f-Divergence Inequalities

Igal Sason, *Senior Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

Abstract—This paper develops systematic approaches to obtain f-divergence inequalities, dealing with pairs of probability measures defined on arbitrary alphabets. Functional domination is one such approach, where special emphasis is placed on finding the best possible constant upper bounding a ratio of f-divergences. Another approach used for the derivation of bounds among f-divergences relies on moment inequalities and the logarithmic-convexity property, which results in tight bounds on the relative entropy and Bhattacharyya distance in terms of χ^2 divergences. A rich variety of bounds are shown to hold under boundedness assumptions on the relative information. Special attention is devoted to the total variation distance and its relation to the relative information and relative entropy, including "reverse Pinsker inequalities," as well as on the E_γ divergence, which generalizes the total variation distance. Pinsker's inequality is extended for this type of f-divergence, a result which leads to an inequality linking the relative entropy and relative information spectrum. Integral expressions of the Rényi divergence in terms of the relative information spectrum are derived, leading to bounds on the Rényi divergence in terms of either the variational distance or relative entropy.

Index Terms—Relative entropy, total variation distance, f-divergence, Rényi divergence, Pinsker's inequality, relative information.

I. INTRODUCTION

THROUGHOUT their development, information theory, and more generally, probability theory, have benefitted from non-negative measures of dissimilarity, or loosely speaking, distances, between pairs of probability measures defined on the same measurable space (see, e.g., [41], [65], [105]). Notable among those measures are (see Section II for definitions):

- total variation distance $|P - Q|$;
- relative entropy $D(P\|Q)$;
- χ^2-divergence $\chi^2(P\|Q)$;
- Hellinger divergence $\mathscr{H}_\alpha(P\|Q)$;
- Rényi divergence $D_\alpha(P\|Q)$.

It is useful, particularly in proving convergence results, to give bounds of one measure of dissimilarity in terms of another. The most celebrated among those bounds is Pinsker's

inequality[1]:

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

proved by Csiszár[2] [23] and Kullback and Leibler [60], with Kemperman [56] independently a bit later. Improved and generalized versions of Pinsker's inequality have been studied, among others, in [39], [44], [46], [75], [85], [91], and [103].

Relationships among measures of distances between probability measures have long been a focus of interest in probability theory and statistics (e.g., for studying the rate of convergence of measures). The reader is referred to surveys in [41, Sec. 3], [65, Ch. 2], [85] and [86, Appendix 3], which provide several relationships among useful f-divergences and other measures of dissimilarity between probability measures. Some notable existing bounds among f-divergences include, in addition to (1):

- [63, Lemma 1], [64, p. 25]

$$\mathscr{H}_{\frac{1}{2}}^2(P\|Q) \le |P - Q|^2 \tag{2}$$

$$\le \mathscr{H}_{\frac{1}{2}}(P\|Q)\big(4 - \mathscr{H}_{\frac{1}{2}}(P\|Q)\big); \tag{3}$$

- [15, eq. (2.2)]

$$\tfrac{1}{4}|P - Q|^2 \le 1 - \exp\big(-D(P\|Q)\big); \tag{4}$$

- [41, Th. 5], [35, Th. 4], [99]

$$D(P\|Q) \le \log\big(1 + \chi^2(P\|Q)\big); \tag{5}$$

- [48, Corollary 5.6] For all $\alpha \ge 2$

$$\chi^2(P\|Q) \le \Big(1 + (\alpha - 1)\mathscr{H}_\alpha(P\|Q)\Big)^{\frac{1}{\alpha-1}} - 1; \tag{6}$$

the inequality in (6) is reversed if $\alpha \in (0, 1) \cup (1, 2]$, and it holds with equality if $\alpha = 2$.

- [44], [45], [85, eq. (58)]

$$\chi^2(P\|Q) \ge \begin{cases} |P - Q|^2, & |P - Q| \in [0, 1] \\ \frac{|P-Q|}{2-|P-Q|}, & |P - Q| \in (1, 2). \end{cases} \tag{7}$$

- [96]

$$\frac{D^2(P\|Q)}{D(Q\|P)} \le \tfrac{1}{2}\chi^2(P\|Q) \log e; \tag{8}$$

$$4\mathscr{H}_{\frac{1}{2}}^2(P\|Q) \log^2 e \le D(P\|Q)D(Q\|P) \tag{9}$$

$$\le \tfrac{1}{4}\chi^2(P\|Q)\chi^2(Q\|P) \log^2 e; \tag{10}$$

---

[1]The folklore in information theory is that (1) is due to Pinsker [77], albeit with a suboptimal constant. As explained in [109], although no such inequality appears in [77], it is possible to put together two of Pinsker's bounds to conclude that $\frac{1}{408}|P - Q|^2 \log e \le D(P\|Q)$.

[2]Csiszár derived (1) in [23, Th. 4.1] after publishing a weaker version in [22, Corollary 1] a year earlier.

$$4 \mathscr{H}_{\frac{1}{2}}(P\|Q) \log e \leq D(P\|Q) + D(Q\|P) \tag{11}$$

$$\leq \tfrac{1}{2}\left(\chi^2(P\|Q) + \chi^2(Q\|P)\right) \log e; \tag{12}$$

- [32, eq. (2.8)]

$$D(P\|Q) \leq \tfrac{1}{2}\left(|P - Q| + \chi^2(P\|Q)\right) \log e; \tag{13}$$

- [46], [85, Corollary 32], [90]

$$D(P\|Q) + D(Q\|P)$$
$$\geq |P - Q| \log \left(\frac{2 + |P - Q|}{2 - |P - Q|}\right), \tag{14}$$

$$\chi^2(P\|Q) + \chi^2(Q\|P)$$
$$\geq \frac{8|P - Q|^2}{4 - |P - Q|^2}; \tag{15}$$

- [53, p. 711] (cf. a generalized form in [86, Lemma A.3.5])

$$\mathscr{H}_{\frac{1}{2}}(P\|Q) \log e \leq D(P\|Q), \tag{16}$$

generalized in [65, Proposition 2.15]:

$$\mathscr{H}_\alpha(P\|Q) \log e \leq D_\alpha(P\|Q) \leq D(P\|Q), \tag{17}$$

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

$$\mathscr{H}_\alpha(P\|Q) \log e \geq D_\alpha(P\|Q) \geq D(P\|Q) \tag{18}$$

for $\alpha \in (1, \infty)$.

- [65, Proposition 2.35] If $\alpha \in (0, 1)$, $\beta \triangleq \max\{\alpha, 1 - \alpha\}$, then

$$1 - \left(1 + \tfrac{1}{2}|P - Q|\right)^\beta \left(1 - \tfrac{1}{2}|P - Q|\right)^{1-\beta}$$
$$\leq (1 - \alpha)\mathscr{H}_\alpha(P\|Q) \tag{19}$$
$$\leq \tfrac{1}{2}|P - Q|. \tag{20}$$

- [37, Ths. 3 and 16]
  - $D_\alpha(P\|Q)$ is monotonically increasing in $\alpha > 0$;
  - $\left(\frac{1}{\alpha} - 1\right) D_\alpha(P\|Q)$ is monotonically decreasing in $\alpha \in (0, 1]$;
  - [65, Proposition 2.7] the same monotonicity properties hold for $\mathscr{H}_\alpha(P\|Q)$.

- [46] If $\alpha \in (0, 1]$, then

$$\frac{\alpha}{2}|P - Q|^2 \log e \leq D_\alpha(P\|Q); \tag{21}$$

- An inequality for $\mathscr{H}_\alpha(P\|Q)$ [61, Lemma 1] becomes (at $\alpha = 1$), the parallelogram identity [26, eq. (2.2)]

$$D(P_0\|Q) + D(P_1\|Q)$$
$$= D(P_0\|P_{\frac{1}{2}}) + D(P_1\|P_{\frac{1}{2}}) + 2D(P_{\frac{1}{2}}\|Q). \tag{22}$$

with $P_{\frac{1}{2}} = \frac{1}{2}(P_0 + P_1)$, and extends a result for the relative entropy in [26, Th. 2.1].

- A "reverse Pinsker inequality", providing an upper bound on the relative entropy in terms of the total variation distance, does not exist in general since we can find distributions which are arbitrarily close in total variation but with arbitrarily high relative entropy. Nevertheless, it is possible to introduce constraints under which such

reverse Pinsker inequalities hold. In the special case of a finite alphabet $\mathcal{A}$, Csiszár and Talata [29, p. 1012] showed that

$$D(P\|Q) \leq \left(\frac{\log e}{Q_{\min}}\right) \cdot |P - Q|^2 \tag{23}$$

when $Q_{\min} \triangleq \min_{a \in \mathcal{A}} Q(a)$ is positive.[3]

- [25, Th. 3.1] if $f : (0, \infty) \to \mathbb{R}$ is a strictly convex function, then there exists a real-valued function ψ_f such that $\lim_{x\downarrow 0} \psi_f(x) = 0$ and[4]

$$|P - Q| \leq \psi_f\left(D_f(P\|Q)\right). \tag{24}$$

which implies

$$\lim_{n\to\infty} D_f(P_n\|Q_n) = 0 \Rightarrow \lim_{n\to\infty} |P_n - Q_n| = 0. \tag{25}$$

The numerical optimization of an f-divergence subject to simultaneous constraints on f_i-divergences ($i = 1, \ldots, L$) was recently studied in [48], which showed that for that purpose it is enough to restrict attention to alphabets of cardinality $L + 2$. Earlier, [50] showed that if $L = 1$, then either the solution is obtained by a pair (P, Q) on a binary alphabet, or it is a deflated version of such a point. Therefore, from a purely numerical standpoint, the minimization of $D_f(P\|Q)$ such that $D_g(P\|Q) \geq d$ can be accomplished by a grid search on $[0, 1]^2$. Occasionally, as in the case where $D_f(P\|Q) = D(P\|Q)$ and $D_g(P\|Q) = |P - Q|$, it is actually possible to determine analytically the locus of $(D_f(P\|Q), D_g(P\|Q))$ (see [39]). In fact, as shown in [104, eq. (22)], a binary alphabet suffices if the single constraint is on the total variation distance. The same conclusion holds when minimizing the Rényi divergence [91].

In this work, we find relationships among the various divergence measures outlined above as well as a number of other measures of dissimilarity between probability measures. The framework of f-divergences, which encompasses the foregoing measures (Rényi divergence is a one-to-one transformation of the Hellinger divergence) serves as a convenient playground.

The rest of the paper is structured as follows:

Section II introduces the basic definitions needed and in particular the various measures of dissimilarity between probability measures used throughout.

Based on *functional domination*, Section III provides a basic tool for the derivation of bounds among f-divergences. Under mild regularity conditions, this approach further enables to prove the optimality of constants in those bounds. In addition, we show instances where such optimality can be shown in the absence of regularity conditions. The basic tool used in Section III is exemplified in obtaining relationships among important f-divergences such as relative entropy, Hellinger divergence and total variation distance. This approach is also useful in strengthening and providing an alternative proof of Samson's inequality [88] (a counterpart to Pinsker's inequality

---

[3]Recent applications of (23) can be found in [57, Appendix D] and [100, Lemma 7] for the analysis of the third-order asymptotics of the discrete memoryless channel with or without cost constraints.

[4]Eq. (24) follows as a special case of [25, Th. 3.1] with $m = 1$ and $w_m = 1$.

using Marton's divergence, useful in proving certain concentration of measure results [13]), whose constant we show cannot be improved. In addition, we show several new results in Section III-D on the maximal ratios of various f-divergences to total variation distance.

Section IV provides an approach for bounding ratios of f-divergences, assuming that the relative information (see Definition 1) is lower and/or upper bounded with probability one. The approach is exemplified in bounding ratios of relative entropy to various f-divergences, and analyzing the local behavior of f-divergence ratios when the reference measure is fixed. We also show that bounded relative information leads to a strengthened version of Jensen's inequality, which, in turn, results in upper and lower bounds on the ratio of the non-negative difference $\log\left(1 + \chi^2(P\|Q)\right) - D(P\|Q)$ to $D(Q\|P)$. A new reverse version of Samson's inequality is another byproduct of the main tool in this section.

The rich structure of the total variation distance as well as its importance in both fundamentals and applications merits placing special attention on bounding the rest of the distance measures in terms of $|P - Q|$. Section V gives several useful identities linking the total variation distance with the relative information spectrum, which result in a number of upper and lower bounds on $|P - Q|$, some of which are tighter than Pinsker's inequality in various ranges of the parameters. It also provides refined bounds on $D(P\|Q)$ as a function of χ^2-divergences and the total variation distance.

Section VI is devoted to proving "reverse Pinsker inequalities," namely, lower bounds on $|P - Q|$ as a function of $D(P\|Q)$ involving either (a) bounds on the relative information, (b) Lipschitz constants, or (c) the minimum mass of the reference measure (in the finite alphabet case). In the latter case, we also examine the relationship between entropy and the total variation distance from the equiprobable distribution, as well as the exponential decay of the probability that an independent identically distributed sequence is not strongly typical.

Section VII focuses on the E_γ divergence. This f-divergence generalizes the total variation distance, and its utility in information theory has been exemplified in [17], [67]–[69], and [77]–[79]. Based on the operational interpretation of the DeGroot statistical information [31] and the integral representation of f-divergences as a function of DeGroot's measure, Section VII provides an integral representation of f-divergences as a function of the E_γ divergence; this representation shows that $\{(E_\gamma(P\|Q), E_\gamma(Q\|P)), \gamma \geq 1\}$ uniquely determines $D(P\|Q)$ and $\mathscr{H}_\alpha(P\|Q)$, as well as any other f-divergence with twice differentiable f. Accordingly, bounds on the E_γ divergence directly translate into bounds on other important f-divergences. In addition, we show an extension of Pinsker's inequality (1) to E_γ divergence, which leads to a relationship between the relative information spectrum and relative entropy.

The Rényi divergence, which has found a plethora of information-theoretic applications, is the focus of Section VIII. Expressions of the Rényi divergence are derived in Section VIII-A as a function of the relative information spectrum. These expressions lead in Section VIII-B to the derivation of bounds on the Rényi divergence as a function of the variational distance under the boundedness assumption of the relative information. Bounds on the Rényi divergence of an arbitrary order are derived in Section VIII-C as a function of the relative entropy when the relative information is bounded.

II. BASIC DEFINITIONS

A. Relative Information and Relative Entropy

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: If $P \ll Q$, the *relative information* provided by $a \in \mathcal{A}$ according to (P, Q) is given by[5]

$$\iota_{P\|Q}(a) \triangleq \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a). \tag{26}$$

When the argument of the relative information is distributed according to P, the resulting real-valued random variable is of particular interest. Its cumulative distribution function and expected value are known as follows.

Definition 2: If $P \ll Q$, the *relative information spectrum* is the cumulative distribution function

$$\mathbb{F}_{P\|Q}(x) = \mathbb{P}\left[\iota_{P\|Q}(X) \leq x\right], \tag{27}$$

with[6] $X \sim P$. The *relative entropy* of P with respect to Q is

$$D(P\|Q) = \mathbb{E}\left[\iota_{P\|Q}(X)\right] \tag{28}$$
$$= \mathbb{E}\left[\iota_{P\|Q}(Y) \exp\left(\iota_{P\|Q}(Y)\right)\right], \tag{29}$$

where $Y \sim Q$.

B. f-Divergences

Introduced by Ali and Silvey [2] and Csiszár [21], [23], a useful generalization of the relative entropy, which retains some of its major properties (and, in particular, the data processing inequality [112]), is the class of f-divergences. A general definition of an f-divergence is given in [66, p. 4398], specialized next to the case where $P \ll Q$.

Definition 3: Let $f : (0, \infty) \to \mathbb{R}$ be a convex function, and suppose that $P \ll Q$. The *f-divergence* from P to Q is given by

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q \tag{30}$$
$$= \mathbb{E}[f(Z)] \tag{31}$$

where

$$Z = \exp\left(\iota_{P\|Q}(Y)\right), \quad Y \sim Q. \tag{32}$$

and in (30), we took the continuous extension[7]

$$f(0) = \lim_{t \downarrow 0} f(t) \in (-\infty, +\infty]. \tag{33}$$

---

[5] $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative (or density) of P with respect to Q. Logarithms have an arbitrary common base, and $\exp(\cdot)$ indicates the inverse function of the logarithm with that base.

[6] $X \sim P$ means that $\mathbb{P}[X \in \mathcal{F}] = P(\mathcal{F})$ for any event $\mathcal{F} \in \mathscr{F}$.

[7] The convexity of $f : (0, \infty) \to \mathbb{R}$ implies its continuity on $(0, \infty)$.

We can also define $D_f(P\|Q)$ without requiring $P \ll Q$. Let $f \colon (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let $f^\star \colon (0, \infty) \to \mathbb{R}$ be given by

$$f^\star(t) = tf\left(\tfrac{1}{t}\right) \tag{34}$$

for all $t > 0$. Note that $f^\star$ is also convex (see, e.g., [14, Section 3.2.6]), $f^\star(1) = 0$, and $D_f(P\|Q) = D_{f^\star}(Q\|P)$ if $P \ll\gg Q$. By definition, we take

$$f^\star(0) = \lim_{t \downarrow 0} f^\star(t) = \lim_{u \to \infty} \frac{f(u)}{u}. \tag{35}$$

If p and q denote, respectively, the densities of P and Q with respect to a σ-finite measure μ (i.e., $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$), then we can write (30) as

$$D_f(P\|Q) = \int qf\left(\frac{p}{q}\right)\mathrm{d}\mu \tag{36}$$

$$= \int_{\{pq>0\}} qf\left(\frac{p}{q}\right)\mathrm{d}\mu$$
$$+ f(0)Q(p=0) + f^\star(0)P(q=0). \tag{37}$$

Remark 1: Different functions may lead to the same f-divergence for all (P, Q): if for an arbitrary $b \in \mathbb{R}$, we have

$$f_b(t) = f_0(t) + b(t-1), \quad t \geq 0 \tag{38}$$

then

$$D_{f_0}(P\|Q) = D_{f_b}(P\|Q). \tag{39}$$

The following key property of f-divergences follows from Jensen's inequality.

Proposition 1: If $f \colon (0, \infty) \to \mathbb{R}$ is convex and $f(1) = 0$, $P \ll Q$, then

$$D_f(P\|Q) \geq 0. \tag{40}$$

If, furthermore, f is strictly convex at $t = 1$, then equality in (40) holds if and only if $P = Q$.

Surveys on general properties of f-divergences can be found in [64], [104], and [105].

The assumptions of Proposition 1 are satisfied by many interesting measures of dissimilarity between probability measures. In particular, the following examples receive particular attention in this paper. As per Definition 3, in each case the function f is defined on $(0, \infty)$.

1) *Relative Entropy [59]:* $f(t) = t \log t$,

$$D(P\|Q) = D_f(P\|Q) \tag{41}$$
$$= D_r(P\|Q) \tag{42}$$

with $r \colon (0, \infty) \to [0, \infty)$ defined as

$$r(t) \triangleq t \log t + (1-t) \log e. \tag{43}$$

2) *Relative Entropy:* $(P \ll\gg Q)$ $f(t) = -\log t$,

$$D(Q\|P) = D_f(P\|Q); \tag{44}$$

3) *Jeffrey's Divergence [54]:* $(P \ll\gg Q)$ $f(t) = (t-1)\log t$,

$$D(P\|Q) + D(Q\|P) = D_f(P\|Q); \tag{45}$$

4) *χ^2-Divergence [76]:* $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$,

$$\chi^2(P\|Q) = D_f(P\|Q) \tag{46}$$

$$\chi^2(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q \tag{47}$$

$$= \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1 \tag{48}$$

$$= \mathbb{E}\left[\exp\left(2\imath_{P\|Q}(Y)\right)\right] - 1 \tag{49}$$

$$= \mathbb{E}\left[\exp\left(\imath_{P\|Q}(X)\right)\right] - 1 \tag{50}$$

with $X \sim P$ and $Y \sim Q$. Note that if $P \ll\gg Q$, then from the right side of (48), we obtain

$$\chi^2(Q\|P) = D_g(P\|Q) \tag{51}$$

with $g(t) = \frac{1}{t} - t = f^\star(t)$, and $f(t) = t^2 - 1$.

5) *Hellinger divergence of order $\alpha \in (0, 1) \cup (1, \infty)$ [54], [65, Definition 2.10]:*

$$\mathscr{H}_\alpha(P\|Q) = D_{f_\alpha}(P\|Q) \tag{52}$$

with

$$f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}. \tag{53}$$

The χ^2-divergence is the Hellinger divergence of order 2, while $\frac{1}{2}\mathscr{H}_{\frac{1}{2}}(P\|Q)$ is usually referred to as the *squared Hellinger distance*. The analytic extension of $\mathscr{H}_\alpha(P\|Q)$ at $\alpha = 1$ yields

$$\mathscr{H}_1(P\|Q) \log e = D(P\|Q). \tag{54}$$

6) *Total Variation Distance:* Setting

$$f(t) = |t-1| \tag{55}$$

results in

$$|P - Q| = D_f(P\|Q) \tag{56}$$

$$= \int \left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right| \mathrm{d}Q \tag{57}$$

$$= 2 \sup_{\mathcal{F} \in \mathscr{F}} \left(P(\mathcal{F}) - Q(\mathcal{F})\right). \tag{58}$$

7) *Triangular Discrimination [62], [108] (a.k.a. Vincze-Le Cam Distance):*

$$\Delta(P\|Q) = D_f(P\|Q) \tag{59}$$

with

$$f(t) = \frac{(t-1)^2}{t+1}. \tag{60}$$

Note that

$$\tfrac{1}{2}\Delta(P\|Q) = \chi^2(P\|\tfrac{1}{2}P + \tfrac{1}{2}Q) \tag{61}$$

$$= \chi^2(Q\|\tfrac{1}{2}P + \tfrac{1}{2}Q). \tag{62}$$

8) *Jensen-Shannon Divergence [67] (a.k.a. Capacitory Discrimination):*

$$\mathrm{JS}(P\|Q) = D\left(P\|\tfrac{1}{2}P + \tfrac{1}{2}Q\right) + D\left(Q\|\tfrac{1}{2}P + \tfrac{1}{2}Q\right) \quad (63)$$
$$= D_f(P\|Q) \quad (64)$$

with

$$f(t) = t\log t - (1+t)\log\left(\frac{1+t}{2}\right). \quad (65)$$

9) E_γ *Divergence (See, e.g., [78, p. 2314]):* For $\gamma \geq 1$,

$$E_\gamma(P\|Q) = D_{f_\gamma}(P\|Q) \quad (66)$$

with

$$f_\gamma(t) = (t-\gamma)^+ \quad (67)$$

where $(x)^+ \triangleq \max\{x, 0\}$. E_γ is sometimes called "hockey-stick divergence" because of the shape of f_γ. If $\gamma = 1$, then

$$E_1(P\|Q) = \tfrac{1}{2}|P - Q|. \quad (68)$$

10) *DeGroot Statistical Information [31]:* For $p \in (0, 1)$,

$$\mathcal{I}_p(P\|Q) = D_{\phi_p}(P\|Q) \quad (69)$$

with

$$\phi_p(t) = \min\{p, 1-p\} - \min\{p, 1-pt\}. \quad (70)$$

Invoking (66)–(70), we get (cf. [66, eq. (77)])

$$\mathcal{I}_{\frac{1}{2}}(P\|Q) = \tfrac{1}{2}E_1(P\|Q) = \tfrac{1}{4}|P - Q|. \quad (71)$$

This measure was first proposed by DeGroot [31] due to its operational meaning in Bayesian statistical hypothesis testing (see Section VII-B), and it was later identified as an f-divergence (see [66, Th. 10]).

11) *Marton's Divergence [72, pp. 558–559]:*

$$d_2^2(P, Q) = \min \mathbb{E}\left[\mathbb{P}^2[X \neq Y|Y]\right] \quad (72)$$
$$= D_s(P\|Q) \quad (73)$$

where the minimum is over all probability measures P_{XY} with respective marginals $P_X = P$ and $P_Y = Q$, and

$$s(t) = (t-1)^2 1\{t < 1\}. \quad (74)$$

Note that Marton's divergence satisfies the triangle inequality [72, Lemma 3.1], and $d_2(P, Q) = 0$ implies $P = Q$; however, due to its asymmetry, it is not a distance measure.

C. Rényi Divergence

Another generalization of relative entropy was introduced by Rényi [87] in the special case of finite alphabets. The general definition (assuming[8] $P \ll Q$) is the following.

Definition 4: Let $P \ll Q$. The *Rényi divergence of order* $\alpha \geq 0$ from P to Q is given as follows:

• If $\alpha \in (0, 1) \cup (1, \infty)$, then

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\left(\mathbb{E}\left[\exp\left(\alpha \iota_{P\|Q}(Y)\right)\right]\right) \quad (75)$$
$$= \frac{1}{\alpha - 1}\log\left(\mathbb{E}\left[\exp\left((\alpha - 1)\iota_{P\|Q}(X)\right)\right]\right) \quad (76)$$

with $X \sim P$ and $Y \sim Q$.
• If $\alpha = 0$, then[9]

$$D_0(P\|Q) = \max_{\mathcal{F} \in \mathscr{F}: P(\mathcal{F})=1}\log\left(\frac{1}{Q(\mathcal{F})}\right). \quad (77)$$

• If $\alpha = 1$, then

$$D_1(P\|Q) = D(P\|Q) \quad (78)$$

which is the analytic extension of $D_\alpha(P\|Q)$ at $\alpha = 1$. If $D(P\|Q) < \infty$, it can be verified by L'Hôpital's rule that $D(P\|Q) = \lim_{\alpha \uparrow 1} D_\alpha(P\|Q)$.
• If $\alpha = +\infty$ then

$$D_\infty(P\|Q) = \log\left(\mathrm{ess\,sup}\,\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right) \quad (79)$$

with $Y \sim Q$. If $P \ll Q$, we take $D_\infty(P\|Q) = \infty$.
Rényi divergence is a one-to-one transformation of Hellinger divergence of the same order $\alpha \in (0, 1) \cup (1, \infty)$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\left(1 + (\alpha - 1)\mathscr{H}_\alpha(P\|Q)\right) \quad (80)$$

which, when particularized to order 2 becomes

$$D_2(P\|Q) = \log\left(1 + \chi^2(P\|Q)\right). \quad (81)$$

Note that (6), (17), (18) follow from (80) and the monotonicity of the Rényi divergence in its order, which in turn yields (16).

Introduced in [8], the Bhattacharyya distance was popularized in the engineering literature in [55].

Definition 5: The *Bhattacharyya distance* between P and Q, denoted by $B(P\|Q)$, is given by

$$B(P\|Q) = \tfrac{1}{2}D_{\frac{1}{2}}(P\|Q) \quad (82)$$
$$= \log\left(\frac{1}{1 - \tfrac{1}{2}\mathscr{H}_{\frac{1}{2}}(P\|Q)}\right). \quad (83)$$

Note that, if $P \lll Q$, then $B(P\|Q) = B(Q\|P)$ and $B(P\|Q) = 0$ if and only if $P = Q$, though $B(P\|Q)$ does not satisfy the triangle inequality.

III. FUNCTIONAL DOMINATION

Let f and g be convex functions on $(0, \infty)$ with $f(1) = g(1) = 0$, and let P and Q be probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$. If there exits $\alpha > 0$ such that $f(t) \leq \alpha g(t)$ for all $t \in (0, \infty)$, then it follows from Definition 3 that

$$D_f(P\|Q) \leq \alpha D_g(P\|Q). \quad (84)$$

This simple observation leads to a proof of, for example, (16) and the left inequality in (2) with the aid of Remark 1.

A. Basic Tool

Theorem 1: Let $P \ll Q$, and assume
- f is convex on $(0, \infty)$ with $f(1) = 0$;
- g is convex on $(0, \infty)$ with $g(1) = 0$;
- $g(t) > 0$ for all $t \in (0, 1) \cup (1, \infty)$.

Denote the function $\kappa : (0, 1) \cup (1, \infty) \to \mathbb{R}$

$$\kappa(t) = \frac{f(t)}{g(t)}, \quad t \in (0, 1) \cup (1, \infty) \qquad (85)$$

and

$$\bar{\kappa} = \sup_{t \in (0,1) \cup (1,\infty)} \kappa(t). \qquad (86)$$

Then,

a)

$$D_f(P \| Q) \leq \bar{\kappa} D_g(P \| Q). \qquad (87)$$

b) If, in addition, $f'(1) = g'(1) = 0$, then

$$\sup_{P \neq Q} \frac{D_f(P \| Q)}{D_g(P \| Q)} = \bar{\kappa}. \qquad (88)$$

Proof:

a) The bound in (87) follows from (84) and $f(t) \leq \bar{\kappa} g(t)$ for all $t > 0$.

b) Since g is positive except at $t = 1$, $D_g(P \| Q) > 0$ if $P \neq Q$. The convexity of f, g on $(0, \infty)$ implies their continuity; and since $g(t) > 0$ for all $t \in (0, 1) \cup (1, \infty)$, $\kappa(\cdot)$ is continuous on both $(0, 1)$ and $(1, \infty)$.

To show (88), we fix an arbitrary $\nu \in (0, 1) \cup (1, \infty)$ and construct a sequence of pairs of probability measures whose ratio of f-divergence to g-divergence converges to $\kappa(\nu)$. To that end, for sufficiently small $\varepsilon > 0$, let P_ε and Q_ε be parametric probability measures defined on the set $\mathcal{A} = \{0, 1\}$ with $P_\varepsilon(0) = \nu\varepsilon$ and $Q_\varepsilon(0) = \varepsilon$. Then,

$$\lim_{\varepsilon \to 0} \frac{D_f(P_\varepsilon \| Q_\varepsilon)}{D_g(P_\varepsilon \| Q_\varepsilon)}$$

$$= \lim_{\varepsilon \to 0} \frac{\varepsilon f(\nu) + (1 - \varepsilon) f\left(\frac{1 - \nu\varepsilon}{1 - \varepsilon}\right)}{\varepsilon g(\nu) + (1 - \varepsilon) g\left(\frac{1 - \nu\varepsilon}{1 - \varepsilon}\right)} \qquad (89)$$

$$= \lim_{\alpha \to 0} \frac{f(\nu) + \frac{\nu - 1}{\alpha} f(1 - \alpha)}{g(\nu) + \frac{\nu - 1}{\alpha} g(1 - \alpha)} \qquad (90)$$

$$= \kappa(\nu) \qquad (91)$$

where (90) holds by change of variable $\varepsilon = \alpha / (\nu - 1 + \alpha)$, and (91) holds by the assumption on the derivatives of f and g at 1, the assumption that $f(1) = g(1) = 0$, and the continuity of $\kappa(\cdot)$ at ν. If $\bar{\kappa} = \kappa(\nu)$ we are done. If the supremum in (86) is not attained on $(0, 1) \cup (1, \infty)$, then the right side of (91) can be made arbitrarily close to $\bar{\kappa}$ by an appropriate choice of ν. ∎

Remark 2: Beyond the restrictions in Theorem 1a), the only operative restriction imposed by Theorem 1b) is the differentiability of the functions f and g at $t = 1$. Indeed, we can invoke Remark 1 and add $f'(1)(1 - t)$ to $f(t)$, without changing D_f (and likewise with g) and thereby satisfying the condition in Theorem 1b); the stationary point at 1 must be a minimum of both f and g because of the assumed convexity, which implies their non-negativity on $(0, \infty)$.

Remark 3: It is useful to generalize Theorem 1b) by dropping the assumption on the existence of the derivatives at 1. To that end, note that the inverse transformation used for the transition to (90) is given by $\nu = 1 + \alpha \left(\frac{1}{\varepsilon} - 1\right)$ where $\varepsilon > 0$ is sufficiently small, so if $\nu > 1$ (resp. $\nu < 1$), then $\alpha > 0$ (resp. $\alpha < 0$). Consequently, it is easy to see from (90) that if $\bar{\kappa} = \sup_{t > 1} \kappa(t)$, the construction in the proof can restrict to $\nu > 1$, in which case it is enough to require that the left derivatives of f and g at 1 be equal to 0. Analogously, if $\bar{\kappa} = \sup_{0 < t < 1} \kappa(t)$, it is enough to require that the right derivatives of f and g at 1 be equal to 0. When neither left nor right derivatives at 1 are 0, then (88) need not hold as the following example shows.

Example 1: Let $f(t) = |t - 1|$ and

$$g(t) = 2f(t) + 1 - t. \qquad (92)$$

Then, $\bar{\kappa} = 1$, while in view of (39) and (56) for all (P, Q),

$$D_g(P \| Q) = 2D_f(P \| Q) = 2|P - Q|. \qquad (93)$$

B. Relationships Among $D(P \| Q)$, $\chi^2(P \| Q)$ and $|P - Q|$

Since the Rényi divergence of order $\alpha > 0$ is monotonically increasing in α, (81) yields

$$D(P \| Q) \leq \log \left(1 + \chi^2(P \| Q)\right) \qquad (94)$$

$$\leq \chi^2(P \| Q) \log e. \qquad (95)$$

Inequality (94), which can be found in [99] and [41, Th. 5], is sharpened in Theorem 11 under the assumption of bounded relative information. In view of (80), an alternative way to sharpen (94) is

$$D(P \| Q) \leq \frac{1}{\alpha - 1} \log \left(1 + (\alpha - 1)\mathcal{H}_\alpha(P \| Q)\right) \qquad (96)$$

for $\alpha \in (1, 2)$, which is tight as $\alpha \to 1$.

Relationships between the relative entropy, total variation distance and χ^2 divergence are derived next.

Theorem 2: a) If $P \ll Q$ and $c_1, c_2 \geq 0$, then

$$D(P \| Q) \leq \left(c_1 |P - Q| + c_2 \chi^2(P \| Q)\right) \log e \qquad (97)$$

holds if $(c_1, c_2) = (0, 1)$ and $(c_1, c_2) = \left(\frac{1}{4}, \frac{1}{2}\right)$. Furthermore, if $c_1 = 0$ then $c_2 = 1$ is optimal, and if $c_2 = \frac{1}{2}$ then $c_1 = \frac{1}{4}$ is optimal.

b)

$$\sup \frac{D(P \| Q) + D(Q \| P)}{\chi^2(P \| Q) + \chi^2(Q \| P)} = \frac{1}{2} \log e \qquad (98)$$

where the supremum is over $P \lll Q$ and $P \neq Q$.

Proof:

a) The satisfiability of (101) with $(c_1, c_2) = (0, 1)$ is equivalent to (95).

Let $P \ll Q$, and $Y \sim Q$. Then,

$$D(P \| Q) = \mathbb{E}\left[r\left(\frac{dP}{dQ}(Y)\right)\right] \qquad (99)$$

$$\leq \tfrac{1}{2}\mathbb{E}\left[\left(1 - \frac{dP}{dQ}(Y)\right)^+ + \left(\frac{dP}{dQ}(Y) - 1\right)^2\right]\log e \tag{100}$$

$$= \left(\tfrac{1}{4}|P - Q| + \tfrac{1}{2}\chi^2(P\|Q)\right)\log e \tag{101}$$

where (99) follows from the definition of relative entropy with the function $r\colon (0, \infty) \to \mathbb{R}$ defined in (43); (100) holds since for $t \in (0, \infty)$

$$r(t) \leq \tfrac{1}{2}\left[(1 - t)^+ + (t - 1)^2\right]\log e \tag{102}$$

and (101) follows from (47), (57), and the identity

$$(1 - t)^+ = \tfrac{1}{2}\big[|1 - t| + (1 - t)\big].$$

This proves (97) with $(c_1, c_2) = \left(\tfrac{1}{4}, \tfrac{1}{2}\right)$.

Next, we show that if $c_1 = 0$ then $c_2 = 1$ is the best possible constant in (97). To that end, let $f_2(t) = (t - 1)^2$, and let $\kappa(t)$ be the continuous extension of $\frac{r(t)}{f_2(t)}$. It can be verified that the function κ is monotonically decreasing on $(0, \infty)$, so

$$\bar{\kappa} = \lim_{t\downarrow 0}\kappa(t) = \log e. \tag{103}$$

Since $D_r(P\|Q) = D(P\|Q)$ and $D_{f_2}(P\|Q) = \chi^2(P\|Q)$, and $r'(1) = f_2'(1) = 0$, the desired result follows from Theorem 1b).

To show that $c_1 = \tfrac{1}{4}$ is the best possible constant in (97) if $c_2 = \tfrac{1}{2}$, we let $g_2(t) = \tfrac{1}{2}\left[(1 - t)^+ + (t - 1)^2\right]$. Theorem 1b) does not apply here since g_2 is not differentiable at $t = 1$. However, we can still construct probability measures for proving the optimality of the point $(c_1, c_2) = \left(\tfrac{1}{4}, \tfrac{1}{2}\right)$. To that end, let $\varepsilon \in (0, 1)$, and define probability measures P_ε and Q_ε on the set $\mathcal{A} = \{0, 1\}$ with $P_\varepsilon(1) = \varepsilon^2$ and $Q_\varepsilon(1) = \varepsilon$. Since $D_r(P\|Q) = D(P\|Q)$ and $D_{g_2}(P\|Q) = \tfrac{1}{4}|P - Q| + \tfrac{1}{2}\chi^2(P\|Q)$,

$$\lim_{\varepsilon\downarrow 0} \frac{D(P_\varepsilon\|Q_\varepsilon)}{\tfrac{1}{4}|P_\varepsilon - Q_\varepsilon| + \tfrac{1}{2}\chi^2(P_\varepsilon\|Q_\varepsilon)}$$
$$= \lim_{\varepsilon\downarrow 0} \frac{(1 - \varepsilon)r(1 + \varepsilon) + \varepsilon r(\varepsilon)}{(1 - \varepsilon)g_2(1 + \varepsilon) + \varepsilon g_2(\varepsilon)} \tag{104}$$
$$= \log e \tag{105}$$

where (105) holds since we can write numerator and denominator in the right side as

$$(1 - \varepsilon)r(1 + \varepsilon) + \varepsilon r(\varepsilon)$$
$$= (1 - \varepsilon^2)\log(1 + \varepsilon) + \varepsilon^2 \log\varepsilon \tag{106}$$
$$= \varepsilon \log e + o(\varepsilon), \tag{107}$$
$$(1 - \varepsilon)g_2(1 + \varepsilon) + \varepsilon g_2(\varepsilon)$$
$$= \varepsilon - \varepsilon^2. \tag{108}$$

b) We have

$$D_f(P\|Q) = D(P\|Q) + D(Q\|P), \tag{109}$$
$$D_g(P\|Q) = \chi^2(P\|Q) + \chi^2(Q\|P) \tag{110}$$

with

$$f(t) \triangleq (t - 1)\log t \tag{111}$$

$$g(t) \triangleq t^2 - t - 1 + \frac{1}{t} \tag{112}$$

$$\kappa(t) = \frac{f(t)}{g(t)} = \frac{t\log t}{t^2 - 1}, \quad t \in (0, 1) \cup (1, \infty) \tag{113}$$

$$\lim_{t\to 1}\kappa(t) = \bar{\kappa} = \tfrac{1}{2}\log e \tag{114}$$

where (114) is easy to verify since κ is monotonically increasing on $(0, 1)$, and monotonically decreasing on $(1, \infty)$. The desired result follows since the conditions of Theorem 1b) apply. ∎

Remark 4: Inequality (97) strengthens the bound claimed in [32, eq. (2.8)],

$$D(P\|Q) \leq \tfrac{1}{2}\left(|P - Q| + \chi^2(P\|Q)\right)\log e, \tag{115}$$

although the short outline of the suggested proof in [32, p. 710] leads to the weaker upper bound $|P - Q| + \tfrac{1}{2}\chi^2(P\|Q)$ nats.

Remark 5: Note that (95) implies a looser result where the constant in the right side of (98) is doubled. Furthermore, (1) and (98) result in the bound $\chi^2(P\|Q) + \chi^2(Q\|P) \geq 2|P - Q|^2$ which, although weaker than (15), has the same behavior for small values of $|P - Q|$.

C. An Alternative Proof of Samson's Inequality

An analog of Pinsker's inequality, which comes in handy for the proof of Marton's conditional transportation inequality [13, Lemma 8.4], is the following bound due to Samson [88, Lemma 2]:

Theorem 3: If $P \ll Q$, then

$$d_2^2(P, Q) + d_2^2(Q, P) \leq \frac{2}{\log e}D(P\|Q) \tag{116}$$

where $d_2(P, Q)$ is the distance measure defined in (72).

We provide next an alternative proof of Theorem 3, in view of Theorem 1b), with the following advantages:

a) This proof yields the optimality of the constant in (116), i.e., we prove that

$$\sup \frac{d_2^2(P, Q) + d_2^2(Q, P)}{D(P\|Q)} = \frac{2}{\log e} \tag{117}$$

where the supremum is over all probability measures P, Q such that $P \neq Q$ and $P \ll\gg Q$.

b) A simple adaptation of this proof enables to derive a reverse inequality to (116), which holds under the boundedness assumption of the relative information (see Section IV-D).

Proof:

$$d_2^2(P, Q) + d_2^2(Q, P) = D_s(P\|Q) + D_{s^\star}(P\|Q) \tag{118}$$
$$= D_f(P\|Q) \tag{119}$$

where, from (74), $s^\star\colon (0, \infty) \to [0, \infty)$ is the convex function

$$s^\star(t) = ts\left(\tfrac{1}{t}\right) = \frac{(t - 1)^2 1\{t > 1\}}{t} \tag{120}$$

and, from (74) and (120), the non-negative and convex function $f\colon (0, \infty) \to \mathbb{R}$ in (119) is given by

$$f(t) = s(t) + s^\star(t) = \frac{(t - 1)^2}{\max\{1, t\}} \tag{121}$$

for all $t > 0$. Let r be the non-negative and convex function $r: (0, \infty) \to \mathbb{R}$ defined in (43), which yields $D_r(P\|Q) = D(P\|Q)$. Note that $f(1) = r(1) = 0$, and the functions f and r are both differentiable at $t = 1$ with $f'(1) = r'(1) = 0$. The desired result follows from Theorem 1b) since in this case

$$\kappa(t) = \frac{(t-1)^2}{r(t) \max\{1, t\}}, \quad t \in (0, 1) \cup (1, \infty) \quad (122)$$

$$\lim_{t \to 1} \kappa(t) = \frac{2}{\log e} = \bar{\kappa}, \quad (123)$$

as can be verified from the monotonicity of κ on $(0, 1)$ (increasing) and $(1, \infty)$ (decreasing). ∎

Remark 6: As mentioned in [88, p. 438], Samson's inequality (116) strengthens the Pinsker-type inequality in [72, Lemma 3.2]:

$$d_2^2(P, Q) \le \frac{2}{\log e} \min\{D(P\|Q), D(Q\|P)\} \quad (124)$$

Nevertheless, similarly to our alternative proof of Theorem 3, one can verify that Theorem 1b) yields the optimality of the constant in (124).

D. Ratio of f-Divergence to Total Variation Distance

Vajda [104, Th. 2] showed that the range of an f-divergence is given by (see (35))

$$0 \le D_f(P\|Q) \le f(0) + f^\star(0) \quad (125)$$

where every value in this range is attainable by a suitable pair of probability measures $P \ll Q$. Recalling Remark 1, note that $f_b(0) + f_b^\star(0) = f(0) + f^\star(0)$ with $f_b(\cdot)$ defined in (38). Basu *et al.* [9, Lemma 11.1] strengthened (125), showing that

$$D_f(P\|Q) \le \frac{1}{2}\big(f(0) + f^\star(0)\big) |P - Q|. \quad (126)$$

Note that, provided $f(0)$ and $f^\star(0)$ are finite, (126) yields a counterpart to (24). Next, we show that the constant in (126) cannot be improved.

Theorem 4: If $f: (0, \infty) \to \mathbb{R}$ is convex with $f(1) = 0$, then

$$\sup \frac{D_f(P\|Q)}{|P - Q|} = \frac{1}{2}\big(f(0) + f^\star(0)\big) \quad (127)$$

where the supremum is over all probability measures P, Q such that $P \ll Q$ and $P \ne Q$.

Proof: As the first step, we give a simplified proof of (126) (cf. [9, pp. 344–345]). In view of Remark 1, it is sufficient to show that for all $t > 0$,

$$f(t) + \frac{1}{2}\big(f(0) - f^\star(0)\big)(t - 1) \le \frac{1}{2}\big(f(0) + f^\star(0)\big)|t - 1|. \quad (128)$$

If $t \in (0, 1)$, (128) reduces to $f(t) \le (1 - t) f(0)$, which holds in view of the convexity of f and $f(1) = 0$. If $t \ge 1$, we can readily check, with the aid of (34), that (128) reduces to $f^\star(\frac{1}{t}) \le (1 - \frac{1}{t}) f^\star(0)$, which, in turn, holds because $f^\star$ is convex and $f^\star(1) = 0$.

For the second part of the proof of (127), we construct a pair of probability measures P_ε and Q_ε such that, for a sufficiently small $\varepsilon > 0$, $\frac{D_f(P_\varepsilon\|Q_\varepsilon)}{|P_\varepsilon - Q_\varepsilon|}$ can be made arbitrarily close to the right side of (127). To that end, let $\varepsilon \in (0, \frac{1}{2}(\sqrt{5} - 1)]$, and let

P_ε and Q_ε be defined on the set $\mathcal{A} = \{0, 1, 2\}$ with $P_\varepsilon(0) = Q_\varepsilon(1) = \varepsilon$ and $P_\varepsilon(1) = Q_\varepsilon(0) = \varepsilon^2$. Then,

$$\lim_{\varepsilon \to 0} \frac{D_f(P_\varepsilon\|Q_\varepsilon)}{|P_\varepsilon - Q_\varepsilon|} = \lim_{\varepsilon \to 0} \frac{\varepsilon f(\varepsilon) + \varepsilon^2 f(\frac{1}{\varepsilon})}{2\varepsilon(1 - \varepsilon)} \quad (129)$$

$$= \frac{1}{2} f(0) + \frac{1}{2} f^\star(0) \quad (130)$$

where (129) holds since $P_\varepsilon(2) = Q_\varepsilon(2)$, and (130) follows from (33) and (35). ∎

Remark 7: Csiszár [24, Th. 2] showed that if $f(0)$ and $f^\star(0)$ are finite and $P \ll Q$, then there exists a constant $C_f > 0$ which depends only on f such that

$$D_f(P\|Q) \le C_f \sqrt{|P - Q|}. \quad (131)$$

If $|P - Q| < 1$, then (131) is superseded by (126) where the constant is not only explicit but is the best possible according to Theorem 4.

A direct application of Theorem 4 yields
Corollary 1:

$$\sup_{P \ne Q} \frac{\mathcal{H}_\alpha(P\|Q)}{|P - Q|} = \frac{1}{2(1 - \alpha)}, \quad \forall \alpha \in (0, 1) \quad (132)$$

$$\sup_{P \ne Q} \frac{\Delta(P\|Q)}{|P - Q|} = 1, \quad (133)$$

$$\sup_{P \ne Q} \frac{\mathrm{JS}(P\|Q)}{|P - Q|} = \log 2, \quad (134)$$

$$\sup_{P \ne Q} \frac{d_2^2(P, Q)}{|P - Q|} = \frac{1}{2}, \quad (135)$$

$$\sup_{P \ne Q} \frac{d_2^2(P, Q) + d_2^2(Q, P)}{|P - Q|} = 1 \quad (136)$$

where the suprema in (132)–(135) are over all $P \ll Q$ with $P \ne Q$, and the supremum in (136) is over all $P \ll\gg Q$ with $P \ne Q$.

Remark 8: The results in (132), (133), and (134) strengthen, respectively, the inequalities in [65, Proposition 2.35], [101, eq. (11)], and [101, Th. 2]. The results in (135) and (136) form counterparts of (117).

IV. BOUNDED RELATIVE INFORMATION

In this section we show that it is possible to find bounds among f-divergences without requiring a strong condition of functional domination (see Section III) as long as the relative information is upper and/or lower bounded almost surely.

A. Definition of β_1 and β_2

The following notation is used throughout the rest of the paper. Given a pair of probability measures (P, Q) on the same measurable space, denote $\beta_1, \beta_2 \in [0, 1]$ by

$$\beta_1 = \exp\big(-D_\infty(P\|Q)\big), \quad (137)$$
$$\beta_2 = \exp\big(-D_\infty(Q\|P)\big) \quad (138)$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$. Note that if $\beta_1 > 0$,

then $P \ll Q$, while $\beta_2 > 0$ implies $Q \ll P$. Furthermore, if $P \ll\gg Q$, then with $Y \sim Q$,

$$\beta_1 = \operatorname{ess\,inf} \frac{dQ}{dP}(Y) = \left(\operatorname{ess\,sup} \frac{dP}{dQ}(Y)\right)^{-1}, \quad (139)$$

$$\beta_2 = \operatorname{ess\,inf} \frac{dP}{dQ}(Y) = \left(\operatorname{ess\,sup} \frac{dQ}{dP}(Y)\right)^{-1}. \quad (140)$$

The following examples illustrate important cases in which β_1 and β_2 are positive.

Example 2 (Gaussian Distributions): Let P and Q be Gaussian probability measures with equal means, and variances σ_0^2 and σ_1^2 respectively. Then,

$$\beta_1 = \frac{\sigma_0}{\sigma_1} 1\{\sigma_0 \leq \sigma_1\}, \quad (141)$$

$$\beta_2 = \frac{\sigma_1}{\sigma_0} 1\{\sigma_1 \leq \sigma_0\}. \quad (142)$$

Example 3 (Shifted Laplace Distributions): Let P and Q be the probability measures whose probability density functions are, respectively, given by $f_\lambda(\cdot - a_0)$ and $f_\lambda(\cdot - a_1)$ with

$$f_\lambda(x) = \tfrac{\lambda}{2} \exp(-\lambda|x|), \quad x \in \mathbb{R} \quad (143)$$

where $\lambda > 0$. In this case, (143) gives

$$\frac{dP}{dQ}(x) = \exp\big(\lambda(|x - a_1| - |x - a_0|)\big), \quad x \in \mathbb{R} \quad (144)$$

which yields

$$\beta_1 = \beta_2 = \exp\big(-\lambda|a_1 - a_0|\big) \in (0, 1]. \quad (145)$$

Example 4 (Cramér Distributions): Suppose that P and Q have Cramér probability density functions f_{θ_1, m_1} and f_{θ_0, m_0}, respectively, with

$$f_{\theta, m}(x) = \frac{\theta}{2(1 + \theta|x - m|)^2}, \quad x \in \mathbb{R} \quad (146)$$

where $\theta > 0$ and $m \in \mathbb{R}$. In this case, we have $\beta_1, \beta_2 \in (0, 1)$ since the ratio of the probability density functions, $\frac{f_{\theta_1, m_1}}{f_{\theta_0, m_0}}$, tends to $\frac{\theta_0}{\theta_1} < \infty$ in the limit $x \to \pm\infty$. In the special case where $m_0 = m_1 = m \in \mathbb{R}$, the ratio of these probability density functions is $\frac{\theta_1}{\theta_0}$ at $x = m$; due also to the symmetry of the probability density functions around m, it can be verified that in this special case

$$\beta_1 = \beta_2 = \min\left\{\frac{\theta_0}{\theta_1}, \frac{\theta_1}{\theta_0}\right\}. \quad (147)$$

Example 5 (Cauchy Distributions): Suppose that P and Q have Cauchy probability density functions g_{γ_1, m_1} and g_{γ_0, m_0}, respectively, $\gamma_0 \neq \gamma_1$ and

$$g_{\gamma, m}(x) = \frac{1}{\pi\gamma}\left[1 + \left(\frac{x - m}{\gamma}\right)^2\right]^{-1}, \quad x \in \mathbb{R} \quad (148)$$

where $\gamma > 0$. In this case, we also have $\beta_1, \beta_2 \in (0, 1)$ since the ratio of the probability density functions tends to $\frac{\gamma_0}{\gamma_1} < \infty$ in the limit $x \to \pm\infty$. In the special case where $m_0 = m_1$,

$$\beta_1 = \beta_2 = \min\left\{\frac{\gamma_1}{\gamma_0}, \frac{\gamma_0}{\gamma_1}\right\}. \quad (149)$$

B. Basic Tool

Since $\beta_1 = 1 \Leftrightarrow \beta_2 = 1 \Leftrightarrow P = Q$, it is advisable to avoid trivialities by excluding that case.

Theorem 5: Let f and g satisfy the assumptions in Theorem 1, and assume that $(\beta_1, \beta_2) \in [0, 1)^2$. Then,

$$D_f(P\|Q) \leq \kappa^\star D_g(P\|Q) \quad (150)$$

where

$$\kappa^\star = \sup_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \quad (151)$$

and $\kappa(\cdot)$ is defined in (85).

Proof: Defining $g(0)$ and $g^\star(0)$ as in (33)-(35), respectively, note that

$$f(0)Q(p = 0) \leq g(0)\kappa^\star Q(p = 0) \quad (152)$$

because if $\beta_2 = 0$ then $f(0) \leq \kappa^\star g(0)$, and if $\beta_2 > 0$ then $Q(p = 0) = 0$. Similarly,

$$f^\star(0)P(q = 0) \leq g^\star(0)\kappa^\star P(q = 0) \quad (153)$$

because if $\beta_1 = 0$ then $f^\star(0) \leq \kappa^\star g^\star(0)$ and if $\beta_1 > 0$ then $P(q = 0) = 0$. Moreover, since $f(1) = g(1) = 0$, we can substitute $pq > 0$ by $\{pq > 0\} \cap \{p \neq q\}$ in the right side of (37) for $D_f(P\|Q)$ and likewise for $D_g(P\|Q)$. In view of the definition of $\kappa^\star$, (152) and (153),

$$D_f(P\|Q) \leq \kappa^\star \int_{pq>0, p\neq q} qg\left(\frac{p}{q}\right) d\mu$$
$$+ \kappa^\star g(0)Q(p = 0) + \kappa^\star g^\star(0)P(q = 0) \quad (154)$$
$$= \kappa^\star D_g(P\|Q). \quad (155)$$

∎

Note that if $\beta_1 = \beta_2 = 0$, then Theorem 5 does not improve upon Theorem 1a).

Remark 9: In the application of Theorem 5, it is often convenient to make use of the freedom afforded by Remark 1 and choose the corresponding offsets such that:

- the positivity property of g required by Theorem 5 is satisfied;
- the lowest $\kappa^\star$ is obtained.

Remark 10: Similarly to the proof of Theorem 1b), under the conditions therein, one can verify that the constants in Theorem 5 are the best possible among all probability measures P, Q with given $(\beta_1, \beta_2) \in [0, 1)^2$.

Remark 11: Note that if we swap the assumptions on f and g in Theorem 5, the same result translates into

$$\inf_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \cdot D_g(P\|Q) \leq D_f(P\|Q). \quad (156)$$

Furthermore, provided both f and g are positive (except at $t = 1$) and κ is monotonically increasing, Theorem 5 and (156) result in

$$\kappa(\beta_2)D_g(P\|Q) \leq D_f(P\|Q) \quad (157)$$
$$\leq \kappa(\beta_1^{-1})D_g(P\|Q). \quad (158)$$

In this case, if $\beta_1 > 0$, sometimes it is convenient to replace $\beta_1 > 0$ with $\beta_1' \in (0, \beta_1)$ at the expense of loosening the bound. A similar observation applies to β_2.

Example 6: If $f(t) = (t-1)^2$ and $g(t) = |t-1|$, we get

$$\chi^2(P\|Q) \leq \max\{\beta_1^{-1} - 1, 1 - \beta_2\}|P - Q|. \quad (159)$$

C. Bounds on $\frac{D(P\|Q)}{D(Q\|P)}$

The remaining part of this section is devoted to various applications of Theorem 5. From this point, we make use of the definition of $r \colon (0, \infty) \to [0, \infty)$ in (43).

An illustrative application of Theorem 5 gives upper and lower bounds on the ratio of relative entropies.

Theorem 6: Let $P \ll\gg Q$, $P \neq Q$, and $(\beta_1, \beta_2) \in (0, 1)^2$. Let $\kappa \colon (0, 1) \cup (1, \infty) \to (0, \infty)$ be defined as

$$\kappa(t) = \frac{t \log t + (1-t) \log e}{(t-1)\log e - \log t}. \quad (160)$$

Then,

$$\kappa(\beta_2) \leq \frac{D(P\|Q)}{D(Q\|P)} \leq \kappa(\beta_1^{-1}). \quad (161)$$

Proof: For $t > 0$, let

$$g(t) = -\log t + (t-1)\log e \quad (162)$$

then $D_g(P\|Q) = D(Q\|P)$, $D_r(P\|Q) = D(P\|Q)$ and the conditions of Theorem 5 are satisfied. The desired result follows from Theorem 5 and the monotonicity of $\kappa(\cdot)$ shown in Appendix A. $\blacksquare$

D. Reverse Samson's Inequality

The next result gives a counterpart to Samson's inequality (116).

Theorem 7: Let $(\beta_1, \beta_2) \in (0, 1)^2$. Then,

$$\inf \frac{d_2^2(P, Q) + d_2^2(Q, P)}{D(P\|Q)} = \min\{\kappa(\beta_1^{-1}), \kappa(\beta_2)\} \quad (163)$$

where the infimum is over all $P \ll Q$ with given (β_1, β_2), and where $\kappa \colon (0, 1) \cup (1, \infty) \to \left(0, \frac{2}{\log e}\right)$ is given in (122).

Proof: Applying Remark 11 to the convex and positive (except at $t = 1$) function $f(t)$ given in (121), and $g(t) = r(t)$, the lower bound on $\frac{d_2^2(P,Q)+d_2^2(Q,P)}{D(P\|Q)}$ in the right side of (163) follows from the fact that (122) is monotonically increasing on $(0, 1)$, and monotonically decreasing on $(1, \infty)$. To verify that this is the best possible lower bound, we recall Remark 10 since in this case $f'(1) = g'(1) = 0$. $\blacksquare$

E. Bounds on $\frac{D(P\|Q)}{\mathscr{H}_\alpha(P\|Q)}$

The following result bounds the ratio of relative entropy to Hellinger divergence of an arbitrary positive order $\alpha \neq 1$. Theorem 8 extends and strengthens a result for $\alpha \in (0, 1)$ by Haussler and Opper [51, Lemma 4 and eq. (6)] (see also [16]), which in turn generalizes the special case for $\alpha = \frac{1}{2}$ obtained simultaneously and independently by Birgé and Massart [11, eq. (7.6)].

Theorem 8: Let $P \ll Q$, $P \neq Q$, $\alpha \in (0, 1) \cup (1, \infty)$ and $(\beta_1, \beta_2) \in [0, 1)^2$. Define the continuous function on $[0, \infty]$:

$$\kappa_\alpha(t) = \begin{cases} \log e & t = 0; \\ \frac{(1-\alpha)r(t)}{1-t^\alpha+\alpha t-\alpha} & t \in (0, 1) \cup (1, \infty); \\ \alpha^{-1}\log e & t = 1; \\ \infty & t = \infty \text{ and } \alpha \in (0, 1); \\ 0 & t = \infty \text{ and } \alpha \in (1, \infty). \end{cases} \quad (164)$$

Then, for $\alpha \in (0, 1)$,

$$\kappa_\alpha(\beta_2) \leq \frac{D(P\|Q)}{\mathscr{H}_\alpha(P\|Q)} \leq \kappa_\alpha(\beta_1^{-1}) \quad (165)$$

and, for $\alpha \in (1, \infty)$,

$$\kappa_\alpha(\beta_1^{-1}) \leq \frac{D(P\|Q)}{\mathscr{H}_\alpha(P\|Q)} \leq \kappa_\alpha(\beta_2). \quad (166)$$

Proof: $D_r(P\|Q) = D(P\|Q)$, and $D_{g_\alpha}(P\|Q) = \mathscr{H}_\alpha(P\|Q)$ with

$$g_\alpha(t) = \frac{1 - t^\alpha + \alpha t - \alpha}{1 - \alpha}, \quad t \in (0, \infty) \quad (167)$$

in view of Remark 1, (42) and (52). Since $g'_\alpha(t) = \frac{\alpha(1-t^{\alpha-1})}{1-\alpha}$ g_α is monotonically decreasing on $(0, 1]$ and monotonically increasing on $[1, \infty)$; hence, $g_\alpha(1) = 0$ implies that g_α is positive except at $t = 1$. The convexity conditions required by Theorem 5 are also easy to check for both $r(\cdot)$ and $g_\alpha(\cdot)$. The function in (164) is the continuous extension of $\frac{r}{g_\alpha}$, which as shown in Appendix B, is monotonically increasing on $[0, \infty]$ if $\alpha \in (0, 1)$, and it is monotonically decreasing on $[0, \infty]$ if $\alpha \in (1, \infty)$. Therefore, Theorem 5 results in (165) and (166) for $\alpha \in (0, 1)$ and $\alpha \in (1, \infty)$, respectively. $\blacksquare$

Remark 12: Theorem 8 is of particular interest for $\alpha = \frac{1}{2}$. In this case since, from (164), $\kappa_{\frac{1}{2}} \colon [0, \infty] \to [0, \infty]$ is monotonically increasing with $\kappa_{\frac{1}{2}}(0) = \log e$, the left inequality in (166) yields (16).

For large arguments, $\kappa_{\frac{1}{2}}(\cdot)$ grows logarithmically. For example, if $\beta_1 > 9.56 \cdot 10^{-9}$, it follows from (165) that

$$D(P\|Q) \leq \left(14 + \frac{2\log_e 2}{(1-e^{-1})^2}\right) \mathscr{H}_{\frac{1}{2}}(P\|Q) \text{ nats} \quad (168)$$

which is strictly smaller than the upper bound on the relative entropy in [111, Th. 5], given not in terms of β_1 but in terms of another more cumbersome quantity that controls the mass that $\frac{dP}{dQ}$ may have at large values.

As mentioned in Section II-B, $\chi^2(P\|Q)$ is equal to the Hellinger divergence of order 2. Specializing Theorem 8 to the case $\alpha = 2$ results in

$$\kappa_2(\beta_1^{-1}) \leq \frac{D(P\|Q)}{\chi^2(P\|Q)} \leq \kappa_2(\beta_2), \quad (169)$$

which improves the upper and lower bounds in [34, Proposition 2]:

$$\tfrac{1}{2}\beta_1 \log e \leq \frac{D(P\|Q)}{\chi^2(P\|Q)} \leq \tfrac{1}{2}\beta_2^{-1} \log e. \quad (170)$$

For example, if $\beta_1 = \beta_2 = \frac{1}{100}$, (169) gives a possible range [0.037, 0.9631] nats for the ratio of relative entropy to χ^2-divergence, while (170) gives a range of [0.005, 50] nats.

Note that if $\beta_2 = 0$, then the upper bound in (169) is $\kappa_2(0) = \log e$ whereas it is ∞ according to [34, Proposition 2]. In view of Remark 10, the bounds in (169) are the best possible among all probability measures P, Q with given $(\beta_1, \beta_2) \in [0, 1)^2$.

F. Local Behavior of f-Divergences

Another application of Theorem 5 shows that the local behavior of f-divergences differs by only a constant, provided that the first distribution approaches the reference measure in a certain strong sense.

Theorem 9: Suppose that $\{P_n\}$, a sequence of probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, converges to Q (another probability measure on the same space) in the sense that, for $Y \sim Q$,

$$\lim_{n \to \infty} \text{ess sup} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y) = 1 \tag{171}$$

where it is assumed that $P_n \ll Q$ for all sufficiently large n. If f and g are convex on $(0, \infty)$ and they are positive except at $t = 1$ (where they are 0), then

$$\lim_{n \to \infty} D_f(P_n \| Q) = \lim_{n \to \infty} D_g(P_n \| Q) = 0, \tag{172}$$

and

$$\min\{\kappa(1^-), \kappa(1^+)\} \le \lim_{n \to \infty} \frac{D_f(P_n \| Q)}{D_g(P_n \| Q)} \le \max\{\kappa(1^-), \kappa(1^+)\} \tag{173}$$

where we have indicated the left and right limits of the function $\kappa(\cdot)$, defined in (85), at 1 by $\kappa(1^-)$ and $\kappa(1^+)$, respectively.

Proof: Since $f(1) = 0$,

$$0 \le D_f(P_n \| Q) = \int f\left(\frac{\mathrm{d}P_n}{\mathrm{d}Q}\right) \mathrm{d}Q \tag{174}$$

$$\le \sup_{\beta \in [\beta_{2,n}, \beta_{1,n}^{-1}]} f(\beta) \tag{175}$$

where we have abbreviated

$$\beta_{1,n}^{-1} \triangleq \text{ess sup} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y), \tag{176}$$

$$\beta_{2,n} \triangleq \text{ess inf} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y). \tag{177}$$

The condition in (171) yields

$$\lim_{n \to \infty} \beta_{1,n} = 1, \tag{178}$$

$$\lim_{n \to \infty} \beta_{2,n} = 1 \tag{179}$$

where (178) is a restatement of (171) (see the notation in (176)), and Appendix C justifies (179). Hence, (172) follows from (175), (178), (179), the continuity of f at 1 (due to its convexity).

Abbreviating $I_n = [\beta_{2,n}, 1) \cup (1, \beta_{1,n}^{-1}]$, (150) and (156) result in

$$\inf_{\beta \in I_n} \kappa(\beta) D_g(P_n \| Q) \le D_f(P_n \| Q) \le \sup_{\beta \in I_n} \kappa(\beta) D_g(P_n \| Q). \tag{180}$$

The right and left continuity of $\kappa(\cdot)$ at 1 together with (178) and (179) imply that $\inf_{\beta \in I_n} \kappa(\beta) \to \min\{\kappa(1^-), \kappa(1^+)\}$ and $\sup_{\beta \in I_n} \kappa(\beta) \to \max\{\kappa(1^-), \kappa(1^+)\}$ by letting $n \to \infty$. ∎

Corollary 2: Let $\{P_n \ll Q\}$ converge to Q in the sense of (171). Then, $D(P_n \| Q)$ and $D(Q \| P_n)$ vanish as $n \to \infty$ with

$$\lim_{n \to \infty} \frac{D(P_n \| Q)}{D(Q \| P_n)} = 1. \tag{181}$$

Corollary 3: Let $\{P_n \ll Q\}$ converge to Q in the sense of (171). Then, $\chi^2(P_n \| Q)$ and $D(P_n \| Q)$ vanish as $n \to \infty$ with

$$\lim_{n \to \infty} \frac{D(P_n \| Q)}{\chi^2(P_n \| Q)} = \tfrac{1}{2} \log e. \tag{182}$$

Note that (182) is known in the finite alphabet case [28, Th. 4.1]).

In Example 1, the ratio in (173) is equal to $\frac{1}{2}$, while the lower and upper bounds are $\frac{1}{3}$ and 1, respectively.

Continuing with Examples 3, 4 and 5, it is easy to check that (171) is satisfied in the following cases.

Example 7: A sequence of Laplacian probability density functions with common variance and converging means:

$$p_n(x) = \frac{\lambda}{2} \cdot \exp\left(-\lambda |x - a_n|\right) \tag{183}$$

$$\lim_{n \to \infty} a_n = a. \tag{184}$$

Example 8: A sequence of converging Cramér probability density functions:

$$p_n(x) = \frac{\theta_n}{2\left(1 + \theta_n |x - m_n|\right)^2}, \quad x \in \mathbb{R} \tag{185}$$

$$\lim_{n \to \infty} m_n = m \in \mathbb{R} \tag{186}$$

$$\lim_{n \to \infty} \theta_n = \theta > 0. \tag{187}$$

Example 9: A sequence of converging Cauchy probability density functions:

$$p_n(x) = \frac{1}{\pi \gamma_n}\left[1 + \left(\frac{x - m_n}{\gamma_n}\right)^2\right]^{-1}, \quad x \in \mathbb{R} \tag{188}$$

$$\lim_{n \to \infty} m_n = m \in \mathbb{R} \tag{189}$$

$$\lim_{n \to \infty} \gamma_n = \gamma > 0. \tag{190}$$

G. Strengthened Jensen's Inequality

Bounding away from zero a certain density between two probability measures enables the following strengthened version of Jensen's inequality, which generalizes a result in [33, Th. 1].

Lemma 1: Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, $P_1 \ll P_0$ be probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, and fix an arbitrary random transformation $P_{Z|X} : \mathcal{A} \to \mathbb{R}$. Denote[10] $P_0 \to P_{Z|X} \to P_{Z_0}$, and $P_1 \to P_{Z|X} \to P_{Z_1}$. Then,

$$\beta\left(\mathbb{E}\left[f\left(\mathbb{E}[Z_0|X_0]\right)\right] - f\left(\mathbb{E}[Z_0]\right)\right) \le \mathbb{E}\left[f\left(\mathbb{E}[Z_1|X_1]\right)\right] - f\left(\mathbb{E}[Z_1]\right) \tag{191}$$

[10]We follow the notation in [110] where $P_0 \to P_{Z|X} \to P_{Z_0}$ means that the marginal probability measures of the joint distribution $P_0 P_{Z|X}$ are P_0 and P_{Z_0}.

where $X_0 \sim P_0$, $X_1 \sim P_1$, and

$$\beta \triangleq \text{ess inf} \frac{\mathrm{d} P_1}{\mathrm{d} P_0}(X_0). \tag{192}$$

Proof: If $\beta = 0$, the claimed result is Jensen's inequality, while if $\beta = 1$, $P_0 = P_1$ and the result is trivial. Hence, we assume $\beta \in (0, 1)$. Note that $P_1 = \beta P_0 + (1 - \beta)P_2$ where P_2 is the probability measure whose density with respect to P_0 is given by

$$\frac{\mathrm{d} P_2}{\mathrm{d} P_0} = \frac{1}{1 - \beta}\left(\frac{\mathrm{d} P_1}{\mathrm{d} P_0} - \beta\right) \geq 0. \tag{193}$$

Letting $P_2 \to P_{Z|X} \to P_{Z_2}$, Jensen's inequality implies

$$f(\mathbb{E}[Z_1]) \leq \beta f(\mathbb{E}[Z_0]) + (1 - \beta)f(\mathbb{E}[Z_2]). \tag{194}$$

Furthermore, we can apply Jensen's inequality again to obtain

$$f(\mathbb{E}[Z_2]) = f(\mathbb{E}[\mathbb{E}[Z_2|X_2]]) \tag{195}$$

$$\leq \mathbb{E}[f(\mathbb{E}[Z_2|X_2])] \tag{196}$$

$$= \frac{\mathbb{E}[f(\mathbb{E}[Z_1|X_1])] - \beta \mathbb{E}[f(\mathbb{E}[Z_0|X_0])]}{1 - \beta}. \tag{197}$$

Substituting this bound on $f(\mathbb{E}[Z_2])$ in (194) we obtain the desired result. ∎

Remark 13: Letting $Z = X$, and choosing P_0 so that $\beta = 0$ (e.g., P_1 is a restriction of P_0 to an event of P_0-probability less than 1), (191) becomes Jensen's inequality $f(\mathbb{E}[X_1]) \leq \mathbb{E}[f(X_1)]$.

Lemma 1 finds the following application to the derivation of f-divergence inequalities.

Theorem 10: Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Fix $P \ll Q$ on the same space with $(\beta_1, \beta_2) \in [0, 1)^2$ and let $X \sim P$. Then,

$$\beta_2 D_f(P \| Q) \leq \mathbb{E}\left[f\left(\exp(\imath_{P\|Q}(X))\right)\right] - f\left(1 + \chi^2(P\|Q)\right) \tag{198}$$

$$\leq \beta_1^{-1} D_f(P\|Q). \tag{199}$$

Proof: We invoke Lemma 1 with $P_{Z|X}$ that is given by the deterministic transformation $\exp\left(\imath_{P\|Q}(\cdot)\right) : \mathcal{A} \to \mathbb{R}$. Then, $\mathbb{E}[Z_0|X_0] = \exp\left(\imath_{P\|Q}(X_0)\right)$. If, moreover, we let $X_0 \sim Q = P_0$, we obtain

$$\mathbb{E}[Z_0] = 1, \tag{200}$$

$$\mathbb{E}[f(\mathbb{E}[Z_0|X_0])] = D_f(P\|Q) \tag{201}$$

and if we let $X_1 \sim P = P_1$, we have (see (50))

$$\mathbb{E}[Z_1] = 1 + \chi^2(P\|Q), \tag{202}$$

$$\mathbb{E}[f(\mathbb{E}[Z_1|X_1])] = \mathbb{E}\left[f\left(\exp(\imath_{P\|Q}(X))\right)\right]. \tag{203}$$

Therefore, (198) follows from Lemma 1. Recalling (139), inequality (199) follows from Lemma 1 as well switching the roles P_0 and P_1, namely, now we take $P = P_0$ and $Q = P_1$. ∎

Specializing Theorem 10 to the convex function on $(0, \infty)$ where $f(t) = -\log t$ sharpens inequality (94) under the assumption of bounded relative information.

Theorem 11: Fix $P \lll Q$ such that $(\beta_1, \beta_2) \in (0, 1)^2$. Then,

$$\beta_2 D(Q\|P) \leq \log\left(1 + \chi^2(P\|Q)\right) - D(P\|Q) \tag{204}$$

$$\leq \beta_1^{-1} D(Q\|P). \tag{205}$$

V. TOTAL VARIATION DISTANCE, RELATIVE INFORMATION SPECTRUM AND RELATIVE ENTROPY

A. Exact Expressions

The following result provides several useful expressions of the total variation distance in terms of the relative information.

Theorem 12: Let $P \ll Q$, and let $X \sim P$ and $Y \sim Q$ be defined on a measurable space $(\mathcal{A}, \mathcal{F})$. Then,[11]

$$|P - Q| = \mathbb{E}\left[\left|1 - \exp(\imath_{P\|Q}(Y))\right|\right] \tag{206}$$

$$= 2\mathbb{E}\left[\left(1 - \exp(\imath_{P\|Q}(Y))\right)^+\right] \tag{207}$$

$$= 2\mathbb{E}\left[\left(1 - \exp(\imath_{P\|Q}(Y))\right)^-\right] \tag{208}$$

$$= 2\mathbb{E}\left[\left(1 - \exp(-\imath_{P\|Q}(X))\right)^+\right] \tag{209}$$

$$= 2\left(\mathbb{P}[\imath_{P\|Q}(X) > 0] - \mathbb{P}[\imath_{P\|Q}(Y) > 0]\right) \tag{210}$$

$$= 2\left(\mathbb{P}[\imath_{P\|Q}(Y) \leq 0] - \mathbb{P}[\imath_{P\|Q}(X) \leq 0]\right) \tag{211}$$

$$= 2\int_0^1 \mathbb{P}[\imath_{P\|Q}(Y) < \log \beta]\mathrm{d}\beta \tag{212}$$

$$= 2\int_0^1 \mathbb{P}\left[\imath_{P\|Q}(X) > \log \frac{1}{\beta}\right]\mathrm{d}\beta \tag{213}$$

$$= 2\int_1^{\beta_1^{-1}} \beta^{-2}\left[1 - \mathbb{F}_{P\|Q}(\log \beta)\right]\mathrm{d}\beta. \tag{214}$$

Furthermore, if $P \lll Q$, then

$$|P - Q| = 2\mathbb{E}\left[\left(1 - \exp(-\imath_{P\|Q}(X))\right)^-\right] \tag{215}$$

$$= \mathbb{E}\left[\left|1 - \exp(-\imath_{P\|Q}(X))\right|\right]. \tag{216}$$

Proof: See Appendix D. ∎

Remark 14: In view of (210), if $P \ll Q$, the supremum in (58) is a maximum which is achieved by the event

$$\mathcal{F}^* = \left\{a \in \mathcal{A} : \imath_{P\|Q}(a) > 0\right\} \in \mathcal{F}. \tag{217}$$

Similarly to Theorem 12, the following theorem provides several expressions of the relative entropy in terms of the relative information spectrum.

Theorem 13: If $D(P\|Q) < \infty$, then

$$D(P\|Q)$$

$$= \int_0^\infty \left(1 - \mathbb{F}_{P\|Q}(\alpha)\right)\mathrm{d}\alpha - \int_{-\infty}^0 \mathbb{F}_{P\|Q}(\alpha)\mathrm{d}\alpha \tag{218}$$

$$= \int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta}\mathrm{d}\beta - \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log \beta)}{\beta}\mathrm{d}\beta \tag{219}$$

$$= \int_0^\infty \mathbb{P}[\imath_{P\|Q}(Y) > \alpha]\alpha e^\alpha \mathrm{d}\alpha$$

$$\quad - \int_{-\infty}^0 \mathbb{P}[\imath_{P\|Q}(Y) < \alpha]\alpha e^\alpha \mathrm{d}\alpha \tag{220}$$

where $Y \sim Q$, and for convenience (219) and (220) assume that the relative information and the resulting relative entropy are in nats.

Proof: The expectation of a real-valued random variable V is equal to

$$\mathbb{E}[V] = \int_0^\infty \mathbb{P}[V > t]\mathrm{d}t - \int_{-\infty}^0 \mathbb{P}[V < t]\mathrm{d}t \tag{221}$$

[11] $(z)^+ \triangleq z1\{z > 0\} = \max\{z, 0\}$, and $(z)^- \triangleq -z1\{z < 0\} = \max\{-z, 0\}$.

where we are free to substitute $>$ by $\geq$, and $<$ by $\leq$. If we let $V = \iota_{P\|Q}(X)$ with $X \sim P$, then (221) yields (218) provided that $D(P\|Q) = \mathbb{E}[V] < \infty$.

Eq. (219) follows from (218) by the substitution $\alpha = \log \beta$ when the relative entropy is expressed in nats.

To prove (220), let $Z = \iota_{P\|Q}(Y)$ with $Y \sim Q$, and let $V = r(Z)$ where $r\colon (0, \infty) \to [0, \infty)$ is given in (43) with natural logarithm. The function r is strictly monotonically increasing on $[1, \infty)$, on which interval we define its inverse by $s_1\colon [0, \infty) \to [1, \infty)$; it is also strictly monotonically decreasing on $(0, 1]$, on which interval we define its inverse by $s_2\colon [0, 1] \to (0, 1]$. Then, only the first integral on the right side of (221) can be non-zero, and we decompose it as

$$D(P\|Q)$$
$$= \int_0^\infty \mathbb{P}[Z \geq 1, r(Z) > t]\mathrm{d}t + \int_0^\infty \mathbb{P}[Z < 1, r(Z) > t]\mathrm{d}t$$
$$= \int_0^\infty \mathbb{P}[Z > s_1(t)]\mathrm{d}t + \int_0^1 \mathbb{P}[Z < s_2(t)]\mathrm{d}t \qquad (222)$$
$$= \int_1^\infty \mathbb{P}[Z > v] \log_e v \, \mathrm{d}v - \int_0^1 \mathbb{P}[Z < v] \log_e v \, \mathrm{d}v \qquad (223)$$

where (223) follows from the change of variable of integration $t = r(v)$. Upon taking $\log_e$ on both sides of the inequalities inside the probabilities in (223), and a further change of the variable of integration $v = e^\alpha$, (223) is seen to be equal to (220). ∎

B. Upper Bounds on $|P - Q|$

In this section, we provide three upper bounds on $|P - Q|$ which complement (1).

Theorem 14: If $P \ll Q$ and $X \sim P$, then

$$|P - Q| \log e \leq D(P\|Q) + \mathbb{E}\big[|\iota_{P\|Q}(X)|\big]. \qquad (224)$$

Proof: For every $z \in [-\infty, \infty]$,

$$\big(1 - \exp(-z)\big)\mathbf{1}\{z > 0\} \leq \left(\frac{z}{\log e}\right)\mathbf{1}\{z > 0\}. \qquad (225)$$

Substituting $z = \iota_{P\|Q}(X)$, taking expectation of both sides of (225), and using (209) give

$$|P - Q| \log e \leq 2\mathbb{E}\big[\iota_{P\|Q}(X)\mathbf{1}\{\iota_{P\|Q}(X) > 0\}\big] \qquad (226)$$
$$= \mathbb{E}\big[\iota_{P\|Q}(X) + |\iota_{P\|Q}(X)|\big] \qquad (227)$$
$$= D(P\|Q) + \mathbb{E}\big[|\iota_{P\|Q}(X)|\big]. \qquad (228)$$
∎

Remark 15: Theorem 14 is tighter than Pinsker's bound in [77, eq. (2.3.14)]:

$$|P - Q| \log e \leq 2\mathbb{E}\big[|\iota_{P\|Q}(X)|\big]. \qquad (229)$$

The second upper bound on $|P - Q|$ is a consequence of Theorem 12.

Theorem 15: Let $P \ll Q$ with $(\beta_1, \beta_2) \in [0, 1)^2$. Then, for every $\beta_0 \in [\beta_1, 1]$,

$$\tfrac{1}{2}|P - Q| \leq (1 - \beta_0)\mathbb{P}[\iota_{P\|Q}(X) > 0]$$
$$+ (\beta_0 - \beta_1)\mathbb{P}\left[\iota_{P\|Q}(X) > \log \tfrac{1}{\beta_0}\right] \qquad (230)$$

where $X \sim P$, and, for every $\beta_0 \in [\beta_2, 1]$,

$$\tfrac{1}{2}|P - Q| \leq (1 - \beta_0)\mathbb{P}[\iota_{P\|Q}(Y) < 0]$$
$$+ (\beta_0 - \beta_2)\mathbb{P}[\iota_{P\|Q}(Y) < \log \beta_0] \qquad (231)$$

where $Y \sim Q$. Furthermore, both upper bounds on $|P - Q|$ in (230) and (231) are tight in the sense that they are achievable by suitable pairs of probability measures defined on a binary alphabet.

Proof: Since the integrand in the right side of (213) is monotonically increasing in β, we may upper bound it by $\mathbb{P}[\iota_{P\|Q}(X) > \log \tfrac{1}{\beta_0}]$ when $\beta \in [\beta_1, \beta_0]$, and by $\mathbb{P}[\iota_{P\|Q}(X) > 0]$ when $\beta \in (\beta_0, 1]$. The same reasoning applied to (212) yields (231).

To check the tightness of (230) for any $\beta_1 \in (0, 1]$, choose an arbitrary $\eta \in (0, 1)$ and a pair of probability measures P and Q defined on the binary alphabet $\mathcal{A} = \{0, 1\}$ with

$$P(0) = \frac{1 - \eta}{1 - \eta\beta_1}, \qquad (232)$$
$$Q(0) = \beta_1 P(0). \qquad (233)$$

Then, we have $\iota_{P\|Q}(0) = \log \tfrac{1}{\beta_1}$, $\iota_{P\|Q}(1) = \log \eta < 0$, and both sides of (230) are readily seen to be equal when $\beta_0 = \beta_1$. The tightness of (231) can be shown in a similar way. ∎

The third upper bound on $|P - Q|$ is a classical inequality [55, eq. (99)], usually given in the context of bounding the error probability of Bayesian binary hypothesis testing in terms of the Bhattacharyya distance.

Theorem 16:

$$\tfrac{1}{4}|P - Q|^2 \leq 1 - \exp\big(-D_{\frac{1}{2}}(P\|Q)\big). \qquad (234)$$

Remark 16: The bound in (234) is tight if P, Q are defined on $\{0, 1\}$ with $P(0) = Q(0)$ or $P(0) = Q(1)$.

In view of the monotonicity of $D_\alpha(P\|Q)$ in α, Theorem 16 yields (4), which is equivalent to the Bretagnole-Huber inequality [15, eq. (2.2)] (see also [107, pp. 30–31]). Note that (4) is tighter than (1) only when $|P - Q| > 1.7853$.

C. Lower Bounds on $|P - Q|$

In this section, we give several lower bounds on $|P - Q|$ in terms of the relative information spectrum. Furthermore, in Section VI, we give lower bounds on $|P - Q|$ in terms of the relative entropy (as well as other features of P and Q).

If, for at least one value of $\beta \in (0, 1)$, either $\mathbb{P}[\iota_{P\|Q}(X) > \log \tfrac{1}{\beta}]$ or $\mathbb{P}[\iota_{P\|Q}(Y) < \log \beta]$ are known then we get the following lower bounds on $|P - Q|$ as a consequence of Theorem 12:

Theorem 17: If $P \ll Q$ then, for every $\beta_0 \in (0, 1)$,

$$|P - Q| \geq 2(1 - \beta_0)\mathbb{P}\big[\iota_{P\|Q}(Y) < \log \beta_0\big], \qquad (235)$$
$$|P - Q| \geq 2(1 - \beta_0)\mathbb{P}\left[\iota_{P\|Q}(X) > \log \frac{1}{\beta_0}\right] \qquad (236)$$

with $X \sim P$ and $Y \sim Q$.

Proof: The lower bounds in (235) and (236) follow from (212) and (213) respectively. For example, from (212), it follows that for an arbitrary $\beta_0 \in (0, 1)$

$$\tfrac{1}{2}|P - Q| = \int_0^1 \mathbb{P}\big[\iota_{P\|Q}(Y) < \log \beta\big]\mathrm{d}\beta \qquad (237)$$

$$\geq \int_{\beta_0}^{1} \mathbb{P}\big[\iota_{P\|Q}(Y) < \log \beta\big] \, d\beta \tag{238}$$

$$\geq (1 - \beta_0) \, \mathbb{P}\big[\iota_{P\|Q}(Y) < \log \beta_0\big] \tag{239}$$

where (239) holds since the integrand in (238) is monotonically increasing in $\beta \in (0, 1]$. ∎

Next we exemplify the utility of Theorem 17 by giving an alternative proof to the tight lower bound on the relative information spectrum, given in [68, Proposition 2] as a function of the total variation distance.

Proposition 2: Let $P \ll Q$, then for every $\beta > 0$

$$\mathbb{F}_{P\|Q}(\log \beta) \geq \begin{cases} 0, & \beta \in \big(0, \frac{2}{2-|P-Q|}\big], \\ 1 - \frac{\beta|P-Q|}{2(\beta-1)}, & \beta \in \big(\frac{2}{2-|P-Q|}, \infty\big). \end{cases} \tag{240}$$

Furthermore, for every $\beta > 0$ and $\delta \in [0, 1)$, the lower bound in (240) is attainable by a pair (P, Q) with $|P - Q| = 2\delta$.

Proof: Since $P \ll Q$, they cannot be mutually singular and therefore $|P - Q| < 2$. From (27) and (236) (see Theorem 17), it follows that for every $\beta_0 \in (0, 1)$

$$\tfrac{1}{2}|P - Q| \geq (1 - \beta_0)\Big[1 - \mathbb{F}_{P\|Q}\Big(\log \frac{1}{\beta_0}\Big)\Big]. \tag{241}$$

Consequently, the substitution $\beta = \frac{1}{\beta_0} > 1$ yields

$$\mathbb{F}_{P\|Q}(\log \beta) \geq 1 - \frac{\beta|P-Q|}{2(\beta-1)} \tag{242}$$

which provides a non-negative lower bound on the relative information spectrum provided that $\beta \geq \frac{2}{2-|P-Q|}$. Having shown (240), we proceed to argue that it is tight. Fix $\delta \in [0, 1)$ and let $|P - Q| = 2\delta$, which yields $\frac{2}{2-|P-Q|} = \frac{1}{1-\delta}$ in the right side of (240).

- If $\beta < \frac{1}{1-\delta}$, let the pair (P, Q) be defined on the binary alphabet $\{0, 1\}$ with $P(1) = 1$ and $Q(1) = 1 - \delta$ (thereby ensuring $2\delta = |P - Q|$). Then, from (27),

$$\mathbb{F}_{P\|Q}(\log \beta) = P(0) = 0. \tag{243}$$

- If $\beta \geq \frac{1}{1-\delta}$, let $\tau > \beta$ and consider the probability measures $P = P_\tau$ and $Q = Q_\tau$ defined on the binary alphabet $\{0, 1\}$ with $P_\tau(1) = \frac{\tau\delta}{\tau-1}$ and $Q_\tau(1) = \frac{\delta}{\tau-1}$ (note that indeed $2\delta = |P_\tau - Q_\tau|$). Since $1 < \beta < \tau$ then

$$\mathbb{F}_{P\|Q}(\log \beta) = P_\tau(0) = 1 - \frac{\tau\delta}{\tau-1} \tag{244}$$

which tends to $1 - \frac{\beta\delta}{\beta-1}$ in the right side of (240) by letting $\tau \downarrow \beta$. ∎

Attained under certain conditions, the following counterpart to Theorem 15 gives a lower bound on the total variation distance based on the distribution of the relative information. It strengthens the bound in [109, Th. 8], which in turn tightens the lower bounds in [77, eq. (2.3.18)] and [98, Lemma 7].

Theorem 18: If $P \ll\gg Q$ then, for any $\eta_1, \eta_2 > 0$,

$$|P - Q| \geq \big(1 - \exp(-\eta_1)\big)\mathbb{P}\big[\iota_{P\|Q}(X) \geq \eta_1\big]$$
$$+ \big(\exp(\eta_2) - 1\big)\mathbb{P}\big[\iota_{P\|Q}(X) \leq -\eta_2\big] \tag{245}$$

with $X \sim P$. Equality holds in (245) if P and Q are probability measures defined on $\{0, 1\}$ and, for an arbitrary

$\eta_1, \eta_2 > 0$,

$$P(0) = \frac{1 - \exp(-\eta_2)}{1 - \exp(-\eta_1 - \eta_2)}, \tag{246}$$

$$Q(0) = \exp(-\eta_1) P(0). \tag{247}$$

Proof: From (216), it follows that for arbitrary $\eta_1, \eta_2 > 0$,

$$|P - Q| \geq \mathbb{E}\big[\big|1 - \exp(-\iota_{P\|Q}(X))\big|1\{\iota_{P\|Q}(X) \geq \eta_1\}\big]$$
$$+ \mathbb{E}\big[\big|1 - \exp(-\iota_{P\|Q}(X))\big|1\{\iota_{P\|Q}(X) \leq -\eta_2\}\big] \tag{248}$$

which is readily loosened to obtain (245). Equality holds in (245) for P and Q in the theorem statement since $\iota_{P\|Q}(X)$ only takes the values $\log \frac{P(0)}{Q(0)} = \eta_1$ and $\log \frac{P(1)}{Q(1)} = -\eta_2$. ∎

The following lower bound on the total variation distance is the counterpart to Theorem 14.

Theorem 19: If $P \ll\gg Q$, and $X \sim P$ then

$$|P - Q| \log e \geq \mathbb{E}\big[|\iota_{P\|Q}(X)|\big] - D(P\|Q). \tag{249}$$

Proof: We reason in parallel to the proof of Theorem 14. For all $z \in [-\infty, \infty]$,

$$\big[1 - \exp(-z)\big]^{-} \geq \frac{(z)^{-}}{\log e}. \tag{250}$$

Substituting $z = \iota_{P\|Q}(X)$, taking expectation of both sides of (250), and using (215) we obtain

$$|P - Q| \log e \geq 2\mathbb{E}\big[\big(\iota_{P\|Q}(X)\big)^{-}\big] \tag{251}$$

$$= \mathbb{E}\big[|\iota_{P\|Q}(X)| - \iota_{P\|Q}(X)\big] \tag{252}$$

$$= \mathbb{E}\big[|\iota_{P\|Q}(X)|\big] - D(P\|Q). \tag{253}$$

∎

Remark 17: The combination of Pinsker's inequality (1) and (249) yields the following inequality due to Barron (see [6, p. 339]) which is useful in establishing convergence results for relative entropy (e.g. [7])

$$\mathbb{E}\big[|\iota_{P\|Q}(X)|\big] \leq D(P\|Q) + \sqrt{2D(P\|Q) \log e} \tag{254}$$

with $X \sim P$.

D. Relative Entropy and Bhattacharyya Distance

The following result refines (5) by using an approach which relies on moment inequalities [94]–[96]. The coverage in this section is self-contained.

Theorem 20: If $P \ll\gg Q$, then

$$D(P\|Q) \leq \log\big(1 + \chi^2(P\|Q)\big)$$
$$- \frac{\frac{3}{2}\big(\chi^2(P\|Q)\big)^2 \log e}{\big(1 + \chi^2(Q\|P)\big)\big(1 + \chi^2(P\|Q)\big)^2 - 1}. \tag{255}$$

Furthermore, if $\{P_n\}$ converges to Q in the sense of (171), then the ratio of $D(P_n\|Q)$ and its upper bound in (255) tends to 1 as $n \to \infty$.

Proof: The derivation of (255) relies on [95, Th. 2.1] which states that if W is a non-negative random variable, then

$$\lambda_\alpha \triangleq \begin{cases} \frac{\big(\mathbb{E}[W^\alpha] - \mathbb{E}^\alpha[W]\big) \log e}{\alpha(\alpha-1)}, & \alpha \neq 0, 1 \\ \log\big(\mathbb{E}[W]\big) - \mathbb{E}[\log W], & \alpha = 0 \\ \mathbb{E}[W \log W] - \mathbb{E}[W] \log\big(\mathbb{E}[W]\big), & \alpha = 1 \end{cases} \tag{256}$$

is log-convex in $\alpha \in \mathbb{R}$.

To prove (255), let $W = \frac{dP}{dQ}(X)$ with $X \sim P$, then (256) yields

$$\lambda_0 = \log\big(1 + \chi^2(P\|Q)\big) - D(P\|Q), \tag{257}$$

$$\lambda_{-\alpha} = \frac{1}{\alpha(\alpha+1)}\Big[1 + (\alpha-1)\mathscr{H}_\alpha(Q\|P) \\ - \big(1 + \chi^2(P\|Q)\big)^{-\alpha}\Big]\log e \tag{258}$$

for all $\alpha > 0$, and specializing (258) yields

$$\lambda_{-1} = \frac{\chi^2(P\|Q)\log e}{2\big(1 + \chi^2(P\|Q)\big)}, \tag{259}$$

$$\lambda_{-2} = \frac{1}{6}\left[1 + \chi^2(Q\|P) - \frac{1}{\big(1 + \chi^2(P\|Q)\big)^2}\right]\log e. \tag{260}$$

In view of the log-convexity of λ_α in $\alpha \in \mathbb{R}$, then

$$\lambda_0\lambda_{-2} \geq \lambda_{-1}^2 \tag{261}$$

which, by assembling (257)–(261), yields (255).

Suppose that $\{P_n\}$ converges to Q in the sense of (171). Then, it follows from Theorem 9 and Corollary 3 that

$$\lim_{n\to\infty} D(P_n\|Q) = 0, \tag{262}$$

$$\lim_{n\to\infty} \chi^2(P_n\|Q) = 0, \tag{263}$$

$$\lim_{n\to\infty} \frac{D(P_n\|Q)}{\chi^2(P_n\|Q)} = \frac{1}{2}\log e, \tag{264}$$

$$\lim_{n\to\infty} \frac{\chi^2(Q\|P_n)}{\chi^2(P_n\|Q)} = 1. \tag{265}$$

Let U_n denote the upper bound on $D(P_n\|Q)$ in (255). Assembling (262)–(265), it can be verified that

$$\lim_{n\to\infty} \frac{U_n}{D(P_n\|Q)} = 1. \tag{266}$$

∎

Remark 18: In view of (262)–(265), while the ratio of the right side of (255) with $P = P_n$ and $D(P_n\|Q)$ tends to 1, the ratio of the looser bound in (5) and $D(P_n\|Q)$ tends to 2.

Remark 19: If $\{P_n\}$ and Q are defined on a finite set $\mathcal{A}$, then the condition in (171) is equivalent to $|P_n - Q| \to 0$ with $Q(a) > 0$ for all $a \in \mathcal{A}$.

Remark 20: An alternative refinement of (5) has been recently obtained in [97] as a function of $\chi^2(P\|Q)$ and the Bhattacharyya distance $B(P\|Q)$ (see Definition 5):

$$D(P\|Q) \\ \leq \log\big(1 + \chi^2(P\|Q)\big) \\ - \frac{32}{9}\frac{\Big[\exp\big(-B(P\|Q)\big)\sqrt{1 + \chi^2(P\|Q)} - 1\Big]^2 \log e}{\chi^2(P\|Q)}. \tag{267}$$

Eq. (267) can be generalized by relying on the log-convexity of λ_α in $\alpha \in \mathbb{R}$, which yields

$$\lambda_0^{1-\alpha}\lambda_{-1}^\alpha \geq \lambda_{-\alpha} \tag{268}$$

for all $\alpha \in (0, 1)$; consequently, assembling (257), (258), (259) and (268) yields

$$D(P\|Q) \\ \leq \log\big(1 + \chi^2(P\|Q)\big) \\ - \left(\frac{2^\alpha}{\alpha(\alpha+1)}\right)^{\frac{1}{1-\alpha}} \big(\chi^2(P\|Q)\big)^{-\frac{\alpha}{1-\alpha}}\log e \\ \cdot \Big[\big(1 - (1-\alpha)\mathscr{H}_\alpha(Q\|P)\big)\big(1 + \chi^2(P\|Q)\big)^\alpha - 1\Big]^{\frac{1}{1-\alpha}} \tag{269}$$

for all $\alpha \in (0, 1)$. Note that in the special case $\alpha = \frac{1}{2}$, (269) becomes (267), as can be readily verified in view of (83) and the symmetry property $\mathscr{H}_{\frac{1}{2}}(P\|Q) = \mathscr{H}_{\frac{1}{2}}(Q\|P)$.

Remark 21: The following lower bound on the relative entropy has been derived in [97], based on the approach of moment inequalities[12]:

$$D(P\|Q) \geq 2B(P\|Q) \\ + \frac{6\big[1 - \exp\big(-2B(P\|Q)\big)\big]^2 \log e}{1 - \exp\big(-4B(P\|Q)\big) + \chi^2(Q\|P)}. \tag{270}$$

Note that from (82)

$$B(P\|Q) \geq \frac{1}{2}\log\left(\frac{1}{1 - \frac{1}{4}|P - Q|^2}\right) \tag{271}$$

and since the right side of (270) is monotonically increasing in $B(P\|Q)$, the replacement of $B(P\|Q)$ in the right side of (270) with its lower bound in (271) yields

$$D(P\|Q) \geq \log\left(\frac{1}{1 - \frac{1}{4}|P - Q|^2}\right) \\ + \frac{\frac{3}{4}|P - Q|^2 \log e}{1 - \frac{1}{8}|P - Q|^2 + \frac{2\chi^2(Q\|P)}{|P-Q|^2}}. \tag{272}$$

Although (272) improves the bound in (4), it is weaker than (270), and it satisfies the tightness property in Theorem 20 only in special cases such as when P and Q are defined on $\mathcal{A} = \{0, 1\}$ with $P(0) = Q(1) = \frac{1}{2} - \varepsilon$ and we let $\varepsilon \to 0$.

Define the binary relative entropy function as the continuous extension to $[0, 1]^2$ of

$$d(x\|y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right). \tag{273}$$

The following result improves the upper bound in (169).

Theorem 21: Let $P \lll\ggg Q$ with $(\beta_1, \beta_2) \in (0, 1)^2$. Then,

a)

$$\chi^2(P\|Q) \leq \big(\beta_1^{-1} - 1\big)(1 - \beta_2), \tag{274}$$

which is attainable for binary alphabets.

b)

$$D(P\|Q) \\ \leq \min\left\{\log(1 + c) - \frac{\frac{3}{2}c\log e}{1 + \big(1 + \beta_2^{-1}\big)(1 + c)}, \right. \tag{275}$$

[12]For the derivation of (270) for a general alphabet, similarly to [97], set $W = \sqrt{\frac{dQ}{dP}(X)}$ in (256) with $X \sim P$, and use the inequality $\lambda_0\lambda_4 \geq \lambda_2^2$ which follows from the log-convexity of λ_α in α.

$$\frac{\left(\sqrt{c^2 + 8\beta_2 c \log_e(1+c)} - c\right)\log e}{4\beta_2}, \qquad (276)$$

$$d\left(\frac{\beta_1^{-1} - 1}{\beta_1^{-1}\beta_2^{-1} - 1} \,\middle\|\, \frac{\beta_2^{-1}(\beta_1^{-1} - 1)}{\beta_1^{-1}\beta_2^{-1} - 1}\right)\Bigg\} \qquad (277)$$

where we have abbreviated $c = \chi^2(P\|Q)$ for typographical convenience.

To prove (274), we first consider the case where P, Q are defined on $\mathcal{A} = \{0, 1\}$ and $\frac{P(0)}{Q(0)} = \beta_2$, $\frac{P(1)}{Q(1)} = \beta_1^{-1}$. Straightforward calculation yields

$$P(0) = \frac{\beta_1^{-1} - 1}{\beta_1^{-1}\beta_2^{-1} - 1}, \quad Q(0) = \beta_2^{-1}P(0) \qquad (278)$$

and

$$\chi^2(P\|Q) = (\beta_1^{-1} - 1)(1 - \beta_2). \qquad (279)$$

In the case of a general alphabet, consider the elementary bound with $a < 0 < b$: $\mathbb{E}[Z^2] \leq -ab$ which holds for any $Z \in [a, b]$, $\mathbb{E}[Z] = 0$, and follows simply by taking expectations of

$$Z^2 = -ab + Z(a+b) - (Z-a)(b-Z) \qquad (280)$$
$$\leq -ab + Z(a+b). \qquad (281)$$

Since $\chi^2(P\|Q) = \mathbb{E}[Z^2]$, (274) follows by letting $a = \beta_2 - 1$, $b = \beta_1^{-1} - 1$ and

$$Z = \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) - 1, \quad Y \sim Q. \qquad (282)$$

To prove (275), note that it follows by combining (255) with the left side of the inequality

$$\beta_2\chi^2(Q\|P) \leq \frac{\chi^2(P\|Q)}{1 + \chi^2(P\|Q)} \leq \beta_1^{-1}\chi^2(Q\|P) \qquad (283)$$

where (283) follows from Theorem 10 with $f(t) = \frac{1}{t} - 1$ for $t > 0$.

To prove (276), note that assembling (8) and (204) yields

$$D(P\|Q) \leq \log\left(1 + \chi^2(P\|Q)\right) - \frac{2\beta_2 D^2(P\|Q)}{\chi^2(P\|Q)\log e} \qquad (284)$$

and solving this quadratic inequality in $D(P\|Q)$, for fixed $\chi^2(P\|Q)$, yields the bound in (276).

Bound (277) holds since the maximal $D(P\|Q)$, for fixed $\chi^2(P\|Q)$, is monotonically increasing in $\chi^2(P\|Q)$. In view of (274), $D(P\|Q)$ cannot be larger than its maximal value when $\chi^2(P\|Q) = (\beta_1^{-1} - 1)(1 - \beta_2)$. In the latter case, the condition of equality in (281) (recall that $\mathbb{E}[Z] = 0$) is

$$\mathbb{P}[Z = a] = \frac{b}{b - a} = 1 - \mathbb{P}[Z = b] \qquad (285)$$

which implies that the maximal relative entropy $D(P\|Q)$ over all $P \lll Q$ with given $(\beta_1, \beta_2) \in (0, 1)^2$ is equal to

$$\mathbb{E}\big[(1+Z)\log(1+Z)\big] \qquad (286)$$

$$= \frac{b(1+a)\log(1+a) - a(1+b)\log(1+b)}{b - a} \qquad (287)$$
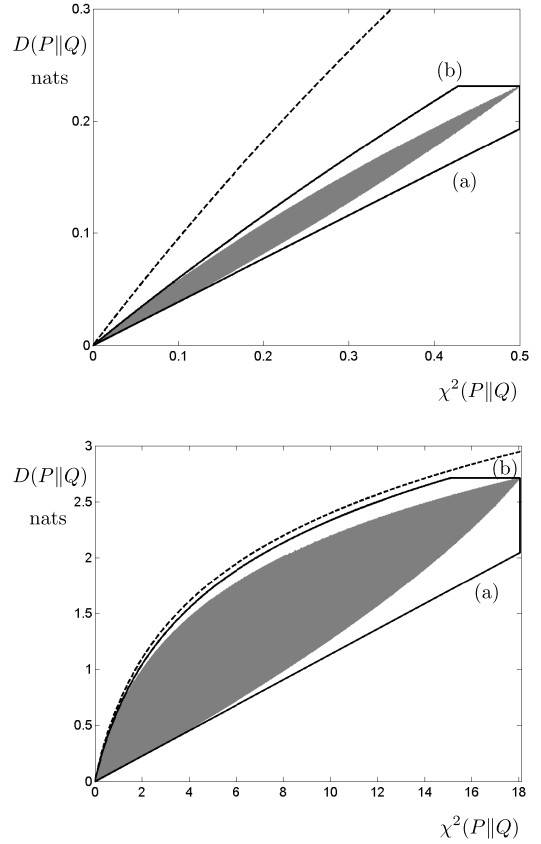

Fig. 1.   Comparison of the locus of the points $(\chi^2(P\|Q), D(P\|Q))$ when P, Q are defined on $\mathcal{A} = \{0, 1\}$ with the bounds in the left side of (169) (a) and Theorem 21 (b); $(\beta_1, \beta_2) = \left(\frac{1}{2}, \frac{1}{2}\right)$ and $(\beta_1, \beta_2) = \left(\frac{1}{20}, \frac{1}{20}\right)$ in the upper and lower plots, respectively. The dashed curve in each plot corresponds to the looser bound in (5).

$$= d\left(\frac{\beta_1^{-1} - 1}{\beta_1^{-1}\beta_2^{-1} - 1} \,\middle\|\, \frac{\beta_2^{-1}(\beta_1^{-1} - 1)}{\beta_1^{-1}\beta_2^{-1} - 1}\right) \qquad (288)$$

where (286)–(288) follow from (273) and (285) with $a = \beta_2 - 1$ and $b = \beta_1^{-1} - 1$.

Remark 22: The proof of (275) relies on the left side of (283); this strengthens the bound which follows from Theorem 5, given by $\chi^2(Q\|P) \leq \beta_2^{-1}\chi^2(P\|Q)$. The bound (276) is typically of similar tightness as the bound in (275), although none of them outperforms the other for all $(\beta_1, \beta_2) \in (0, 1)^2$ and $\chi^2(P\|Q) \in \left[0, (\beta_1^{-1} - 1)(1 - \beta_2)\right]$ (see (274)).

Remark 23: The left inequality in (169) and Theorem 21 provide an analytical outer bound on the locus of the points $(\chi^2(P\|Q), D(P\|Q))$ where $P \lll Q$ and $\beta_2 \leq \frac{\mathrm{d}P}{\mathrm{d}Q} \leq \beta_1^{-1}$ for given $(\beta_1, \beta_2) \in (0, 1)^2$.

Example 10: In continuation to Remark 23, for given $(\beta_1, \beta_2) \in (0, 1)^2$, Figure 1 compares the locus of the points $(\chi^2(P\|Q), D(P\|Q))$ when P, Q are restricted to binary alphabets, and $\frac{P}{Q}$ is bounded between β_2 and β_1^{-1}, with an outer bound constructed with the left inequality in (169) and Theorem 21 (recall that the outer bound is valid for an arbitrary alphabet).

The following result relies on the earlier analysis to provide bounds on the Bhattacharyya distance, expressed in terms of χ^2 divergences and relative entropy.

Theorem 22: If $P \ll\gg Q$, then the following bounds on the Bhattacharyya distance hold:

$$\frac{1}{2}\log(1 + \chi^2(P\|Q))$$

$$- \log\left(1 + \frac{3}{4}\sqrt{\frac{\chi^2(P\|Q)}{2\log e}\left[\log(1 + \chi^2(P\|Q)) - D(P\|Q)\right]}\right)$$

$$\leq B(P\|Q) \tag{289}$$

$$\leq \frac{1}{2}\log(1 + \chi^2(P\|Q))$$

$$- \log\left(1 + \frac{\left(\frac{3}{4}\chi^2(P\|Q)\right)^{\frac{3}{2}}}{\sqrt{(1 + \chi^2(P\|Q))^2(1 + \chi^2(Q\|P)) - 1}}\right). \tag{290}$$

Furthermore, if $\{P_n\}$ converges to Q in the sense of (171), then the ratio of the bounds on $B(P_n\|Q)$ in (289) and (290) tends to 1 as $n \to \infty$.

Proof: In view of the log-convexity of λ_α in $\alpha \in \mathbb{R}$,

$$\lambda_0\lambda_{-1} \geq \lambda_{-\frac{1}{2}}^2, \quad \lambda_{-\frac{1}{2}}^2\lambda_{-2} \geq \lambda_{-1}^3 \tag{291}$$

for any choice of the random variable W in (256). Consequently, assembling (83), (257), (258) and (291) yield the bounds on $B(P\|Q)$ in (289) and (290).

Suppose that $\{P_n\}$ converges to Q in the sense of (171). Let L_n and U_n denote, respectively, the lower and upper bounds on $B(P_n\|Q)$ in (289) and (290). Assembling (262)–(265), it can be easily verified that

$$\lim_{n\to\infty} \frac{L_n}{\chi^2(P_n\|Q)} = \lim_{n\to\infty} \frac{U_n}{\chi^2(P_n\|Q)} = \frac{1}{8}\log e \tag{292}$$

which yields that $\lim_{n\to\infty} \frac{U_n}{L_n} = 1$. ∎

Remark 24: Note that (290) refines the bound

$$B(P\|Q) \leq \frac{1}{2}\log(1 + \chi^2(P\|Q)) \tag{293}$$

which is equivalent to $\lambda_{-\frac{1}{2}} \geq 0$ (in view of Jensen's inequality, (83) and (258)).

Remark 25: Let $\{P_n\}$ converge to Q in the sense of (171). In view of (292), it follows that

$$\lim_{n\to\infty} \frac{B(P_n\|Q)}{\chi^2(P_n\|Q)} = \frac{1}{8}\log e, \tag{294}$$

from which we can surmise that both upper bounds in (255) and (267) are tight under the condition in (171) (see Theorem 20), although (255) only depends on χ^2-divergences. In view of (294), the lower bound in (270) is also tight under the condition in (171), in the sense that the ratio of $D(P_n\|Q)$ and its lower bound in (270) tends to 1 as $n \to \infty$; this sufficient condition for the tightness of (270) strengthens the result in [97, Sec. 4].

VI. REVERSE PINSKER INEQUALITIES

It is not possible to lower bound $|P - Q|$ solely in terms of $D(P\|Q)$ since for any arbitrarily small $\epsilon > 0$ and arbitrarily large $\lambda > 0$, we can construct examples with $|P - Q| < \epsilon$ and $\lambda < D(P\|Q) < \infty$. Therefore, each of the bounds in this section involves not only $D(P\|Q)$ but another feature of the pair (P, Q).

A. Bounded Relative Information

As in Section IV, the following result involves the bounds on the relative information.

Theorem 23: If $\beta_1 \in (0, 1)$ and $\beta_2 \in [0, 1)$, then,

$$D(P\|Q) \leq \frac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)|P - Q| \tag{295}$$

where $\varphi: [0, \infty) \to [0, \infty)$ is given by

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \frac{t\log t}{t - 1} & t \in (0, 1) \cup (1, \infty) \\ \log e & t = 1. \end{cases} \tag{296}$$

Proof: Let $X \sim P$, $Y \sim Q$, and Z be defined in (32). The function $\varphi: [0, \infty) \to [0, \infty)$ is continuous, monotonically increasing and non-negative; the monotonicity property holds since $(t - 1)^2\varphi'(t) = (t - 1)\log e - \log t \geq 0$ for all $t > 0$, and its non-negativity follows from the fact that φ is monotonically increasing on $[0, \infty)$ and $\varphi(0) = 0$. Accordingly,

$$\varphi(\beta_2) \leq \varphi(Z) \leq \varphi(\beta_1^{-1}) \tag{297}$$

since (32) and (139)–(140) imply that $Z \in [\beta_2, \beta_1^{-1}]$ with probability one. The relative entropy satisfies

$$D(P\|Q) = \mathbb{E}[Z\log Z] \tag{298}$$

$$= \mathbb{E}[\varphi(Z)(Z - 1)] \tag{299}$$

$$= \mathbb{E}[\varphi(Z)(Z - 1)1\{Z > 1\}]$$
$$+ \mathbb{E}[\varphi(Z)(Z - 1)1\{Z < 1\}]. \tag{300}$$

We bound each of the summands in the right side of (300) separately. Invoking (297), we have

$$\mathbb{E}[\varphi(Z)(Z - 1)1\{Z > 1\}]$$
$$\leq \varphi(\beta_1^{-1})\mathbb{E}[(Z - 1)1\{Z > 1\}] \tag{301}$$
$$= \varphi(\beta_1^{-1})\mathbb{E}[(1 - Z)^-] \tag{302}$$
$$= \frac{1}{2}\varphi(\beta_1^{-1})|P - Q| \tag{303}$$

where (302) holds since $(x)^- \triangleq -x1\{x < 0\}$, and (303) follows from (208) with Z in (32). Similarly, (297) yields

$$\mathbb{E}[\varphi(Z)(Z - 1)1\{Z < 1\}]$$
$$\leq \varphi(\beta_2)\mathbb{E}[(Z - 1)1\{Z < 1\}] \tag{304}$$
$$= -\varphi(\beta_2)\mathbb{E}[(1 - Z)^+] \tag{305}$$
$$= -\frac{1}{2}\varphi(\beta_2)|P - Q| \tag{306}$$

where (306) follows from (207). Assembling (300), (303) and (306), we obtain (295). ∎

Remark 26: By dropping the negative term in (295), we can get the weaker version in [109, Th. 7]:

$$D(P\|Q) \leq \left(\frac{\log\frac{1}{\beta_1}}{2(1 - \beta_1)}\right)|P - Q|. \tag{307}$$

The coefficient of $|P - Q|$ in the right side of (307) is monotonically decreasing in β_1 and it tends to $\frac{1}{2}\log e$ by letting $\beta_1 \to 1$. The improvement over (307) afforded by (295) is exemplified in Appendix E. The bound in (307) has been recently used in the context of the optimal quantization of probability measures [12, Proposition 4].

Remark 27: The proof of Theorem 23 hinges on the fact that the function φ is monotonically increasing. It can be verified that φ is also concave and differentiable. Taking into account these additional properties of φ, the bound in Theorem 23 can be tightened as (see Appendix F):

$$D(P\|Q) \leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2) - \varphi'(\beta_1^{-1})\beta_1^{-1}\right)|P - Q|$$
$$+ \varphi'(\beta_1^{-1})\mathbb{E}\big[Z(Z-1)1\{Z > 1\}\big] \tag{308}$$

which is expressed in terms of the distribution of the relative information. The second summand in the right side of (308) satisfies

$$\chi^2(P\|Q) + \tfrac{\beta_2}{2}|P - Q| \leq \mathbb{E}\big[Z(Z-1)1\{Z > 1\}\big] \tag{309}$$
$$\leq \chi^2(P\|Q) + \tfrac{1}{2}|P - Q|. \tag{310}$$

From (32), (208) and $Z \in [\beta_2, \beta_1^{-1}]$, the gap between the upper and lower bounds in (310) satisfies

$$\tfrac{1}{2}(1 - \beta_2)|P - Q| = (1 - \beta_2)\mathbb{E}\big[(1 - Z)^+\big] \tag{311}$$
$$\leq (1 - \beta_2)^2 \tag{312}$$

which is upper bounded by 1, and it is close to zero if $\beta_2 \approx 1$. The combination of (308) and (310) leads to

$$D(P\|Q)$$
$$\leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2) + \varphi'(\beta_1^{-1})\big(1 - \beta_1^{-1}\big)\right)|P - Q|$$
$$+ \varphi'(\beta_1^{-1}) \cdot \chi^2(P\|Q). \tag{313}$$

Remark 28: A special case of (307), where Q is the uniform distribution over a set of a finite size, was recently rediscovered in [58, Corollary 13] based on results from [52].

Remark 29: For $\varepsilon \in [0, 2]$ and a fixed probability measure Q, define

$$D^*(\varepsilon, Q) = \inf_{P : |P-Q| \geq \varepsilon} D(P\|Q). \tag{314}$$

From Sanov's theorem (see [20, Th. 11.4.1]), $D^*(\varepsilon, Q)$ is equal to the asymptotic exponential decay of the probability that the total variation distance between the empirical distribution of a sequence of i.i.d. random variables and the true distribution Q is more than a specified value ε. Bounds on $D^*(\varepsilon, Q)$ have been shown in [10, Th. 1], which, locally, behave quadratically in ε. Although this result was classified in [10] as a reverse Pinsker inequality, note that it differs from the scope of this section which provides, under suitable conditions, lower bounds on the total variation distance as a function of the relative entropy.

B. Lipschitz Constraints

Definition 6: A function $f : \mathcal{B} \to \mathbb{R}$, where $\mathcal{B} \subseteq \mathbb{R}$, is L-Lipschitz if for all $x, y \in \mathcal{B}$

$$|f(x) - f(y)| \leq L|x - y|. \tag{315}$$

The following bound generalizes [35, Th. 6] to the non-discrete setting.

Theorem 24: Let $P \ll Q$ with $\beta_1 \in (0, 1)$ and $\beta_2 \in [0, 1)$, and $f : [0, \infty) \to \mathbb{R}$ be continuous and convex with $f(1) = 0$, and L-Lipschitz on $[\beta_2, \beta_1^{-1}]$. Then,

$$D_f(P\|Q) \leq L|P - Q|. \tag{316}$$

Proof: If $Y \sim Q$, and Z is given by (32) then $f(1) = 0$ yields

$$D_f(P\|Q) = \mathbb{E}\big[f(Z)\big] \tag{317}$$
$$\leq \mathbb{E}\big[|f(Z) - f(1)|\big] \tag{318}$$
$$\leq L\mathbb{E}\big[|Z - 1|\big] \tag{319}$$
$$= L|P - Q| \tag{320}$$

where (320) holds due to (206). ∎

Note that if f has a bounded derivative on $[\beta_2, \beta_1^{-1}]$, we can choose

$$L = \sup_{t \in [\beta_2, \beta_1^{-1}]} |f'(t)| < \infty. \tag{321}$$

Remark 30: In the case $f(t) = t \log t$, $f(0) = 0$, (321) particularizes to

$$L = \max\{|\log(e\beta_2)|, \log(e\beta_1^{-1})\} \tag{322}$$

resulting in a reverse Pinsker inequality which is weaker than that in (295) by at least a factor of 2.

C. Finite Alphabet

Throughout this subsection, we assume that P and Q are probability measures defined on a common finite set $\mathcal{A}$, and Q is strictly positive on $\mathcal{A}$, which has more than one element.

The bound in (307) strengthens the finite-alphabet bound in [38, Lemma 3.10]:

$$D(P\|Q) \leq \log\left(\frac{1}{Q_{\min}}\right) \cdot |P - Q| \tag{323}$$

with

$$Q_{\min} = \min_{a \in \mathcal{A}} Q(a) \leq \tfrac{1}{2}. \tag{324}$$

To verify this, notice that $\beta_1 \geq Q_{\min}$. Let $v : (0, 1) \to (0, \infty)$ be defined by $v(t) = \frac{1}{1-t} \cdot \log \frac{1}{t}$; since v is a monotonically decreasing and non-negative function, we can weaken (307) to write

$$D(P\|Q) \leq \left(\frac{\log \frac{1}{Q_{\min}}}{2(1 - Q_{\min})}\right)|P - Q| \tag{325}$$

$$\leq \log\left(\frac{1}{Q_{\min}}\right) \cdot |P - Q| \tag{326}$$

where (326) follows from (324).

The main result in this subsection is the following bound.

Theorem 25:

$$D(P\|Q) \leq \log\left(1 + \frac{|P - Q|^2}{2Q_{\min}}\right). \tag{327}$$

Furthermore, if $Q \ll P$ and β_2 is defined as in (138), then the following tightened bound holds:

$$D(P\|Q) \leq \log\left(1 + \frac{|P - Q|^2}{2Q_{\min}}\right) - \frac{\beta_2 \log e}{2} \cdot |P - Q|^2. \tag{328}$$

Proof: Combining (5) and the following finite-alphabet upper bound on $\chi^2(P\|Q)$ yields (327):

$$Q_{\min}\chi^2(P\|Q) = \sum_{a \in \mathcal{A}} \frac{(P(a) - Q(a))^2}{Q(a)/Q_{\min}} \tag{329}$$

$$\leq \sum_{a \in \mathcal{A}} \big(P(a) - Q(a)\big)^2 \tag{330}$$

$$\leq \max_{x \in \mathcal{A}} |P(x) - Q(x)| \sum_{a \in \mathcal{A}} |P(a) - Q(a)|$$

$$= |P - Q| \max_{a \in \mathcal{A}} |P(a) - Q(a)| \tag{331}$$

$$\leq \tfrac{1}{2} |P - Q|^2. \tag{332}$$

If $P \lll Q$, then (328) follows by combining (332) and

$$\chi^2(P \| Q) \geq \exp\big(D(P\|Q) + \beta_2 D(Q\|P)\big) - 1 \tag{333}$$

$$\geq \exp\big(D(P\|Q) + \tfrac{1}{2}|P - Q|^2 \beta_2 \log e\big) - 1 \tag{334}$$

where (333) follows by rearranging (204), and (334) follows from (1). ∎

Remark 31: It is easy to check that Theorem 25 strengthens the bound by Csiszár and Talata (23) by at least a factor of 2 since upper bounding the logarithm in (327) gives

$$D(P\|Q) \leq \frac{\log e}{2 Q_{\min}} \cdot |P - Q|^2. \tag{335}$$

Remark 32: In the finite-alphabet case, similarly to (327), one can obtain another upper bound on $D(P\|Q)$ as a function of the ℓ_2 norm $\|P - Q\|_2$:

$$D(P\|Q) \leq \frac{1}{Q_{\min}} \cdot \|P - Q\|_2^2 \log e \tag{336}$$

which appears in the proof of Property 4 of [100, Lemma 7], and also used in [57, eq. (174)]. Furthermore, similarly to (328), the following tightened bound holds if $P \lll Q$:

$$D(P\|Q) \leq \log\left(1 + \frac{\|P - Q\|_2^2}{Q_{\min}}\right) - \frac{\beta_2 \log e}{2} \cdot \|P - Q\|_2^2 \tag{337}$$

which follows by combining (330), (334), and the inequality $\|P - Q\|_2 \leq |P - Q|$.

Remark 33: Combining (1) and (332) yields that if $P \neq Q$ are defined on a common finite set, then

$$\frac{D(P\|Q)}{\chi^2(P\|Q)} \geq Q_{\min} \log e \tag{338}$$

which at least doubles the lower bound in [71, Lemma 6]. This, in turn, improves the tightened upper bound on the strong data processing inequality constant in [71, Th. 10] by a factor of 2.

Remark 34: Reverse Pinsker inequalities have been also derived in quantum information theory ([3], [4]), providing upper bounds on the relative entropy of two quantum states as a function of the trace norm distance when the minimal eigenvalues of the states are positive (c.f. [3, Th. 6] and [4, Th. 1]). When the variational distance is much smaller than the minimal eigenvalue (see [3, eq. (57)]), the latter bounds have a quadratic scaling in this distance, similarly to (327); they are also inversely proportional to the minimal eigenvalue, similarly to the dependence of (327) in $Q_{\min}$.

Remark 35: Let P and Q be probability distributions defined on an arbitrary alphabet $\mathcal{A}$. Combining Theorems 6 and 23 leads to a derivation of an upper bound on the difference $D(P\|Q) - D(Q\|P)$ as a function of $|P - Q|$ as long as $P \lll Q$ and the relative information $\imath_{P\|Q}$

is bounded away from $-\infty$ and $+\infty$. Furthermore, another upper bound on the difference of the relative entropies can be readily obtained by combining Theorems 6 and 25 when P and Q are probability measures defined on a finite alphabet. In the latter case, combining Theorem 6 and (337) also yields another upper bound on $D(P\|Q) - D(Q\|P)$ which scales quadratically with $\|P - Q\|_2$. All these bounds form a counterpart to [5, Th. 1] and Theorem 6, providing measures of the asymmetry of the relative entropy when the relative information is bounded.

D. Distance From the Equiprobable Distribution

If P is a distribution on a finite set $\mathcal{A}$, $H(P)$ gauges the "distance" from U, the equiprobable distribution defined on $\mathcal{A}$, since $H(P) = \log |\mathcal{A}| - D(P\|\mathsf{U})$. Thus, it is of interest to explore the relationship between $H(P)$ and $|P - \mathsf{U}|$. Next, we determine the exact locus of the points $(H(P), |P - \mathsf{U}|)$ among all probability measures P defined on $\mathcal{A}$, and this region is compared to upper and lower bounds on $|P - \mathsf{U}|$ as a function of $H(P)$. As usual, $h(x)$ denotes the continuous extension of $-x \log x - (1 - x) \log(1 - x)$ to $x \in [0, 1]$ and $d(x\|y)$ denotes the binary relative entropy in (273).

Theorem 26: Let U be the equiprobable distribution on a $\{1, \ldots, |\mathcal{A}|\}$, with $1 < |\mathcal{A}| < \infty$.

a) For $\Delta \in (0, 2(1 - |\mathcal{A}|^{-1})]$,[13]

$$\max_{P : |P - \mathsf{U}| = \Delta} H(P) = \log |\mathcal{A}| - \min_m d\left(\tfrac{m}{|\mathcal{A}|} + \tfrac{1}{2}\Delta \,\middle\|\, \tfrac{m}{|\mathcal{A}|}\right) \tag{339}$$

where the minimum in the right side of (339) is over

$$m \in \{1, \ldots, |\mathcal{A}| - \lceil \tfrac{1}{2}\Delta|\mathcal{A}| \rceil\}. \tag{340}$$

Denoting such an integer by m_Δ, the maximum in the left side of (339) is attained by

$$P_\Delta(\ell) = \begin{cases} |\mathcal{A}|^{-1} + \dfrac{\Delta}{2m_\Delta}, & \ell \in \{1, \ldots, m_\Delta\}, \\[2mm] |\mathcal{A}|^{-1} - \dfrac{\Delta}{2(|\mathcal{A}| - m_\Delta)}, & \ell \in \{m_\Delta + 1, \ldots, |\mathcal{A}|\}. \end{cases} \tag{341}$$

b) Let

$$h_k = \begin{cases} 0, & k = 0 \\ h(|\mathcal{A}|^{-1}k) + |\mathcal{A}|^{-1}k \log k, & k \in \{1, \ldots, |\mathcal{A}| - 2\} \\ \log |\mathcal{A}|, & k = |\mathcal{A}| - 1. \end{cases} \tag{342}$$

If $H \in [h_{k-1}, h_k)$ for $k \in \{1, \ldots, |\mathcal{A}| - 1\}$, then

$$\min_{P : H(P) = H} |P - \mathsf{U}| = 2\big(1 - (k + \theta)|\mathcal{A}|^{-1}\big) \tag{343}$$

which is achieved by

$$P_\theta^{(k)}(\ell)$$

$$= \begin{cases} 1 - (k - 1 + \theta)|\mathcal{A}|^{-1}, & \ell = 1 \\ |\mathcal{A}|^{-1}, & \ell \in \{2, \ldots, k\}, \\ \theta|\mathcal{A}|^{-1}, & \ell = k + 1 \\ 0, & \ell \in \{k + 2, \ldots, |\mathcal{A}|\} \end{cases} \tag{344}$$

[13] There is no P with $|P - \mathsf{U}| > 2(1 - |\mathcal{A}|^{-1})$.

where $\theta \in [0, 1)$ is chosen so that $H(P_\theta^{(k)}) = H$.

Proof: See Appendix G. ∎

Remark 36: For probability measures defined on a 2-element set $\mathcal{A}$, the maximal and minimal values of $|P - \mathsf{U}|$ in Theorem 26 coincide. This can be verified since, if $P(1) = p$ for $p \in [0, 1]$, then $|P - \mathsf{U}| = |1 - 2p|$ and $H(P) = h(p)$. Hence, if $|\mathcal{A}| = 2$ and $H(P) = H \in [0, \log 2]$, then

$$|P - \mathsf{U}| = 1 - 2h^{-1}(H) \qquad (345)$$

where $h^{-1}: [0, \log 2] \rightarrow [0, \frac{1}{2}]$ denotes the inverse of the binary entropy function.

Results on the more general problem of finding bounds on $|H(P) - H(Q)|$ based on $|P - Q|$ can be found in [20, Th. 17.3.3], [52], [89], and [113]. Most well-known among them is

$$|H(P) - H(Q)| \leq |P - Q| \log \left(\frac{|\mathcal{A}|}{|P - Q|} \right) \qquad (346)$$

which holds if P, Q are probability measures defined on a finite set $\mathcal{A}$ with $\left| P(a) - Q(a) \right| \leq \frac{1}{2}$ for all $a \in \mathcal{A}$ (see [110], and [27, Lemma 2.7] with a stronger sufficient condition). Particularizing (346) to the case where $Q = \mathsf{U}$, and $\left| P(a) - \frac{1}{|\mathcal{A}|} \right| \leq \frac{1}{2}$ for all $a \in \mathcal{A}$ yields

$$H(P) \geq \log |\mathcal{A}| - |P - \mathsf{U}| \log \left(\frac{|\mathcal{A}|}{|P - \mathsf{U}|} \right), \qquad (347)$$

a bound which finds use in information-theoretic security [30].

Particularizing (1), (4), and (327) we obtain

$$H(P) \leq \log |\mathcal{A}| - \frac{1}{2}|P - \mathsf{U}|^2 \log e, \qquad (348)$$

$$H(P) \leq \log |\mathcal{A}| + \log \left(1 - \frac{1}{4}|P - \mathsf{U}|^2 \right), \qquad (349)$$

$$H(P) \geq \log |\mathcal{A}| - \log \left(1 + \frac{|\mathcal{A}|}{2}|P - \mathsf{U}|^2 \right). \qquad (350)$$

If either $|\mathcal{A}| = 2$ or $8 \leq |\mathcal{A}| \leq 102$, it can be checked that the lower bound on $H(P)$ in (347) is worse than (350), irrespectively of $|P - \mathsf{U}|$ (note that $0 \leq |P - \mathsf{U}| \leq 2(1 - |\mathcal{A}|^{-1})$).

The exact locus of $(H(P), |P - \mathsf{U}|)$ among all the probability measures P defined on a finite set $\mathcal{A}$ (see Theorem 26), and the bounds in (348)–(350) are illustrated in Figure 2 for $|\mathcal{A}| = 4$ and $|\mathcal{A}| = 256$. For $|\mathcal{A}| = 4$, the lower bound in (350) is tighter than (347). For $|\mathcal{A}| = 256$, we only show (350) in Figure 2 as in this case (347) offers a very minor improvement in a small range. As the cardinality of the set $\mathcal{A}$ increases, the gap between the exact locus (shaded region) and the upper bound obtained from (348) and (349) (Curves (a) and (b), respectively) decreases, whereas the gap between the exact locus and the lower bound in (350) (Curve (c)) increases.

E. The Exponential Decay of the Probability of Non-Strongly Typical Sequences

The objective is to bound the function

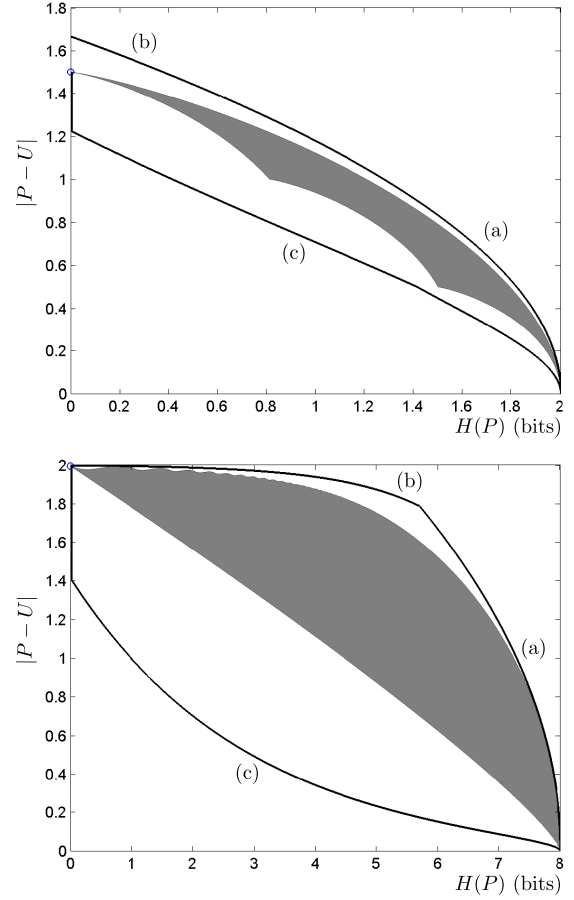$$L_\delta(Q) = \min_{P \notin \mathcal{T}_\delta(Q)} D(P \| Q) \qquad (351)$$

Fig. 2. The exact locus of $(H(P), |P - \mathsf{U}|)$ among all the probability measures P defined on a finite set $\mathcal{A}$, and bounds on $|P - \mathsf{U}|$ as a function of $H(P)$ for $|\mathcal{A}| = 4$ (left plot), and $|\mathcal{A}| = 256$ (right plot). The point $(H(P), |P - \mathsf{U}|) = (0, 2(1 - |\mathcal{A}|^{-1}))$ is depicted on the y-axis. In the two plots, Curves (a), (b) and (c) refer, respectively, to (348), (349) and (350); the exact locus (shaded region) refers to Theorem 26.

where the subset of probability measures on $(\mathcal{A}, \mathscr{F})$ which are δ-close to Q is given by

$$\mathcal{T}_\delta(Q) = \left\{ P : \forall a \in \mathcal{A}, |P(a) - Q(a)| \leq \delta Q(a) \right\}. \qquad (352)$$

Note that $(a_1, \ldots, a_n)$ is strongly δ-typical according to Q if its empirical distribution belongs to $\mathcal{T}_\delta(Q)$. According to Sanov's theorem (e.g. [20, Th. 11.4.1]), if the random variables are independent and distributed according to Q, then the probability that $(Y_1, \ldots, Y_n)$, is not δ-typical vanishes exponentially with exponent $L_\delta(Q)$.

To state the next result, we invoke the following notions from [75]. Given a probability measure Q, its *balance coefficient* is given by

$$\beta_Q = \inf_{\mathcal{F} \in \mathscr{F} : Q(\mathcal{F}) \geq \frac{1}{2}} Q(\mathcal{F}). \qquad (353)$$

The function $\phi: (0, \frac{1}{2}] \rightarrow [\frac{1}{2} \log e, \infty)$ is a monotonically decreasing and convex function, which is given by

$$\phi(p) = \begin{cases} \frac{1}{4(1-2p)} \log \left(\frac{1-p}{p} \right), & p \in \left(0, \frac{1}{2}\right), \\ \frac{1}{2} \log e, & p = \frac{1}{2}. \end{cases} \qquad (354)$$

Theorem 27: If $Q_{\min} > 0$, then

$$\phi(1 - \beta_Q) Q_{\min}^2 \delta^2 \le L_\delta(Q) \tag{355}$$

$$\le \log\left(1 + 2 Q_{\min}\delta^2\right) \tag{356}$$

where (356) holds if $\delta < \frac{1 - Q_{\max}}{Q_{\min}}$.

Proof: The following refinement of Pinsker's inequality (1) was derived in [75, Sec. 4]:

$$\phi(1 - \beta_Q)|P - Q|^2 \le D(P\|Q). \tag{357}$$

Note that if $Q_{\min} > 0$ then $\beta_Q \le 1 - Q_{\min} < 1$, and $\phi(1 - \beta_Q)$ is well defined and finite. If $P \notin \mathcal{T}_\delta(Q)$, the simple bound

$$|P - Q| > \delta Q_{\min} \tag{358}$$

together with (351) and (357) yields (355).

The upper bound (356) follows from (327) and the fact that if $\delta < \frac{1 - Q_{\max}}{Q_{\min}}$, then

$$\inf_{P \notin \mathcal{T}_\delta(Q)} |P - Q| = 2\delta Q_{\min}. \tag{359}$$

To verify (359), note that for every $P \notin \mathcal{T}_\delta(Q)$, there exists $a \in \mathcal{A}$ such that $|P(a) - Q(a)| > \delta Q(a)$, which implies that $|P - Q| > 2\delta Q(a) \ge 2\delta Q_{\min}$, thereby establishing $\ge$ in (359). To show equality, let $a_0 \in \mathcal{A}$ be such that $Q(a_0) = Q_{\min}$, and let $a_1 \ne a_0$; since by assumption $\delta < \frac{1 - Q_{\max}}{Q_{\min}}$, we have $Q(a_1) + \delta Q_{\min} \le Q_{\max} + \delta Q_{\min} < 1$. Let

$$P(a) = \begin{cases} (1 - \delta - \varepsilon) Q_{\min} & a = a_0 \\ Q(a_1) + (\delta + \varepsilon) Q_{\min} & a = a_1 \\ Q(a) & \text{otherwise} \end{cases} \tag{360}$$

for a sufficiently small $\varepsilon > 0$ so that (360) is a probability measure. Then, $P \notin \mathcal{T}_\delta(Q)$ and $|P - Q| = 2(\delta + \varepsilon) Q_{\min}$, which verifies the equality in (359) by letting $\varepsilon \downarrow 0$. ∎

Remark 37: If $\delta < \frac{1 - Q_{\max}}{Q_{\min}}$, the ratio between the upper and lower bounds in (356), satisfies

$$\frac{1}{Q_{\min}} \cdot \frac{\log e}{2\phi(1 - \beta_Q)} \cdot \frac{\log\left(1 + 2 Q_{\min}\delta^2\right)}{\frac{1}{2} Q_{\min}\delta^2 \log e} \le \frac{4}{Q_{\min}} \tag{361}$$

where (361) follows from the fact that its second and third factors are less than or equal to 1 and 4, respectively. Note that both bounds in (356) scale like δ^2 for $\delta \approx 0$.

VII. THE E_γ DIVERGENCE

A. Basic Properties

Generalizing the total variation distance, the E_γ divergence in (66) is an f-divergence whose utility in information theory has been exemplified in [17], [67]–[69], and [77]–[79].

In this subsection, we provide some basic properties of the E_γ divergence, which are essential to Sections VII-B–VII-D. The reader is referred to [68, Secs. 2.B and 2.C] for some additional basic properties of the E_γ divergence. We assume throughout this section that $\gamma \ge 1$.

Let $P \ll Q$. The E_γ divergence in (66) can be expressed in the form

$$E_\gamma(P\|Q)$$
$$= \mathbb{P}[\imath_{P\|Q}(X) > \log\gamma] - \gamma \, \mathbb{P}[\imath_{P\|Q}(Y) > \log\gamma] \tag{362}$$

$$= \max_{\mathcal{F} \in \mathscr{F}} \left(P(\mathcal{F}) - \gamma \, Q(\mathcal{F})\right) \tag{363}$$

where $X \sim P$ and $Y \sim Q$, and (363) follows from the Neyman-Pearson lemma.

Although the E_γ divergence generalizes the total variation distance, $E_\gamma(P\|Q) = 0$ for $\gamma > 1$ does not imply $P = Q$ since in that case (67) is not strictly convex at $t = 1$ (see Proposition 1). This is illustrated in the following example.

Example 11: Let $\gamma > 1$, and let P and Q be probability measures defined on $\mathcal{A} = \{0, 1\}$:

$$P(0) = \frac{1 + \gamma}{2\gamma}, \quad Q(0) = \frac{1}{\gamma}. \tag{364}$$

Since $\imath_{P\|Q}(x) = \log\gamma \, 1\{x = 0\} - \log 2 < \log\gamma$ for all $x \in \mathcal{A}$, (362) implies that $E_\gamma(P\|Q) = 0$.

The monotonicity of the E_γ divergence in $\gamma \in [1, \infty)$ holds since $f_{\gamma_1}(t) \le f_{\gamma_2}(t)$ for all $t > 0$ with $f_\gamma(t) = (t - \gamma)^+$ and $\gamma_1 \ge \gamma_2 \ge 1$. Therefore,

$$\frac{E_{\gamma_1}(P\|Q)}{E_{\gamma_2}(P\|Q)} \le 1. \tag{365}$$

Although Theorem 1b) does not apply in order to prove that 1 is the best constant in (365), we can verify it by defining P and Q on $\mathcal{A} = \{0, 1\}$ with $P(0) = \frac{1}{2}$ and $Q(0) = \varepsilon > 0$. This yields that if $\gamma_1 \ge \gamma_2 \ge 1$, then for all $\varepsilon \in (0, \frac{1}{2\gamma_1})$,

$$\frac{E_{\gamma_1}(P\|Q)}{E_{\gamma_2}(P\|Q)} = \frac{1 - 2\varepsilon\gamma_1}{1 - 2\varepsilon\gamma_2}, \tag{366}$$

yielding the optimality of the constant in the right side of (365) by letting $\varepsilon \downarrow 0$ in (366).

From (363), the following inequality holds: If $P \ll R \ll Q$, and $\gamma_1, \gamma_2 \ge 1$ then

$$E_{\gamma_1\gamma_2}(P\|Q) \le E_{\gamma_1}(P\|R) + \gamma_1 E_{\gamma_2}(R\|Q). \tag{367}$$

Letting $\gamma_1 = 1$ in (367) (see (68)) and $\gamma_2 = \gamma$ yield

$$E_\gamma(P\|Q) - E_\gamma(R\|Q) \le \tfrac{1}{2}|P - R|. \tag{368}$$

Generalizing the fact that $E_1(P\|Q) = \frac{1}{2}|P - Q|$, the following identity is a special case of [47, Corollary 2.3]:

$$\min_{R: P, Q \ll R} \left\{ E_\gamma(P\|R) + E_\gamma(Q\|R) \right\} = \left(1 - \gamma + \tfrac{1}{2}|P - Q|\right)^+ \tag{369}$$

while [68, eq. (21)] states that

$$\min_{R \ll P, Q} \left\{ E_\gamma(R\|P) + E_\gamma(R\|Q) \right\} \ge \left(1 - \gamma + \tfrac{\gamma}{2}|P - Q|\right)^+, \tag{370}$$

which implies, by taking $R = P$,

$$\left(1 - \gamma + \tfrac{\gamma}{2}|P - Q|\right)^+ \le E_\gamma(P\|Q). \tag{371}$$

We end this subsection with the following result.

Theorem 28: If $P \ll Q$ and $Y \sim Q$, then

$$\mathbb{E}\left[\left|\exp(\imath_{P\|Q}(Y)) - \gamma\right|\right] = 2E_\gamma(P\|Q) + \gamma - 1, \tag{372}$$

$$\mathbb{E}\left[\max\{\gamma, \exp(\imath_{P\|Q}(Y))\}\right] = \gamma + E_\gamma(P\|Q), \tag{373}$$

$$\mathbb{E}\left[\min\{\gamma, \exp(\imath_{P\|Q}(Y))\}\right] = 1 - E_\gamma(P\|Q). \tag{374}$$

Proof: The identity $|z| = 2(z)^+ - z$, for all $z \in \mathbb{R}$, and (66) are used to prove (372):

$$\mathbb{E}\left[\left|\exp\left(\iota_{P\|Q}(Y)\right) - \gamma\right|\right] \tag{375}$$

$$= 2\mathbb{E}\left[\left(\exp\left(\iota_{P\|Q}(Y)\right) - \gamma\right)^+\right] - \mathbb{E}\left[\exp\left(\iota_{P\|Q}(Y)\right) - \gamma\right]$$

$$= 2E_\gamma(P\|Q) + \gamma - 1. \tag{376}$$

Eqs. (373) and (374) follow from (372), and the identities

$$\max\{x_1, x_2\} = \tfrac{1}{2}\left[x_1 + x_2 + |x_1 - x_2|\right], \tag{377}$$

$$\min\{x_1, x_2\} = \tfrac{1}{2}\left[x_1 + x_2 - |x_1 - x_2|\right] \tag{378}$$

for all $x_1, x_2 \in \mathbb{R}$. ∎

Remark 38: In view of (68), it follows that (372) and (373) are specialized respectively to (206) and [49, eq. (20)] by letting $\gamma = 1$.

B. An Integral Representation of f-Divergences

In this subsection we show that

$$\left\{(E_\gamma(P\|Q), E_\gamma(Q\|P)), \gamma \geq 1\right\}$$

uniquely determines $D(P\|Q)$, $\mathcal{H}_\alpha(P\|Q)$, as well as any other f-divergence with twice differentiable f.

Proposition 3: Let $P \lll Q$, and let $f : (0, \infty) \to \mathbb{R}$ be convex and twice differentiable with $f(1) = 0$. Then,

$$D_f(P\|Q) = \int_1^\infty \left(f''(\gamma)E_\gamma(P\|Q) + \gamma^{-3}f''(\gamma^{-1})E_\gamma(Q\|P)\right)\mathrm{d}\gamma. \tag{379}$$

Proof: From [66, Th. 11], if $f : (0, \infty) \to \mathbb{R}$ is a convex function with $f(1) = 0$, then[14]

$$D_f(P\|Q) = \int_0^1 \mathcal{I}_p(P\|Q)\mathrm{d}\Gamma_f(p) \tag{380}$$

where Γ_f is the σ-finite measure defined on Borel subsets of $(0, 1)$ by

$$\Gamma_f((p_1, p_2]) = \int_{p_1}^{p_2} \frac{1}{p}\mathrm{d}g_f(p) \tag{381}$$

for the non-decreasing function

$$g_f(p) = -f'_+\left(\frac{1-p}{p}\right), \quad p \in (0, 1) \tag{382}$$

where f'_+ denotes the right derivative of f.

The DeGroot statistical information in (69) has the following operational role [31], which is used in this proof. Assume hypotheses H_0 and H_1 have a-priori probabilities p and $1-p$, respectively, and let P and Q be the conditional probability measures of an observation Y given H_0 or H_1. Then, $\mathcal{I}_p(P\|Q)$ is equal to the difference between the minimum error probabilities when the most likely *a-priori* hypothesis is selected, and when the most likely *a posteriori* hypothesis is selected. This measure therefore quantifies the value of the observations for the task of discriminating between the hypotheses. From the operational role of this measure, it follows that if $P \lll Q$

$$\mathcal{I}_p(P\|Q) = \mathcal{I}_{1-p}(Q\|P). \tag{383}$$

The E_γ divergence and DeGroot statistical information are related by

$$\mathcal{I}_p(P\|Q) = \begin{cases} pE_{\frac{1-p}{p}}(P\|Q), & p \in \left(0, \tfrac{1}{2}\right] \\ (1-p)E_{\frac{p}{1-p}}(Q\|P), & p \in \left[\tfrac{1}{2}, 1\right). \end{cases} \tag{384}$$

The expression for $p \in \left(0, \tfrac{1}{2}\right]$ follows from the fact that the functions that yield E_γ and $\mathcal{I}_p$ in (67) and (70), respectively, satisfy

$$\phi_p = pf_{\frac{1-p}{p}}. \tag{385}$$

The remainder of (384) follows in view of (383).

Specializing (380) to a twice differentiable f gives

$$D_f(P\|Q)$$

$$= \int_0^1 \mathcal{I}_p(P\|Q) \cdot \frac{1}{p^3}f''\left(\frac{1-p}{p}\right)\mathrm{d}p \tag{386}$$

$$= \int_0^{\frac{1}{2}} \mathcal{I}_p(P\|Q) \cdot \frac{1}{p^3}f''\left(\frac{1-p}{p}\right)\mathrm{d}p$$

$$+ \int_{\frac{1}{2}}^1 \mathcal{I}_p(P\|Q) \cdot \frac{1}{p^3}f''\left(\frac{1-p}{p}\right)\mathrm{d}p \tag{387}$$

$$= \int_0^{\frac{1}{2}} E_{\frac{1-p}{p}}(P\|Q) \cdot \frac{1}{p^2}f''\left(\frac{1-p}{p}\right)\mathrm{d}p$$

$$+ \int_{\frac{1}{2}}^1 E_{\frac{p}{1-p}}(Q\|P) \cdot \frac{1-p}{p^3}f''\left(\frac{1-p}{p}\right)\mathrm{d}p \tag{388}$$

$$= \int_1^\infty E_\gamma(P\|Q)f''(\gamma)\mathrm{d}\gamma + \int_0^1 \gamma E_{\gamma^{-1}}(Q\|P)f''(\gamma)\mathrm{d}\gamma \tag{389}$$

$$= \int_1^\infty \left[f''(\gamma)E_\gamma(P\|Q) + \gamma^{-3}f''(\gamma^{-1})E_\gamma(Q\|P)\right]\mathrm{d}\gamma \tag{390}$$

where (386) follows from (380)–(382); (387) follows by splitting the interval of integration into two parts; (388) follows from (384); (389) follows by the substitution $\gamma = \frac{1-p}{p}$, and (390) follows by changing the variable of integration $t = \frac{1}{\gamma}$ in the second integral in (389). ∎

Particularizing Proposition 3 to the most salient f-divergences we obtain (cf. [66, eqs. (84)–(86)] for alternative integral representations as a function of DeGroot statistical information)

$$D(P\|Q)$$

$$= \log e \int_1^\infty \left(\gamma^{-1}E_\gamma(P\|Q) + \gamma^{-2}E_\gamma(Q\|P)\right)\mathrm{d}\gamma, \tag{391}$$

$$\mathcal{H}_\alpha(P\|Q)$$

$$= \alpha \int_1^\infty \left(\gamma^{\alpha-2}E_\gamma(P\|Q) + \gamma^{-\alpha-1}E_\gamma(Q\|P)\right)\mathrm{d}\gamma, \tag{392}$$

and specializing (392) to $\alpha = 2$ yields

$$\chi^2(P\|Q) = 2\int_1^\infty \left(E_\gamma(P\|Q) + \gamma^{-3}E_\gamma(Q\|P)\right)\mathrm{d}\gamma. \tag{393}$$

Accordingly, bounds on the E_γ divergence, such as those presented in Section VII-C, directly translate into bounds on other important f-divergences.

Remark 39: Proposition 3 can be derived also from the integral representation of f-divergences in [18, Corollary 3.7].

[14]See also [74, Th. 1] for an earlier representation of f-divergence as an averaged DeGroot statistical information.

C. Extension of Pinsker's Inequality to E_γ Divergence

This subsection upper bounds E_γ divergence in terms of the relative entropy.

Theorem 29:

$$E_\gamma(P\|Q)\log\gamma \le D(P\|Q) + 2e^{-1}\log e, \quad (394)$$

$$2E_\gamma^2(P\|Q)\log e \le D(P\|Q). \quad (395)$$

Proof: The bound in (394) appears in [68, Proposition 13]. For $\gamma = 1$, (395) reduces to (1). Since E_γ is monotonically decreasing in γ, (395) also holds for $\gamma > 1$. ∎

For $\gamma = 1$, (395) becomes Pinsker inequality (1), for which there is no tighter constant. Moreover, in view of (68), for small $E_1(P\|Q)$, the minimum achievable $D(P\|Q)$ is indeed quadratic in $E_1(P\|Q)$ [39]. This 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 $D(P\|Q)$.

Theorem 30: For every $\gamma > 1$,

$$\sup \frac{E_\gamma(P\|Q)}{D(P\|Q)} = c_\gamma \quad (396)$$

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

$$c_\gamma = \frac{t_\gamma - \gamma}{r(t_\gamma)}, \quad (397)$$

$$t_\gamma = -\gamma W_{-1}\left(-\frac{1}{\gamma}e^{-\frac{1}{\gamma}}\right) \quad (398)$$

with r in (397) is given in (43), and W_{-1} in (398) denotes the secondary real branch of the Lambert W function [19].

Proof: The functions $f_\gamma(t) = (t - \gamma)^+$ and r (see (43)) satisfy the sufficient conditions of Theorem 1. Their ratio is

$$\kappa_\gamma(t) = \begin{cases} \frac{t-\gamma}{r(t)} & t \in [\gamma, \infty) \\ 0 & t \in (0, \gamma]. \end{cases} \quad (399)$$

For $t > \gamma$

$$\kappa_\gamma'(t) = \frac{\gamma\log t + (1 - t)\log e}{r^2(t)}. \quad (400)$$

Since $\gamma > 1$, it follows from (400) that there exists $t_\gamma \in (\gamma, \infty)$ such that κ_γ is monotonically increasing on $[\gamma, t_\gamma]$, and it is monotonically decreasing on $[t_\gamma, \infty)$. The value t_γ is the unique solution of the equation $\kappa_\gamma'(t) = 0$ in (γ, ∞). From (400), $t_\gamma \in (\gamma, \infty)$ solves the equation

$$\gamma\log t = (t - 1)\log e \quad (401)$$

which, after exponentiating both sides of (401) and making the substitution $x = -\frac{t}{\gamma}$, gives

$$xe^x = -\frac{1}{\gamma}e^{-\frac{1}{\gamma}}. \quad (402)$$

The trivial solution of (402) $x = -\frac{1}{\gamma}$ corresponds to $t = 1$, which is an improper solution of (401) since $t < \gamma$. The proper solution of (402) is its second real solution given by

$$x = W_{-1}\left(-\frac{1}{\gamma}e^{-\frac{1}{\gamma}}\right); \quad (403)$$

consequently, $t = -\gamma x$ and (403) give (398). In conclusion, for $t \ge 0$ and $\gamma > 1$,

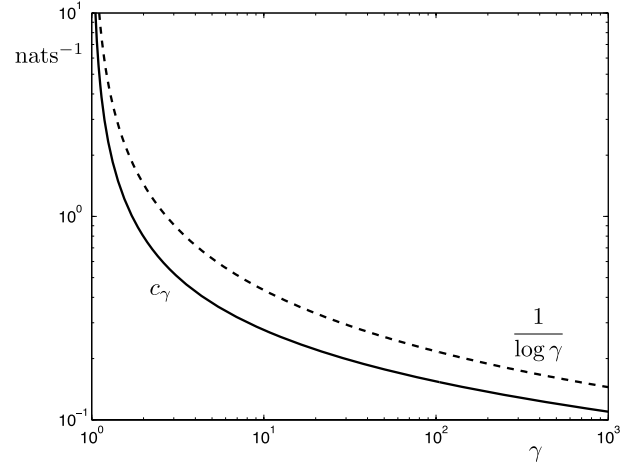$$0 \le \kappa_\gamma(t) \le \kappa_\gamma(t_\gamma) = c_\gamma \quad (404)$$

Fig. 3. The coefficient c_γ in (397) (solid line) compared to $\frac{1}{\log\gamma}$ (cf. (394)) (dashed line).

where the equality in (404) follows from (397), (399), and $t_\gamma > \gamma$. Theorem 1a) yields

$$E_\gamma(P\|Q) \le c_\gamma D(P\|Q). \quad (405)$$

To show (396), or in other words that there is no better constant in (405) than c_γ, it is enough to restrict to binary alphabets: Let $\mathcal{A} = \{0, 1\}$, $\varepsilon \in (0, 1)$, and $P_\varepsilon(0) = \varepsilon$, $Q_\varepsilon(0) = \frac{\varepsilon}{t_\gamma}$. A straightforward calculation (the details are provided in [93, (443)–(451) and Appendix H].) results in

$$\lim_{\varepsilon \to 0} \frac{E_\gamma(P_\varepsilon\|Q_\varepsilon)}{D(P_\varepsilon\|Q_\varepsilon)} = c_\gamma. \quad (406)$$

∎

Remark 40: The value of c_γ given in (397) can be approximated by

$$c_\gamma \approx \frac{\delta}{(\delta + \gamma)\log\left(\frac{\delta+\gamma}{e}\right) + \log e} \quad (407)$$

$$\delta = \frac{\alpha\gamma\log\gamma}{\log e}, \quad \alpha = 1.1791 \quad (408)$$

with a relative error of less than 1% for all $\gamma > 1$, and no more than 10^{-3} for $\gamma \ge 2$.

It can be verified that the bound in Theorem 30 is tighter than (394) since $c_\gamma < \frac{1}{\log\gamma}$ for $\gamma > 1$, and the additional positive summand $\frac{2\log e}{e\log\gamma}$ in the right side of (394) further loosens the bound (394) in comparison to (396). According to the approximation of c_γ in (407) and (408), we have for large values of γ

$$c_\gamma \approx \frac{1}{\log\left(\frac{\alpha\gamma\log\gamma}{e\log e}\right)}. \quad (409)$$

Remark 41: The impossibility of a general lower bound on $E_\gamma(P\|Q)$, for $\gamma > 1$, in terms of the relative entropy $D(P\|Q)$ is evident from Example 11.

Remark 42: The fact that $\{c_\gamma\}_{\gamma \ge 1}$ in (397) is monotonically decreasing in γ (see Figure 3) is consistent with (396) and the fact that the E_γ divergence is monotonically decreasing in γ.

Remark 43: The fact that the behavior of $D(P\|Q)$ for small $|P - Q|$ is quadratic rather than linear does not contradict Theorem 30 because $\lim_{\gamma \downarrow 1} c_\gamma = +\infty$ (see Figure 3).

In view of (363) and (396) we obtain

Corollary 4: If $P \ll Q$, $\gamma > 1$ and $\mathcal{F} \in \mathscr{F}$, then

$$P(\mathcal{F}) \leq \gamma\, Q(\mathcal{F}) + c_\gamma\, D(P\|Q). \tag{410}$$

Corollary 5: If $P \ll Q$ and $\gamma > 1$, then

$$E_\gamma(P\|Q)$$
$$\leq \min_{\lambda \in [0,1]} \left\{ \tfrac{\lambda}{2}|P - Q| + c_\gamma\, D\big((1 - \lambda)P + \lambda Q\|Q\big) \right\}. \tag{411}$$

Proof: For $\lambda \in [0, 1]$, let $R = (1 - \lambda)P + \lambda Q$. Then, we have for $\gamma > 1$,

$$E_\gamma(P\|Q) \leq \tfrac{1}{2}|P - R| + E_\gamma(R\|Q) \tag{412}$$
$$= \tfrac{\lambda}{2}|P - Q| + E_\gamma\big((1 - \lambda)P + \lambda Q\|Q\big) \tag{413}$$
$$\leq \tfrac{\lambda}{2}|P - Q| + c_\gamma\, D\big((1 - \lambda)P + \lambda Q\|Q\big) \tag{414}$$

where (412) is (368); and (414) follows from (396). ∎

Remark 44: Note that the upper bounds in the right sides of (371) and (405) follow from (411) by setting $\lambda = 1$ or $\lambda = 0$, respectively.

Remark 45: Further upper bounding the right side of (414) by invoking Pinsker's inequality and the convexity of relative entropy, followed by an optimization over the free parameter $\lambda \in [0, 1]$, does not lead to an improvement beyond the minimum of the bounds in (395) and (405).

D. Lower Bound on $\mathbb{F}_{P\|Q}$ as a Function of $D(P\|Q)$

The E_γ divergence proves to be instrumental in the proof of the following bound on the complementary relative information spectrum for positive arguments.

Theorem 31: If $P \ll Q$, $P \neq Q$, and $\beta > 1$, then

$$\frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{D(P\|Q)} \leq u(\beta) \triangleq \min_{\gamma \in (1, \beta)} \left(\frac{\beta c_\gamma}{\beta - \gamma} \right), \tag{415}$$

where c_γ is given in (397). Furthermore, the function $u \colon (1, \infty) \to \mathbb{R}^+$ is monotonically decreasing with

$$u(\beta) \leq \frac{2}{\log\left(\frac{\beta}{2e}\right)}, \qquad \forall \beta > 2e. \tag{416}$$

Proof: For $\beta > 1$, denote the event

$$\mathcal{F}_\beta \triangleq \{x \in \mathcal{A} \colon \imath_{P\|Q}(x) > \log \beta\} \tag{417}$$

which satisfies

$$P(\mathcal{F}_\beta) > \beta\, Q(\mathcal{F}_\beta). \tag{418}$$

Then,

$$1 - \mathbb{F}_{P\|Q}(\log \beta) = P(\mathcal{F}_\beta) \tag{419}$$
$$\leq \inf_{\gamma \in (1, \beta)} \frac{P(\mathcal{F}_\beta) - \gamma\, Q(\mathcal{F}_\beta)}{1 - \frac{\gamma}{\beta}} \tag{420}$$
$$\leq \inf_{\gamma \in (1, \beta)} \frac{\beta\, E_\gamma(P\|Q)}{\beta - \gamma} \tag{421}$$
$$\leq \inf_{\gamma \in (1, \beta)} \left(\frac{\beta c_\gamma}{\beta - \gamma} \right) D(P\|Q) \tag{422}$$

where (419) holds by Definition 2; (420) follows from (418); (421) is satisfied by (363), and (422) is due to (396). Note that the infimum in (422) is attained because c_γ is continuous and for $\beta > 1$, $\frac{\beta c_\gamma}{\beta - \gamma}$ tends to $+\infty$ at both extremes of the interval $(1, \beta)$. The monotonicity of $u(\beta)$ and the bound in (416) are proved in Appendix H. ∎

VIII. RÉNYI DIVERGENCE

The Rényi divergence (Definition 4) admits a variational representation in terms of the relative entropy [94, Th. 1]. Let $P_1 \ll P_0$ then, for $\alpha > 0$,

$$(1 - \alpha)D_\alpha(P_1\|P_0)$$
$$= \min_{P \ll P_1} \{\alpha D(P\|P_1) + (1 - \alpha)D(P\|P_0)\}. \tag{423}$$

In this section, integral expressions for the Rényi divergence are derived in terms of the relative information spectrum (Definition 2). These expressions are used to obtain bounds on the Rényi divergence as a function of the variational distance under the assumption of bounded relative information.

A. Expressions in Terms of the Relative Information Spectrum

To state the results in this section, it is convenient to introduce $\zeta_\alpha \colon (0, \infty) \to [0, \infty)$

$$\zeta_\alpha(\beta) = \beta^{\alpha-2}\big(1 - \mathbb{F}_{P\|Q}(\log \beta)\big). \tag{424}$$

The Rényi divergence admits the following representation in terms of the relative information spectrum and the relative information bounds $(\beta_1, \beta_2) \in [0, 1]^2$ in (137)–(138).

Theorem 32: Let $P \ll Q$.
- If $\beta_1 > 0$ and $\alpha \in (0, 1) \cup (1, \infty)$, then

$$D_\alpha(P\|Q)$$
$$= \tfrac{1}{\alpha-1} \log\left(\beta_1^{1-\alpha} + (1 - \alpha)\int_{\beta_2}^{\beta_1^{-1}} \big(\beta^{\alpha-2} - \zeta_\alpha(\beta)\big)\mathrm{d}\beta \right). \tag{425}$$

- If $\beta_1 = 0$ and $\alpha \in (0, 1)$, then

$$D_\alpha(P\|Q)$$
$$= \frac{1}{\alpha - 1} \log\left((1 - \alpha)\int_{\beta_2}^{\infty} \big(\beta^{\alpha-2} - \zeta_\alpha(\beta)\big)\mathrm{d}\beta \right). \tag{426}$$

- If $\alpha \in (1, \infty)$, then

$$D_\alpha(P\|Q)$$
$$= \frac{1}{\alpha - 1} \log\left((\alpha - 1)\int_0^\infty \zeta_\alpha(\beta)\mathrm{d}\beta \right) \tag{427}$$
$$= \frac{1}{\alpha - 1} \log\left(\beta_2^{\alpha-1} + (\alpha - 1)\int_{\beta_2}^\infty \zeta_\alpha(\beta)\mathrm{d}\beta \right). \tag{428}$$

Proof: If $\alpha > 1$, (76) implies that $D_\alpha(P\|Q)$ is given by

$$\tfrac{1}{\alpha-1} \log\big(\mathbb{E}\big[\exp\big((\alpha - 1)\imath_{P\|Q}(X)\big)\big]\big)$$
$$= \tfrac{1}{\alpha-1} \log\left(\int_0^\infty \mathbb{P}\big[\exp\big((\alpha - 1)\imath_{P\|Q}(X)\big) > t\big]\,\mathrm{d}t \right) \tag{429}$$
$$= \tfrac{1}{\alpha-1} \log\left(\int_0^\infty \mathbb{P}\left[\imath_{P\|Q}(X) > \frac{\log t}{\alpha - 1}\right]\mathrm{d}t \right) \tag{430}$$

where (429) follows from (221) for an arbitrary non-negative random variable V, and we use $\alpha > 1$ to write (430). Then, (427) holds by the definition of the relative information spectrum in (27) and by changing the integration variable $t = \beta^{\alpha-1}$. If $\beta_1 > 0$, the integrand in the right side of (427) is zero in $[\beta_1^{-1}, \infty)$ and the expression in (425) is readily verified (for $\alpha > 1$). More generally (without requiring $\beta_1 > 0$), we split the integral in the right side of (427) into $[0, \beta_2) \cup [\beta_2, \infty)$, and (428) follows since the integral over the leftmost interval is $\beta_2^{\alpha-1}$ considering that $\mathbb{F}_{P\|Q}(\log \beta) = 0$ therein.

If $\alpha \in (0,1)$, we write $D_\alpha(P\|Q)$ as

$$\frac{1}{\alpha-1} \log\Big(\mathbb{E}\big[\exp\big((\alpha-1)\imath_{P\|Q}(X)\big)\big]\Big)$$

$$= \frac{1}{\alpha-1} \log\left(\int_0^\infty \mathbb{P}\big[\exp\big((\alpha-1)\imath_{P\|Q}(X)\big) \geq t\big]\,dt\right) \quad (431)$$

$$= \frac{1}{\alpha-1} \log\left(\int_0^\infty \mathbb{P}\Big[\imath_{P\|Q}(X) \leq \frac{\log t}{\alpha-1}\Big]\,dt\right) \quad (432)$$

$$= \frac{1}{\alpha-1} \log\left(\int_\infty^0 \mathbb{P}\big[\imath_{P\|Q}(X) \leq \log \beta\big](\alpha-1)\beta^{\alpha-2}d\beta\right) \quad (433)$$

$$= \frac{1}{\alpha-1} \log\left((1-\alpha)\int_{\beta_2}^\infty \beta^{\alpha-2}\mathbb{F}_{P\|Q}\big(\log \beta\big)d\beta\right) \quad (433)$$

which is the expression in (426). If $\beta_1 > 0$, then we can further split the integral in the right side of (433) into the intervals $[\beta_2, \beta_1^{-1}) \cup [\beta_1^{-1}, \infty)$. Over the rightmost interval, $\mathbb{F}_{P\|Q}(\log \beta) = 1$ and the integral is seen to be $\beta_1^{1-\alpha}$, thereby verifying (425) for $\alpha \in (0,1)$. ∎

The close relationship between the Rényi and Hellinger divergences in (80) results in the following integral representations for the Hellinger divergence.

Corollary 6: Let $P \ll Q$.
• If $\beta_1 > 0$ and $\alpha \in (0,1) \cup (1,\infty)$, then

$$\mathscr{H}_\alpha(P\|Q) = \frac{\beta_1^{1-\alpha}-1}{\alpha-1} - \int_{\beta_2}^{\beta_1^{-1}} \big(\beta^{\alpha-2} - \zeta_\alpha(\beta)\big)d\beta. \quad (434)$$

• If $\beta_1 = 0$ and $\alpha \in (0,1)$, then

$$\mathscr{H}_\alpha(P\|Q) = \frac{1}{1-\alpha} - \int_{\beta_2}^\infty \big(\beta^{\alpha-2} - \zeta_\alpha(\beta)\big)d\beta. \quad (435)$$

• If $\alpha \in (1,\infty)$, then

$$\mathscr{H}_\alpha(P\|Q) = \int_0^\infty \zeta_\alpha(\beta)d\beta - \frac{1}{\alpha-1} \quad (436)$$

$$= \frac{\beta_2^{\alpha-1}-1}{\alpha-1} + \int_{\beta_2}^\infty \zeta_\alpha(\beta)d\beta. \quad (437)$$

Proof: Combining (80) with (425), (426), (427), (428) yields (434)–(437), respectively. ∎

Particularizing (424), (427) and (436) to $\alpha = 2$, we obtain

$$D_2(P\|Q) = \log\left(\int_0^\infty \big(1 - \mathbb{F}_{P\|Q}(\log \beta)\big)\,d\beta\right), \quad (438)$$

$$\chi^2(P\|Q) = \int_0^\infty \big(1 - \mathbb{F}_{P\|Q}(\log \beta)\big)\,d\beta - 1 \quad (439)$$

$$= \int_1^\infty \big(1 - \mathbb{F}_{P\|Q}(\log \beta)\big)\,d\beta$$

$$- \int_0^\infty \mathbb{F}_{P\|Q}(\log \beta)\,d\beta. \quad (440)$$

Note the resemblance of the integral expressions in (219) and (440) for $D(P\|Q)$ and $\chi^2(P\|Q)$, respectively.

We conclude this subsection by proving three properties of the Hellinger divergence as a function of its order. The first two monotonicity properties are analogous to [37, Ths. 3 and 16] for the Rényi divergence; these monotonicity properties have been originally stated in [65, Proposition 2.7], though the following alternative proof is more transparent.

Theorem 33: The Hellinger divergence satisfies the following properties:

a) $\mathscr{H}_\alpha(P\|Q)$ is monotonically increasing in $\alpha \in (0,\infty)$;
b) $\big(\frac{1}{\alpha}-1\big)\mathscr{H}_\alpha(P\|Q)$ is monotonically decreasing in $\alpha \in (0,1)$;
c) $\frac{1}{\alpha}\mathscr{H}_\alpha(P\|Q)$ is log-convex in $\alpha \in (0,\infty)$, which implies that for every $\alpha, \beta > 0$

$$\mathscr{H}_{\frac{\alpha+\beta}{2}}^2(P\|Q) \leq \frac{(\alpha+\beta)^2}{4\alpha\beta}\mathscr{H}_\alpha(P\|Q)\mathscr{H}_\beta(P\|Q). \quad (441)$$

Proof:
a) From (39) and (52), we have

$$\mathscr{H}_\alpha(P\|Q) = D_{f_\alpha}(P\|Q) \quad (442)$$

with

$$f_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha-1}, \quad t > 0, \quad (443)$$

whose derivative is

$$\frac{\partial}{\partial\alpha} f_\alpha(t) = \frac{tr(t^{\alpha-1})}{(\alpha-1)^2 \log e} > 0 \quad (444)$$

where the function $r \colon (0,\infty) \to \mathbb{R}$ is defined in (43). Since it is strictly positive except at $t = 1$, f_α is monotonically increasing in $\alpha \in (0,\infty)$. Hence, Part a) follows from (442).

b) From (442), for $\alpha \in (0,1)$, we have

$$\Big(\frac{1}{\alpha}-1\Big)\mathscr{H}_\alpha(P\|Q) = D_{g_\alpha}(P\|Q) \quad (445)$$

where $g_\alpha \colon (0,\infty) \to \mathbb{R}$ is the convex function

$$g_\alpha(t) = t - 1 - \frac{t^\alpha-1}{\alpha}, \quad t > 0. \quad (446)$$

with derivative

$$\frac{\partial}{\partial\alpha} g_\alpha(t) = -\frac{r(t^\alpha)}{\alpha^2 \log e} < 0, \quad (447)$$

so g_α is monotonically decreasing in $\alpha \in (0,1)$. Hence, Part b) follows from (445).

c) To prove the log-convexity of $\frac{1}{\alpha}\mathscr{H}_\alpha(P\|Q)$ in $\alpha \in (0,\infty)$, we rely on [95, Th. 2.1] which states that if W is a non-negative random variable, then λ_α in (256) is log-convex in α. The claim now follows from (256) by setting $W = \frac{dP}{dQ}(Y)$ with $Y \sim Q$, which yields that $\lambda_\alpha = \frac{1}{\alpha}\mathscr{H}_\alpha(P\|Q)\log e$ for $\alpha \in (0,\infty)$. ∎

B. Bounds as a Function of the Total Variation Distance

Just as with Pinsker's inequality, for any $\varepsilon \in (0, 2]$, the minimum value of $D_\alpha(P\|Q)$ compatible with $|P - Q| \geq \varepsilon$, is achieved with distributions on a binary alphabet [91, Proposition 1]:

$$\min_{P, Q:\, |P-Q|\geq\varepsilon} D_\alpha(P\|Q) = \min_{p, q:\, |p-q|\geq\frac{\varepsilon}{2}} d_\alpha(p\|q) \quad (448)$$

where the binary order-α Rényi divergence is defined as

$$d_\alpha(p\|q)$$
$$\triangleq \begin{cases} \frac{1}{\alpha-1}\log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha(1-q)^{1-\alpha}\Big), & \text{if } \alpha \neq 1 \\ p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, & \text{if } \alpha = 1. \end{cases}$$
$$(449)$$

We proceed to use Theorem 32 to get an upper bound on $D_\alpha(P\|Q)$ expressed in terms of $|P - Q|$.

Theorem 34: If $\beta_1 \in (0, 1)$ and $\alpha \in (0, 1) \cup (1, \infty)$, then

$$D_\alpha(P\|Q) \leq \frac{1}{\alpha-1}\log\left(1 + \frac{|P-Q|}{2}\frac{\beta_1^{1-\alpha}-1}{1-\beta_1}\right). \quad (450)$$

Proof: Regardless of whether $\alpha < 1$ or $\alpha > 1$, we can only get an upper bound if, in view of (424), in the integral in (425) we drop the interval $[\beta_2, 1]$:

$$D_\alpha(P\|Q)$$
$$\leq \frac{1}{\alpha-1}\log\left(\beta_1^{1-\alpha} + (1-\alpha)\int_1^{\beta_1^{-1}} \beta^{\alpha-2}\mathbb{F}_{P\|Q}(\log\beta)\mathrm{d}\beta\right)$$
$$= \frac{1}{\alpha-1}\log\left(1 - (1-\alpha)\int_1^{\beta_1^{-1}} \beta^{\alpha-2}\big(1 - \mathbb{F}_{P\|Q}(\log\beta)\big)\mathrm{d}\beta\right)$$
$$= \frac{1}{\alpha-1}\log\big(1 + (\alpha-1)\delta\mathbb{E}[W^\alpha]\big) \quad (451)$$

where (451) holds with $\delta = \frac{1}{2}|P - Q|$ and $W \sim p_1$ where p_1 is the probability density function supported on $[1, \beta_1^{-1}]$:

$$p_1(\beta) = \frac{1}{\delta\beta^2}\big(1 - \mathbb{F}_{P\|Q}(\log\beta)\big). \quad (452)$$

Note that p_1 is indeed a probability density function due to (214). In order to proceed, we derive an upper bound on $\mathbb{E}[W^\alpha]$ expressed in terms of $|P - Q|$ by invoking Lemma 3 in Appendix I. To that end, denote the monotonically increasing and non-negative function $g(x) = x^\alpha$ for $x \geq 0$, and let p_2 be the probability density function supported on $[1, \beta_1^{-1}]$:

$$p_2(\beta) = \frac{1}{1-\beta_1}\frac{1}{\beta^2}. \quad (453)$$

Note that, on their support, $\beta^2 p_1(\beta)$ is monotonically decreasing while $\beta^2 p_2(\beta)$ is constant. Therefore, we can apply Lemma 3 to $W \sim p_1$ and $V \sim p_2$ to obtain

$$\mathbb{E}[W^\alpha] \leq \mathbb{E}[V^\alpha] = \frac{\beta_1^{1-\alpha}-1}{(1-\beta_1)(\alpha-1)}. \quad (454)$$

which gives the desired result upon substituting in (451). ∎

Corollary 7: If $\beta_1 \in (0, 1)$ and $\alpha \in (0, 1) \cup (1, \infty)$, then

$$\mathscr{H}_\alpha(P\|Q) \leq \frac{\beta_1^{1-\alpha}-1}{2(\alpha-1)(1-\beta_1)}\cdot|P - Q|. \quad (455)$$

Proof: Combining (80) and (450) yields (455). ∎

Particularizing (455) to $\alpha = 2$ yields

$$\chi^2(P\|Q) \leq \tfrac{1}{2}\beta_1^{-1}|P - Q| \quad (456)$$

which improves the bound in (159) if either $\beta_1 \leq \frac{1}{2}$ or $\beta_2 = 0$.

The combination of (1), (4) and (456) yields the following bound:

Corollary 8: If $\beta_1 > 0$, then

$$\chi^2(P\|Q) \leq \frac{1}{\beta_1}\sqrt{\min\left\{\frac{D(P\|Q)}{2\log e}, 1 - \exp\big(-D(P\|Q)\big)\right\}}.$$
$$(457)$$

Example 12: Let P, Q be defined on $\mathcal{A} = \{0, 1\}$ with $P(0) = Q(1) = \frac{1}{100}$, which implies that $\beta_1 = \frac{1}{99}$. Then $\chi^2(P\|Q) = 97.01$, and the bound in (457) is equal to 98.45 in contrast to the upper bound in (169) whose value is 121.17.

Remark 46: By letting $\alpha \to \infty$ in (450), we obtain $D_\infty(P\|Q) \leq \log\frac{1}{\beta_1}$, which shows that the bound in (450) is asymptotically tight (cf. (79)).

Remark 47: By letting $\alpha \to 1$ in (450), we get (307). Therefore, Theorem 34 generalizes [109, Th. 7].

Remark 48: By letting $\alpha \to 0$, it follows from (450) that

$$D_0(P\|Q) \leq \log\left(\frac{1}{1 - \frac{1}{2}|P - Q|}\right), \quad (458)$$

a bound which, in view of (77), is achieved with equality in the case of a finite alphabet with

$$P(a) = \begin{cases} \frac{Q(a)}{1-\delta}, & a \in \mathcal{F} \\ 0, & a \in \mathcal{F}^c \end{cases} \quad (459)$$

with the event $\mathcal{F}$ selected to satisfy $Q(\mathcal{F}) = 1 - \delta$.

Remark 49: Another upper bound on the Rényi divergence can be obtained by the simpler bound

$$\mathbb{E}[W^\alpha] \leq \beta_1^{-\alpha}, \quad (460)$$

which holds because $W \in [1, \beta_1^{-1}]$. Combining (451) and (460) yields

$$D_\alpha(P\|Q) \leq \frac{1}{\alpha-1}\log\big(1 + (\alpha-1)\delta\beta_1^{-\alpha}\big). \quad (461)$$

Note that, in the limit $\alpha \to 0$, the bounds in (450) and (461) coincide and are equal to the tight bound $-\log(1 - \delta)$.

Remark 50: Alternatively, we have the bound

$$D_\alpha(P\|Q)$$
$$\leq \frac{1}{\alpha-1}\log\left((1-\delta)^{1-\alpha} + \delta(\alpha-1)\int_{1-\delta}^{\frac{1}{\beta_1}}\frac{\beta^{\alpha-1}}{\beta-1}\mathrm{d}\beta\right)$$
$$(462)$$

obtained from (451) and

$$\mathbb{E}[W^\alpha] \leq \frac{(1-\delta)^{1-\alpha}-1}{\delta(\alpha-1)} + \int_{1-\delta}^{\frac{1}{\beta_1}}\frac{\beta^{\alpha-1}}{\beta-1}\mathrm{d}\beta. \quad (463)$$

which holds since, in view of (452) and (236),

$$p_1(\beta) \le \begin{cases} \frac{1}{\delta\beta^2}, & \beta \in \left[0, \frac{1}{1-\delta}\right] \\ \frac{1}{\beta(\beta-1)}, & \beta \in \left[\frac{1}{1-\delta}, \beta_1^{-1}\right] \\ 0 & \text{otherwise.} \end{cases} \quad (464)$$

The upper bounds in (450) and (462) asymptotically coincide in the limit where $\alpha \to \infty$, giving the common limit of $\log\left(\frac{1}{\beta_1}\right)$ which is a tight upper bound (cf. Remark 46).

C. Bounds as a Function of the Relative Entropy

In this section, we provide upper and lower bounds on the Rényi divergence $D_\alpha(P\|Q)$, for an arbitrary order $\alpha \in (0,1) \cup (1,\infty)$, expressed in terms of the relative entropy $D(P\|Q)$ and β_1, β_2.

Theorem 35: Let $(\beta_1, \beta_2) \in [0,1)^2$, $\alpha \in (0,1) \cup (1,\infty)$, and $u_\alpha : [0,\infty] \to [0,\infty]$ be

$$u_\alpha = \frac{\alpha - 1}{\kappa_\alpha(t)} \quad (465)$$

with κ_α defined in (164).

a) If $\alpha \in (0,1)$, then

$$\frac{1}{\alpha-1} \log\left(1 + u_\alpha(\beta_1^{-1})D(P\|Q)\right)$$
$$\le D_\alpha(P\|Q) \quad (466)$$
$$\le \min\left\{D(P\|Q), \frac{1}{\alpha-1}\log\left(1 + u_\alpha(\beta_2)D(P\|Q)\right)^+\right\}. \quad (467)$$

b) If $\alpha \in (1,\infty)$, then

$$\max\left\{D(P\|Q), \frac{1}{\alpha-1}\log\left(1 + u_\alpha(\beta_2)D(P\|Q)\right)\right\}$$
$$\le D_\alpha(P\|Q) \quad (468)$$
$$\le \min\left\{\log\frac{1}{\beta_1}, \frac{1}{\alpha-1}\log\left(1 + u_\alpha(\beta_1^{-1})D(P\|Q)\right)\right\}. \quad (469)$$

c) Furthermore, if $\alpha \in (0,1) \cup (1,\infty)$, then

$$D_\alpha(P\|Q) \le \frac{1}{\alpha-1}\log\left(1 + \frac{\bar\delta(\beta_1^{1-\alpha}-1)}{1-\beta_1}\right) \quad (470)$$

where

$$\bar\delta^2 = \min\left\{\frac{D(P\|Q)}{2\log e}, 1 - \exp(-D(P\|Q))\right\}. \quad (471)$$

Proof: Parts a) and b) follow from Theorem 8 in view of (80), from the fact that $D_\alpha(P\|Q)$ is monotonically increasing in $\alpha > 0$, and from $D_\infty(P\|Q) = \log\frac{1}{\beta_1}$.

Part c) follows from Theorem 34 replacing $\delta \triangleq \frac{1}{2}|P - Q|$ by its upper bound $\bar\delta$ obtained from (1) and (4). ∎

The next three remarks address the tightness of the bounds (466)–(467) and (468)–(469).

Remark 51: The constants $u_\alpha(\beta_1^{-1})$ and $u_\alpha(\beta_2)$ in (466)–(467) and (468)–(469) are the best possible among all probability measures P, Q with given $(\beta_1, \beta_2) \in [0,1)^2$. This follows from (80), and in view of the tightness of the constants in Theorem 8 (Remark 10).

Remark 52: Let $P, Q = Q_\varepsilon$ be defined on a binary alphabet with $P(0) = \frac{1}{2}$ and $Q(0) = \frac{1}{2} - \varepsilon$. Then, it is easy to
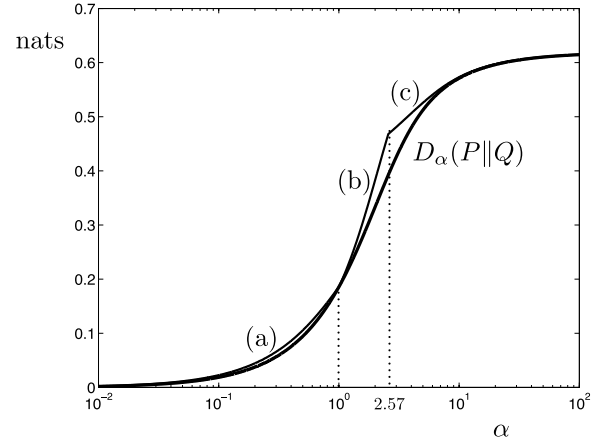

Fig. 4. The Rényi divergence $D_\alpha(P\|Q)$ for $\mathcal{A} = \{0,1\}$ with $P(0) = Q(1) = 0.65$, compared to the tightest upper bound in Theorem 35: (a) (467) for $\alpha \in (0,1)$; (b): (469) for $\alpha \in [1, 2.57]$; (c): (470) for $\alpha > 2.57$.

verify that the ratio of the upper to lower bounds in Parts a) and b) converges to 1 as $\varepsilon \to 0$.

Remark 53: Let P and Q be defined on a binary alphabet with $P(0) = \frac{1}{2}$, $Q(0) = \varepsilon \in (0,1)$. Then, in the limit $\varepsilon \to 0$, the ratio of $D_\alpha(P\|Q)$ to the left side of (466) is equal to $\frac{\alpha}{\log_2\left(\frac{2}{2-\alpha}\right)} \in (1, \log_e(4))$ for $\alpha \in (0,1)$. Moreover, if $\varepsilon \to 0$, the ratio of $D_\alpha(P\|Q)$ and the right side of (469) tends to 1 for $\alpha \in (1,\infty)$. The reader is referred to [93, (518)–(524)] for the details of the proofs.

Example 13: Figure 4 illustrates the upper bounds on $D_\alpha(P\|Q)$ in Theorem 35 for binary alphabets.

Remark 54: [40, Proposition 11] shows an upper bound on $D_\alpha(P\|Q)$ for $\alpha \in [1, \frac{5}{4}]$, which is expressed in terms of $D(P\|Q)$ and the finite cardinalities of the alphabets over which P and Q are defined. Although the bound in [40, eq. (9)] is not tight, it leads to a strong converse for a certain class of discrete memoryless networks.

IX. SUMMARY

Since many distance measures of interest fall under the common paradigm of an f-divergence, it is not surprising that bounds on the ratios of various f-divergences are useful in many instances such as proving convergence of probability measures according to various metrics, analysis of rates of convergence and concentration of measure bounds [13], [41], [75], [81], [88], [111], hypothesis testing [31], testing goodness of fit [50], [84], minimax risk in estimation and modeling [47], [51], [85], [107], strong data processing inequality constants and maximal correlation [1], [80], [83], transportation-cost inequalities [13], [73], [81], [82], contiguity [64], [65], etc.

While the derivation of f-divergence inequalities has received considerable attention in the literature, the proof techniques have been tailored to the specific instances. In contrast, we have proposed several systematic approaches to the derivation of f-divergence inequalities. Introduced in Section III-A, functional domination emerges as a basic tool to obtain f-divergence inequalities. Another basic tool that capitalizes on many cases of interest (including the finite

alphabet one) is introduced in Section IV-B, where not only one of the distributions is absolutely continuous with respect to the other but their relative information is almost surely bounded.

Section V-D illustrates the use of moment inequalities and the log-convexity property, while the utility of Lipschitz constraints in deriving bounds is highlighted in Section VI-B.

In addition, new f-divergence inequalities (frequently with optimal constants) arise from:

- integral representation of f-divergences, expressed in terms of the E_γ divergence (Section VII-B);
- extension of Pinsker's inequality to E_γ divergence (Section VII-C);
- a relation between the relative information and the relative entropy (Section VII-D);
- exact expressions of Rényi divergence in terms of the relative information spectrum (Section VIII-A);
- the exact locus of the entropy and the variational distance from the equiprobable probability mass function (Section VI-D).

APPENDIX A
COMPLETION OF THE PROOF OF THEOREM 6

Lemma 2: The function $\kappa : (0, \infty) \to (0, \infty)$ which is the continuous extension of the function in (160) with $\kappa(1) = 1$ is strictly monotonically increasing.

Proof: From (160) and (162),

$$\kappa'(t) = \frac{(t-1)^2 \log^2 e - t \log^2 t}{t g^2(t)}, \qquad (472)$$

if $t \in (0, 1) \cup (1, \infty)$, while $\kappa'(1) = \frac{1}{3}$. To show

$$(t-1)^2 \log^2 e - t \log^2 t > 0, \quad \forall t \in (0, 1) \cup (1, \infty). \quad (473)$$

we substitute $t = \exp(x)$ to obtain that if $x \neq 0$, then

$$s(x) = \exp(2x) - (2 + x^2) \exp(x) + 1 > 0, \qquad (474)$$

which holds since $s(0) = 0$ and the derivative

$$s'(x) = 2 \exp(x) \left[\exp(x) - \left(1 + x + \frac{x^2}{2} \right) \right] \qquad (475)$$

is negative on $(-\infty, 0)$ and positive on $(0, \infty)$. ∎

APPENDIX B
PROOF OF THE MONOTONICITY OF κ_α IN (164)

To show that the function $\kappa_\alpha : [0, \infty] \to [0, \infty]$ in (164) is monotonically increasing on $[0, \infty]$ if $\alpha \in (0, 1)$ and monotonically decreasing on $[0, \infty]$ if $\alpha \in (1, \infty)$ it is sufficient to show that

$$\frac{d}{dt} \left(\frac{r(t)}{1 - t^\alpha + \alpha(t-1)} \right) > 0 \qquad (476)$$

for $(\alpha, t) \in \mathcal{F} = ((0, 1) \cup (1, \infty))^2$. From (43), straightforward calculus gives

$$[1 - t^\alpha + \alpha(t-1)]^2 \frac{d}{dt} \left(\frac{r(t)}{1 - t^\alpha + \alpha(t-1)} \right) \qquad (477)$$

$$= (1 - \alpha)(1 - t^\alpha) \log t - \alpha(1 - t)(1 - t^{\alpha-1}) \log e \qquad (478)$$

$$\triangleq g_\alpha(t) \qquad (479)$$

so the desired result will follow upon showing

$$g_\alpha(t) > 0, \quad (\alpha, t) \in \mathcal{F}. \qquad (480)$$

The details of the proof of (480) are omitted, and we refer the reader to a proof in [93, (534)–(565)].

APPENDIX C
PROOF OF (179)

To verify that (179) follows from (171), fix arbitrarily small $\varepsilon > 0$ and $\rho > 0$. Consider the partition $\mathcal{A} = \mathcal{A}_1^{(n)} \cup \mathcal{A}_2^{(n)} \cup \mathcal{A}_3^{(n)}$ with

$$\mathcal{A}_1^{(n)} \triangleq \left\{ a \in \mathcal{A} : \frac{dP_n}{dQ}(a) \in [0, 1-\rho] \right\}, \qquad (481)$$

$$\mathcal{A}_2^{(n)} \triangleq \left\{ a \in \mathcal{A} : \frac{dP_n}{dQ}(a) \in (1-\rho, 1+\varepsilon] \right\}, \qquad (482)$$

$$\mathcal{A}_3^{(n)} \triangleq \left\{ a \in \mathcal{A} : \frac{dP_n}{dQ}(a) \in (1+\varepsilon, \infty) \right\}, \qquad (483)$$

then

$$I_1^{(n)} + I_2^{(n)} + I_3^{(n)} = 1 \qquad (484)$$

where

$$I_j^{(n)} \triangleq \int_{\mathcal{A}_j^{(n)}} \frac{dP_n}{dQ}(a) dQ(a). \qquad (485)$$

From the assumption in (171), $I_3^{(n)} \to 0$ when $n \to \infty$ since for all sufficiently large n

$$Q(\mathcal{A}_3^{(n)}) = 0. \qquad (486)$$

Let

$$d_n \triangleq Q(\mathcal{A}_1^{(n)}) \qquad (487)$$

then, from (486), for all sufficiently large n

$$1 - d_n = Q(\mathcal{A}_2^{(n)}). \qquad (488)$$

Consequently, from (481), (482), (484), (486), (487) and (488), it follows that for all sufficiently large n,

$$1 = I_1^{(n)} + I_2^{(n)} \qquad (489)$$

$$\leq d_n(1-\rho) + (1-d_n)(1+\varepsilon) \triangleq \mu_n. \qquad (490)$$

If $\liminf d_n = 0$ for an arbitrarily small $\rho > 0$ then (179) holds by the definition in (487). Assuming otherwise, namely,

$$\liminf d_n = \theta \in (0, 1) \qquad (491)$$

leads to the following contradiction:

$$1 \leq \liminf \mu_n \qquad (492)$$

$$\leq (1-\rho) \liminf d_n + (1+\varepsilon) \limsup(1-d_n) \qquad (493)$$

$$= \theta(1-\rho) + (1-\theta)(1+\varepsilon) \qquad (494)$$

$$= 1 - \frac{\theta\rho}{2} \qquad (495)$$

where $\varepsilon = \frac{\theta\rho}{2(1-\theta)}$; (492) follows from (489), (490); (493) holds by (490); (494) is due to (491).

APPENDIX D
PROOF OF THEOREM 12

Eq. (206) follows from the definitions in (26) and (57). Since $z^+ = \frac{1}{2}(|z| + z)$ and $z^- = \frac{1}{2}(|z| - z)$, for all $z \in \mathbb{R}$, (207) and (208) follow from (206) and

$$\mathbb{E}[1 - \exp(\imath_{P\|Q}(Y))] = \int \left(1 - \frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q = 0. \quad (496)$$

By change of measure, for every measurable function $f : \mathcal{A} \to \mathbb{R}$ with $\mathbb{E}[f(X)] < \infty$ and $\mathbb{E}[f(Y)] < \infty$,

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)f(Y)\right]$$
$$= \mathbb{E}[\exp(\imath_{P\|Q}(Y))f(Y)]. \quad (497)$$

Hence, it follows from (497) that

$$\mathbb{P}[\imath_{P\|Q}(X) > 0] = \mathbb{E}[1\{\imath_{P\|Q}(X) > 0\}] \quad (498)$$
$$= \mathbb{E}[\exp(\imath_{P\|Q}(Y))1\{\imath_{P\|Q}(Y) > 0\}] \quad (499)$$

and

$$\mathbb{P}[\imath_{P\|Q}(X) \leq 0] = \mathbb{E}[1\{\imath_{P\|Q}(X) \leq 0\}] \quad (500)$$
$$= \mathbb{E}[\exp(\imath_{P\|Q}(Y))1\{\imath_{P\|Q}(Y) \leq 0\}]. \quad (501)$$

To show (209), note that from (208) and the change of measure in (497), we get

$$\tfrac{1}{2}|P - Q|$$
$$= \mathbb{E}[(1 - \exp(\imath_{P\|Q}(Y)))^-] \quad (502)$$
$$= \mathbb{E}[(\exp(\imath_{P\|Q}(Y)) - 1)1\{\imath_{P\|Q}(Y) > 0\}] \quad (503)$$
$$= \mathbb{E}[(1 - \exp(-\imath_{P\|Q}(X)))1\{\imath_{P\|Q}(X) > 0\}] \quad (504)$$
$$= \mathbb{E}[(1 - \exp(-\imath_{P\|Q}(X)))^+]. \quad (505)$$

To show (210) and (211), we get from (208) and (499)

$$\tfrac{1}{2}|P - Q|$$
$$= \mathbb{E}[(1 - \exp(\imath_{P\|Q}(Y)))^-] \quad (506)$$
$$= \mathbb{E}\left[(\exp(\imath_{P\|Q}(Y)) - 1)1\{\imath_{P\|Q}(Y) > 0\}\right] \quad (507)$$
$$= \mathbb{P}[\imath_{P\|Q}(X) > 0] - \mathbb{P}[\imath_{P\|Q}(Y) > 0] \quad (508)$$

where (508) is (210), and (211) is equivalent to (210).

To show (212), we use (208) and the notation in (32) in order to write

$$\tfrac{1}{2}|P - Q| = \mathbb{E}[(1 - Z)^-] \quad (509)$$
$$= \mathbb{E}[(Z - 1)1\{Z > 1\}] \quad (510)$$
$$= \int_0^\infty \mathbb{P}[(Z - 1)1\{Z > 1\} \geq \beta]\mathrm{d}\beta \quad (511)$$
$$= \int_1^\infty \mathbb{P}[Z \geq \beta]\mathrm{d}\beta \quad (512)$$
$$= \int_0^1 \mathbb{P}[Z < \beta]\mathrm{d}\beta \quad (513)$$

where (509) follows from (208) with Z in (32); (511) exploits the fact that the expectation of a non-negative random variable is the integral of its complementary cumulative distribution

function; and (513) is satisfied since Z is non-negative with $\mathbb{E}[Z] = 1$.

To show (213), we use (209) to write

$$\tfrac{1}{2}|P - Q|$$
$$= \mathbb{E}[(1 - \exp(-\imath_{P\|Q}(X)))^+] \quad (514)$$
$$= \int_0^\infty \mathbb{P}[(1 - \exp(-\imath_{P\|Q}(X)))^+ > \beta]\mathrm{d}\beta \quad (515)$$
$$= \int_0^1 \mathbb{P}[(1 - \exp(-\imath_{P\|Q}(X)))^+ > \beta]\mathrm{d}\beta \quad (516)$$
$$= \int_0^1 \mathbb{P}\left[\imath_{P\|Q}(X) > \log\frac{1}{1 - \beta}\right]\mathrm{d}\beta \quad (517)$$
$$= \int_0^1 \mathbb{P}\left[\imath_{P\|Q}(X) > \log\frac{1}{\beta}\right]\mathrm{d}\beta. \quad (518)$$

To prove (214), a change of variable of integration in (213), and the fact that $\mathbb{F}_{P\|Q}(\log \beta) = 1$ for $\beta > \beta_1^{-1}$ give

$$\tfrac{1}{2}|P - Q| = \int_0^1 \mathbb{P}\left[\imath_{P\|Q}(X) > \log\frac{1}{t}\right]\mathrm{d}t \quad (519)$$
$$= \int_0^1 \left[1 - \mathbb{F}_{P\|Q}\left(\log\frac{1}{t}\right)\right]\mathrm{d}t \quad (520)$$
$$= \int_1^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\mathrm{d}\beta \quad (521)$$
$$= \int_1^{\beta_1^{-1}} \frac{1 - \mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\mathrm{d}\beta \quad (522)$$

with the convention that $\beta_1^{-1} = \infty$ if $\beta_1 = 0$.

Assume that $P \ll\gg Q$. To show (215) simply note that (206), the symmetry of the total variation distance, and the anti-symmetry of the relative information where $\imath_{Q\|P} = -\imath_{P\|Q}$ enable to conclude that

$$|P - Q| = \mathbb{E}[|1 - \exp(\imath_{Q\|P}(X))|] \quad (523)$$
$$= \mathbb{E}[|1 - \exp(-\imath_{P\|Q}(X))|]. \quad (524)$$

Similarly, switching P and Q in (208) results in

$$\tfrac{1}{2}|P - Q| = \mathbb{E}[(1 - \exp(\imath_{Q\|P}(X)))^-] \quad (525)$$
$$= \mathbb{E}[(1 - \exp(-\imath_{P\|Q}(X)))^-] \quad (526)$$

which proves (216).

APPENDIX E
(295) VS. (307)

A. Example for the Strengthened Inequality in Theorem 23

We exemplify the improvement obtained by (295), in comparison to (307), due to the introduction of the additional parameter β_2 in (138). Note that when β_2 is replaced by zero (i.e., no information on the infimum of $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is available or $\beta_2 = 0$), inequalities (295) and (307) coincide.

Let P and Q be two probability measures, defined on $(\mathcal{A}, \mathscr{F})$, $P \ll Q$, and assume that

$$1 - \eta \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \leq 1 + \eta, \quad \forall a \in \mathcal{A} \quad (527)$$

for a fixed $\eta \in (0, 1)$.

In (295), one can replace β_1 and β_2 with lower bounds on these constants. Since $\beta_1 \geq \frac{1}{1+\eta}$ and $\beta_2 \geq 1 - \eta$ it follows from (295) that

$$D(P\|Q)$$
$$\leq \tfrac{1}{2}\left(\frac{(1+\eta)\log(1+\eta)}{\eta} + \frac{(1-\eta)\log(1-\eta)}{\eta}\right)|P - Q| \tag{528}$$

$$\leq \eta \log e \cdot |P - Q|. \tag{529}$$

From (527)

$$\left|\exp(\iota_{P\|Q}(a)) - 1\right| \leq \eta, \quad \forall a \in \mathcal{A} \tag{530}$$

so, from (206), the total variation distance satisfies (recall that $Y \sim Q$)

$$|P - Q| = \mathbb{E}\left[\left|\exp(\iota_{P\|Q}(Y)) - 1\right|\right] \leq \eta. \tag{531}$$

Combining (531) with (529) yields

$$D(P\|Q) \leq \eta^2 \log e, \quad \forall \eta \in (0, 1). \tag{532}$$

For comparison, it follows from (307) (see [109, Th. 7]) that

$$D(P\|Q) \leq \frac{\log \frac{1}{\beta_1}}{2(1-\beta_1)} \cdot |P - Q| \tag{533}$$

$$\leq \frac{(1+\eta)\log(1+\eta)}{2\eta} \cdot |P - Q| \tag{534}$$

$$\leq \tfrac{1}{2}(1+\eta)\log(1+\eta) \tag{535}$$

$$\leq \tfrac{1}{2}\eta(1+\eta)\log e. \tag{536}$$

The upper bound on the relative entropy in (533) scales like η, for small η, whereas the tightened bound in (532) scales like η^2, which is tight according to Pinsker's inequality (1). For example, consider the probability measures defined on a two-element set $\mathcal{A} = \{a, b\}$ with

$$P(a) = Q(b) = \tfrac{1}{2} - \tfrac{\eta}{4}, \quad P(b) = Q(a) = \tfrac{1}{2} + \tfrac{\eta}{4}. \tag{537}$$

Condition (527) is satisfied for $\eta \approx 0$, and Pinsker's inequality (1) yields

$$D(P\|Q) \geq \tfrac{1}{2}\eta^2 \log e \tag{538}$$

so the ratio of the upper and lower bounds in (532) and (538) is 2, and both provide the true quadratic scaling in η whereas the weaker upper bound in (533) scales linearly in η for $\eta \approx 0$.

APPENDIX F
DERIVATION OF (308)–(313)

Similarly to the proof of Theorem 23, let $X \sim P$, $Y \sim Q$, and $Z = \exp(\iota_{P\|Q}(Y))$. We rely on the concavity of $\varphi : [0, \infty) \to [0, \infty)$, defined to be the continuous extension of $\frac{t \log t}{t-1}$, for tightening the upper bound in (303). The combination of this tightened bound with (300) and (306) serves to derive a tighter bound on the relative entropy in comparison to (295).

Since $Z \leq \beta_1^{-1}$, and φ is concave, monotonically increasing and differentiable, we can write

$$\varphi(Z) \leq \varphi(\beta_1^{-1}) - \varphi'(\beta_1^{-1})(\beta_1^{-1} - Z) \leq \varphi(\beta_1^{-1}) \tag{539}$$

which improves the upper bound on $\varphi(Z)$ in (297). Consequently, from (539), the first summand in the right side of (300) is upper bounded as follows:

$$\mathbb{E}\left[\varphi(Z)(Z - 1)1\{Z > 1\}\right] \tag{540}$$

$$\leq \mathbb{E}\left[\left(\varphi(\beta_1^{-1}) - \varphi'(\beta_1^{-1})(\beta_1^{-1} - Z)\right)(Z - 1)1\{Z > 1\}\right] \tag{541}$$

$$= \left(\varphi(\beta_1^{-1}) - \varphi'(\beta_1^{-1})\beta_1^{-1}\right)\mathbb{E}\left[(Z - 1)1\{Z > 1\}\right]$$
$$+ \varphi'(\beta_1^{-1})\mathbb{E}\left[Z(Z - 1)1\{Z > 1\}\right] \tag{542}$$

$$= \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi'(\beta_1^{-1})\beta_1^{-1}\right)|P - Q|$$
$$+ \varphi'(\beta_1^{-1})\mathbb{E}\left[Z(Z - 1)1\{Z > 1\}\right] \tag{543}$$

where (543) follows from (32) and (207). Combining (300), (306) and (543) gives the upper bound on the relative entropy in (308).

The second term in the right side of (543) depends on the distribution of the relative information. To circumvent this dependence, we derive upper and lower bounds in terms of f-divergences.

$$\mathbb{E}\left[Z(Z - 1)1\{Z > 1\}\right]$$
$$= \mathbb{E}\left[(Z - 1)^2 1\{Z > 1\}\right] + \mathbb{E}\left[(Z - 1)1\{Z > 1\}\right] \tag{544}$$
$$= \mathbb{E}\left[(Z - 1)^2 1\{Z > 1\}\right] + \tfrac{1}{2}|P - Q| \tag{545}$$

where (545) follows from (207), and consequently the following upper and lower bounds on (544) are derived:

$$\mathbb{E}\left[Z(Z - 1)1\{Z > 1\}\right] \leq \mathbb{E}\left[(Z - 1)^2\right] + \tfrac{1}{2}|P - Q| \tag{546}$$
$$= \chi^2(P\|Q) + \tfrac{1}{2}|P - Q| \tag{547}$$

where (547) follows from (32) and (46). Furthermore, from (208), (297) and (544)

$$\mathbb{E}\left[Z(Z - 1)1\{Z > 1\}\right]$$
$$= \mathbb{E}\left[(Z - 1)^2 1\{Z > 1\}\right] + \tfrac{1}{2}|P - Q| \tag{548}$$
$$= \mathbb{E}\left[(Z - 1)^2\right] - \mathbb{E}\left[(Z - 1)^2 1\{\beta_2 \leq Z \leq 1\}\right]$$
$$+ \tfrac{1}{2}|P - Q| \tag{549}$$
$$= \chi^2(P\|Q) + \tfrac{1}{2}|P - Q|$$
$$- \mathbb{E}\left[(Z - 1)^2 1\{\beta_2 \leq Z \leq 1\}\right] \tag{550}$$
$$\geq \chi^2(P\|Q) + \tfrac{1}{2}|P - Q|$$
$$- (1 - \beta_2)\mathbb{E}\left[(1 - Z)1\{\beta_2 \leq Z \leq 1\}\right] \tag{551}$$
$$= \chi^2(P\|Q) + \tfrac{1}{2}|P - Q| - (1 - \beta_2)\mathbb{E}\left[(Z - 1)^-\right] \tag{552}$$
$$= \chi^2(P\|Q) + \tfrac{\beta_2}{2}|P - Q|. \tag{553}$$

Combining (546) and (553) gives the inequality in (310), and combining (300), (543) and (546) gives the upper bound on the relative entropy in (313).

APPENDIX G
PROOF OF THEOREM 26

B. Proof of Theorem 26a)

The concavity of the entropy functional implies that given a probability mass function P on a finite set $\{1, \ldots, |\mathcal{A}|\}$, and

given any subset $\mathcal{S} \subset \mathcal{A}$, $H(P) \leq H(P_{\mathcal{S}})$ with

$$
P_{\mathcal{S}}(k) = \begin{cases} \frac{P(\mathcal{S})}{|\mathcal{S}|} & k \in \mathcal{S} \\ \frac{P(\mathcal{S}^c)}{|\mathcal{S}^c|} & k \notin \mathcal{S}. \end{cases} \tag{554}
$$

Applying this fact with $\mathcal{S}$ given by the indices of the masses below $|\mathcal{A}|^{-1}$, we conclude that $H(P) \leq H(\bar{P})$ with

$$
\bar{P}(k) = \begin{cases} \frac{\sum_{a \in \mathcal{A}} P(a)1\{P(a) \geq |\mathcal{A}|^{-1}\}}{\sum_{a \in \mathcal{A}} 1\{P(a) \geq |\mathcal{A}|^{-1}\}} & k: P(k) \geq |\mathcal{A}|^{-1}, \\ \frac{\sum_{a \in \mathcal{A}} P(a)1\{P(a) < |\mathcal{A}|^{-1}\}}{\sum_{a \in \mathcal{A}} 1\{P(a) < |\mathcal{A}|^{-1}\}} & k: P(k) < |\mathcal{A}|^{-1}. \end{cases} \tag{555}
$$

Moreover, if U is the equiprobable distribution on $\mathcal{A}$, then

$$
|P - \mathsf{U}| = |\bar{P} - \mathsf{U}|. \tag{556}
$$

Consequently, in order to maximize entropy subject to a given (positive) total variation distance from the equiprobable distribution on $\mathcal{A}$, it is enough to restrict attention to distributions whose masses take two distinct values only, i.e., of the form (341). The only remaining optimization is to determine m_Δ, the number of masses larger than $|\mathcal{A}|^{-1}$. The requirement that m_Δ satisfy (340) is made so that (341) is a valid probability distribution. The solution is as given in Part a) since $H(P) = \log|\mathcal{A}| - D(P\|\mathsf{U})$, and

$$
D(P_\Delta\|\mathsf{U}) = d\left(\frac{m_\Delta}{|\mathcal{A}|} + \frac{\Delta}{2} \,\Big\|\, \frac{m_\Delta}{|\mathcal{A}|}\right). \tag{557}
$$

C. Proof of Theorem 26b)

The minimizing probability measure in (344) is a special case of [52, Th. 3], which gives the general solution of minimizing the entropy subject to a constraint on the maximal total variation distance from a fixed discrete distribution Q (here, $Q = \mathsf{U}$).

APPENDIX H
COMPLETION OF THE PROOF OF THEOREM 31

Proof of monotonicity and boundedness of (415): Substituting $\gamma = \beta x$ into the right side of (415) gives that, for $\beta > 1$,

$$
u(\beta) = \min_{x \in \left(\frac{1}{\beta}, 1\right)} \left(\frac{c_{\beta x}}{1 - x}\right). \tag{558}
$$

The function u in (558) is indeed monotonically decreasing on $(1, \infty)$ since, if $\beta_2 > \beta_1 > 1$,

$$
u(\beta_1) = \min_{x \in \left(\frac{1}{\beta_1}, 1\right)} \frac{c_{\beta_1 x}}{1 - x} \tag{559}
$$

$$
\geq \min_{x \in \left(\frac{1}{\beta_1}, 1\right)} \frac{c_{\beta_2 x}}{1 - x} \tag{560}
$$

$$
\geq \min_{x \in \left(\frac{1}{\beta_2}, 1\right)} \frac{c_{\beta_2 x}}{1 - x} \tag{561}
$$

$$
= u(\beta_2) \tag{562}
$$

where (560) holds since c_γ is monotonically decreasing in $\gamma \in (1, \infty)$ (see Theorem 30).

Proof of (416): From (396) and (397), we obtain $t_\gamma > \gamma$ for $\gamma > 1$. Furthermore, since $t/r(t)$ is monotonically decreasing on $(1, \infty)$, if $\gamma > e$, then

$$
c_\gamma = \frac{t_\gamma - \gamma}{r(t)} < \frac{\gamma}{r(\gamma)} = \frac{\gamma}{\log e + \gamma \log \frac{\gamma}{e}} < \frac{1}{\log \frac{\gamma}{e}}. \tag{563}
$$

Hence, for $\beta > 2e$,

$$
u(\beta) = \min_{\gamma \in (1, \beta)} \frac{\beta c_\gamma}{\beta - \gamma} \tag{564}
$$

$$
\leq 2c_{\beta/2} \tag{565}
$$

$$
< \frac{2}{\log \frac{\beta}{2e}} \tag{566}
$$

where (565) follows by choosing $\gamma = \frac{\beta}{2}$ in the minimization, and (566) follows from (563).

APPENDIX I
A LEMMA USED FOR PROVING (454)

Lemma 3: Let g be a monotonically increasing and non-negative function on $[a, b]$, and let p_1, p_2 be probability density functions supported on $[a, b]$. Assume that there exists $c \in (a, b)$ such that

$$
\begin{aligned}
p_1(\beta) &\geq p_2(\beta), \quad \forall \beta \in [a, c], \\
p_1(\beta) &< p_2(\beta), \quad \forall \beta \in (c, b].
\end{aligned} \tag{567}
$$

Let $W \sim p_1$ and $V \sim p_2$, then

$$
\mathbb{E}[g(W)] \leq \mathbb{E}[g(V)]. \tag{568}
$$

Proof: The function $d \triangleq p_2 - p_1$, defined on $[a, b]$, satisfies

$$
d(\beta) \leq 0, \quad \forall \beta \in [a, c] \tag{569}
$$

$$
d(\beta) \geq 0, \quad \forall \beta \in [c, b] \tag{570}
$$

$$
\int_a^b d(\beta)\mathrm{d}\beta = 0. \tag{571}
$$

Consequently, we get

$$
\mathbb{E}[g(V)] - \mathbb{E}[g(W)]
$$

$$
= \int_a^c d(\beta)g(\beta)\mathrm{d}\beta + \int_c^b d(\beta)g(\beta)\mathrm{d}\beta \tag{572}
$$

$$
\geq g(c)\int_a^c d(\beta)\mathrm{d}\beta + g(c)\int_c^b d(\beta)\mathrm{d}\beta \tag{573}
$$

$$
= 0 \tag{574}
$$

where (573) follows from (570), (571) and the monotonicity of g, and (574) is due to (571). ∎

REFERENCES

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 1976.

[2] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 28, no. 1, pp. 131–142, 1966.

[3] K. M. R. Audenaert and J. Eisert, "Continuity bounds on the quantum relative entropy," *J. Math. Phys.*, vol. 46, no. 10, p. 102104, Oct. 2005.

[4] K. M. R. Audenaert and J. Eisert, "Continuity bounds on the quantum relative entropy—II," *J. Math. Phys.*, vol. 52, no. 11, p. 112201, Nov. 2011.

[5] K. M. R. Audenaert, "On the asymmetry of the relative entropy," *J. Math. Phys.*, vol. 54, no. 7, p. 073506, Jul. 2013.

[6] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336–342, Jan. 1986.

[7] A. Barron, "Information theory and martingales," presented at the IEEE Int. Symp. Inf. Theory, Budapest, Hungary, Jun. 1991. [Online]. Available: http://www.stat.yale.edu/~arb4/publications_files/informationtheoryandmartingales.pdf

[8] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation (contd.)," *Sankhyā, Indian J. Statist.*, vol. 8, no. 3, pp. 201–218, Oct. 1947.

[9] A. Basu, H. Shioya, and C. Park, *Statistical Inference: The Minimum Distance Approach* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), vol. 120. Boca Raton, FL, USA: CRC Press, Jun. 2011.

[10] D. Berend, P. Harremoës and A. Kontorovich, "Minimum KL-divergence on complements of L_1 balls," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3172–3177, Jun. 2014.

[11] L. Birgé and P. Massart, "Minimum contrast estimators on sieves: Exponential bounds and rates of convergence," *Bernoulli*, vol. 4, no. 3, pp. 329–375, 1998.

[12] G. Böcherer and B. C. Geiger. (Mar. 2015). "Optimal quantization for distribution synthesis." [Online]. Available: https://arxiv.org/abs/1307.6843

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. New-York, NY, USA: Cambridge Univ. Press, 2004.

[15] J. Bretagnolle and C. Huber, "Estimation des densités: Risque minimax," *Probab. Theory Rel. Fields*, vol. 47, no. 2, pp. 119–137, 1979.

[16] X. Chen, A. Guntuboyina, and Y. Zhang. (Oct. 2014). "On Bayes risk lower bounds." [Online]. Available: https://arxiv.org/abs/1410.0503

[17] C. D. Charalambous, I. Tzortzis, S. Loyka, and T. Charalambous, "Extremum problems with total variation distance and their applications," *IEEE Trans. Autom. Control*, vol. 59, no. 9, pp. 2353–2368, Sep. 2014.

[18] J. E. Cohen, J. H. B. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices With Applications in Information Theory, Statistics, Economics and Population*. Boston, MA, USA: Springer, 1998.

[19] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[21] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungarian Acad. Sci.*, vol. 8, pp. 85–108, Jan. 1963.

[22] I. Csiszár, "A note on Jensen's inequality," *Studia Sci. Math. Hungarica*, vol. 1, pp. 185–188, Jan. 1966.

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

[24] I. Csiszár, "On topological properties of f-divergences," *Studia Scientiarum Math. Hungarica*, vol. 2, pp. 329–339, Jan. 1967.

[25] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Math. Hungarica*, vol. 2, no. 1, pp. 191–213, Mar. 1972.

[26] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, Feb. 1975.

[27] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[28] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 4, pp. 417–528, 2004.

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

[30] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.

[31] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 404–419, 1962.

[32] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *Ann. Appl. Probab.*, vol. 6, no. 3, pp. 695–750, 1996.

[33] S. S. Dragomir, "Bounds for the normalised Jensen functional," *Bull. Austral. Math. Soc.*, vol. 74, no. 3, pp. 471–478, Dec. 2006.

[34] S. S. Dragomir, "Upper and lower bounds for Csiszár f-divergence in terms of the Kullback-Leibler distance and applications," in *Inequalities for Csiszár f-Divergence in Information Theory, RGMIA Monographs*, S. S. Dragomir and T. M. Rassias, Eds. Melbourne VIC, Australia: Victoria University, Australia, 2000.

[35] S. S. Dragomir, "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*, S. S. Dragomir and T. M. Rassias, Eds. Melbourne VIC, Australia: Victoria University, Australia, 2000.

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

[37] T. van Erven and P. Harremoës, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[38] E. Even-Dar, S. M. Kakade, and Y. Mansour, "The value of observation for monitoring dynamic systems," *Proc. Int. Joint Conf. Artif. Intell.*, Hyderabad, India, Jan. 2007, pp. 2474–2479.

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

[40] S. L. Fong and V. Y. F. Tan, "Strong converse theorems for classes of multimessage multicast networks: A Rényi divergence approach," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 4953–4967, Sep. 2016.

[41] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Statist. Rev.*, vol. 70, no. 3, pp. 419–435, Dec. 2002.

[42] M. Gil, "On Rényi divergence measures for continuous alphabet sources," M.S. thesis, Dept. Math. Statist., Queen's Univ., Kingston, ON, Canada, Aug. 2011.

[43] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Inf. Sci.*, vol. 249, pp. 124–131, Nov. 2013.

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

[45] G. L. Gilardoni, "Corrigendum to the note on the minimum f-divergence for given total variation," *Comp. Rendus Math.*, vol. 348, nos. 5–6, p. 299, 2010.

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

[47] A. Guntuboyina, "Lower bounds for the minimax risk using f-divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, Apr. 2011.

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

[49] P. Harremoës, "Some new maximal inequalities," *Statist. Probab. Lett.*, vol. 78, no. 16, pp. 2776–2780, Nov. 2008.

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

[51] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, pp. 2451–2492, Dec. 1997.

[52] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.

[53] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions," *Ann. Math. Statist.*, vol. 29, no. 3, pp. 700–718, Sep. 1958.

[54] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 186, no. 1007, pp. 453–461, Sep. 1946.

[55] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[56] J. H. B. Kemperman, "On the optimum rate of transmitting information," *Ann. Math. Statist.*, vol. 40, no. 6, pp. 2156–2177, Dec. 1969.

[57] V. Kostina and S. Verdú, "Channels with cost constraints: Strong converse and dispersion," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2415–2429, May 2015.

[58] M. Krajči, C. F. Liu, L. Mikeš, and S. M. Moser, "Performance analysis of Fano coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015, pp. 1746–1750.

[59] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[60] S. Kullback, "A lower bound for discrimination information in terms of variation (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 126–127, Jan. 1967.

[61] M. A. Kumar and I. Sason, "Projection theorems for the Rényi divergence on α-convex sets," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 4924–4935, Sep. 2016.

[62] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York, NY, USA: Springer, 1986.

[63] L. LeCam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, no. 1, pp. 38–53, Jan. 1973.

[64] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*. New York, NY, USA: Springer, 1990.

[65] F. Liese and I. Vajda, *Convex Statistical Distances* (Teubner-Texte Zur Mathematik), vol. 95. Leipzig, Germany: Teubner, 1987.

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

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

[68] J. Liu, P. Cuff, and S. Verdú. (Nov. 2015). "E_γ-resolvability." [Online]. Available: http://arxiv.org/abs/1511.07829

[69] J. Liu, P. Cuff, and S. Verdú, "Resolvability in E_γ with applications to lossy compression and wiretap channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015, pp. 755–759.

[70] J. Liu, P. Cuff, and S. Verdú, "One-shot mutual covering lemma and Marton's inner bound with a common message," in *Proc. IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015, pp. 1457–1461.

[71] A. Makur and L. Zheng. (Oct. 2015). "Bounds between contraction coefficients." [Online]. Available: https://arxiv.org/abs/1510.01844

[72] K. Marton, "A measure concentration inequality for contracting Markov chains," *Geometric Funct. Anal.*, vol. 6, no. 3, pp. 556–571, May 1996.

[73] K. Marton, "Distance-divergence inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013, pp. 1849–1853.

[74] F. Österreicher and I. Vajda, "Statistical information and discrimination," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 1036–1039, May 1993.

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

[76] K. Pearson, "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 reasonably be supposed to have arisen from random sampling," *Philos. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.

[77] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA, USA: Holden-Day, 1964.

[78] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[79] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf.*, Monticello, IL, USA, Sep. 2010, pp. 1327–1333.

[80] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan. 2016.

[81] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–248, 2013.

[82] M. Raginsky and I. Sason, "Concentration of measure inequalities and their communication and information-theoretic applications," *IEEE Inf. Theory Soc. Newslett.*, vol. 65, no. 4, pp. 24–34, Dec. 2015.

[83] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, Jun. 2016.

[84] T. R. C. Read and N. A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York, NY, USA: Springer, 1988.

[85] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 731–817, Mar. 2011.

[86] R.-D. Reiss, *Approximate Distributions of Order Statistics: With Applications to Non-Parametric Statistics*. New York, NY, USA: Springer, 1989.

[87] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berekely Symp. Probab. Theory Math. Statist.*, Berkeley, CA, USA, 1961, pp. 547–561.

[88] P.-M. Samson, "Concentration of measure inequalities for Markov chains and ϕ-mixing processes," *Ann. Probab.*, vol. 28, no. 1, pp. 416–461, Jan. 2000.

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

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

[91] I. Sason, "On the Rényi divergence, joint range of relative entropies, and a channel coding theorem," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 23–34, Jan. 2016.

[92] I. Sason and S. Verdú, "Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets," in *Proc. IEEE Inf. Theory Workshop*, Jeju Island, South Korea, Oct. 2015, pp. 214–218.

[93] I. Sason and S. Verdú, "f-divergence inequalities." [Online]. Available: https://arxiv.org/abs/1508.00335

[94] O. Shayevitz, "On Rényi measures and hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Aug. 2011, pp. 800–804.

[95] S. Simic, "On logarithmic convexity for differences of power means," *J. Inequalities Appl.*, vol. 2007, Oct. 2007, Art. no. 037359.

[96] S. Simic, "On a new moments inequality," *Statist. Probab. Lett.*, vol. 78, no. 16, pp. 2671–2678, Nov. 2008.

[97] S. Simic. (Sep. 2015). "Refinement of some moment inequalities." [Online]. Available: http://arxiv.org/abs/1509.0851

[98] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 63–86, Jan. 1996.

[99] F. E. Su, "Methods for quantifying rates of convergence for random walks on groups," Ph.D. dissertation, Dept. Math., Harvard Univ., Cambridge, MA, USA, 1995.

[100] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7041–7051, Nov. 2013.

[101] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602–1609, Jul. 2000.

[102] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer, 2009.

[103] I. Vajda, "Note on discrimination information and variation (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 16, no. 6, pp. 771–773, Nov. 1970.

[104] I. Vajda, "On the f-divergence and singularity of probability measures," *Periodica Math. Hungarica*, vol. 2, no. 1, pp. 223–234, May 1972.

[105] I. Vajda, *Theory of Statistical Inference and Information*. Dordrecht, The Netherlands: Kluwer, 1989.

[106] I. Vajda, "On metric divergences of probability measures," *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.

[107] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[108] I. Vincze, "On the concept and measure of information contained in an observation," in *Contributions to Probability*, J. Gani and V. K. Rohatgi, Eds. New York, NY, USA: Academic, 1981, pp. 207–214.

[109] S. Verdú, "Total variation distance and the distribution of relative information," in *Proc. Inf. Theory Appl. Workshop*, San Diego, CA, USA, Feb. 2014, pp. 499–501.

[110] S. Verdú, *Information Theory*, in preparation.

[111] W. H. Wong and X. Shen, "Probability inequalities for likelihood ratios and convergence rates of sieve MLES," *Ann. Statist.*, vol. 23, no. 2, pp. 339–362, Apr. 1995.

[112] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," *Information Theory New Trends and Open Problems*, G. Longo, Ed. New York, NY, USA: Springer, 1975, pp. 87–123.

[113] Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3280–3282, Sep. 2007.

Igal Sason (S'98–M'02–SM'11) was born in Israel in 1969. He received the B.Sc. and Ph.D. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1992 and 2001, respectively. During 1993-1997, he worked in Israel as a communication engineer.

During 2001-2003, he held a position of a scientific collaborator at the School of Computer and Communication Sciences of EPFL, Lausanne, Switzerland. Since 2003, he has been with the Department of Electrical Engineering at the Technion where he is currently an Associate Professor.

Dr. Sason is a co-recipient of the 2003 IEEE Communications/Information Theory Joint Paper Award. During 2003-2006, he was a recipient of the Alon fellowship for young faculty members, and the Horev fellowship for leaders in science and technology. He has served as an Associate Editor for the IEEE Transactions on Information Theory: Coding Theory, 2008-2011; at Large, 2014-present. He also served as a program co-chair of the 2013 IEEE International Symposium on Information Theory.

His research interests are in information theory and coding. He is especially interested in information measures, concentration of measure inequalities in information theory, converse techniques in channel coding, interference channels, and performance/complexity analysis of error-correcting codes with an emphasis on codes defined on graphs.

Sergio Verdú (S'80–M'84–SM'88–F'93) received the Telecommunications Engineering degree from the Universitat Politècnica de Barcelona in 1980, and the Ph.D. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1984. Since 1984 he has been a member of the faculty of Princeton University, where he is the Eugene Higgins Professor of Electrical Engineering, and is a member of the Program in Applied and Computational Mathematics. Sergio Verdú is the recipient of the 2007 Claude E. Shannon Award, and the 2008 IEEE Richard W. Hamming Medal. He is a member of both the National Academy of Engineering and the National Academy of Sciences, and a corresponding member of the Royal Academy of Engineering of Spain. He is a recipient of the 2016 National Academy of Sciences Award for Scientific Reviewing. Verdú is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 and 2012 Information Theory Paper Awards, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, the 2006 Joint Communications/Information Theory Paper Award, and the 2009 Stephen O. Rice Prize from the IEEE Communications Society. In 1998, Cambridge University Press published his book *Multiuser Detection*, for which he received the 2000 Frederick E. Terman Award from the American Society for Engineering Education. He was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005. Sergio Verdú served as President of the IEEE Information Theory Society in 1997, and on its Board of Governors (1988-1999, 2009-2014). He has also served in various editorial capacities for the IEEE Transactions on Information Theory: Associate Editor (Shannon Theory, 1990-1993; Book Reviews, 2002-2006), Guest Editor of the Special Fiftieth Anniversary Commemorative Issue (published by IEEE Press as Information Theory: Fifty Years of Discovery), and member of the Executive Editorial Board (2010-2013). He is the founding Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*. He served as a general co-chair of the 2000 and 2016 IEEE International Symposia on Information Theory.