(t) = t\widetilde{f}\left(\frac{1}{t}\right) = tf\left(\frac{1}{t}\right) + f'_+(1)\,(t-1), \tag{57}$$

and the substitution $t := \exp(-x)$ for $x \in \mathbb{R}$ yields (56) in view of (54).

By assumption, $f \in \mathcal{C}$ is strictly convex at 1, and therefore these properties are inherited to $\widetilde{f}$. Since also $\widetilde{f}(1) = \widetilde{f}'(1) = 0$, it follows from [3] Theorem 3 that both $\widetilde{f}$ and $\widetilde{f}^*$ are non-negative on $(0,\infty)$, and

they are also strictly monotonically decreasing on $(0,1]$. Hence, from (12), it follows that the function $(\widetilde{f})^*$ is strictly monotonically increasing on $[1,\infty)$. Finally, the claimed properties of the function g follow from (56), and in view of the fact that the function $(\widetilde{f})^*$ is non-negative with $(\widetilde{f})^*(1)=0$, strictly monotonically decreasing on $(0,1]$ and strictly monotonically increasing on $[1,\infty)$. □

Lemma 2. *Let $f \in \mathcal{C}$ be a strictly convex function at 1, and let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be as in (54). Let*

$$a := \lim_{x\to\infty} g(x) \in (0,\infty], \tag{58}$$

$$b := \lim_{x\to-\infty} g(x) \in (0,\infty], \tag{59}$$

and let $\ell_1\colon [0,a) \mapsto [0,\infty)$ and $\ell_2\colon [0,b) \mapsto (-\infty,0]$ be the two inverse functions of g. Then,

$$D_f(P\|Q) = \int_0^a \big[1-\mathbb{F}_{P\|Q}\big(\ell_1(t)\big)\big]\,\mathrm{d}t + \int_0^b \mathbb{F}_{P\|Q}\big(\ell_2(t)\big)\,\mathrm{d}t. \tag{60}$$

Proof. In view of Lemma 1, it follows that $\ell_1\colon [0,a) \mapsto [0,\infty)$ is strictly monotonically increasing and $\ell_2\colon [0,b) \mapsto (-\infty,0]$ is strictly monotonically decreasing with $\ell_1(0)=\ell_2(0)=0$.

Let $X\sim P$, and let $V := \exp\big(\imath_{P\|Q}(X)\big)$. Then, we have

$$\begin{aligned}
D_f(P\|Q) &= D_{\widetilde{f}}(P\|Q) && (61)\\
&= D_{(\widetilde{f})^*}(Q\|P) && (62)\\
&= \int (\widetilde{f})^*\big(\exp\big(\imath_{Q\|P}(x)\big)\big)\,\mathrm{d}P(x) && (63)\\
&= \int (\widetilde{f})^*\big(\exp\big(-\imath_{P\|Q}(x)\big)\big)\,\mathrm{d}P(x) && (64)\\
&= \int g\big(\imath_{P\|Q}(x)\big)\,\mathrm{d}P(x) && (65)\\
&= \mathbb{E}\big[g(V)\big] && (66)\\
&= \int_0^\infty \mathbb{P}\big[g(V)>t\big]\,\mathrm{d}t && (67)\\
&= \int_0^a \mathbb{P}\big[V\ge 0,\, g(V)>t\big]\,\mathrm{d}t + \int_0^b \mathbb{P}\big[V<0,\, g(V)>t\big]\,\mathrm{d}t && (68)\\
&= \int_0^a \mathbb{P}\big[V>\ell_1(t)\big]\,\mathrm{d}t + \int_0^b \mathbb{P}\big[V\le\ell_2(t)\big]\,\mathrm{d}t && (69)\\
&= \int_0^a \big[1-\mathbb{F}_{P\|Q}\big(\ell_1(t)\big)\big]\,\mathrm{d}t + \int_0^b \mathbb{F}_{P\|Q}\big(\ell_2(t)\big)\,\mathrm{d}t && (70)
\end{aligned}$$

where (61) relies on Proposition 1; (62) relies on Proposition 2; (64) follows from (3); (65) follows from (56); (66) holds by the definition of the random variable V; (67) holds since, in view of Lemma 1, $Z := g(V) \ge 0$, and $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}[Z>t]\,\mathrm{d}t$ for any non-negative random variable Z; (68) holds in view of the monotonicity properties of g in Lemma 1, the definition of a and b in (58) and (59), and by expressing the event $\{g(V)>t\}$ as a union of two disjoint events; (69) holds again by the monotonicity properties of g in Lemma 1, and by the definition of its two inverse functions ℓ_1 and ℓ_2 as above; in (67)–(69) we are free to substitute $>$ by $\ge$, and $<$ by $\le$; finally, (70) holds by the definition of the relative information spectrum in (4). □

Remark 1. *The function $g\colon \mathbb{R} \mapsto \mathbb{R}$ in (54) is invariant to the mapping $f(t) \mapsto f(t) + c\,(t-1)$, for $t>0$, with an arbitrary $c\in\mathbb{R}$. This invariance of g (and, hence, also the invariance of its inverse functions ℓ_1 and ℓ_2) is well expected in view of Proposition 1 and Lemma 2.*

Example 1. *For the chi-squared divergence in (26), letting f be as in (27), it follows from (54) that*

$$g(x) = 4\sinh^2\left(\tfrac{1}{2\log e}\, x\right), \quad x \in \mathbb{R}, \tag{71}$$

which yields, from (58) and (59), $a = b = \infty$. Calculation of the two inverse functions of g, as defined in Lemma 2, yields the following closed-form expression:

$$\ell_{1,2}(u) = \pm 2\log\left(\frac{u+\sqrt{u+4}}{2}\right), \quad u \geq 0. \tag{72}$$

Substituting (72) into (60) provides an integral representation of $\chi^2(P\|Q)$.

Lemma 3.

$$\int_0^\infty \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta = 1. \tag{73}$$

Proof. Let $X \sim P$. Then, we have

$$\int_0^\infty \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta = \int_0^\infty \frac{1}{\beta^2}\,\mathbb{P}[\imath_{P\|Q}(X) \leq \log\beta]\,\mathrm{d}\beta \tag{74}$$

$$= \int_0^\infty \frac{1}{\beta^2}\,\mathbb{P}\left[\exp\big(\imath_{Q\|P}(X)\big) \geq \frac{1}{\beta}\right]\mathrm{d}\beta \tag{75}$$

$$= \int_0^\infty \mathbb{P}\big[\exp\big(\imath_{Q\|P}(X)\big) \geq u\big]\,\mathrm{d}u \tag{76}$$

$$= \mathbb{E}\big[\exp\big(\imath_{Q\|P}(X)\big)\big] \tag{77}$$

$$= 1, \tag{78}$$

where (74) holds by (4); (75) follows from (3); (76) holds by the substitution $u := \frac{1}{\beta}$; (77) holds since $\exp\big(\imath_{Q\|P}(X)\big) \geq 0$, and finally (78) holds since $X \sim P$. □

Remark 2. *Unlike Example 1, in general, the inverse functions ℓ_1 and ℓ_2 in Lemma 2 are not expressible in closed form, motivating our next integral representation in Theorem 1.*

The following theorem provides our main result in this section.

Theorem 1. *The following integral representations of an f-divergence, by means of the relative information spectrum, hold:*

(1) Let

- *$f \in \mathcal{C}$ be differentiable on $(0,\infty)$;*
- *$w_f\colon (0,\infty) \mapsto [0,\infty)$ be the non-negative weight function given, for $\beta > 0$, by*

$$w_f(\beta) := \frac{1}{\beta}\left|f'(\beta) - \frac{f(\beta)+f'(1)}{\beta}\right|; \tag{79}$$

- *the function $G_{P\|Q}\colon (0,\infty) \mapsto [0,1]$ be given by*

$$G_{P\|Q}(\beta) := \begin{cases} 1 - \mathbb{F}_{P\|Q}(\log\beta), & \beta \in [1,\infty), \\ \mathbb{F}_{P\|Q}(\log\beta), & \beta \in (0,1). \end{cases} \tag{80}$$

Then,

$$D_f(P\|Q) = \langle w_f,\, G_{P\|Q}\rangle = \int_0^\infty w_f(\beta)\, G_{P\|Q}(\beta)\, \mathrm{d}\beta. \tag{81}$$

(2) *More generally, for an arbitrary $c \in \mathbb{R}$, let $\widetilde{w}_{f,c}\colon (0,\infty) \mapsto \mathbb{R}$ be a modified real-valued function defined as*

$$\widetilde{w}_{f,c}(\beta) := w_f(\beta) + \frac{c}{\beta^2}\left(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\right). \tag{82}$$

Then,

$$D_f(P\|Q) = \langle \widetilde{w}_{f,c},\, G_{P\|Q}\rangle. \tag{83}$$

Proof. We start by proving the special integral representation in (81), and then extend our proof to the general representation in (83).

(1) We first assume an additional requirement that f is strictly convex at 1. In view of Lemma 2,

$$\ell_1\big(g(u)\big) = u, \quad u \in [0,\infty), \tag{84}$$
$$\ell_2\big(g(u)\big) = u, \quad u \in (-\infty, 0]. \tag{85}$$

Since by assumption $f \in \mathcal{C}$ is differentiable on $(0,\infty)$ and strictly convex at 1, the function g in (54) is differentiable on $\mathbb{R}$. In view of (84) and (85), substituting $t := g\big(\log\beta\big)$ in (60) for $\beta > 0$ implies that

$$D_f(P\|Q) = \int_1^\infty \big[1 - \mathbb{F}_{P\|Q}\big(\log\beta\big)\big]\, \overline{w}_f(\beta)\, \mathrm{d}\beta - \int_0^1 \mathbb{F}_{P\|Q}\big(\log\beta\big)\, \overline{w}_f(\beta)\, \mathrm{d}\beta, \tag{86}$$

where $\overline{w}_f\colon (0,\infty) \mapsto \mathbb{R}$ is given by

$$\overline{w}_f(\beta) := \frac{g'\big(\log\beta\big)}{\beta}\, \log \mathrm{e} \tag{87}$$
$$= \frac{1}{\beta}\left[f'(\beta) - \frac{f(\beta) + f'(1)}{\beta}\right] \tag{88}$$

for $\beta > 0$, where (88) follows from (54). Due to the monotonicity properties of g in Lemma 1, (87) implies that $\overline{w}_f(\beta) \geq 0$ for $\beta \geq 1$, and $\overline{w}_f(\beta) < 0$ for $\beta \in (0,1)$. Hence, the weight function w_f in (79) satisfies

$$w_f(\beta) = \big|\overline{w}_f(\beta)\big| = \overline{w}_f(\beta)\left(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\right), \quad \beta > 0. \tag{89}$$

The combination of (80), (86) and (89) gives the required result in (81).

We now extend the result in (81) when $f \in \mathcal{C}$ is differentiable on $(0,\infty)$, but not necessarily strictly convex at 1. To that end, let $s\colon (0,\infty) \mapsto \mathbb{R}$ be defined as

$$s(t) := f(t) + (t^2 - 1), \quad t > 0. \tag{90}$$

This implies that $s \in \mathcal{C}$ is differentiable on $(0,\infty)$, and it is also strictly convex at 1. In view of the proof of (81) when f is strict convexity of f at 1, the application of this result to the function s in (90) yields

$$D_s(P\|Q) = \langle w_s,\, G_{P\|Q}\rangle. \tag{91}$$

In view of (6), (22), (23), (25) and (90),

$$D_s(P\|Q) = D_f(P\|Q) + \chi^2(P\|Q); \tag{92}$$

from (79), (89), (90) and the convexity and differentiability of $f \in \mathcal{C}$, it follows that the weight function $w_s \in (0,\infty) \mapsto [0,\infty)$ satisfies

$$w_s(\beta) = w_f(\beta) + \left(1 - \frac{1}{\beta^2}\right)\left(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\right) \tag{93}$$

for $\beta > 0$. Furthermore, by applying the result in (81) to the chi-squared divergence $\chi^2(P\|Q)$ in (25) whose corresponding function $f_2(t) := t^2 - 1$ for $t > 0$ is strictly convex at 1, we obtain

$$\chi^2(P\|Q) = \int_0^\infty \left(1 - \frac{1}{\beta^2}\right)\left(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\right) G_{P\|Q}(\beta)\, \mathrm{d}\beta. \tag{94}$$

Finally, the combination of (91)–(94), yields $D_f(P\|Q) = \langle w_f, G_{P\|Q}\rangle$; this asserts that (81) also holds by relaxing the condition that f is strictly convex at 1.

(2) In view of (80)–(82), in order to prove (83) for an arbitrary $c \in \mathbb{R}$, it is required to prove the identity

$$\int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\, \mathrm{d}\beta = \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\, \mathrm{d}\beta. \tag{95}$$

Equality (95) can be verified by Lemma 3: by rearranging terms in (95), we get the identity in (73) (since $\int_1^\infty \frac{\mathrm{d}\beta}{\beta^2} = 1$).

□

Remark 3. *Due to the convexity of f, the absolute value in the right side of (79) is only needed for $\beta \in (0,1)$ (see (88) and (89)). Also, $w_f(1) = 0$ since $f(1) = 0$.*

Remark 4. *The weight function w_f only depends on f, and the function $G_{P\|Q}$ only depends on the pair of probability measures P and Q. In view of Proposition 1, it follows that, for $f, g \in \mathcal{C}$, the equality $w_f = w_g$ holds on $(0,\infty)$ if and only if (11) is satisfied with an arbitrary constant $c \in \mathbb{R}$. It is indeed easy to verify that (11) yields $w_f = w_g$ on $(0,\infty)$.*

Remark 5. *An equivalent way to write $G_{P\|Q}$ in (80) is*

$$G_{P\|Q}(\beta) = \begin{cases} \mathbb{P}\left[\dfrac{dP}{dQ}(X) > \beta\right], & \beta \in [1,\infty) \\ \mathbb{P}\left[\dfrac{dP}{dQ}(X) \leq \beta\right], & \beta \in (0,1) \end{cases} \tag{96}$$

where $X \sim P$. Hence, the function $G_{P\|Q}\colon (0,\infty) \mapsto [0,1]$ is monotonically increasing in $(0,1)$, and it is monotonically decreasing in $[1,\infty)$; note that this function is in general discontinuous at 1 unless $\mathbb{F}_{P\|Q}(0) = \frac{1}{2}$. If $P \ll\gg Q$, then

$$\lim_{\beta\downarrow 0} G_{P\|Q}(\beta) = \lim_{\beta\to\infty} G_{P\|Q}(\beta) = 0. \tag{97}$$

Note that if $P = Q$, then $G_{P\|Q}$ is zero everywhere, which is consistent with the fact that $D_f(P\|Q) = 0$.

Remark 6. *In the proof of Theorem 1-(1), the relaxation of the condition of strict convexity at 1 for a differentiable function $f \in \mathcal{C}$ is crucial, e.g., for the χ^s divergence with $s > 2$. To clarify this claim, note that in view of (32),*

the function $f_s\colon (0,\infty) \mapsto \mathbb{R}$ is differentiable if $s > 1$, and $f_s \in \mathcal{C}$ with $f_s'(1) = 0$; however, $f_s''(1) = 0$ if $s > 2$, so f_s in not strictly convex at 1 unless $s \in [1,2]$.

Remark 7. *Theorem 1-(1) with $c \neq 0$ enables, in some cases, to simplify integral representations of f-divergences. This is next exemplified in the proof of Theorem 2.*

Theorem 1 yields integral representations for various f-divergences and related measures; some of these representations were previously derived by Sason and Verdú in [22] in a case by case basis, without the unified approach of Theorem 1. We next provide such integral representations. Note that, for some f-divergences, the function $f \in \mathcal{C}$ is not differentiable on $(0,\infty)$; hence, Theorem 1 is not necessarily directly applicable.

Theorem 2. *The following integral representations hold as a function of the relative information spectrum:*

(1) Relative entropy [22] (219):

$$\tfrac{1}{\log e}\, D(P\|Q) = \int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{\beta}\,\mathrm{d}\beta - \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta}\,\mathrm{d}\beta. \tag{98}$$

(2) Hellinger divergence of order $\alpha \in (0,1)\cup(1,\infty)$ [22] (434) and (437):

$$\mathscr{H}_\alpha(P\|Q) = \begin{cases} \dfrac{1}{1-\alpha} - \displaystyle\int_0^\infty \beta^{\alpha-2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta, & \alpha \in (0,1) \\ \displaystyle\int_0^\infty \beta^{\alpha-2}\left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\,\mathrm{d}\beta - \dfrac{1}{\alpha-1}, & \alpha \in (1,\infty). \end{cases} \tag{99}$$

In particular, the chi-squared divergence, squared Hellinger distance and Bhattacharyya distance satisfy

$$\chi^2(P\|Q) = \int_0^\infty \left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\,\mathrm{d}\beta - 1; \tag{100}$$

$$\mathscr{H}^2(P\|Q) = 1 - \tfrac{1}{2}\int_0^\infty \beta^{-\frac{3}{2}}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta; \tag{101}$$

$$B(P\|Q) = \log 2 - \log\left(\int_0^\infty \beta^{-\frac{3}{2}}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta\right), \tag{102}$$

where (100) appears in [22] (439).

(3) Rényi divergence [22] (426) and (427): For $\alpha \in (0,1)\cup(1,\infty)$,

$$D_\alpha(P\|Q) = \begin{cases} \dfrac{1}{\alpha-1}\log\left((1-\alpha)\displaystyle\int_0^\infty \beta^{\alpha-2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta\right), & \alpha \in (0,1) \\ \dfrac{1}{\alpha-1}\log\left((\alpha-1)\displaystyle\int_0^\infty \beta^{\alpha-2}\left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\,\mathrm{d}\beta\right), & \alpha \in (1,\infty). \end{cases} \tag{103}$$

(4) χ^s divergence: For $s \geq 1$

$$\begin{aligned}\chi^s(P\|Q) = &\int_1^\infty \frac{1}{\beta}\left(s - 1 + \frac{1}{\beta}\right)(\beta-1)^{s-1}\left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\,\mathrm{d}\beta \\ &+ \int_0^1 \frac{1}{\beta}\left(s - 1 + \frac{1}{\beta}\right)(1-\beta)^{s-1}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta.\end{aligned} \tag{104}$$

In particular, the following identities hold for the total variation distance:

$$|P-Q| = 2\int_1^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta \tag{105}$$

$$= 2\int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta, \tag{106}$$

where (105) *appears in [22] (214).*

(5) *DeGroot statistical information:*

$$\mathcal{I}_w(P\|Q) = \begin{cases} (1-w)\displaystyle\int_0^{\frac{1-w}{w}} \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta, & w\in\left(\frac{1}{2},1\right) \\ (1-w)\displaystyle\int_{\frac{1-w}{w}}^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta, & w\in\left(0,\frac{1}{2}\right]. \end{cases} \tag{107}$$

(6) *Triangular discrimination:*

$$\Delta(P\|Q) = 4\int_0^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{(\beta+1)^2}\,\mathrm{d}\beta - 2. \tag{108}$$

(7) *Lin's measure: For* $\theta\in[0,1]$,

$$L_\theta(P\|Q) = h(\theta) - (1-\theta)\int_0^\infty \frac{\log\left(1+\frac{\theta\beta}{1-\theta}\right)}{\beta^2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta, \tag{109}$$

where $h\colon [0,1]\mapsto[0,\log 2]$ *denotes the binary entropy function. Specifically, the Jensen-Shannon divergence admits the integral representation:*

$$\mathrm{JS}(P\|Q) = \log 2 - \int_0^\infty \frac{\log(\beta+1)}{2\beta^2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta. \tag{110}$$

(8) *Jeffrey's divergence:*

$$J(P\|Q) = \int_1^\infty \left(1-\mathbb{F}_{P\|Q}(\log\beta)\right)\left(\frac{\log e}{\beta}+\frac{\log\beta}{\beta^2}\right)\mathrm{d}\beta - \int_0^1 \mathbb{F}_{P\|Q}(\log\beta)\left(\frac{\log e}{\beta}+\frac{\log\beta}{\beta^2}\right)\mathrm{d}\beta. \tag{111}$$

(9) E_γ *divergence: For* $\gamma\geq 1$,

$$E_\gamma(P\|Q) = \gamma\int_\gamma^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta. \tag{112}$$

Proof. See Appendix A. □

An application of (112) yields the following interplay between the E_γ divergence and the relative information spectrum.

Theorem 3. *Let* $X\sim P$, *and let the random variable* $\imath_{P\|Q}(X)$ *have no probability masses. Denote*

$$\mathcal{A}_1 := \left\{E_\gamma(P\|Q)\colon \gamma\geq 1\right\}, \tag{113}$$

$$\mathcal{A}_2 := \left\{E_\gamma(Q\|P)\colon \gamma> 1\right\}. \tag{114}$$

Then,

- *$E_\gamma(P\|Q)$ is a continuously differentiable function of γ on $(1,\infty)$, and $E'_\gamma(P\|Q) \le 0$;*
- *the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ determine, respectively, the relative information spectrum $\mathbb{F}_{P\|Q}(\cdot)$ on $[0,\infty)$ and $(-\infty,0)$;*
- *for $\gamma > 1$,*

$$\mathbb{F}_{P\|Q}(+\log\gamma) = 1 - E_\gamma(P\|Q) + \gamma E'_\gamma(P\|Q), \tag{115}$$

$$\mathbb{F}_{P\|Q}(-\log\gamma) = -E'_\gamma(Q\|P), \tag{116}$$

$$\mathbb{F}_{P\|Q}(0) = 1 - E_1(P\|Q) + \lim_{\gamma\downarrow 1} E'_\gamma(P\|Q) \tag{117}$$

$$= -\lim_{\gamma\downarrow 1} E'_\gamma(Q\|P). \tag{118}$$

Proof. We start by proving the first item. By our assumption, $\mathbb{F}_{P\|Q}(\cdot)$ is continuous on $\mathbb{R}$. Hence, it follows from (112) that $E_\gamma(P\|Q)$ is continuously differentiable in $\gamma \in (1,\infty)$; furthermore, (45) implies that $E_\gamma(P\|Q)$ is monotonically decreasing in γ, which yields $E'_\gamma(P\|Q) \le 0$.

We next prove the second and third items together. Let $X \sim P$ and $Y \sim Q$. From (112), for $\gamma > 1$,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\left(\frac{E_\gamma(P\|Q)}{\gamma}\right) = -\frac{1 - \mathbb{F}_{P\|Q}(\log\gamma)}{\gamma^2}, \tag{119}$$

which yields (115). Due to the continuity of $\mathbb{F}_{P\|Q}(\cdot)$, it follows that the set $\mathcal{A}_1$ determines the relative information spectrum on $[0,\infty)$.

To prove (116), we have

$$E_\gamma(Q\|P) = \mathbb{P}[\imath_{Q\|P}(Y) > \log\gamma] - \gamma\,\mathbb{P}[\imath_{Q\|P}(X) > \log\gamma] \tag{120}$$

$$= 1 - \mathbb{F}_{Q\|P}(\log\gamma) - \gamma\,\mathbb{P}[\imath_{Q\|P}(X) > \log\gamma] \tag{121}$$

$$= E_\gamma(Q\|P) - \gamma E'_\gamma(Q\|P) - \gamma\,\mathbb{P}[\imath_{Q\|P}(X) > \log\gamma] \tag{122}$$

$$= E_\gamma(Q\|P) - \gamma E'_\gamma(Q\|P) - \gamma\,\mathbb{P}[\imath_{P\|Q}(X) < -\log\gamma] \tag{123}$$

$$= E_\gamma(Q\|P) - \gamma E'_\gamma(Q\|P) - \gamma\,\mathbb{F}_{P\|Q}(-\log\gamma) \tag{124}$$

where (120) holds by switching P and Q in (46); (121) holds since $Y \sim Q$; (122) holds by switching P and Q in (115) (correspondingly, also $X \sim P$ and $Y \sim Q$ are switched); (123) holds since $\imath_{Q\|P} = -\imath_{P\|Q}$; (124) holds by the assumption that $\frac{\mathrm{d}P}{\mathrm{d}Q}(X)$ has no probability masses, which implies that the sign $<$ can be replaced with $\le$ at the term $\mathbb{P}[\imath_{P\|Q}(X) < -\log\gamma]$ in the right side of (123). Finally, (116) readily follows from (120)–(124), which implies that the set $\mathcal{A}_2$ determines $\mathbb{F}_{P\|Q}(\cdot)$ on $(-\infty,0)$.

Equalities (117) and (117) finally follows by letting $\gamma \downarrow 1$, respectively, on both sides of (115) and (116). □

A similar application of (107) yields an interplay between DeGroot statistical information and the relative information spectrum.

Theorem 4. *Let $X \sim P$, and let the random variable $\imath_{P\|Q}(X)$ have no probability masses. Denote*

$$\mathcal{B}_1 := \left\{\mathcal{I}_\omega(P\|Q) : \omega \in \left(0, \tfrac{1}{2}\right]\right\}, \tag{125}$$

$$\mathcal{B}_2 := \left\{\mathcal{I}_\omega(P\|Q) : \omega \in \left(\tfrac{1}{2}, 1\right)\right\}. \tag{126}$$

Then,

- $\mathcal{I}_\omega(P\|Q)$ *is a continuously differentiable function of* ω *on* $(0,\frac{1}{2})\cup(\frac{1}{2},1)$,

$$\lim_{\omega\uparrow\frac{1}{2}} \mathcal{I}'_\omega(P\|Q) - \lim_{\omega\downarrow\frac{1}{2}} \mathcal{I}'_\omega(P\|Q) = 2, \tag{127}$$

and $\mathcal{I}'_\omega(P\|Q)$ *is, respectively, non-negative or non-positive on* $\left(0,\frac{1}{2}\right)$ *and* $\left(\frac{1}{2},1\right)$;

- *the sets* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *determine, respectively, the relative information spectrum* $\mathbb{F}_{P\|Q}(\cdot)$ *on* $[0,\infty)$ *and* $(-\infty,0)$;
- *for* $\omega\in\left(0,\frac{1}{2}\right)$

$$\mathbb{F}_{P\|Q}\left(\log\tfrac{1-\omega}{\omega}\right) = 1-\mathcal{I}_\omega(P\|Q)-(1-\omega)\,\mathcal{I}'_\omega(P\|Q), \tag{128}$$

for $\omega\in\left(\frac{1}{2},1\right)$

$$\mathbb{F}_{P\|Q}\left(\log\tfrac{1-\omega}{\omega}\right) = -\mathcal{I}_\omega(P\|Q)-(1-\omega)\,\mathcal{I}'_\omega(P\|Q), \tag{129}$$

and

$$\mathbb{F}_{P\|Q}(0) = -\mathcal{I}_{\frac{1}{2}}(P\|Q) - \tfrac{1}{2}\lim_{\omega\downarrow\frac{1}{2}} \mathcal{I}'_\omega(P\|Q). \tag{130}$$

Remark 8. *By relaxing the condition in Theorems 3 and 4 where* $\frac{\mathrm{d}P}{\mathrm{d}Q}(X)$ *has no probability masses with* $X\sim P$, *it follows from the proof of Theorem 3 that each one of the sets*

$$\mathcal{A} := \mathcal{A}_1\cup\mathcal{A}_2 = \Big\{\big(E_\gamma(P\|Q), E_\gamma(Q\|P)\big) : \gamma\geq 1\Big\}, \tag{131}$$

$$\mathcal{B} := \mathcal{B}_1\cup\mathcal{B}_2 = \Big\{\mathcal{I}_\omega(P\|Q) : \omega\in(0,1)\Big\} \tag{132}$$

determines $\mathbb{F}_{P\|Q}(\cdot)$ *at every point on* $\mathbb{R}$ *where this relative information spectrum is continuous. Note that, as a cumulative distribution function,* $\mathbb{F}_{P\|Q}(\cdot)$ *is discontinuous at a countable number of points. Consequently, under the condition that* $f\in\mathcal{C}$ *is differentiable on* $(0,\infty)$, *the integral representations of* $D_f(P\|Q)$ *in Theorem 1 are not affected by the countable number of discontinuities for* $\mathbb{F}_{P\|Q}(\cdot)$.

In view of Theorems 1, 3 and 4 and Remark 8, we get the following result.

Corollary 1. *Let* $f\in\mathcal{C}$ *be a differentiable function on* $(0,\infty)$, *and let* $P\ll\gg Q$ *be probability measures. Then, each one of the sets* $\mathcal{A}$ *and* $\mathcal{B}$ *in* (131) *and* (132), *respectively, determines* $D_f(P\|Q)$.

Remark 9. *Corollary 1 is supported by the integral representation of* $D_f(P\|Q)$ *in [3] Theorem 11, expressed as a function of the set of values in* $\mathcal{B}$, *and its analogous representation in [22] Proposition 3 as a function of the set of values in* $\mathcal{A}$. *More explicitly, [3] Theorem 11 states that if* $f\in\mathcal{C}$, *then*

$$D_f(P\|Q) = \int_0^1 \mathcal{I}_\omega(P\|Q)\,d\Gamma_f(\omega) \tag{133}$$

where Γ_f *is a certain* σ*-finite measure defined on the Borel subsets of* $(0,1)$; *it is also shown in [3] (80) that if* $f\in\mathcal{C}$ *is twice differentiable on* $(0,\infty)$, *then*

$$D_f(P\|Q) = \int_0^1 \mathcal{I}_\omega(P\|Q)\,\frac{1}{\omega^3}\,f''\left(\frac{\omega}{1-\omega}\right)d\omega. \tag{134}$$

4. New f-Divergence Inequalities

Various approaches for the derivation of f-divergence inequalities were studied in the literature (see Section 1 for references). This section suggests a new approach, leading to a lower bound on an arbitrary f-divergence by means of the E_γ divergence of an arbitrary order $\gamma \geq 1$ (see (45)) or the DeGroot statistical information (see (50)). This approach leads to generalizations of the Bretagnole-Huber inequality [58], whose generalizations are later motivated in this section. The utility of the f-divergence inequalities in this section is exemplified in the setup of Bayesian binary hypothesis testing.

In the following, we provide the first main result in this section for the derivation of new f-divergence inequalities by means of the E_γ divergence. Generalizing the total variation distance, the E_γ divergence in (45)–(47) is an f-divergence whose utility in information theory has been exemplified in [17] Chapter 3, [54], [57] p. 2314 and [69]; the properties of this measure were studied in [22] Section 7 and [54] Section 2.B.

Theorem 5. *Let $f \in \mathcal{C}$, and let $f^* \in \mathcal{C}$ be the conjugate convex function as defined in (12). Let P and Q be probability measures. Then, for all $\gamma \in [1, \infty)$,*

$$D_f(P\|Q) \geq f^*\left(1 + \tfrac{1}{\gamma}\, E_\gamma(P\|Q)\right) + f^*\left(\tfrac{1}{\gamma}\left(1 - E_\gamma(P\|Q)\right)\right) - f^*\left(\tfrac{1}{\gamma}\right). \tag{135}$$

Proof. Let $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ be the densities of P and Q with respect to a dominating measure μ ($P, Q \ll \mu$). Then, for an arbitrary $a \in \mathbb{R}$,

$$\begin{aligned}
D_f(P\|Q) &= D_{f^*}(Q\|P) && (136)\\
&= \int p\, f^*\left(\frac{q}{p}\right) \mathrm{d}\mu && (137)\\
&= \int p \left[f^*\left(\max\left\{a, \frac{q}{p}\right\}\right) + f^*\left(\min\left\{a, \frac{q}{p}\right\}\right) - f^*(a)\right] \mathrm{d}\mu && (138)\\
&\geq f^*\left(\int p \max\left\{a, \frac{q}{p}\right\} \mathrm{d}\mu\right) + f^*\left(\int p \min\left\{a, \frac{q}{p}\right\} \mathrm{d}\mu\right) - f^*(a) && (139)
\end{aligned}$$

where (139) follows from the convexity of f^* and by invoking Jensen's inequality.

Setting $a := \frac{1}{\gamma}$ with $\gamma \in [1, \infty)$ gives

$$\begin{aligned}
\int p \max\left\{a, \frac{q}{p}\right\} \mathrm{d}\mu &= \int \max\left\{\frac{p}{\gamma}, q\right\} \mathrm{d}\mu && (140)\\
&= \int q\, \mathrm{d}\mu + \int \max\left\{\frac{p}{\gamma} - q, 0\right\} \mathrm{d}\mu && (141)\\
&= 1 + \frac{1}{\gamma} \int q \max\left\{\frac{p}{q} - \gamma, 0\right\} \mathrm{d}\mu && (142)\\
&= 1 + \tfrac{1}{\gamma}\, E_\gamma(P\|Q), && (143)
\end{aligned}$$

and

$$\begin{aligned}
\int p \min\left\{a, \frac{q}{p}\right\} \mathrm{d}\mu &= \int p \left(a + \frac{q}{p} - \max\left\{a, \frac{q}{p}\right\}\right) \mathrm{d}\mu && (144)\\
&= a + 1 - \int p \max\left\{a, \frac{q}{p}\right\} \mathrm{d}\mu && (145)\\
&= \tfrac{1}{\gamma}\left(1 - E_\gamma(P\|Q)\right) && (146)
\end{aligned}$$

where (146) follows from (143) by setting $a := \frac{1}{\gamma}$. Substituting (143) and (146) into the right side of (139) gives (135). □

An application of Theorem 5 gives the following lower bounds on the Hellinger and Rényi divergences with arbitrary positive orders, expressed as a function of the E_γ divergence with an arbitrary order $\gamma \geq 1$.

Corollary 2. *For all $\alpha > 0$ and $\gamma \geq 1$,*

$$\mathscr{H}_\alpha(P\|Q) \geq \begin{cases} \frac{1}{\alpha-1}\left[\left(1+\frac{1}{\gamma}\,E_\gamma(P\|Q)\right)^{1-\alpha} + \left(\frac{1-E_\gamma(P\|Q)}{\gamma}\right)^{1-\alpha} - 1 - \gamma^{\alpha-1}\right], & \alpha \neq 1 \\ -\log_{\mathrm{e}}\left(\left(1+\frac{1}{\gamma}\,E_\gamma(P\|Q)\right)\left(1-E_\gamma(P\|Q)\right)\right), & \alpha = 1, \end{cases} \tag{147}$$

and

$$D_\alpha(P\|Q) \geq \begin{cases} \frac{1}{\alpha-1}\log\left(\left(1+\frac{1}{\gamma}\,E_\gamma(P\|Q)\right)^{1-\alpha} + \gamma^{\alpha-1}\left[\left(1-E_\gamma(P\|Q)\right)^{1-\alpha} - 1\right]\right), & \alpha \neq 1 \\ -\log\left(\left(1+\frac{1}{\gamma}\,E_\gamma(P\|Q)\right)\left(1-E_\gamma(P\|Q)\right)\right), & \alpha = 1. \end{cases} \tag{148}$$

Proof. Inequality (147), for $\alpha \in (0,1) \cup (1,\infty)$, follows from Theorem 5 and (22); for $\alpha = 1$, it holds in view of Theorem 5, and equalities (17) and (24). Inequality (148), for $\alpha \in (0,1) \cup (1,\infty)$, follows from (30) and (147); for $\alpha = 1$, it holds in view of (24), (147) and since $D_1(P\|Q) = D(P\|Q)$. □

Specialization of Corollary 2 for $\alpha = 2$ in (147) and $\alpha = 1$ in (148) gives the following result.

Corollary 3. *For $\gamma \in [1,\infty)$, the following upper bounds on E_γ divergence hold as a function of the relative entropy and χ^2 divergence:*

$$E_\gamma(P\|Q) \leq \tfrac{1}{2}\left[1-\gamma+\sqrt{(\gamma-1)^2 + \frac{4\gamma\,\chi^2(P\|Q)}{1+\gamma+\chi^2(P\|Q)}}\,\right], \tag{149}$$

$$E_\gamma(P\|Q) \leq \tfrac{1}{2}\left[1-\gamma+\sqrt{(\gamma-1)^2 + 4\gamma\left(1-\exp(-D(P\|Q))\right)}\right]. \tag{150}$$

Remark 10. *From [4] (58),*

$$\chi^2(P\|Q) \geq \begin{cases} |P-Q|^2, & |P-Q| \in [0,1) \\ \frac{|P-Q|}{2-|P-Q|}, & |P-Q| \in [1,2) \end{cases} \tag{151}$$

is a tight lower bound on the chi-squared divergence as a function of the total variation distance. In view of (49), we compare (151) with the specialized version of (149) when $\gamma = 1$. The latter bound is expected to be looser than the tight bound in (151), as a result of the use of Jensen's inequality in the proof of Theorem 5; however, it is interesting to examine how much we loose in the tightness of this specialized bound with $\gamma = 1$. From (49), the substitution of $\gamma = 1$ in (149) gives

$$\chi^2(P\|Q) \geq \frac{2|P-Q|^2}{4-|P-Q|^2}, \qquad |P-Q| \in [0,2), \tag{152}$$

and, it can be easily verified that

- *if* $|P-Q| \in [0,1)$, *then the lower bound in the right side of* (152) *is at most twice smaller than the tight lower bound in the right side of* (151);
- *if* $|P-Q| \in [1,2)$, *then the lower bound in the right side of* (152) *is at most* $\frac{3}{2}$ *times smaller than the tight lower bound in the right side of* (151).

Remark 11. *Setting* $\gamma = 1$ *in* (150), *and using* (49), *specializes to the Bretagnole-Huber inequality [58]:*

$$|P-Q| \leq 2\sqrt{1-\exp\bigl(-D(P\|Q)\bigr)}. \tag{153}$$

Inequality (153) forms a counterpart to Pinsker's inequality:

$$\tfrac{1}{2}|P-Q|^2 \log e \leq D(P\|Q), \tag{154}$$

proved by Csiszár [12] and Kullback [70], with Kemperman [71] independently a bit later. As upper bounds on the total variation distance, (154) outperforms (153) if $D(P\|Q) \leq 1.594$ nats, and (153) outperforms (154) for larger values of $D(P\|Q)$.

Remark 12. *In [59] (8), Vajda introduced a lower bound on the relative entropy as a function of the total variation distance:*

$$D(P\|Q) \geq \log\left(\frac{2+|P-Q|}{2-|P-Q|}\right) - \frac{2|P-Q|\log e}{2+|P-Q|}, \quad |P-Q| \in [0,2). \tag{155}$$

The lower bound in the right side of (155) *is asymptotically tight in the sense that it tends to* ∞ *if* $|P-Q| \uparrow 2$, *and the difference between* $D(P\|Q)$ *and this lower bound is everywhere upper bounded by* $\frac{2|P-Q|^3}{(2+|P-Q|)^2} \leq 4$ *(see [59] (9)). The Bretagnole-Huber inequality in* (153), *on the other hand, is equivalent to*

$$D(P\|Q) \geq -\log\left(1-\tfrac{1}{4}|P-Q|^2\right), \quad |P-Q| \in [0,2). \tag{156}$$

Although it can be verified numerically that the lower bound on the relative entropy in (155) *is everywhere slightly tighter than the lower bound in* (156) *(for* $|P-Q| \in [0,2)$*), both lower bounds on* $D(P\|Q)$ *are of the same asymptotic tightness in a sense that they both tend to* ∞ *as* $|P-Q| \uparrow 2$ *and their ratio tends to 1. Apart of their asymptotic tightness, the Bretagnole-Huber inequality in* (156) *is appealing since it provides a closed-form simple upper bound on* $|P-Q|$ *as a function of* $D(P\|Q)$ *(see* (153)*), whereas such a closed-form simple upper bound cannot be obtained from* (155). *In fact, by the substitution* $v := -\frac{2-|P-Q|}{2+|P-Q|}$ *and the exponentiation of both sides of* (155), *we get the inequality* $ve^v \geq -\frac{1}{e}\exp\bigl(-D(P\|Q)\bigr)$ *whose solution is expressed by the Lambert W function [72]; it can be verified that* (155) *is equivalent to the following upper bound on the total variation distance as a function of the relative entropy:*

$$|P-Q| \leq \frac{2\bigl(1+W(z)\bigr)}{1-W(z)}, \tag{157}$$

$$z := -\tfrac{1}{e}\exp\bigl(-D(P\|Q)\bigr), \tag{158}$$

where W *in the right side of* (157) *denotes the principal real branch of the Lambert W function. The difference between the upper bounds in* (153) *and* (157) *can be verified to be marginal if* $D(P\|Q)$ *is large (e.g., if* $D(P\|Q) = 4$ *nats, then the upper bounds on* $|P-Q|$ *are respectively equal to 1.982 and 1.973), though the former upper bound in* (153) *is clearly more simple and amenable to analysis.*

The Bretagnole-Huber inequality in (153) *is proved to be useful in the context of lower bounding the minimax risk (see, e.g., [5] pp. 89–90, 94), and the problem of density estimation (see, e.g., [6] Section 1.6). The utility of this inequality motivates its generalization in this section (see Corollaries 2 and 3, and also see later Theorem 7 followed by Example 2).*

In [22] Section 7.C, Sason and Verdú generalized Pinsker's inequality by providing an upper bound on the E_γ divergence, for $\gamma > 1$, as a function of the relative entropy. In view of (49) and the optimality of the constant in Pinsker's inequality (154), it follows that the minimum achievable $D(P\|Q)$ is quadratic in $E_1(P\|Q)$ for small values of $E_1(P\|Q)$. It has been proved in [22] Section 7.C that this situation ceases to be the case for $\gamma > 1$, in which case it is possible to upper bound $E_\gamma(P\|Q)$ as a constant times $D(P\|Q)$ where this constant tends to infinity as we let $\gamma \downarrow 1$. We next cite the result in [22] Theorem 30, extending (154) by means of the E_γ divergence for $\gamma > 1$, and compare it numerically to the bound in (150).

Theorem 6. *([22] Theorem 30) For every $\gamma > 1$,*

$$\sup \frac{E_\gamma(P\|Q)}{D(P\|Q)} = c_\gamma \tag{159}$$

where the supremum is over $P \ll Q, P \neq Q$, and c_γ is a universal function (independent of (P, Q)), given by

$$c_\gamma = \frac{t_\gamma - \gamma}{t_\gamma \log t_\gamma + (1 - t_\gamma) \log e}, \tag{160}$$

$$t_\gamma = -\gamma W_{-1}\left(-\tfrac{1}{\gamma} e^{-\frac{1}{\gamma}}\right) \tag{161}$$

where W_{-1} in (161) denotes the secondary real branch of the Lambert W function [72].

As an immediate consequence of (159), it follows that

$$E_\gamma(P\|Q) \leq c_\gamma D(P\|Q), \tag{162}$$

which forms a straight-line bound on the E_γ divergence as a function of the relative entropy for $\gamma > 1$. Similarly to the comparison of the Bretagnole-Huber inequality (153) and Pinsker's inequality (154), we exemplify numerically that the extension of Pinsker's inequality to the E_γ divergence in (162) forms a counterpart to the generalized version of the Bretagnole-Huber inequality in (150).

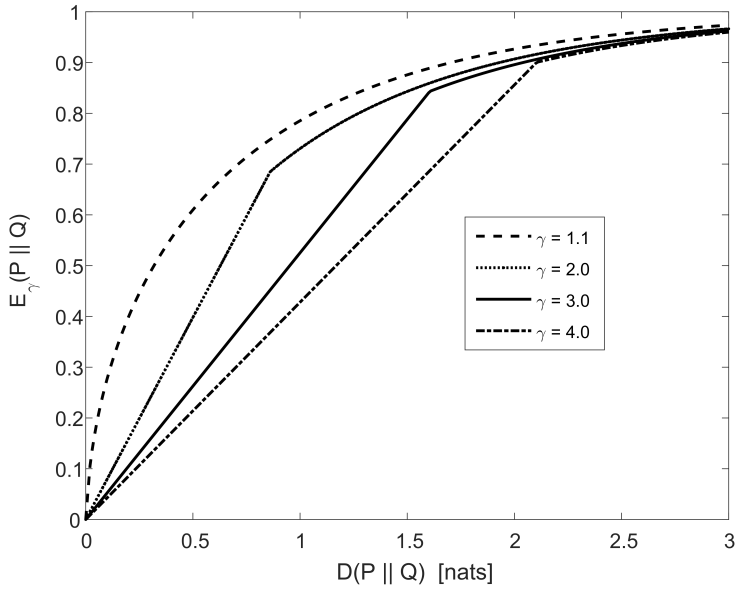

Figure 1. Upper bounds on the E_γ divergence, for $\gamma > 1$, as a function of the relative entropy (the curvy and straight lines follow from (150) and (162), respectively).

Figure 1 plots an upper bound on the E_γ divergence, for $\gamma \in \{1.1, 2.0, 3.0, 4.0\}$, as a function of the relative entropy (or, alternatively, a lower bound on the relative entropy as a function of the E_γ divergence). The upper bound on $E_\gamma(P\|Q)$ for $\gamma > 1$, as a function of $D(P\|Q)$, is composed of the following two components:

- the straight-line bound, which refers to the right side of (162), is tighter than the bound in the right side of (150) if the relative entropy is below a certain value that is denoted by $d(\gamma)$ in nats (it depends on γ);
- the curvy line, which refers to the bound in the right side of (150), is tighter than the straight-line bound in the right side of (162) for larger values of the relative entropy.

It is supported by Figure 1 that $d\colon (1,\infty) \mapsto (0,\infty)$ is positive and monotonically increasing, and $\lim_{\gamma \downarrow 1} d(\gamma) = 0$; e.g., it can be verified that $d(1.1) \approx 0.02$, $d(2) \approx 0.86$, $d(3) \approx 1.61$, and $d(4) \approx 2.10$ (see Figure 1).

Bayesian Binary Hypothesis Testing

The DeGroot statistical information [16] has the following meaning: consider two hypotheses H_0 and H_1, and let $\mathbb{P}[H_0] = \omega$ and $\mathbb{P}[H_1] = 1 - \omega$ with $\omega \in (0,1)$. Let P and Q be probability measures, and consider an observation Y where $Y|H_0 \sim P$, and $Y|H_1 \sim Q$. Suppose that one wishes to decide which hypothesis is more likely given the observation Y. The operational meaning of the DeGroot statistical information, denoted by $\mathcal{I}_\omega(P\|Q)$, is that this measure is equal to the minimal difference between the *a-priori* error probability (without side information) and *a posteriori* error probability (given the observation Y). This measure was later identified as an f-divergence by Liese and Vajda [3] (see (50) here).

Theorem 7. *The DeGroot statistical information satisfies the following upper bound as a function of the chi-squared divergence:*

$$\mathcal{I}_\omega(P\|Q) \leq \begin{cases} \omega - \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{\omega(1-\omega)}{1+\omega\,\chi^2(P\|Q)}}\,, & \omega \in \left(0, \frac{1}{2}\right], \\ \frac{1}{2} - \omega + \sqrt{\frac{1}{4} - \frac{\omega(1-\omega)}{1+\omega\,\chi^2(Q\|P)}}\,, & \omega \in \left(\frac{1}{2}, 1\right), \end{cases} \tag{163}$$

and the following bounds as a function of the relative entropy:

(1)

$$\mathcal{I}_\omega(P\|Q) \leq \begin{cases} \omega\, c_{\frac{1-\omega}{\omega}}\, D(P\|Q)\,, & \omega \in \left(0, \frac{1}{2}\right), \\ \sqrt{\frac{1}{8 \log e}\, \min\{D(P\|Q), D(Q\|P)\}}\,, & \omega = \frac{1}{2}, \\ (1-\omega)\, c_{\frac{\omega}{1-\omega}}\, D(Q\|P)\,, & \omega \in \left(\frac{1}{2}, 1\right), \end{cases} \tag{164}$$

where c_γ for $\gamma > 1$ is introduced in (160);

(2)

$$\mathcal{I}_\omega(P\|Q) \leq \begin{cases} \omega - \frac{1}{2} + \sqrt{\frac{1}{4} - \omega(1-\omega)\, \exp\left(-D(P\|Q)\right)}\,, & \omega \in \left(0, \frac{1}{2}\right], \\ \frac{1}{2} - \omega + \sqrt{\frac{1}{4} - \omega(1-\omega)\, \exp\left(-D(Q\|P)\right)}\,, & \omega \in \left(\frac{1}{2}, 1\right). \end{cases} \tag{165}$$

Proof. The first bound in (163) holds by combining (53) and (149); the second bound in (164) follows from (162) and (53) for $\omega \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right)$, and it follows from (52) and (154) when $\omega = \frac{1}{2}$; finally, the third bound in (165) follows from (150) and (53). □

Remark 13. *The bound in* (164) *forms an extension of Pinsker's inequality* (154) *when* $\omega \neq \frac{1}{2}$ *(i.e., in the asymmetric case where the hypotheses* H_0 *and* H_1 *are not equally probable). Furthermore, in view of* (52)*, the bound in* (165) *is specialized to the Bretagnole-Huber inequality in* (153) *by letting* $\omega = \frac{1}{2}$.

Remark 14. *Numerical evidence shows that none of the bounds in* (163)–(165) *supersedes the others.*

Remark 15. *The upper bounds on* $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *in* (163) *and* (165) *are asymptotically tight when we let* $D(P\|Q)$ *and* $D(Q\|P)$ *tend to infinity. To verify this, first note that (see [23] Theorem 5)*

$$D(P\|Q) \leq \log\left(1 + \chi^2(P\|Q)\right), \tag{166}$$

which implies that also $\chi^2(P\|Q)$ *and* $\chi^2(Q\|P)$ *tend to infinity. In this case, it can be readily verified that the bounds in* (163) *and* (165) *are specialized to* $\mathcal{I}_\omega(P\|Q) \leq \min\{\omega, 1-\omega\}$*; this upper bound, which is equal to the* a-priori *error probability, is also equal to the DeGroot statistical information since the a-posterior error probability tends to zero in the considered extreme case where* P *and* Q *are sufficiently far from each other, so that* H_0 *and* H_1 *are easily distinguishable in high probability when the observation* Y *is available.*

Remark 16. *Due to the one-to-one correspondence between the* E_γ *divergence and DeGroot statistical information in* (53)*, which shows that the two measures are related by a multiplicative scaling factor, the numerical results shown in Figure 1 also apply to the bounds in* (164) *and* (165)*; i.e., for* $\omega \neq \frac{1}{2}$*, the first bound in* (164) *is tighter than the second bound in* (165) *for small values of the relative entropy, whereas* (165) *becomes tighter than* (164) *for larger values of the relative entropy.*

Corollary 4. *Let* $f \in \mathcal{C}$*, and let* $f^* \in \mathcal{C}$ *be as defined in* (12)*. Then,*

(1) for $w \in \left(0, \frac{1}{2}\right]$,

$$D_f(P\|Q) \geq f^*\left(1 + \frac{\mathcal{I}_w(P\|Q)}{1-w}\right) + f^*\left(\frac{w - \mathcal{I}_w(P\|Q)}{1-w}\right) - f^*\left(\frac{w}{1-w}\right); \tag{167}$$

(2) for $w \in \left(\frac{1}{2}, 1\right)$,

$$D_f(P\|Q) \geq f^*\left(1 + \frac{\mathcal{I}_w(Q\|P)}{w}\right) + f^*\left(\frac{1-w-\mathcal{I}_w(Q\|P)}{w}\right) - f^*\left(\frac{1-w}{w}\right). \tag{168}$$

Proof. Inequalities (167) and (168) follow by combining (135) and (53). □

We end this section by exemplifying the utility of the bounds in Theorem 7.

Example 2. *Let* $\mathbb{P}[H_0] = \omega$ *and* $\mathbb{P}[H_1] = 1 - \omega$ *with* $\omega \in (0,1)$*, and assume that the observation* Y *given that the hypothesis is* H_0 *or* H_1 *is Poisson distributed with the positive parameter* μ *or* λ*, respectively:*

$$Y|H_0 \sim P_\mu, \tag{169}$$

$$Y|H_1 \sim P_\lambda \tag{170}$$

where

$$P_\lambda[k] = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \{0, 1, \ldots\}. \tag{171}$$

Without any loss of generality, let $\omega \in \left(0, \frac{1}{2}\right]$. *The bounds on the DeGroot statistical information* $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *in Theorem 7 can be expressed in a closed form by relying on the following identities:*

$$D(P_\mu \| P_\lambda) = \mu \log\left(\frac{\mu}{\lambda}\right) + (\lambda - \mu) \log e, \tag{172}$$

$$\chi^2(P_\mu \| P_\lambda) = e^{\frac{(\mu-\lambda)^2}{\lambda}} - 1. \tag{173}$$

In this example, we compare the simple closed-form bounds on $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *in* (163)–(165) *with its exact value*

$$\mathcal{I}_\omega(P_\mu \| P_\lambda) = \min\{\omega, 1-\omega\} - \sum_{k=0}^{\infty} \min\Big\{\omega P_\mu[k], (1-\omega) P_\lambda[k]\Big\}. \tag{174}$$

To simplify the right side of (174), *let* $\mu > \lambda$, *and define*

$$k_0 = k_0(\lambda, \mu, \omega) := \left\lfloor \frac{\mu - \lambda + \ln \frac{1-\omega}{\omega}}{\ln \frac{\mu}{\lambda}} \right\rfloor, \tag{175}$$

where for $x \in \mathbb{R}$, $\lfloor x \rfloor$ *denotes the largest integer that is smaller than or equal to* x. *It can be verified that*

$$\begin{cases} \omega P_\mu[k] \leq (1-\omega) P_\lambda[k], & \text{for } k \leq k_0 \\ \omega P_\mu[k] > (1-\omega) P_\lambda[k], & \text{for } k > k_0. \end{cases} \tag{176}$$

Hence, from (174)–(176),

$$\mathcal{I}_\omega(P_\mu \| P_\lambda) = \min\{\omega, 1-\omega\} - \omega \sum_{k=0}^{k_0} P_\mu[k] - (1-\omega) \sum_{k=k_0+1}^{\infty} P_\lambda[k] \tag{177}$$

$$= \min\{\omega, 1-\omega\} - \omega \sum_{k=0}^{k_0} P_\mu[k] - (1-\omega) \left(1 - \sum_{k=0}^{k_0} P_\lambda[k]\right). \tag{178}$$

To exemplify the utility of the bounds in Theorem 7, suppose that μ *and* λ *are close, and we wish to obtain a guarantee on how small* $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *is. For example, let* $\lambda = 99$, $\mu = 101$, *and* $\omega = \frac{1}{10}$. *The upper bounds on* $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *in* (163)–(165) *are, respectively, equal to* $4.6 \cdot 10^{-4}$, $5.8 \cdot 10^{-4}$ *and* $2.2 \cdot 10^{-3}$; *we therefore get an informative guarantee by easily calculable bounds. The exact value of* $\mathcal{I}_\omega(P_\mu \| P_\lambda)$ *is, on the other hand, hard to compute since* $k_0 = 209$ *(see* (175)), *and the calculation of the right side of* (178) *appears to be sensitive to the selected parameters in this setting.*

5. Local Behavior of f-Divergences

This section studies the local behavior of f-divergences; the starting point relies on [56] Section 3 which studies the asymptotic properties of f-divergences. The reader is also referred to a related study in [22] Section 4.F.

Lemma 4. *Let*

- $\{P_n\}$ *be a sequence of probability measures on a measurable space* $(\mathcal{A}, \mathscr{F})$;
- *the sequence* $\{P_n\}$ *converge to a probability measure* Q *in the sense that*

$$\lim_{n \to \infty} ess\,sup \frac{dP_n}{dQ}(Y) = 1, \quad Y \sim Q \tag{179}$$

where $P_n \ll Q$ *for all sufficiently large* n;
- $f, g \in \mathcal{C}$ *have continuous second derivatives at 1 and* $g''(1) > 0$.

Then

$$\lim_{n\to\infty} \frac{D_f(P_n\|Q)}{D_g(P_n\|Q)} = \frac{f''(1)}{g''(1)}. \tag{180}$$

Proof. The result in (180) follows from [56] Theorem 3, even without the additional restriction in [56] Section 3 which would require that the second derivatives of f and g are locally Lipschitz at a neighborhood of 1. More explicitly, in view of the analysis in [56] p. 1863, we get by relaxing the latter restriction that (cf. [56] (31))

$$\left|D_f(P_n\|Q) - \tfrac{1}{2}\, f''(1)\, \chi^2(P_n\|Q)\right| \le \tfrac{1}{2} \sup_{y\in[1-\varepsilon_n, 1+\varepsilon_n]} \left|f''(y) - f''(1)\right| \chi^2(P_n\|Q), \tag{181}$$

where $\varepsilon_n \downarrow 0$ as we let $n \to \infty$, and also

$$\lim_{n\to\infty} \chi^2(P_n\|Q) = 0. \tag{182}$$

By our assumption, due to the continuity of f'' and g'' at 1, it follows from (181) and (182) that

$$\lim_{n\to\infty} \frac{D_f(P_n\|Q)}{\chi^2(P_n\|Q)} = \tfrac{1}{2}\, f''(1), \tag{183}$$

$$\lim_{n\to\infty} \frac{D_g(P_n\|Q)}{\chi^2(P_n\|Q)} = \tfrac{1}{2}\, g''(1), \tag{184}$$

which yields (180) (recall that, by assumption, $g''(1) > 0$). □

Remark 17. *Since f and g in Lemma 4 are assumed to have continuous second derivatives at 1, the left and right derivatives of the weight function w_f in (79) at 1 satisfy, in view of Remark 3,*

$$w_f'(1^+) = -w_f'(1^-) = f''(1). \tag{185}$$

Hence, the limit in the right side of (180) is equal to $\frac{w_f'(1^+)}{w_g'(1^+)}$ or also to $\frac{w_f'(1^-)}{w_g'(1^-)}$.

Lemma 5.

$$\chi^2(\lambda P + (1-\lambda)Q \,\|\, Q) = \lambda^2\, \chi^2(P\|Q), \quad \forall\, \lambda \in [0,1]. \tag{186}$$

Proof. Let $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ be the densities of P and Q with respect to an arbitrary probability measure μ such that $P, Q \ll \mu$. Then,

$$\chi^2(\lambda P + (1-\lambda)Q \,\|\, Q) = \int \frac{\big((\lambda p + (1-\lambda)q) - q\big)^2}{q}\, \mathrm{d}\mu \tag{187}$$

$$= \lambda^2 \int \frac{(p-q)^2}{q}\, \mathrm{d}\mu \tag{188}$$

$$= \lambda^2\, \chi^2(P\|Q). \tag{189}$$

□

Remark 18. *The result in Lemma 5, for the chi-squared divergence, is generalized to the identity*

$$\chi^s(\lambda P + (1-\lambda)Q \,\|\, Q) = \lambda^s\, \chi^s(P\|Q), \quad \forall\, \lambda \in [0,1], \tag{190}$$

for all $s \geq 1$ (see (33)). The special case of $s = 2$ is required in the continuation of this section.

Remark 19. *The result in Lemma 5 can be generalized as follows: let P, Q, R be probability measures, and $\lambda \in [0,1]$. Let $P, Q, R \ll \mu$ for an arbitrary probability measure μ, and $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$, and $r := \frac{\mathrm{d}R}{\mathrm{d}\mu}$ be the corresponding densities with respect to μ. Calculation shows that*

$$\chi^2(\lambda P + (1-\lambda)Q \,\|\, R) - \chi^2(Q\|R) = c\lambda + \left[\chi^2(P\|R) - \chi^2(Q\|R) - c\right]\lambda^2 \tag{191}$$

with

$$c := \int \frac{(p-q)q}{r}\,\mathrm{d}\mu. \tag{192}$$

If $Q = R$, then $c = 0$ in (192), and (191) is specialized to (186). However, if $Q \neq R$, then c may be non-zero. This shows that, for small $\lambda \in [0,1]$, the left side of (191) scales linearly in λ if $c \neq 0$, and it has a quadratic scaling in λ if $c = 0$ and $\chi^2(P\|R) \neq \chi^2(Q\|R)$ (e.g., if $Q = R$, as in Lemma 5). The identity in (191) yields

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\chi^2(\lambda P + (1-\lambda)Q \,\|\, R)\Big|_{\lambda=0} = \lim_{\lambda \downarrow 0} \frac{\chi^2(\lambda P + (1-\lambda)Q \,\|\, R) - \chi^2(Q\|R)}{\lambda} = c. \tag{193}$$

We next state the main result in this section.

Theorem 8. *Let*

- *P and Q be probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, $Y \sim Q$, and suppose that*

$$\operatorname{ess\,sup} \frac{dP}{dQ}\,(Y) < \infty; \tag{194}$$

- *$f \in \mathcal{C}$, and f'' be continuous at 1.*

Then,

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2}\, D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2}\, D_f(Q \,\|\, \lambda P + (1-\lambda)Q) \tag{195}$$

$$= \tfrac{1}{2}\, f''(1)\, \chi^2(P\|Q). \tag{196}$$

Proof. Let $\{\lambda_n\}_{n \in \mathbb{N}}$ be a sequence in $[0,1]$, which tends to zero. Define the sequence of probability measures

$$R_n := \lambda_n P + (1-\lambda_n)Q, \qquad n \in \mathbb{N}. \tag{197}$$

Note that $P \ll Q$ implies that $R_n \ll Q$ for all $n \in \mathbb{N}$. Since

$$\frac{\mathrm{d}R_n}{\mathrm{d}Q} = \lambda_n \frac{\mathrm{d}P}{\mathrm{d}Q} + (1-\lambda_n), \tag{198}$$

it follows from (194) that

$$\lim_{n \to \infty} \operatorname{ess\,sup} \frac{\mathrm{d}R_n}{\mathrm{d}Q}\,(Y) = 1. \tag{199}$$

Consequently, (183) implies that

$$\lim_{n \to \infty} \frac{D_f(R_n\|Q)}{\chi^2(R_n\|Q)} = \tfrac{1}{2}\, f''(1) \tag{200}$$

where $\{\lambda_n\}$ in (197) is an arbitrary sequence which tends to zero. Hence, it follows from (197) and (200) that

$$\lim_{\lambda \downarrow 0} \frac{D_f(\lambda P + (1-\lambda)Q \,\|\, Q)}{\chi^2(\lambda P + (1-\lambda)Q \,\|\, Q)} = \tfrac{1}{2} f''(1), \tag{201}$$

and, by combining (186) and (201), we get

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = \tfrac{1}{2} f''(1)\, \chi^2(P\|Q). \tag{202}$$

We next prove the result for the limit in the right side of (195). Let $f^*\colon (0,\infty) \mapsto \mathbb{R}$ be the conjugate function of f, which is given in (12). By the assumption that f has a second continuous derivative, so is f^* and it is easy to verify that the second derivatives of f and f^* coincide at 1. Hence, from (13) and (202),

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_f(Q \,\|\, \lambda P + (1-\lambda)Q) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_{f^*}(\lambda P + (1-\lambda)Q \,\|\, Q) \tag{203}$$

$$= \tfrac{1}{2} f''(1)\, \chi^2(P\|Q). \tag{204}$$

□

Remark 20. *Although an f-divergence is in general not symmetric, in the sense that the equality $D_f(P\|Q) = D_f(Q\|P)$ does not necessarily hold for all pairs of probability measures (P, Q), the reason for the equality in (195) stems from the fact that the second derivatives of f and f^* coincide at 1 when f is twice differentiable.*

Remark 21. *Under the conditions in Theorem 8, it follows from (196) that*

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} D_f(\lambda P + (1-\lambda)Q \,\|\, Q)\Big|_{\lambda=0} = \lim_{\lambda \downarrow 0} \frac{1}{\lambda} D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = 0, \tag{205}$$

$$\lim_{\lambda \downarrow 0} \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = 2 \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = f''(1)\, \chi^2(P\|Q) \tag{206}$$

where (206) relies on L'Hôpital's rule. The convexity of $D_f(P\|Q)$ in (P, Q) also implies that, for all $\lambda \in [0,1]$,

$$D_f(\lambda P + (1-\lambda)Q \,\|\, Q) \le \lambda D_f(P\|Q). \tag{207}$$

The following result refers to the local behavior of Rényi divergences of an arbitrary non-negative order.

Corollary 5. *Under the condition in (194), for every $\alpha \in [0,\infty]$,*

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_\alpha(\lambda P + (1-\lambda)Q \,\|\, Q) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_\alpha(Q \,\|\, \lambda P + (1-\lambda)Q) \tag{208}$$

$$= \tfrac{1}{2}\, \alpha\, \chi^2(P\|Q)\, \log e. \tag{209}$$

Proof. Let $\alpha \in (0,1) \cup (1,\infty)$. In view of (23) and Theorem 8, it follows that the local behavior of the Hellinger divergence of order α satisfies

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} \mathscr{H}_\alpha(\lambda P + (1-\lambda)Q \,\|\, Q) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} \mathscr{H}_\alpha(Q \,\|\, \lambda P + (1-\lambda)Q) \tag{210}$$

$$= \tfrac{1}{2}\, \alpha\, \chi^2(P\|Q). \tag{211}$$

The result now follows from (30), which implies that

$$\lim_{\lambda \downarrow 0} \frac{D_\alpha(\lambda P + (1-\lambda)Q \,\|\, Q)}{\mathscr{H}_\alpha(\lambda P + (1-\lambda)Q \,\|\, Q)} = \lim_{\lambda \downarrow 0} \frac{D_\alpha(Q \,\|\, \lambda P + (1-\lambda)Q)}{\mathscr{H}_\alpha(Q \,\|\, \lambda P + (1-\lambda)Q)} \tag{212}$$

$$= \frac{1}{\alpha - 1} \lim_{u \to 0} \frac{\log\big(1 + (\alpha - 1)u\big)}{u} \tag{213}$$

$$= \log e. \tag{214}$$

The result in (208) and (209), for $\alpha \in (0,1) \cup (1,\infty)$, follows by combining the equalities in (210)–(214). Finally, the result in (208) and (209) for $\alpha \in \{0, 1, \infty\}$ follows from its validity for all $\alpha \in (0,1) \cup (1,\infty)$, and also due to the property where $D_\alpha(\cdot\|\cdot)$ is monotonically increasing in α (see [73] Theorem 3). □

Acknowledgments: The author is grateful to Sergio Verdú and the two anonymous reviewers, whose suggestions improved the presentation in this paper.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 2

We prove in the following the integral representations of f-divergences and related measures in Theorem 2.

(1) Relative entropy: The function $f \in \mathcal{C}$ in (18) yields the following weight function in (79):

$$w_f(\beta) = \left(\frac{1}{\beta} - \frac{1}{\beta^2}\right) \big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big) \log e, \quad \beta > 0. \tag{A1}$$

Consequently, setting $c := \log e$ in (82) yields

$$\widetilde{w}_{f,c}(\beta) = \frac{1}{\beta} \big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big) \log e, \tag{A2}$$

for $\beta > 0$. Equality (98) follows from the substitution of (A2) into the right side of (83).

(2) Hellinger divergence: In view of (22), for $\alpha \in (0,1) \cup (1,\infty)$, the weight function $w_{f_\alpha} \colon (0,\infty) \mapsto [0,\infty)$ in (79) which corresponds to $f_\alpha \colon (0,\infty) \mapsto \mathbb{R}$ in (23) can be verified to be equal to

$$w_{f_\alpha}(\beta) = \left(\beta^{\alpha-2} - \frac{1}{\beta^2}\right) \big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big) \tag{A3}$$

for $\beta > 0$. In order to simplify the integral representation of the Hellinger divergence $\mathscr{H}_\alpha(P\|Q)$, we apply Theorem 1-(1). From (A3), setting $c := 1$ in (82) implies that $\widetilde{w}_{f_\alpha,1} \colon (0,\infty) \to \mathbb{R}$ is given by

$$\widetilde{w}_{f_\alpha,1}(\beta) = \beta^{\alpha-2} \big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big) \tag{A4}$$

for $\beta > 0$. Hence, substituting (80) and (A4) into (83) yields

$$\mathscr{H}_\alpha(P\|Q) = \int_1^\infty \beta^{\alpha-2} \big(1 - \mathbb{F}_{P\|Q}(\log\beta)\big) \, \mathrm{d}\beta - \int_0^1 \beta^{\alpha-2} \, \mathbb{F}_{P\|Q}(\log\beta) \, \mathrm{d}\beta. \tag{A5}$$

For $\alpha > 1$, (A5) yields

$$\mathscr{H}_\alpha(P\|Q) = \int_0^\infty \beta^{\alpha-2}\left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\mathrm{d}\beta - \int_0^1 \beta^{\alpha-2}\,\mathrm{d}\beta \tag{A6}$$

$$= \int_0^\infty \beta^{\alpha-2}\left(1 - \mathbb{F}_{P\|Q}(\log\beta)\right)\mathrm{d}\beta - \frac{1}{\alpha-1}, \tag{A7}$$

and, for $\alpha \in (0,1)$, (A5) yields

$$\mathscr{H}_\alpha(P\|Q) = \int_1^\infty \beta^{\alpha-2}\,\mathrm{d}\beta - \int_0^\infty \beta^{\alpha-2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta \tag{A8}$$

$$= \frac{1}{1-\alpha} - \int_0^\infty \beta^{\alpha-2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta. \tag{A9}$$

This proves (99). We next consider the following special cases:

- In view of (25), equality (100) readily follows from (99) with $\alpha = 2$.
- In view of (28), equality (101) readily follows from (99) with $\alpha = \frac{1}{2}$.
- In view of (29), equality (102) readily follows from (101).

(3) Rényi divergence: In view of the one-to-one correspondence in (30) between the Rényi divergence and the Hellinger divergence of the same order, (103) readily follows from (99).

(4) χ^s divergence with $s \geq 1$: We first consider the case where $s > 1$. From (33), the function $f_s\colon (0,\infty) \mapsto \mathbb{R}$ in (32) is differentiable and $f_s'(1) = 0$. Hence, the respective weight function $w_{f_s}\colon (0,\infty) \mapsto (0,\infty)$ can be verified from (79) to be given by

$$w_{f_s}(\beta) = \frac{1}{\beta}\left(s - 1 + \frac{1}{\beta}\right)|\beta - 1|^{s-1}, \quad \beta > 0. \tag{A10}$$

The result in (104), for $s > 1$, follows readily from (33), (80), (81) and (A10).

We next prove (104) with $s = 1$. In view of (32), (34), (35) and the dominated convergence theorem,

$$|P - Q| = \lim_{s\downarrow 1} \chi^s(P\|Q) \tag{A11}$$

$$= \int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta + \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta. \tag{A12}$$

This extends (104) for all $s \geq 1$, although $f_1(t) = |t-1|$ for $t > 0$ is not differentiable at 1. For $s = 1$, in view of (95), the integral representation in the right side of (A12) can be simplified to (105) and (106).

(5) DeGroot statistical information: In view of (50)–(51), since the function $\phi_w\colon (0,\infty) \mapsto \mathbb{R}$ is not differentiable at the point $\frac{1-\omega}{\omega} \in (0,\infty)$ for $\omega \in (0,1)$, Theorem 1 cannot be applied directly to get an integral representation of the DeGroot statistical information. To that end, for $(\omega,\alpha) \in (0,1)^2$, consider the family of convex functions $f_{\omega,\alpha}\colon (0,\infty) \mapsto \mathbb{R}$ given by (see [3] (55))

$$f_{\omega,\alpha}(t) = \frac{1}{1-\alpha}\left(\left[(\omega t)^{\frac{1}{\alpha}} + (1-\omega)^{\frac{1}{\alpha}}\right]^\alpha - \left[\omega^{\frac{1}{\alpha}} + (1-\omega)^{\frac{1}{\alpha}}\right]^\alpha\right), \tag{A13}$$

for $t > 0$. These differentiable functions also satisfy

$$\lim_{\alpha\downarrow 0} f_{\omega,\alpha}(t) = \phi_w(t), \tag{A14}$$

which holds due to the identities

$$\lim_{\alpha \downarrow 0} \left(a^{\frac{1}{\alpha}} + b^{\frac{1}{\alpha}} \right)^{\alpha} = \max\{a, b\}, \qquad a, b \geq 0; \tag{A15}$$

$$\min\{a, b\} = a + b - \max\{a, b\}, \quad a, b \in \mathbb{R}. \tag{A16}$$

The application of Theorem 1-(1) to the set of functions $f_{\omega,\alpha} \in \mathcal{C}$ with

$$c := \frac{(1-\omega)^{\frac{1}{\alpha}}}{\alpha - 1} \left[\omega^{\frac{1}{\alpha}} + (1-\omega)^{\frac{1}{\alpha}} \right]^{\alpha - 1} \tag{A17}$$

yields

$$\widetilde{w}_{f_{\omega,\alpha},c}(\beta) = \frac{1-\omega}{1-\alpha} \frac{1}{\beta^2} \left[1 + \left(\frac{\omega\beta}{1-\omega} \right)^{\frac{1}{\alpha}} \right]^{\alpha - 1} \left[1\{0 < \beta < 1\} - 1\{\beta \geq 1\} \right], \tag{A18}$$

for $\beta > 0$, and

$$D_{f_{\omega,\alpha}}(P\|Q) = \int_0^{\infty} \widetilde{w}_{f_{\omega,\alpha},c}(\beta)\, G_{P\|Q}(\beta)\, \mathrm{d}\beta \tag{A19}$$

with $G_{P\|Q}(\cdot)$ as defined in (80), and $(\omega, \alpha) \in (0,1)^2$. From (A15) and (A18), it follows that

$$\lim_{\alpha \downarrow 0} \widetilde{w}_{f_{\omega,\alpha},c}(\beta) = \frac{1-\omega}{\beta^2} \left[1\{0 < \beta < 1\} - 1\{\beta \geq 1\} \right] \left[\tfrac{1}{2} 1\{\beta = \tfrac{1-\omega}{\omega}\} + 1\{0 < \beta < \tfrac{1-\omega}{\omega}\} \right], \tag{A20}$$

for $\beta > 0$. In view of (50), (51), (80), (A14), (A19) and (A20), and the monotone convergence theorem,

$$\begin{aligned}
&\mathcal{I}_{\omega}(P\|Q) \\
&= D_{\phi_{\omega}}(P\|Q) && \text{(A21)} \\
&= \lim_{\alpha \downarrow 0} D_{f_{\omega,\alpha}}(P\|Q) && \text{(A22)} \\
&= (1-\omega) \int_0^{\min\{1, \frac{1-\omega}{\omega}\}} \frac{\mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta - (1-\omega) \int_1^{\max\{1, \frac{1-\omega}{\omega}\}} \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta, && \text{(A23)}
\end{aligned}$$

for $\omega \in (0,1)$. We next simplify (A23) as follows:

- if $\omega \in \left(1, \frac{1-\omega}{\omega}\right)$, then $\frac{1-\omega}{\omega} < 1$ and (A23) yields

$$\mathcal{I}_{\omega}(P\|Q) = (1-\omega) \int_0^{\frac{1-\omega}{\omega}} \frac{\mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta; \tag{A24}$$

- if $\omega \in \left(0, \frac{1}{2}\right]$, then $\frac{1-\omega}{\omega} \geq 1$ and (A23) yields

$$\begin{aligned}
\mathcal{I}_{\omega}(P\|Q) &= (1-\omega) \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta - (1-\omega) \int_1^{\frac{1-\omega}{\omega}} \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta && \text{(A25)} \\
&= (1-\omega) \int_1^{\infty} \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta - (1-\omega) \int_1^{\frac{1-\omega}{\omega}} \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta && \text{(A26)} \\
&= (1-\omega) \int_{\frac{1-\omega}{\omega}}^{\infty} \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\, \mathrm{d}\beta, && \text{(A27)}
\end{aligned}$$

where (A26) follows from (95) (or its equivalent from in (73)).

This completes the proof of (107). Note that, due to (95), the integral representation of $\mathcal{I}_\omega(P\|Q)$ in (107) is indeed continuous at $\omega = \frac{1}{2}$.

(6) Triangular discrimination: In view of (36)–(37), the corresponding function $\widetilde{w}_{f,1}\colon (0,\infty) \mapsto \mathbb{R}$ in (82) (i.e., with $c := 1$) can be verified to be given by

$$\widetilde{w}_{f,1}(\beta) = \frac{4}{(\beta+1)^2}\,\big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big) \tag{A28}$$

for $\beta > 0$. Substituting (80) and (A28) into (83) proves (108) as follows:

$$\Delta(P\|Q) = 4\left(\int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{(\beta+1)^2}\,\mathrm{d}\beta - \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{(\beta+1)^2}\,\mathrm{d}\beta\right) \tag{A29}$$

$$= 4\left(\int_0^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{(\beta+1)^2}\,\mathrm{d}\beta - \int_0^1 \frac{1}{(\beta+1)^2}\,\mathrm{d}\beta\right) \tag{A30}$$

$$= 4\int_0^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{(\beta+1)^2}\,\mathrm{d}\beta - 2. \tag{A31}$$

(7) Lin's measure and the Jensen-Shannon divergence: Let $\theta \in (0,1)$ (if $\theta \in \{0,1\}$, then (39) and (40) imply that $L_\theta(P\|Q) = 0$). In view of (41), the application of Theorem 1-(1) with the function $f_\theta\colon (0,\infty) \mapsto \mathbb{R}$ in (42) yields the weight function $w_{f_\theta}\colon (0,\infty) \mapsto [0,\infty)$ defined as

$$w_{f_\theta}(\beta) = \frac{(1-\theta)\,\log(\theta\beta + 1 - \theta)}{\beta^2}\,\Big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\Big). \tag{A32}$$

Consequently, we get

$$L_\theta(P\|Q) = (1-\theta)\left(\int_1^\infty \tfrac{\log(\theta\beta+1-\theta)}{\beta^2}\,\Big(1 - \mathbb{F}_{P\|Q}(\log\beta)\Big)\,\mathrm{d}\beta - \int_0^1 \tfrac{\log(\theta\beta+1-\theta)}{\beta^2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta\right) \tag{A33}$$

$$= (1-\theta)\left(\int_1^\infty \frac{\log(\theta\beta+1-\theta)}{\beta^2}\,\mathrm{d}\beta - \int_0^\infty \frac{\log(\theta\beta+1-\theta)}{\beta^2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta\right) \tag{A34}$$

$$= \theta\log\frac{1}{\theta} - (1-\theta)\int_0^\infty \frac{\log(\theta\beta+1-\theta)}{\beta^2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta \tag{A35}$$

$$= h(\theta) - (1-\theta)\int_0^\infty \frac{1}{\beta^2}\,\log\left(\frac{\theta\beta}{1-\theta}+1\right)\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta \tag{A36}$$

where (A33) follows from (80), (81) and (A32); for $\theta \in (0,1)$, equality (A35) holds since

$$\int_1^\infty \frac{\log(\theta\beta+1-\theta)}{\beta^2}\,\mathrm{d}\beta = \frac{\theta}{1-\theta}\,\log\frac{1}{\theta}; \tag{A37}$$

finally, (A36) follows from (73) where $h\colon [0,1] \mapsto [0,\log 2]$ denotes the binary entropy function. This proves (109). In view of (43), the identity in (110) for the Jensen-Shannon divergence follows from (109) with $\theta = \frac{1}{2}$.

(8) Jeffrey's divergence: In view of (20)–(21), the corresponding weight function $w_f\colon (0,\infty) \mapsto [0,\infty)$ in (79) can be verified to be given by

$$w_f(\beta) = \left(\frac{\log e}{\beta} + \frac{1}{\beta^2}\,\log\frac{\beta}{e}\right)\big(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\big)\,. \tag{A38}$$

Hence, setting $c := \log e$ in (82) implies that

$$\widetilde{w}_{f,c}(\beta) = \left(\frac{\log e}{\beta} + \frac{\log \beta}{\beta^2}\right)\left(1\{\beta \geq 1\} - 1\{0 < \beta < 1\}\right) \tag{A39}$$

for $\beta > 0$. Substituting (80) and (A39) into (83) yields (111).

(9) E_γ divergence: Let $\gamma \geq 1$, and let $\omega \in \left(0, \frac{1}{2}\right]$ satisfy $\frac{1-\omega}{\omega} = \gamma$; hence, $\omega = \frac{1}{1+\gamma}$. From (53), we get

$$E_\gamma(P\|Q) = (1+\gamma)\,\mathcal{I}_{\frac{1}{1+\gamma}}(P\|Q). \tag{A40}$$

The second line in the right side of (107) yields

$$\mathcal{I}_{\frac{1}{1+\gamma}}(P\|Q) = \frac{\gamma}{1+\gamma}\int_\gamma^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2}\,\mathrm{d}\beta. \tag{A41}$$

Finally, substituting (A41) into the right side of (A40) yields (112).

Remark A1. *In view of* (95), *the integral representation for the* χ^s *divergence in* (104) *specializes to* (100), (105) *and* (106) *by letting* $s = 2$ *and* $s = 1$, *respectively.*

Remark A2. *In view of* (49), *the first identity for the total variation distance in* (105) *follows readily from* (112) *with* $\gamma = 1$. *The second identity in* (106) *follows from* (73) *and* (105), *and since* $\int_1^\infty \frac{\mathrm{d}\beta}{\beta^2} = 1$.

References

1. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
2. Liese, F.; Vajda, I. Convex Statistical Distances. In *Teubner-Texte Zur Mathematik*; Springer: Leipzig, Germany, 1987; Volume 95.
3. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [CrossRef]
4. Reid, M.D.; Williamson, R.C. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.* **2011**, *12*, 731–817.
5. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: New York, NY, USA, 2009.
6. Vapnik, V.N. *Statistical Learning Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1998.
7. Verdú, S. Information Theory. Unpublished work, 2018.
8. Csiszár, I. Axiomatic characterization of information measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
9. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 131–142.
10. Csiszár, I. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* **1963**, *8*, 85–108.
11. Csiszár, I. A note on Jensen's inequality. *Stud. Sci. Math. Hung.* **1966**, *1*, 185–188.
12. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hung.* **1967**, *2*, 299–318.
13. Csiszár, I. On topological properties of f-divergences. *Stud. Sci. Math. Hung.* **1967**, *2*, 329–339.
14. Morimoto, T. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331. [CrossRef]
15. Liese, F. ϕ-divergences, sufficiency, Bayes sufficiency, and deficiency. *Kybernetika* **2012**, *48*, 690–713.
16. DeGroot, M.H. Uncertainty, information and sequential experiments. *Ann. Math. Stat.* **1962**, *33*, 404–419. [CrossRef]
17. Cohen, J.E.; Kemperman, J.H.B.; Zbăganu, G. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*; Springer: Berlin, Germany, 1998.
18. Feldman, D.; Österreicher, F. A note on f-divergences. *Stud. Sci. Math. Hung.* **1989**, *24*, 191–200.

19. Guttenbrunner, C. On applications of the representation of f-divergences as averaged minimal Bayesian risk. In Proceedings of the Transactions of the 11th Prague Conferences on Information Theory, Statistical Decision Functions, and Random Processes, Prague, Czechoslovakia, 26–31 August 1992; pp. 449–456.
20. Österreicher, F.; Vajda, I. Statistical information and discrimination. *IEEE Trans. Inf. Theory* **1993**, *39*, 1036–1039. [CrossRef]
21. Torgersen, E. *Comparison of Statistical Experiments*; Cambridge University Press: Cambridge, UK, 1991.
22. Sason, I.; Verdú, S. f-divergence inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
23. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435. [CrossRef]
24. Anwar, M.; Hussain, S.; Pecaric, J. Some inequalities for Csiszár-divergence measures. *Int. J. Math. Anal.* **2009**, 3, 1295–1304.
25. Simic, S. On logarithmic convexity for differences of power means. *J. Inequal. Appl.* **2007**, *2007*, 37359. [CrossRef]
26. Simic, S. On a new moments inequality. *Stat. Probab. Lett.* **2008**, *78*, 2671–2678. [CrossRef]
27. Simic, S. On certain new inequalities in information theory. *Acta Math. Hung.*, **2009**, *124*, 353–361. [CrossRef]
28. Simic, S. Moment Inequalities of the Second and Third Orders. Preprint. Available online: http://arxiv.org/abs/1509.0851 (accessed on 13 May 2016).
29. Harremoës, P.; Vajda, I. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory* **2011**, *57*, 3230–3235. [CrossRef]
30. Sason, I.; Verdú, S. f-divergence inequalities via functional domination. In Proceedings of the 2016 IEEE International Conference on the Science of Electrical Engineering, Eilat, Israel, 16–18 November 2016; pp. 1–5.
31. Taneja, I.J. Refinement inequalities among symmetric divergence measures. *Aust. J. Math. Anal. Appl.* **2005**, *2*, 1–23.
32. Taneja, I.J. Seven means, generalized triangular discrimination, and generating divergence measures. *Information* **2013**, *4*, 198—239. [CrossRef]
33. Guntuboyina, A.; Saha, S.; Schiebinger, G. Sharp inequalities for f-divergences. *IEEE Trans. Inf. Theory* **2014**, *60*, 104–121. [CrossRef]
34. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
35. Kafka, P.; Östreicher, F.; Vincze, I. On powers of f-divergences defining a distance. *Stud. Sci. Math. Hung.* **1991**, *26*, 415–422.
36. Lu, G.; Li, B. A class of new metrics based on triangular discrimination. *Information* **2015**, *6*, 361–374. [CrossRef]
37. Vajda, I. On metric divergences of probability measures. *Kybernetika* **2009**, *45*, 885–900.
38. Gilardoni, G.L. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Trans. Inf. Theory* **2010**, *56*, 5377–5386. [CrossRef]
39. Topsøe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609. [CrossRef]
40. Sason, I.; Verdú, S. Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets. In Proceedings of the 2015 IEEE Information Theory Workshop, Jeju Island, Korea, 11–15 October 2015; pp. 214–218.
41. Dragomir, S.S. Upper and lower bounds for Csiszár f-divergence in terms of the Kullback-Leibler divergence and applications. In *Inequalities for Csiszár f-Divergence in Information Theory, RGMIA Monographs*; Victoria University: Footscray, VIC, Australia, 2000.
42. Dragomir, S.S. Upper and lower bounds for Csiszár f-divergence in terms of Hellinger discrimination and applications. In *Inequalities for Csiszár f-Divergence in Information Theory, RGMIA Monographs*; Victoria University: Footscray, VIC, Australia, 2000.
43. Dragomir, S.S. An upper bound for the Csiszár f-divergence in terms of the variational distance and applications. In *Inequalities for Csiszár f-Divergence in Information Theory, RGMIA Monographs*; Victoria University: Footscray, VIC, Australia, 2000.
44. Dragomir, S.S.; Gluščević V. Some inequalities for the Kullback-Leibler and χ^2-distances in information theory and applications. *Tamsui Oxf. J. Math. Sci.* **2001**, *17*, 97–111.
45. Dragomir, S.S. Bounds for the normalized Jensen functional. *Bull. Aust. Math. Soc.* **2006**, *74*, 471–478. [CrossRef]
46. Kumar, P.; Chhina, S. A symmetric information divergence measure of the Csiszár's f-divergence class and its bounds. *Comp. Math. Appl.* **2005**, *49*, 575–588. [CrossRef]

47. Taneja, I.J. Bounds on non-symmetric divergence measures in terms of symmetric divergence measures. *J. Comb. Inf. Syst. Sci.* **2005**, *29*, 115—134.
48. Binette, O. A note on reverse Pinsker inequalities. Preprint. Available online: http://arxiv.org/abs/1805.05135 (accessed on 14 May 2018).
49. Gilardoni, G.L. On the minimum f-divergence for given total variation. *C. R. Math.* **2006**, *343*, 763–766. [CrossRef]
50. Gilardoni, G.L. Corrigendum to the note on the minimum f-divergence for given total variation. *C. R. Math.* **2010**, *348*, 299. [CrossRef]
51. Gushchin, A.A. The minimum increment of f-divergences given total variation distances. *Math. Methods Stat.* **2016**, *25*, 304–312. [CrossRef]
52. Sason, I. Tight bounds on symmetric divergence measures and a refined bound for lossless source coding. *IEEE Trans. Inf. Theory* **2015**, *61*, 701–707. [CrossRef]
53. Sason, I. On the Rényi divergence, joint range of relative entropies, and a channel coding theorem. *IEEE Trans. Inf. Theory* **2016**, *62*, 23–34. [CrossRef]
54. Liu, J.; Cuff, P.; Verdú, S. E_γ-resolvability. *IEEE Trans. Inf. Theory* **2017**, *63*, 2629–2658.
55. Csiszár, I.; Shields, P.C. Information Theory and Statistics: A Tutorial. *Found. Trends Commun. Inf. Theory* **2004**, *1*, 417–528. [CrossRef]
56. Pardo, M.C.; Vajda, I. On asymptotic properties of information-theoretic divergences. *IEEE Trans. Inf. Theory* **2003**, *49*, 1860–1868. [CrossRef]
57. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
58. Bretagnolle, J.; Huber, C. Estimation des densités: Risque minimax. *Probab. Theory Relat. Fields* **1979**, *47*, 119–137.
59. Vajda, I. Note on discrimination information and variation. *IEEE Trans. Inf. Theory* **1970**, *16*, 771–773. [CrossRef]
60. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1946**, *186*, 453–461. [CrossRef]
61. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
62. Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*; Springer: New York, NY, USA, 1986.
63. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
64. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: New York, NY, USA, 2000.
65. Cichocki, A.; Amari, S.I. Families of Alpha- Beta- and Gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568. [CrossRef]
66. Cichocki, A.; Cruces, S.; Amari, S.I. Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170. [CrossRef]
67. Cichocki, A.; Cruces, S.; Amari, S.I. Log-determinant divergences revisited: Alpha-Beta and Gamma log-det divergences. *Entropy* **2015**, *17*, 2988–3034. [CrossRef]
68. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
69. Polyanskiy, Y.; Wu, Y. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory* **2016**, *62*, 35–55. [CrossRef]
70. Kullback, S. A lower bound for discrimination information in terms of variation. *IEEE Trans. Inf. Theory* **1967**, *13*, 126–127. [CrossRef]
71. Kemperman, J.H.B. On the optimal rate of transmitting information. *Ann. Math. Stat.* **1969**, *40*, 2156–2177. [CrossRef]
72. Corless, R.M.; Gonnet, G.H.; Hare, D.E.; Jeffrey, D.J.; Knuth, D.E. On the Lambert W function. *Adv. Comput. Math.* **1996**, *5*, 329–359. [CrossRef]
73. Van Erven, T.; Harremoës, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]