On Reverse Pinsker Inequalities

Igal Sason

On Reverse Pinsker Inequalities

Igal Sason

Abstract

New upper bounds on the relative entropy are derived as a function of the total variation distance. One bound refines an inequality by Verdú for general probability measures. A second bound improves the tightness of an inequality by Csiszár and Talata for arbitrary probability measures that are defined on a common finite set. The latter result is further extended, for probability measures on a finite set, leading to an upper bound on the Rényi divergence of an arbitrary non-negative order (including ∞) as a function of the total variation distance. Another lower bound by Verdú on the total variation distance, expressed in terms of the distribution of the relative information, is tightened and it is attained under some conditions. The effect of these improvements is exemplified.

Keywords: Pinsker's inequality, relative entropy, relative information, Rényi divergence, total variation distance, typical sequences.

I. INTRODUCTION

Consider two probability measures P and Q defined on a common measurable space $(\mathcal{A}, \mathcal{F})$. The Csiszár-Kemperman-Kullback-Pinsker inequality states that

$$D(P\|Q) \geq \frac{\log e}{2} \cdot |P - Q|^2 \tag{1}$$

where

$$D(P\|Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right] = \int_{\mathcal{A}} \mathrm{d}P \, \log \frac{\mathrm{d}P}{\mathrm{d}Q} \tag{2}$$

designates the relative entropy from P to Q (a.k.a. the Kullback-Leibler divergence), and

$$|P - Q| = 2 \sup_{A \in \mathcal{F}} \big|P(A) - Q(A)\big| \tag{3}$$

designates the total variation distance (or L_1 distance) between P and Q. One of the implications of inequality (1) is that convergence in relative entropy implies convergence in total variation distance. The total variation distance is bounded $|P - Q| \leq 2$, in contrast to the relative entropy.

Inequality (1) is a.k.a. Pinsker's inequality, although the analysis made by Pinsker [15] leads to a significantly looser bound where $\frac{\log e}{2}$ on the RHS of (1) is replaced by $\frac{\log e}{408}$ (see [25, Eq. (51)]). Improved and generalized versions of Pinsker's inequality have been studied in [7], [8], [9], [14], [18], [24].

For any $\varepsilon > 0$, there exists a pair of probability measures P and Q such that $|P - Q| \leq \varepsilon$ while $D(P\|Q) = \infty$. Consequently, a reverse Pinsker inequality which provides an upper bound on the relative entropy in terms of the total variation distance does not exist in general. Nevertheless, under some conditions, such inequalities hold [4], [25], [26] (to be addressed later in this section).

If $P \ll Q$, the relative information in $a \in \mathcal{A}$ according to (P, Q) is defined to be

$$i_{P\|Q}(a) \triangleq \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a). \tag{4}$$

From (2), the relative entropy can be expressed in terms of the relative information as follows:

$$D(P\|Q) = \mathbb{E}\big[i_{P\|Q}(X)\big] = \mathbb{E}\big[i_{P\|Q}(Y)\exp\big(i_{P\|Q}(Y)\big)\big] \tag{5}$$

where $X \sim P$ and $Y \sim Q$ (i.e., X and Y are distributed according to P and Q, respectively). The total variation distance is also expressible in terms of the relative information [25]. If $Q \ll P$

$$|P - Q| = \mathbb{E}\Big[\big|1 - \exp\big(i_{P\|Q}(Y)\big)\big|\Big] \tag{6}$$

and if, in addition, $P \ll Q$, then

$$|P - Q| = \mathbb{E}\Big[\big|1 - \exp\big(-i_{P\|Q}(X)\big)\big|\Big]. \tag{7}$$

Let

$$\beta_1^{-1} \triangleq \sup_{a \in \mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \tag{8}$$

with the convention, implied by continuity, that $\beta_1 = 0$ if $i_{P\|Q}$ is unbounded from above. With $\beta_1 \leq 1$, as it is defined in (8), the following inequality holds (see [25, Theorem 7]):

$$\frac{1}{2}|P - Q| \geq \left(\frac{1 - \beta_1}{\log \frac{1}{\beta_1}}\right) D(P\|Q). \tag{9}$$

From (9), if the relative information is bounded from above, a reverse Pinsker inequality holds. This inequality has been recently used in the context of the optimal quantization of probability measures when the distortion is either characterized by the total variation distance or the relative entropy between the approximating and the original probability measures [2, Proposition 4].

Inequality (9) is refined in this work, and the improvement that is obtained by this refinement is exemplified (see Section II).

In the special case where P and Q are defined on a common discrete set (i.e., a finite or countable set) $\mathcal{A}$, the relative entropy and total variation distance are simplified to

$$D(P\|Q) = \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)},$$

$$|P - Q| = \sum_{a \in \mathcal{A}} \big|P(a) - Q(a)\big| \triangleq |P - Q|_1.$$

A restriction to probability measures on a finite set $\mathcal{A}$ has led in [4, p. 1012 and Lemma 6.3] to the following upper bound on the relative entropy in terms of the total variation distance:

$$D(P\|Q) \leq \left(\frac{\log e}{Q_{\min}}\right) \cdot |P - Q|^2, \tag{10}$$

where $Q_{\min} \triangleq \min_{a \in \mathcal{A}} Q(a)$, suggesting a kind of a reverse Pinsker inequality for probability measures on a finite set. A recent application of this bound has been exemplified in [13, Appendix D] and [23, Lemma 7] for the analysis of the third-order asymptotics of the discrete memoryless channel with or without cost constraints.

The present paper also considers generalized reverse Pinsker inequalities for Rényi divergences. In the discrete setting, the Rényi divergence of order α from P to Q is defined as

$$D_\alpha(P\|Q) \triangleq \frac{1}{\alpha - 1} \log \left(\sum_{a \in \mathcal{A}} P^\alpha(a)\, Q^{1-\alpha}(a)\right), \quad \forall \alpha \in (0, 1) \cup (1, \infty). \tag{11}$$

Recall that $D_1(P\|Q) \triangleq D(P\|Q)$ is defined to be the analytic extension of $D_\alpha(P\|Q)$ at $\alpha = 1$ (if $D(P\|Q) < \infty$, it can be verified with L'Hôpital's rule that $D(P\|Q) = \lim_{\alpha \to 1^-} D_\alpha(P\|Q)$). The extreme cases of $\alpha = 0, \infty$ are defined as follows:

- If $\alpha = 0$ then $D_0(P\|Q) = -\log Q(\text{Support}(P))$ where $\text{Support}(P) = \{x \in \mathcal{X} \colon P(x) > 0\}$ denotes the support of P,
- If $\alpha = +\infty$ then $D_\infty(P\|Q) = \log \left(\text{ess sup} \frac{P}{Q} \right)$ where $\text{ess sup} f$ denotes the essential supremum of a function f.

Pinsker's inequality has been generalized by Gilardoni [9] for Rényi divergences of order $\alpha \in (0, 1]$ (see also [6, Theorem 30]), and it gets the form

$$D_\alpha(P\|Q) \geq \frac{\alpha \log e}{2} \cdot |P - Q|^2.$$

An improved bound, providing the best lower bound on the Rényi divergence of order $\alpha > 0$ in terms of the total variation distance, has been recently introduced in [20, Section 2].

Motivated by these findings, the analysis in this paper suggests an improvement over the upper bound on the relative entropy in (10) for probability measures defined on a common finite set. The improved version of the bound in (10) is further generalized to provide an upper bound on the Rényi divergence of orders $\alpha \in [0, \infty]$ in terms of the total variation distance.

Note that the issue addressed in this paper of deriving, under suitable conditions, upper bounds on the relative entropy as a function of the total variation distance has some similarity to the issue of deriving upper bounds on the difference between entropies as a function of the total variation distance. Note also that in the special case where Q is a Gaussian distribution and P is a probability distribution with the same covariance matrix, then $D(P\|Q) = h(Q) - h(P)$ where $h(\cdot)$ denotes the differential entropy of a specified distribution (see [3, Eq. (8.76)]). Bounds on the entropy difference in terms of the total variation distance have been studied, e.g., in [3, Theorem 17.3.3], [11], [16], [17], [19], [26, Section 1.7], [27].

This paper is structured as follows: Section II refers to [25], deriving a refined version of inequality (9) for general probability measures, and improving another lower bound on the total variation distance which is expressed in terms of the distribution of the relative information. Section III derives a reverse Pinsker inequality for probability measures on a finite set, improving inequality (10) that follows from [4, Lemma 6.3]. Section IV extends the analysis to Rényi divergences of arbitrary non-negative orders. Section V exemplifies the utility of a reverse Pinsker inequality in the context of typical sequences.

II. A Refined Reverse Pinsker Inequality for General Probability Measures

The present section derives a reverse Pinsker inequality for general probability measures, suggesting a refined version of [25, Theorem 7]. The utility of this new inequality is exemplified. This section also provides a lower bound on the total variation distance which is based on the distribution of the relative information; the latter inequality is based on a modification of the proof of [25, Theorem 8], and it has the advantage of being tight for a double-parameter family of probability measures which are defined on an arbitrary set of 2 elements.

A. Main Result and Proof

Inequality (9) provides an upper bound on the relative entropy $D(P\|Q)$ as a function of the total variation distance when $P \ll Q$, and the relative information $i_{P\|Q}$ is bounded from above (this implies that β_1 in (8) is positive). The following theorem tightens this upper bound.

Theorem 1: Let P and Q be probability measures on a measurable space $(\mathcal{A}, \mathcal{F})$, $P \ll Q$, and let $\beta_1, \beta_2 \in [0, 1]$ be given by

$$\beta_1^{-1} \triangleq \sup_{a \in \mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(a), \quad \beta_2 \triangleq \inf_{a \in \mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(a). \tag{12}$$

Then, the following inequality holds:

$$D(P\|Q) \le \frac{1}{2} \left(\frac{\log \frac{1}{\beta_1}}{1 - \beta_1} - \beta_2 \log e \right) |P - Q|. \tag{13}$$

Proof: Let $X \sim P$, $Y \sim Q$, and

$$\mathcal{B} \triangleq \{ a \in \mathcal{A} : i_{P\|Q}(a) > 0 \}. \tag{14}$$

From (5), the relative entropy is equal to

$$D(P\|Q) = \int_{\mathcal{A}} \mathrm{d}Q \, \exp(i_{P\|Q}) \, i_{P\|Q}$$

$$= \int_{\mathcal{B}} \mathrm{d}Q \, \exp(i_{P\|Q}) \, i_{P\|Q} + \int_{\mathcal{A} \setminus \mathcal{B}} \mathrm{d}Q \, \exp(i_{P\|Q}) \, i_{P\|Q}. \tag{15}$$

In the following, the two integrals on the RHS of (15) are upper bounded. The upper bound on the first integral on the RHS of (15) is based on the proof of [25, Theorem 7]; it is provided in the following for completeness, and with more details in order to clarify the way that this bound is refined here. Let $z(a) \triangleq \exp(i_{P\|Q}(a))$ for $a \in \mathcal{A}$. By assumption $1 < z(a) \le \frac{1}{\beta_1}$ for all $a \in \mathcal{B}$. The function $f(z) = \frac{z \log(z)}{z-1}$ is monotonic increasing over the interval $(1, \infty)$ since we have $(z-1)^2 f'(z) = (z-1) \log e - \log z > 0$ for $z > 1$. Consequently, we have

$$\frac{z(a) \log z(a)}{z(a) - 1} \le \frac{\log \frac{1}{\beta_1}}{1 - \beta_1}, \quad \forall a \in \mathcal{B} \tag{16}$$

and

$$\int_{\mathcal{B}} \mathrm{d}Q \, \exp(i_{P\|Q}) \, i_{P\|Q}$$

$$\overset{(a)}{\le} \frac{\log \frac{1}{\beta_1}}{1 - \beta_1} \int_{\mathcal{B}} \mathrm{d}Q \, (\exp(i_{P\|Q}) - 1)$$

$$\overset{(b)}{=} \frac{\log \frac{1}{\beta_1}}{1 - \beta_1} \int_{\mathcal{A}} \mathrm{d}Q(a) \, (1 - \exp(i_{P\|Q}(a)))^-$$

$$\overset{(c)}{=} \left(\frac{\log \frac{1}{\beta_1}}{1 - \beta_1} \right) \mathbb{E} \left[(1 - \exp(i_{P\|Q}(Y)))^- \right]$$

$$\overset{(d)}{=} \left(\frac{\log \frac{1}{\beta_1}}{2(1 - \beta_1)} \right) |P - Q| \tag{17}$$

where inequality (a) follows from (16), equality (b) is due to (14) and the definition $(a)^- \triangleq -a \, 1\{a < 0\}$, equality (c) holds since $Y \sim Q$, and equality (d) follows from [25, Eq. (14)].

At this point, we deviate from the analysis in [25] where the second integral on the RHS of (15) has been upper bounded by zero (since $i_{P\|Q}(a) \leq 0$ for all $a \in \mathcal{A} \setminus \mathcal{B}$). If $\beta_2 > 0$, we provide in the following a strictly negative upper bound on this integral. Since $P \ll Q$, we have

$$\int_{\mathcal{A}\setminus\mathcal{B}} \mathrm{d}Q \, \exp(i_{P\|Q}) \, i_{P\|Q}$$

$$\overset{(a)}{=} \int_{\{a\in\mathcal{A}\,:\, i_{P\|Q}(a)<0\}} \mathrm{d}Q(a) \, \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \, i_{P\|Q}(a)$$

$$\overset{(b)}{\leq} \beta_2 \int_{\{a\in\mathcal{A}\,:\, i_{P\|Q}(a)<0\}} \mathrm{d}Q(a) \, i_{P\|Q}(a)$$

$$\overset{(c)}{\leq} \beta_2 \log e \int_{\{a\in\mathcal{A}\,:\, i_{P\|Q}(a)<0\}} \mathrm{d}Q(a) \, \Big(\exp\big(i_{P\|Q}(a)\big) - 1 \Big)$$

$$\overset{(d)}{=} -\beta_2 \log e \int_{\mathcal{A}\setminus\mathcal{B}} \mathrm{d}Q(a) \, \Big(1 - \exp\big(i_{P\|Q}(a)\big) \Big)$$

$$\overset{(e)}{=} -\beta_2 \log e \int_{\mathcal{A}} \mathrm{d}Q(a) \, \Big(1 - \exp\big(i_{P\|Q}(a)\big) \Big)^{+}$$

$$\overset{(f)}{=} -\beta_2 \log e \cdot \mathbb{E}\Big[\big(1 - \exp(i_{P\|Q}(Y)) \big)^{+} \Big]$$

$$\overset{(g)}{=} -\frac{\beta_2 \log e}{2} \cdot |P - Q| \tag{18}$$

where equality (a) holds due to (4), (14) and since the integrand is zero if $i_{P\|Q} = 0$, inequality (b) follows from the definition of β_2 in (12) and since $i_{P\|Q}$ is negative over the domain of integration, inequality (c) holds since the inequality $x \leq \log e \big(\exp(x) - 1 \big)$ is satisfied for all $x \in \mathbb{R}$, equalities (d) and (e) follow from the definition of the set $\mathcal{B}$ in (14), equality (f) holds since $Y \sim Q$, and equality (g) follows from [25, Eq. (15)].

Inequality (13) finally follows by combining (15), (17) and (18). ∎

B. Example for the Refined Inequality in Theorem 1

We exemplify in the following the improvement obtained by (13), in comparison to (9), due to the introduction of the additional parameter β_2 in (12). 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 (9) and (13) coincide.

Let P and Q be two probability measures, defined on a set $\mathcal{A}$, where $P \ll Q$ and assume that

$$1 - \eta \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \leq 1 + \eta, \quad \forall\, a \in \mathcal{A} \tag{19}$$

for a fixed constant $\eta \in (0, 1)$.

In (13), one can replace β_1 and β_2 with lower bounds on these constants. From (12), we have $\beta_1 \geq \frac{1}{1+\eta}$ and $\beta_2 \geq 1 - \eta$, and it follows from (13) that

$$D(P\|Q) \leq \frac{1}{2} \left(\frac{(1+\eta)\,\log(1+\eta)}{\eta} - (1-\eta)\log e \right) |P - Q|$$

$$\leq \frac{1}{2}\Big((1+\eta)\log e - (1-\eta)\log e \Big) |P - Q|$$

$$= \big(\eta \log e\big)|P - Q|. \tag{20}$$

From (19)

$$\left|\exp\left(i_{P\|Q}(a)\right) - 1\right| \leq \eta, \quad \forall\, a \in \mathcal{A}$$

so, from (6), the total variation distance satisfies (recall that $Y \sim Q$)

$$|P - Q| = \mathbb{E}\left[\left|\exp\left(i_{P\|Q}(Y)\right) - 1\right|\right] \leq \eta.$$

Combining the last inequality with (20) gives that

$$D(P\|Q) \leq \eta^2 \log e, \quad \forall\, \eta \in (0, 1). \tag{21}$$

For comparison, it follows from (9) (see [25, Theorem 7]) that

$$\begin{aligned}
D(P\|Q) &\leq \frac{\log \frac{1}{\beta_1}}{2(1 - \beta_1)} \cdot |P - Q| \\
&\leq \frac{(1 + \eta)\,\log(1 + \eta)}{2\eta} \cdot |P - Q| \\
&\leq \frac{1}{2}\,(1 + \eta)\log(1 + \eta) \\
&\leq \left(\frac{\log e}{2}\right) \eta(1 + \eta).
\end{aligned} \tag{22}$$

Let $\eta \approx 0$. The upper bound on the relative entropy in (22) scales like η whereas the tightened bound in (21) scales like η^2. The scaling in (21) is correct, as it follows from Pinsker's inequality. For example, consider the probability measures defined on a two-element set $\mathcal{A} = \{a, b\}$ with

$$P(a) = Q(b) = \frac{1}{2} - \frac{\eta}{4}, \quad P(b) = Q(a) = \frac{1}{2} + \frac{\eta}{4}.$$

Condition (19) is satisfied for $\eta \approx 0$, and Pinsker's inequality yields that

$$D(P\|Q) \geq \left(\frac{\log e}{2}\right) \eta^2 \tag{23}$$

so the ratio of the upper and lower bounds in (21) and (23) is 2, and both provide the true quadratic scaling in η whereas the weaker upper bound in (22) scales linearly in η for $\eta \approx 0$.

C. Another Lower Bound on the Total Variation Distance

The following lower bound on the total variation distance is based on the distribution of the relative information, and it improves the lower bounds in [15, Eq. (2.3.18)], [22, Lemma 7] and [25, Theorem 8] by modifying the proof of the latter theorem in [25]. Besides of improving the tightness of the bound, the motivation for the derivation of the following lower bound is that it is achieved under some conditions.

Theorem 2: If P and Q are mutually absolutely continuous probability measures, then

$$|P - Q| \geq \sup_{\eta > 0}\left\{\left(1 - \exp(-\eta)\right)\left(\mathbb{P}\left[i_{P\|Q}(X) \geq \eta\right] + \exp(\eta)\,\mathbb{P}\left[i_{P\|Q}(X) \leq -\eta\right]\right)\right\} \tag{24}$$

where $X \sim P$. This lower bound is attained if P and Q are probability measures on a 2-element set $\mathcal{A} = \{a, b\}$ where, for an arbitrary $\eta > 0$,

$$P(a) = \frac{\exp(\eta) - 1}{2\sinh(\eta)}, \quad Q(a) = \frac{1 - \exp(-\eta)}{2\sinh(\eta)}. \tag{25}$$

Proof: Since $P \ll\gg Q$, we have

$$|P - Q| = \mathbb{E}\big[\big|1 - \exp\big(-i_{P\|Q}(X)\big)\big|\big]$$
$$\geq \mathbb{E}\big[\big|1 - \exp\big(-i_{P\|Q}(X)\big)\big| \, 1\big\{\big|i_{P\|Q}(X)\big| \geq \eta\big\}\big], \quad \forall \eta > 0$$

where $1\{\cdot\}$ is the indicator function of the specified event (it is equal to 1 if the event occurs, and it is zero otherwise). At this point we deviate from the proof of [25, Theorem 8], and write

$$|P - Q| \geq \mathbb{E}\big[\big|1 - \exp\big(-i_{P\|Q}(X)\big)\big| \, 1\big\{i_{P\|Q}(X) \geq \eta\big\}\big]$$
$$+ \mathbb{E}\big[\big|1 - \exp\big(-i_{P\|Q}(X)\big)\big| \, 1\big\{i_{P\|Q}(X) \leq -\eta\big\}\big]$$
$$\overset{(a)}{\geq} \big(1 - \exp(-\eta)\big) \, \mathbb{E}\big[1\big\{i_{P\|Q}(X) \geq \eta\big\}\big] + \big(\exp(\eta) - 1\big) \, \mathbb{E}\big[1\big\{i_{P\|Q}(X) \leq -\eta\big\}\big]$$
$$= \big(1 - \exp(-\eta)\big) \, \Big(\mathbb{P}\big[i_{P\|Q}(X) \geq \eta\big] + \exp(\eta) \, \mathbb{P}\big[i_{P\|Q}(X) \leq -\eta\big]\Big) \tag{26}$$

where step (a) follows from the inequality $\big|1 - \exp(-z)\big| \geq 1 - \exp(-\eta)$ if $z \geq \eta$, and $\big|1 - \exp(-z)\big| \geq \exp(\eta) - 1$ if $z \leq -\eta$. Taking the supremum on the right-hand side of (26), w.r.t. the free parameter $\eta > 0$, gives the lower bound on $|P - Q|$ in (24).

The condition (25) for the tightness of the lower bound in (24) follows from the fact that, for an arbitrary $\eta > 0$, we have $\log\big(\frac{P(a)}{Q(a)}\big) = \eta$ and $\log\big(\frac{1-P(a)}{1-Q(a)}\big) = -\eta$. This yields that the inequalities in the derivation of the lower bound (24) turn to be satisfied with equalities. ∎

Remark 1: One can further tighten the lower bound in (24) by writing, for arbitrary $\eta_1, \eta_2 > 0$,

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

and proceeding similarly to (26) to get the following lower bound on the total variation distance:

$$|P - Q| \geq \sup_{\eta_1, \eta_2 > 0} \bigg\{ \big(1 - \exp(-\eta_1)\big) \, \Big(\mathbb{P}\big[i_{P\|Q}(X) \geq \eta_1\big]$$
$$+ \Big(\frac{\exp(\eta_2) - 1}{1 - \exp(-\eta_1)}\Big) \, \mathbb{P}\big[i_{P\|Q}(X) \leq -\eta_2\big]\Big) \bigg\}.$$

This lower bound is achieved if P and Q are probability measures on a 2-element set $\mathcal{A} = \{a, b\}$ where, for an arbitrary $\eta_1, \eta_2 > 0$,

$$P(a) = \frac{\exp(\eta_1) - \exp(\eta_1 - \eta_2)}{\exp(\eta_1) - \exp(-\eta_2)}, \quad Q(a) = \frac{1 - \exp(-\eta_2)}{\exp(\eta_1) - \exp(-\eta_2)} \tag{27}$$

which implies that $\log\big(\frac{P(a)}{Q(a)}\big) = \eta_1$ and $\log\big(\frac{1-P(a)}{1-Q(a)}\big) = -\eta_2$. Condition (27) is specialized to the condition in (25) when $\eta_1 = \eta_2 = \eta > 0$.

III. A REVERSE PINSKER INEQUALITY FOR PROBABILITY MEASURES ON A FINITE SET

The present section introduces a strengthened version of inequality (10) (see Theorem 3) as a reverse Pinsker inequality for probability measures on a finite set, followed by a discussion and an example.

A. Main Result and Proof

Theorem 3: Let P and Q be probability measures defined on a common finite set $\mathcal{A}$, and assume that Q is strictly positive on $\mathcal{A}$. Then, the following inequality holds:

$$D(P\|Q) \leq \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}}\right) - \frac{\beta_2 \log e}{2} \cdot |P-Q|^2 \tag{28}$$

where

$$Q_{\min} \triangleq \min_{a \in \mathcal{A}} Q(a) > 0, \quad \beta_2 \triangleq \min_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \in [0,1]. \tag{29}$$

Remark 2: The upper bound on the relative entropy in Theorem 3 improves the bound in (10). The improvement in (28) is demonstrated as follows: let $V \triangleq |P-Q|$, then the RHS of (28) satisfies

$$\log\left(1 + \frac{V^2}{2Q_{\min}}\right) - \frac{\beta_2 \log e}{2} \cdot V^2 \leq \log\left(1 + \frac{V^2}{2Q_{\min}}\right) \leq \frac{V^2 \log e}{2Q_{\min}} \leq \frac{V^2 \log e}{Q_{\min}}.$$

Hence, the upper bound on $D(P\|Q)$ in Theorem 3 can be loosened to (10).

Proof: Theorem 3 is proved by obtaining upper and lower bounds on the χ^2-divergence from P to Q

$$\chi^2(P,Q) \triangleq \sum_{a \in \mathcal{A}} \frac{(P(a) - Q(a))^2}{Q(a)} = \sum_{a \in \mathcal{A}} \frac{P(a)^2}{Q(a)} - 1. \tag{30}$$

A lower bound follows by invoking Jensen's inequality:

$$\begin{aligned}
\chi^2(P,Q) &= \sum_{a \in \mathcal{A}} \frac{P(a)^2}{Q(a)} - 1 \\
&= \sum_{a \in \mathcal{A}} P(a) \exp\left(\log \frac{P(a)}{Q(a)}\right) - 1 \\
&\geq \exp\left(\sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)}\right) - 1 \\
&= \exp\big(D(P\|Q)\big) - 1. \tag{31}
\end{aligned}$$

A refined version of (31) is derived in the following. The starting point of its derivation relies on a refined version of Jensen's inequality from [5, Theorem 1], which enables to get the inequality

$$\min_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \cdot D(Q\|P) \leq \log\big(1 + \chi^2(P,Q)\big) - D(P\|Q) \leq \max_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \cdot D(Q\|P). \tag{32}$$

Inequality (32) is proved in the appendix. From the LHS of (32) and the definition of β_2 in (29), we have

$$\begin{aligned}
\chi^2(P,Q) &\geq \exp\Big(D(P\|Q) + \beta_2\, D(Q\|P)\Big) - 1 \\
&\geq \exp\left(D(P\|Q) + \frac{\beta_2 \log e}{2} \cdot |P-Q|^2\right) - 1 \tag{33}
\end{aligned}$$

where the last inequality relies on Pinsker's lower bound on $D(Q\|P)$. Inequality (33) refines the lower bound in (31) since $\beta_2 \in [0,1]$, and it coincides with (31) in the worst case where $\beta_2 = 0$.

An upper bound on $\chi^2(P,Q)$ is derived as follows:

$$\chi^2(P,Q) = \sum_{a \in \mathcal{A}} \frac{(P(a) - Q(a))^2}{Q(a)}$$

$$\leq \frac{\sum_{a \in \mathcal{A}}(P(a) - Q(a))^2}{\min_{a \in \mathcal{A}} Q(a)}$$

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

$$= \frac{|P - Q| \ \max_{a \in \mathcal{A}}|P(a) - Q(a)|}{Q_{\min}} \tag{34}$$

and, from (3),

$$|P - Q| \geq 2 \max_{a \in \mathcal{A}}|P(a) - Q(a)| \tag{35}$$

since, for every $a \in \mathcal{A}$, the 1-element set $\{a\}$ is included in the σ-algebra $\mathcal{F}$. Combining (34) and (35) gives that

$$\chi^2(P,Q) \leq \frac{|P - Q|^2}{2Q_{\min}}. \tag{36}$$

Inequality (28) finally follows from the bounds on the χ^2-divergence in (33) and (36). ∎

Corollary 1: Under the same setting as in Theorem 3, we have

$$D(P\|Q) \leq \log\left(1 + \frac{|P - Q|^2}{2Q_{\min}}\right). \tag{37}$$

Proof: This inequality follows from (28) and since $\beta_2 \geq 0$. ∎

B. Discussion

In the following, we discuss Theorem 3 and its proof, and link it to some related results.

Remark 3: The combination of (31) with the second line of (34), without further loosening the upper bound on the χ^2-divergence as is done in the third line of (34) and inequality (35), gives the following tighter upper bound on the relative entropy in terms of the Euclidean norm $|P - Q|_2$:

$$D(P\|Q) \leq \log\left(1 + \frac{|P - Q|_2^2}{Q_{\min}}\right). \tag{38}$$

This improves the upper bound on the relative entropy in the proofs of Property 4 of [23, Lemma 7] and [13, Appendix D]:

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

Furthermore, avoiding the use of Jensen's inequality in (31), gives the equality (see [6, Eq. (6)])

$$\chi^2(P,Q) = \exp\big(D_2(P\|Q)\big) - 1 \tag{39}$$

whose combination with the second line of (34) gives

$$D_2(P\|Q) \leq \log\left(1 + \frac{|P - Q|_2^2}{Q_{\min}}\right). \tag{40}$$

Inequality (40) improves the tightness of inequality (38). Note that (40) is satisfied with equality when Q is an equiprobable distribution over a finite set.

Remark 4: Inequality (31) improves the lower bound on the χ^2-divergence in [4, Lemma 6.3] which states that $\chi^2(P,Q) \geq D(P\|Q)$; this improvement also follows from [6, Eqs. (6), (7)].

Remark 5: The upper bound on the relative entropy in (28) involves the parameter $\beta_2 \in [0,1]$ as defined in (29). A non-trivial lower bound on β_2 can be used in conjunction with (28) for improving the upper bound in Corollary 1. We derive in the following a lower bound on β_2 for a given probability measure Q and a given total variation distance $|P-Q|$, which can be used in conjunction with (28), to get an upper bound on the relative entropy $D(P\|Q)$. We have

$$\beta_2 = \min_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \geq \frac{P_{\min}}{Q_{\max}}$$

where

$$P_{\min} \triangleq \min_{a \in \mathcal{A}} P(a), \quad Q_{\max} \triangleq \max_{a \in \mathcal{A}} Q(a).$$

Note that, if $|P-Q| < Q_{\min}$ then $P_{\min} \geq Q_{\min} - |P-Q| > 0$. Let $(x)^+ \triangleq \max\{x, 0\}$, then

$$\beta_2 \geq \frac{\left(Q_{\min} - |P-Q|\right)^+}{Q_{\max}}. \tag{41}$$

Remark 6: In an attempt to extend the concept of proof of Theorem 3 to general probability measures, we have

$$\begin{aligned}
\chi^2(P,Q) &= \int_{\mathcal{A}} \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q \\
&= \mathbb{E}\left[\left(\exp\left(i_{P\|Q}(Y)\right) - 1\right)^2\right] \qquad (Y \sim Q) \\
&\leq \sup_{a \in \mathcal{A}} \left|\exp\left(i_{P\|Q}(a)\right) - 1\right| \cdot \mathbb{E}\left[\left|\exp\left(i_{P\|Q}(Y)\right) - 1\right|\right] \\
&\overset{(a)}{=} \sup_{a \in \mathcal{A}} \left|\exp\left(i_{P\|Q}(a)\right) - 1\right| \cdot |P-Q| \\
&= \sup_{a \in \mathcal{A}} \left|\frac{\mathrm{d}P}{\mathrm{d}Q}(a) - 1\right| \cdot |P-Q| \tag{42}
\end{aligned}$$

where equality (a) holds due to (6). Let $\beta_1, \beta_2 \in [0,1]$ be defined as in Theorem 1 (see (12)). Since we have $\beta_2 \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \leq \beta_1^{-1}$ for all $a \in \mathcal{A}$ then

$$\sup_{a \in \mathcal{A}} \left|\frac{\mathrm{d}P}{\mathrm{d}Q}(a) - 1\right| \leq \max\{\beta_1^{-1} - 1, 1 - \beta_2\}. \tag{43}$$

A combination of (42) and (43) leads to the following upper bound on the χ^2-divergence:

$$\chi^2(P,Q) \leq \max\{\beta_1^{-1} - 1, 1 - \beta_2\} \cdot |P-Q|. \tag{44}$$

A combination of (39) (see [6, Eq. (6)]) and (44) gives

$$D_2(P\|Q) \leq \log\left(1 + \max\{\beta_1^{-1} - 1, 1 - \beta_2\} \cdot |P-Q|\right) \tag{45}$$

and since the Rényi divergence is monotonic non-decreasing in its order (see, e.g., [6, Theorem 3]) and $D(P\|Q) = D_1(P\|Q)$, it follows that

$$D(P\|Q) \leq \log\left(1 + \max\{\beta_1^{-1} - 1, 1 - \beta_2\} \cdot |P-Q|\right). \tag{46}$$

A comparison of the upper bound on the relative entropy in (46) and the bound of Theorem 1 in (13) yields that the latter bound is superior. Hence, the extension of the concept of proof of Theorem 3 to general probability measures does not improve the bound in Theorem 1.

Remark 7: The second inequality in (33) relies on Pinsker's inequality as a lower bound on $D(Q\|P)$. This lower bound can be slightly improved by invoking higher-order Pinsker's-type inequalities (see [9, Section 5] and references therein). In [9, Section 6], Gilardoni derived a lower bound on the relative entropy which is tight for both large and small total variation distances. Hence, the second inequality in (33) can instead rely on the inequality (see [9, Eq. (2)]):

$$D(Q\|P) \geq -\log\left(1 - \frac{|P - Q|}{2}\right) - \left(1 - \frac{|P - Q|}{2}\right)\log\left(1 + \frac{|P - Q|}{2}\right).$$

Note that although the latter lower bound on the relative entropy is tight for both large and small total variation distances, it is not uniformly tighter than Pinsker's inequality. For this reason and for the simplicity of the bound, we rely on Pinsker's inequality in the second inequality of (33). An exact parametrization of the minimum of the relative entropy in terms of the total variation distance was introduced in [7, Theorem 1], expressed in terms of hyperbolic functions; the bound, however, is not expressed in closed form in terms of the total variation distance.

Remark 8: A related problem to Theorem 3 has been recently studied in [1]. Consider an arbitrary probability measure Q, and an arbitrary $\varepsilon \in [0, 2]$. The problem studied in [1] is the characterization of $D^*(\varepsilon, Q)$, defined to be the infimum of $D(P\|Q)$ over all probability measures P that are at least ε-far away from Q in total variation, i.e.,

$$D^*(\varepsilon, Q) = \inf_{P:\, |P-Q| \geq \varepsilon} D(P\|Q), \quad \varepsilon \in [0, 2].$$

Note that $D(P\|Q) < \infty$ yields that $\mathrm{Supp}(P) \subseteq \mathrm{Supp}(Q)$. From Sanov's theorem (see [3, Theorem 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. Upper and lower bounds on $D^*(\varepsilon, Q)$ have been introduced in [1, Theorem 1], in terms of the balance coefficient $\beta \geq \frac{1}{2}$ that is defined as

$$\beta \triangleq \inf\left\{x \in \{Q(A)\colon A \in \mathcal{F}\}\colon x \geq \frac{1}{2}\right\}.$$

It has been demonstrated in [1, Theorem 1] that

$$D^*(\varepsilon, Q) = C\varepsilon^2 + O(\varepsilon^3) \tag{47}$$

where

$$\frac{1}{4(2\beta - 1)}\log\left(\frac{\beta}{1 - \beta}\right) \leq C \leq \frac{\log e}{8\beta(1 - \beta)}.$$

If the support of Q is a finite set $\mathcal{A}$, Theorem 3 and (41) yield that

$$D^*(\varepsilon, Q) \leq \log\left(1 + \frac{\varepsilon^2}{2Q_{\min}}\right) - \frac{\log e}{2} \cdot \frac{1}{Q_{\max}} \cdot \left(Q_{\min} - \varepsilon\right)^+.$$

Hence, it follows that $D^*(\varepsilon, Q) \leq C_1\varepsilon^2 + O(\varepsilon^3)$ where

$$C_1 = \frac{\log e}{2}\left(\frac{1}{Q_{\min}} - \frac{Q_{\min}}{Q_{\max}}\right).$$

Similarly to (47), the same quadratic scaling of $D^*(\varepsilon, Q)$ holds for small values of ε, but with different coefficients.

C. Example: Total Variation Distance From the Equiprobable Distribution

Let $\mathcal{A}$ be a finite set, and let U be the equiprobable probability measure on $\mathcal{A}$ (i.e., $U(a) = \frac{1}{|\mathcal{A}|}$ for every $a \in \mathcal{A}$). The relative entropy of an arbitrary distribution P on $\mathcal{A}$ with respect to the equiprobable distribution satisfies

$$D(P\|U) = \log|\mathcal{A}| - H(P). \tag{48}$$

From Pinsker's inequality (1), the following upper bound on the total variation distance holds:

$$|P - U| \leq \sqrt{\frac{2}{\log e} \cdot \big(\log|\mathcal{A}| - H(P)\big)}. \tag{49}$$

From [26, Theorem 2.51], for all probability measures P and Q,

$$|P - Q| \leq 2\sqrt{1 - \exp\big(-D(P\|Q)\big)}$$

which gives the second upper bound

$$|P - U| \leq 2\sqrt{1 - \frac{1}{|\mathcal{A}|} \cdot \exp\big(H(P)\big)}. \tag{50}$$

From Theorem 3 and (41), we have

$$D(P\|U) \leq \log\left(1 + \frac{|\mathcal{A}|}{2} \cdot |P - U|^2\right) - \left(\frac{|\mathcal{A}| \log e}{2}\right) \cdot \left(\frac{1}{|\mathcal{A}|} - |P - U|\right)^+ \cdot |P - U|^2.$$

A loosening of the latter bound by removing its second non-negative term on the RHS of this inequality, in conjunction with (48), leads to the following closed-form expression for the lower bound on the total variation distance:

$$|P - U| \geq \sqrt{2\left(\exp\big(-H(P)\big) - \frac{1}{|\mathcal{A}|}\right)}. \tag{51}$$

Let $H(P) = \beta \log|\mathcal{A}|$, so $\beta \in [0, 1]$. From (49), (50) and (51), it follows that

$$\sqrt{2\left[\left(\frac{1}{|\mathcal{A}|}\right)^\beta - \frac{1}{|\mathcal{A}|}\right]} \leq |P - U| \leq \min\left\{\sqrt{2(1 - \beta)\ln|\mathcal{A}|}, \; 2\sqrt{1 - |\mathcal{A}|^{\beta - 1}}\right\}. \tag{52}$$

As expected, if $\beta = 1$, both upper and lower bounds are equal to zero (since $D(P\|U) = 0$). The lower bound on the LHS of (52) improves the lower bound on the total variation distance which follows from (10):

$$|P - U| \geq \sqrt{\frac{(1 - \beta)\ln|\mathcal{A}|}{|\mathcal{A}|}} \tag{53}$$

For example, for a set of size $|\mathcal{A}| = 1024$ and $\beta = 0.5$, the improvement in the new lower bound on the total variation distance is from 0.0582 to 0.2461.

Note that if $\beta \to 0$ (i.e., P is far in relative entropy from the equiprobable distribution), and the set $\mathcal{A}$ stays fixed, the ratio between the upper and lower bounds in (52) tends to $\sqrt{2}$. On the other hand, in this case, the ratio between the upper and the looser lower bound in (53) tends to

$$2\sqrt{\frac{|\mathcal{A}| - 1}{\ln|\mathcal{A}|}},$$

which can be made arbitrarily large for a sufficiently large set $\mathcal{A}$.

IV. EXTENSION OF THEOREM 3 TO RÉNYI DIVERGENCES

The present section extends Theorem 3 to Rényi divergences of an arbitrary order $\alpha \in [0, \infty]$ (i.e., it relies on Theorem 3 to provide a generalization of the special case where $\alpha = 1$), and the use of this generalized inequality is exemplified.

A. Main Result and Proof

The following theorem provides a kind of a generalized reverse Pinsker inequality where the Rényi divergence of an arbitrary order $\alpha \in [0, \infty]$ is upper bounded in terms of the total variation distance for probability measures defined on a common finite set.

Theorem 4: Let P and Q be probability measures on a common finite set $\mathcal{A}$, and assume that P, Q are strictly positive. Let $\varepsilon \triangleq |P - Q|$ (recall that $\varepsilon \in [0, 2]$), $\varepsilon' \triangleq \min\{1, \varepsilon\}$, and

$$P_{\min} \triangleq \min_{a \in \mathcal{A}} P(a), \quad Q_{\min} \triangleq \min_{a \in \mathcal{A}} Q(a).$$

Then, the Rényi divergence of order $\alpha \in [0, \infty]$ satisfies

$$D_\alpha(P \| Q) \leq \begin{cases} \log\left(1 + \frac{\varepsilon}{2Q_{\min}}\right), & \text{if } \alpha \in (2, \infty] \\ \log\left(1 + \frac{\varepsilon \varepsilon'}{2Q_{\min}}\right), & \text{if } \alpha \in [1, 2] \\ \min\{f_1, f_2\}, & \text{if } \alpha \in \left(\frac{1}{2}, 1\right) \\ \min\left\{-2\log\left(1 - \frac{\varepsilon}{2}\right), f_1, f_2\right\}, & \text{if } \alpha \in \left[0, \frac{1}{2}\right] \end{cases} \tag{54}$$

where, for $\alpha \in [0, 1)$,

$$f_1 \equiv f_1(\alpha, P_{\min}, Q_{\min}, \varepsilon) \triangleq \left(\frac{\alpha}{1 - \alpha}\right)\left[\log\left(1 + \frac{\varepsilon^2}{2P_{\min}}\right) - \left(\frac{Q_{\min} \log e}{2}\right)\varepsilon^2\right], \tag{55}$$

$$f_2 \equiv f_2(P_{\min}, Q_{\min}, \varepsilon, \varepsilon') \triangleq \log\left(1 + \frac{\varepsilon \varepsilon'}{2Q_{\min}}\right) - \left(\frac{P_{\min} \log e}{2}\right)\varepsilon^2. \tag{56}$$

Proof: The Rényi divergence of order ∞ satisfies (see, e.g., [6, Theorem 6])

$$D_\infty(P \| Q) = \log\left(\operatorname{ess\,sup} \frac{P}{Q}\right).$$

Since, by assumption, the probability measures P and Q are defined on a common finite set $\mathcal{A}$

$$\begin{aligned} D_\infty(P \| Q) &= \log\left(\max_{a \in \mathcal{A}} \frac{P(a)}{Q(a)}\right) \\ &= \log\left(1 + \max_{a \in \mathcal{A}} \frac{P(a) - Q(a)}{Q(a)}\right) \\ &\leq \log\left(1 + \frac{\max_{a \in \mathcal{A}} |P(a) - Q(a)|}{\min_{a \in \mathcal{A}} Q(a)}\right) \\ &\leq \log\left(1 + \frac{|P - Q|}{2Q_{\min}}\right) \end{aligned} \tag{57}$$

where the last inequality follows from (35). Since the Rényi divergence of order $\alpha \in [0, \infty]$ is monotonic non-decreasing in α (see, e.g., [6, Theorem 3]), it follows from (57) that

$$D_\alpha(P\|Q) \le D_\infty(P\|Q) \le \log\left(1 + \frac{\varepsilon}{2Q_{\min}}\right), \quad \forall \alpha \in [0, \infty] \tag{58}$$

which proves the first line in (54) when the validity of the bound is restricted to $\alpha \in (2, \infty]$.

For proving the second line in (54), it is shown that the bound in (37) can be sharpened by replacing $D(P\|Q)$ on the LHS of (37) with the quadratic Rényi divergence $D_2(P\|Q)$ (note that $D_2(P\|Q) \ge D(P\|Q)$), leading to

$$D_2(P\|Q) \le \log\left(1 + \frac{|P - Q|^2}{2Q_{\min}}\right). \tag{59}$$

The strengthened inequality in (59), in comparison to (37), follows by replacing inequality (31) with the equality in (39). Combining (36) and (39) gives inequality (59), and

$$D_\alpha(P\|Q) \le D_2(P\|Q) \le \log\left(1 + \frac{\varepsilon^2}{2Q_{\min}}\right), \quad \forall \alpha \in [0, 2]. \tag{60}$$

The combination of (58) with (60) gives the second line in (54) (note that $\varepsilon\varepsilon' = \min\{\varepsilon, \varepsilon^2\}$) while the validity of the bound is restricted to $\alpha \in [1, 2]$.

For $\alpha \in (0, 1)$, $D_\alpha(P\|Q)$ satisfies the skew-symmetry property $D_\alpha(P\|Q) = \frac{\alpha}{1-\alpha}\cdot D_{1-\alpha}(Q\|P)$ (see, e.g., [6, Proposition 2]). Consequently, we have

$$
\begin{aligned}
D_\alpha(P\|Q) &= \left(\frac{\alpha}{1-\alpha}\right) D_{1-\alpha}(Q\|P) \\
&\le \left(\frac{\alpha}{1-\alpha}\right) D(Q\|P) \\
&\le \left(\frac{\alpha}{1-\alpha}\right)\left[\log\left(1 + \frac{\varepsilon^2}{2P_{\min}}\right) - \left(\frac{Q_{\min}\log e}{2}\right)\varepsilon^2\right], \quad \forall \alpha \in (0, 1)
\end{aligned}
\tag{61}
$$

where the first inequality holds since the Rényi divergence is monotonic non-decreasing in its order, and the second inequality follows from Theorem 3 which implies that

$$
\begin{aligned}
D(Q\|P) &\le \log\left(1 + \frac{\varepsilon^2}{2P_{\min}}\right) - \frac{\log e}{2}\cdot \min_{a \in \mathcal{A}} \frac{Q(a)}{P(a)}\cdot \varepsilon^2 \\
&\le \log\left(1 + \frac{\varepsilon^2}{2P_{\min}}\right) - \left(\frac{Q_{\min}\log e}{2}\right)\varepsilon^2.
\end{aligned}
$$

The third line in (54) follows from (58), (60) and (61) while restricting the validity of the bound to $\alpha \in \left(\frac{1}{2}, 1\right)$.

For proving the fourth line in (54), note that from (11) $D_{1/2}(P\|Q) = -2\log Z(P, Q)$ where $Z(P, Q) \triangleq \sum_{a \in \mathcal{A}} \sqrt{P(a)Q(a)}$ is the Bhattacharyya coefficient between P and Q [12]. The Bhattacharyya distance is defined as minus the logarithm of the Bhattacharyya coefficient, which is non-negative in general and it is zero if and only if $P = Q$ (since $0 \le Z(P, Q) \le 1$, and $Z(P, Q) = 1$ if and only if $P = Q$). Hence, the Rényi divergence of order $\frac{1}{2}$ is twice the Bhattacharyya distance. Based on the inequality $Z(P, Q) \ge 1 - \frac{|P-Q|}{2}$, which follows from [10, Example 6.2] (see also [21, Proposition 1]), we have

$$D_\alpha(P\|Q) \le D_{1/2}(P\|Q) \le -2\log\left(1 - \frac{\varepsilon}{2}\right), \quad \forall \alpha \in \left[0, \frac{1}{2}\right] \tag{62}$$

where $\varepsilon \triangleq |P - Q| \in [0, 2]$. Finally, the last case in (54) follows from (58), (60), (61) and (62). ∎

B. Example: Rényi Divergence for Multinomial Distributions

Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with $X_i \sim \text{Bernoulli}(p_i)$, and let $Y_1, Y_2, \ldots$ be independent Bernoulli random variables with $Y_i \sim \text{Bernoulli}(q_i)$ (assume w.l.o.g. that $q_i \leq \frac{1}{2}$). Let U_n and V_n be the partial sums $U_n = \sum_{i=1}^{n} X_i$ and $V_n = \sum_{i=1}^{n} Y_i$, and let P_{U_n}, P_{V_n} denote their multinomial distributions. For all $\alpha \in [0, 2]$ and $n \in \mathbb{N}$, we have

$$
D_\alpha(P_{U_n} \| P_{V_n})
$$
$$
\overset{(a)}{\leq} D_\alpha(P_{X_1, \ldots, X_n} \| P_{Y_1, \ldots, Y_n})
$$
$$
\overset{(b)}{=} \sum_{i=1}^{n} D_\alpha(P_{X_i} \| P_{Y_i})
$$
$$
\overset{(c)}{\leq} \log \left(1 + \frac{|P_{X_i} - P_{Y_i}|^2}{2 \left(P_{Y_i} \right)_{\min}} \right)
$$
$$
\overset{(d)}{=} \sum_{i=1}^{n} \log \left(1 + 2q_i \left(\frac{p_i}{q_i} - 1 \right)^2 \right) \tag{63}
$$

where inequality (a) follows from the data processing inequality for the Rényi divergence (see [6, Theorem 9]), equality (b) follows from the additivity property of the Rényi divergence under the independence assumption for $\{X_i\}$ and for $\{Y_i\}$ (see [6, Theorem 28]), inequality (c) follows from Theorem 4, and equality (d) holds since $|P_{X_i} - P_{Y_i}| = 2|p_i - q_i|$ for Bernoulli random variables, and $(P_{Y_i})_{\min} = \min\{q_i, 1 - q_i\} = q_i \ (q_i \leq \frac{1}{2})$. Similarly, for all $\alpha > 2$ and $n \in \mathbb{N}$,

$$
D_\alpha(P_{U_n} \| P_{V_n}) \leq \sum_{i=1}^{n} \log \left(1 + 2 \left| \frac{p_i}{q_i} - 1 \right| \right). \tag{64}
$$

The only difference in the derivation of (64) is in inequality (c) of (63) where the bound in the first line of (54) is used this time.

Let $\{\varepsilon_n\}_{n=1}^{\infty}$ be a non-negative sequence such that

$$
(1 - \varepsilon_n) q_n \leq p_n \leq (1 + \varepsilon_n) q_n, \quad \forall n \in \mathbb{N}
$$

and

$$
\sum_{n=1}^{\infty} \varepsilon_n^2 < \infty.
$$

Then, from (63), it follows that $D_\alpha(P_{U_n} \| P_{V_n}) \leq K_1$ for all $\alpha \in [0, 2]$ and $n \in \mathbb{N}$ where

$$
K_1 \triangleq \sum_{n=1}^{\infty} \log \left(1 + \varepsilon_n^2 \right) < \infty.
$$

Furthermore, if $\sum_{n=1}^{\infty} \varepsilon_n < \infty$, it follows from (64) that $D_\alpha(P_{U_n} \| P_{V_n}) \leq K_2$ for all $\alpha > 2$ and $n \in \mathbb{N}$ where

$$
K_2 \triangleq \sum_{n=1}^{\infty} \log \left(1 + 2\varepsilon_n \right) < \infty.
$$

Note that although $D_\alpha(P_{X_i} \| P_{Y_i})$ in equality (b) of (63) is equal to the binary Rényi divergence

$$
d_\alpha(p_i \| q_i) \triangleq \begin{cases} \left(\frac{1}{\alpha - 1} \right) \log \left(p_i^\alpha q_i^{1-\alpha} + (1 - p_i)^\alpha (1 - q_i)^{1-\alpha} \right), & \text{if } \alpha \in (0, 1) \cup (1, \infty) \\ p_i \log \left(\frac{p_i}{q_i} \right) + (1 - p_i) \log \left(\frac{1 - p_i}{1 - q_i} \right), & \text{if } \alpha = 1 \end{cases}
$$

the reason for the use of the upper bounds in step (c) of (63) and (64) is to state sufficient conditions, in terms of $\{\varepsilon_n\}_{n=1}^{\infty}$, for the boundedness of the Rényi divergence $D_\alpha(P_{U_n} \| P_{V_n})$.

V. THE EXPONENTIAL DECAY OF THE PROBABILITY FOR A NON-TYPICAL SEQUENCE

Let $U^N = (U_1, \ldots, U_N)$ be a sequence of i.i.d. symbols that are emitted by a memoryless and stationary source with distribution Q and a finite alphabet $\mathcal{A}$. Let $|\mathcal{A}| = r < \infty$ denote the cardinality of the source alphabet, and assume that all symbols are emitted with positive probability (i.e., $Q_{\min} \triangleq \min_{a \in \mathcal{A}} Q(a) > 0$). The empirical probability distribution of the emitted sequence $\hat{P}_{U^N}$ is given by

$$\hat{P}_{U^N}(a) \triangleq \frac{1}{N} \sum_{k=1}^{N} 1\{U_k = a\}, \quad \forall a \in \mathcal{A}.$$

For an arbitrary $\delta > 0$, let the δ-typical set be defined as

$$T_Q(\delta) \triangleq \left\{ u^N \in \mathcal{A}^N : \left| \hat{P}_{u^N}(a) - Q(a) \right| < \delta \, Q(a), \quad \forall a \in \mathcal{A} \right\}, \tag{65}$$

i.e., the empirical distribution of every symbol in an N-length δ-typical sequence deviates from the true distribution of this symbol by a fraction of less than δ. Consequently, the complementary of (65) is given by

$$T_Q(\delta)^c = \left\{ u^N \in \mathcal{A}^N : \exists a \in \mathcal{A}, \ \left| \hat{P}_{u^N}(a) - Q(a) \right| \geq \delta \, Q(a) \right\}.$$

From Sanov's theorem (see [3, Theorem 11.4.1]), the asymptotic exponential decay of the probability that a sequence U^N is not δ-typical, for a specified $\delta > 0$, is given by

$$\lim_{N \to \infty} -\frac{1}{N} \log Q^N \big(T_Q(\delta)^c \big) = \min_{P \in \mathcal{P}_Q} D(P \| Q) \tag{66}$$

where

$$\mathcal{P}_Q \triangleq \left\{ P \text{ is a probability measure on } (\mathcal{A}, \mathcal{F}) : \exists a \in \mathcal{A}, \ |P(a) - Q(a)| \geq \delta \, Q(a) \right\}. \tag{67}$$

We obtain in the following explicit upper and lower bounds on the exponential decay rate on the RHS of (66). The emphasis is on the upper bound, which is based on Theorem 3, and we first introduce the lower bound for completeness. The derivation of the lower bound is similar to the analysis in [14, Section 4]; note, however, that there is a difference between the δ-typicality in [14, Eq. (19)] and the way it is defined in (65). The probability-dependent refinement of Pinsker's inequality (see [14, Theorem 2.1]) states that

$$D(P \| Q) \geq \varphi(\pi_Q) \, |P - Q|^2 \tag{68}$$

where

$$\pi_Q \triangleq \max_{A \in \mathcal{F}} \min \big\{ Q(A), 1 - Q(A) \big\} \leq \frac{1}{2} \tag{69}$$

and

$$\varphi(p) = \begin{cases} \frac{1}{4(1-2p)} \log\left(\frac{1-p}{p}\right), & \text{if } p \in \left[0, \frac{1}{2}\right), \\ \frac{\log e}{2}, & \text{if } p = \frac{1}{2} \end{cases} \tag{70}$$

is a monotonic decreasing and continuous function. Hence, $\varphi(\pi_Q) \geq \frac{\log e}{2}$, and (68) forms a probability-dependent refinement of Pinsker's inequality [14]. From (67) and (68), we have

$$\min_{P \in \mathcal{P}_Q} D(P\|Q)$$

$$\geq \varphi(\pi_Q) \min_{P \in \mathcal{P}_Q} |P - Q|^2$$

$$= \varphi(\pi_Q) \left(\min_{a \in \mathcal{A}} \delta\, Q(a) \right)^2$$

$$= \varphi(\pi_Q)\, Q_{\min}^2\, \delta^2 \triangleq E_{\mathrm{L}} \tag{71}$$

$$\geq \left(\frac{Q_{\min}^2 \log e}{2} \right) \delta^2 \tag{72}$$

where the transition from (71) to (72) follows from the global lower bound on $\varphi(\pi_Q)$.

We derive in the following an upper bound on the asymptotic exponential decay rate in (66):

$$\min_{P \in \mathcal{P}_Q} D(P\|Q)$$

$$\overset{(a)}{\leq} \min_{P \in \mathcal{P}_Q} \left\{ \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}} \right) \right\}$$

$$= \log\left(1 + \frac{\left(\min_{P \in \mathcal{P}_Q} |P-Q| \right)^2}{2Q_{\min}} \right)$$

$$\overset{(b)}{=} \log\left(1 + \frac{\left(\min_{a \in \mathcal{A}} (\delta\, Q(a)) \right)^2}{2Q_{\min}} \right)$$

$$= \log\left(1 + \frac{Q_{\min} \delta^2}{2} \right) \triangleq E_{\mathrm{U}} \tag{73}$$

where inequality (a) follows from (37), and equality (b) follows from (67).

The ratio between the upper and lower bounds on the asymptotic exponent in (66), as given in (71) and (73) respectively, satisfies

$$1 \leq \frac{E_{\mathrm{U}}}{E_{\mathrm{L}}}$$

$$= \frac{1}{Q_{\min}} \cdot \frac{\log e}{2\, \varphi(\pi_Q)} \cdot \frac{\log\left(1 + \frac{Q_{\min} \delta^2}{2} \right)}{\frac{\log e \cdot Q_{\min} \delta^2}{2}} \tag{74}$$

$$\leq \frac{1}{Q_{\min}}$$

where inequality (74) follows from the fact that the second and third multiplicands in (74) are both less than or equal to 1. Note that both bounds in (71) and (73) scale like δ^2 for $\delta \approx 0$.

APPENDIX: A PROOF OF INEQUALITY (32)

This appendix proves inequality (32), which provides upper and lower bounds on the difference $\log(1 + \chi^2(P,Q)) - D(P\|Q)$ in terms of the dual relative entropy $D(Q\|P)$. To this end, we first prove a new inequality relating f-divergences [21], and the bounds in (32) then follow as a special case.

Recall the following definition of an f-divergence:

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

$$D_f(P\|Q) \triangleq \sum_{a \in \mathcal{A}} Q(a) f\left(\frac{P(a)}{Q(a)}\right) \tag{75}$$

with the convention that

$$0f\left(\frac{0}{0}\right) = 0, \quad f(0) = \lim_{t \to 0^+} f(t),$$

$$0f\left(\frac{b}{0}\right) = \lim_{t \to 0^+} tf\left(\frac{b}{t}\right) = b \lim_{u \to \infty} \frac{f(u)}{u}, \quad \forall\, b > 0. \tag{76}$$

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

$$\min_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \cdot D_f(P\|Q) \leq -D_g(P\|Q) - f\big(1 + \chi^2(P, Q)\big) \leq \max_{a \in \mathcal{A}} \frac{P(a)}{Q(a)} \cdot D_f(P\|Q). \tag{77}$$

Proof: Let $\mathcal{A} = \{a_1, \ldots, a_n\}$, and $\underline{u} = (u_1, \ldots, u_n) \in \mathbb{R}_+^n$ be an arbitrary n-tuple with positive entries. Define

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

$$J_n(f, \underline{u}, Q) \triangleq \sum_{i=1}^{n} Q(a_i) f(u_i) - f\left(\sum_{i=1}^{n} Q(a_i) u_i\right). \tag{78}$$

The following refinement of Jensen's inequality has been introduced in [5, Theorem 1] for a convex function $f \colon (0, \infty) \to \mathbb{R}$:

$$\min_{i \in \{1, \ldots, n\}} \frac{P(a_i)}{Q(a_i)} \cdot J_n(f, \underline{u}, Q) \leq J_n(f, \underline{u}, P) \leq \max_{i \in \{1, \ldots, n\}} \frac{P(a_i)}{Q(a_i)} \cdot J_n(f, \underline{u}, Q). \tag{79}$$

Let $u_i \triangleq \frac{P(a_i)}{Q(a_i)}$ for $i \in \{1, \ldots, n\}$. Calculation of (78) gives that

$$J_n(f, \underline{u}, Q) = \sum_{i=1}^{n} Q(a_i) f\left(\frac{P(a_i)}{Q(a_i)}\right) - f\left(\sum_{i=1}^{n} Q(a_i) \cdot \frac{P(a_i)}{Q(a_i)}\right)$$

$$= \sum_{a \in \mathcal{A}} Q(a) f\left(\frac{P(a)}{Q(a)}\right) - f(1)$$

$$= D_f(P\|Q), \tag{80}$$

$$J_n(f, \underline{u}, P) = \sum_{i=1}^{n} P(a_i) f\left(\frac{P(a_i)}{Q(a_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(a_i)^2}{Q(a_i)}\right)$$

$$\overset{(a)}{=} -\sum_{i=1}^{n} Q(a_i) g\left(\frac{P(a_i)}{Q(a_i)}\right) - f\left(\sum_{i=1}^{n} \frac{P(a_i)^2}{Q(a_i)}\right)$$

$$\overset{(b)}{=} -D_g(P\|Q) - f\big(1 + \chi^2(P, Q)\big) \tag{81}$$

where equality (a) holds by the definition of g, and equality (b) follows from equalities (30) and (75). The substitution of (80) and (81) into (79) gives inequality (77). ∎

As a consequence of Proposition 1, we prove inequality (32). Let $f(t) = -\log(t)$ for $t > 0$. The function $f : (0, \infty) \to \mathbb{R}$ is convex with $f(1) = 0$, and $g(t) = -tf(t) = t\log(t)$ for $t > 0$ is also convex with $g(1) = 0$. Inequality (32) follows by substituting f, g into (77) where $D_f(P||Q) = D(Q||P)$ and $D_g(P||Q) = D(P||Q)$. Inequality (32) also holds in the case where P is not strictly positive on $\mathcal{A}$ with the convention in (76) where $0\log 0 = \lim_{t \to 0^+} g(t) = 0$.

ACKNOWLEDGMENT

REFERENCES

[1] D. Berend, P. Harremoës and A. Kontorovich, "Minimum KL-divergence on complements of L_1 balls," *IEEE Trans. on Information Theory*, vol. 60, no. 6, pp. 3172–3177, June 2014.

[2] G. Böcherer and B. C. Geiger, "Optimal quantization for distribution synthesis," March 2015. [Online]. Available: http://arxiv.org/abs/1307.6843.

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

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

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

[6] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

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

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

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

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

[11] S. W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 5906–5929, December 2010.

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

[13] V. Kostina and S. Verdú, "Channels with cost constraints: strong converse and dispersion," to appear in the *IEEE Trans. on Information Theory*, vol. 61, no. 5, May 2015.

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

[15] M. S. Pinsker, *Information and Information Stability of Random Variables and Random Processes*, San-Fransisco: Holden-Day, 1964, originally published in Russian in 1960.

[16] V. V. Prelov, "On inequalities between mutual information and variation," *Problems of Information Transmission*, vol. 43, no. 1, pp. 12–23, March 2007.

[17] V. V. Prelov and E. C. van der Meulen, "Mutual information, variation, and Fano's inequality," *Problems of Information Transmission*, vol. 44, no. 3, pp. 185–197, September 2008.

[18] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *Journal of Machine Learning Research*, vol. 12, no. 3, pp. 731–817, March 2011.

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

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

[21] I. Sason, "Tight bounds on symmetric divergence measures and a new inequality relating f-divergences," accepted to the *IEEE 2015 Information Theory Workshop*, Jerusalem, Israel, April 26–May 1, 2015. [Online]. Available at `http://arxiv.org/abs/1502.06428`.

[22] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. on Information Theory*, vol. 42, no. 1, pp. 63–86, January 1996.

[23] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. on Information Theory*, vol. 59, no. 11, pp. 7041–7051, November 2013.

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

[25] S. Verdú, "Total variation distance and the distribution of the relative information," *Proceedings of the 2014 Information Theory and Applications (ITA) Workshop*, pp. 499–501, San-Diego, California, USA, February 2014.

[26] S. Verdú, *Information Theory*, in preparation.

[27] Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. on Information Theory*, vol. 53, no. 9, pp. 3280–3282, September 2007.