Entropy Bounds for Discrete Random Variables via Maximal Coupling

Igal Sason, *Senior Member, IEEE*

Abstract

This paper derives new bounds on the difference of the entropies of two discrete random variables in terms of the local and total variation distances between their probability mass functions. The derivation of the bounds relies on maximal coupling, and they apply to discrete random variables which are defined over finite or countably infinite alphabets. Loosened versions of these bounds are demonstrated to reproduce some previously reported results. The use of the new bounds is exemplified for the Poisson approximation, where bounds on the local and total variation distances follow from Stein's method.

Keywords

Coupling, entropy, local distance, Stein's method, total variation distance.

©2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The manuscript was submitted to the *IEEE Trans. on Information Theory* in September 20, 2012, and accepted in July 18, 2013. This research work was supported by the Israel Science Foundation (grant no. 12/12). The material in this paper was presented in part at the 2013 International Symposium on Information Theory (ISIT '13), Istanbul, Turkey, July 2013.

Communicated by Ioannis Kontoyiannis, Associate Editor At Large.

The author is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. His e-mail address is sason@ee.technion.ac.il.

I. Introduction

The question of quantifying the continuity (or lack of it) of entropy, with respect to natural topologies on discrete probability distributions is fundamental. This question has been studied in the literature for the topology induced by the total variation distance, and there it is well known that the entropy is continuous when the alphabet is finite, but it is not necessarily continuous when the alphabet is countably infinite. The interplay between the difference of the entropies of two discrete random variables and their total variation distance has been extensively studied (see, e.g., [8, Theorem 17.3.3], [9], [10, Lemma 1], [15]–[17], [21], [25]–[29], [33], [37], [38]).

New bounds on the difference of the entropies of two discrete random variables are derived in this work. The bounds apply to random variables with finite or countably infinite alphabet, and they improve some previously reported bounds. The derivation of the new bounds relies on the notion of *maximal coupling*, which is also known to be useful for the derivation of error bounds via Stein's method (see, e.g., [31, Chapter 2] and [32]). Stein's method also serves to exemplify the use of the new bounds in the context of the Poisson approximation. The link between Stein's method and information theory was pioneered in [6] in the context of the compound Poisson approximation, and a recent work [22] (that was done independently and in parallel to this work) further links between information theory and Stein's method for discrete probability distributions.

To set definitions and notation, we introduce essential terms that serve to derive the new bounds in this paper.

Definition 1: A *coupling* of a pair of two random variables (X, Y) is a pair of two random variables $(\hat{X}, \hat{Y})$ with the same marginal probability distributions as of (X, Y).

Definition 2: For a pair of random variables (X, Y), a coupling $(\hat{X}, \hat{Y})$ is called a *maximal coupling* if $\mathbb{P}(\hat{X} = \hat{Y})$ gets its maximal value among all the couplings of (X, Y).

Definition 3: Let X and Y be discrete random variables that take values in a set $\mathcal{A}$, and let P_X and P_Y be their respective probability mass functions. The *local distance* and *total variation distance* between X and Y are, respectively,

$$d_{\text{loc}}(X, Y) \triangleq \sup_{u \in \mathcal{A}} |P_X(u) - P_Y(u)| \tag{1}$$

$$d_{\text{TV}}(X, Y) \triangleq \frac{1}{2} \sum_{u \in \mathcal{A}} |P_X(u) - P_Y(u)|. \tag{2}$$

The local distance is the l^∞ distance between the probability mass functions, and the total variation distance is half the l^1 distance. The factor of one-half on the right-hand side of (2) normalizes the total variation distance to get values between zero and one. It is noted that the notation in the literature is not consistent, with a factor 2 on the right-hand side of (2) often being present or not. It is easy to show (see, e.g., [13, Lemma 5.4 on pp. 133–134]) that with this definition

$$d_{\text{TV}}(X, Y) = \sup_{\mathcal{B} \subseteq \mathcal{A}} |\mathbb{P}(X \in \mathcal{B}) - \mathbb{P}(Y \in \mathcal{B})|.$$

From the last equality and the definition of the local distance in (1), it follows that $d_{\text{loc}}(X, Y) \leq d_{\text{TV}}(X, Y)$.

A basic property that links between maximal coupling and the total variation distance is that if $(\hat{X}, \hat{Y})$ is a maximal coupling of (X, Y) then $\mathbb{P}(\hat{X} \neq \hat{Y}) = d_{\text{TV}}(X, Y)$. Throughout this paper, the term 'distribution' refers to the probability mass function of a discrete random variable defined over a finite or countably infinite alphabet.

The following theorem is a basic result on maximal coupling that also suggests, as part of its proof, a construction for maximal coupling (see, e.g., [31, Chapter 2]). We later rely on

this particular construction to derive in Section III some new bounds on the entropy of discrete random variables.

Theorem 1: Let X and Y be discrete random variables that take values in a set $\mathcal{A}$, and let their respective probability mass functions be

$$P_X(x) = \mathbb{P}(X = x), \quad P_Y(y) = \mathbb{P}(Y = y), \quad \forall\, x, y \in \mathcal{A}.$$

Then, the maximal coupling of (X, Y) satisfies

$$\mathbb{P}(\hat{X} = \hat{Y}) = \sum_{u \in \mathcal{A}} \min\{P_X(u), P_Y(u)\}. \tag{3}$$

Proof: Let $\mathcal{B} \triangleq \{u \in \mathcal{A} : \, P_X(u) < P_Y(u)\}$, and let $\mathcal{B}^{\mathrm{c}} \triangleq \mathcal{A} \setminus \mathcal{B}$. Then, for every coupling $(\hat{X}, \hat{Y})$ of (X, Y),

$$\begin{aligned}
&\mathbb{P}(\hat{X} = \hat{Y}) \\
&= \mathbb{P}(\hat{X} = \hat{Y}, \, \hat{Y} \in \mathcal{B}) + \mathbb{P}(\hat{X} = \hat{Y}, \, \hat{Y} \in \mathcal{B}^{\mathrm{c}}) \\
&\leq \mathbb{P}(\hat{X} \in \mathcal{B}) + \mathbb{P}(\hat{Y} \in \mathcal{B}^{\mathrm{c}}) \\
&= \mathbb{P}(X \in \mathcal{B}) + \mathbb{P}(Y \in \mathcal{B}^{\mathrm{c}}) \\
&= \sum_{u \in \mathcal{B}} P_X(u) + \sum_{u \in \mathcal{B}^{\mathrm{c}}} P_Y(u) \\
&= \sum_{u \in \mathcal{B}} \min\{P_X(u), P_Y(u)\} + \sum_{u \in \mathcal{B}^{\mathrm{c}}} \min\{P_X(u), P_Y(u)\} \\
&= \sum_{u \in \mathcal{A}} \min\{P_X(u), P_Y(u)\} \triangleq p.
\end{aligned} \tag{4}$$

The following provides a construction of a coupling $(\hat{X}, \hat{Y})$ that achieves the bound in (4) with equality, so it forms a maximal coupling of (X, Y). Let U, V, W and J be independent discrete random variables, where

$$\mathbb{P}(J = 0) = 1 - p, \quad \mathbb{P}(J = 1) = p \tag{5}$$

so $J \sim \text{Bernoulli}(p)$, and let U, V, W have the following probability mass functions:

$$P_U(u) = \frac{\min\{P_X(u), P_Y(u)\}}{p}, \quad \forall\, u \in \mathcal{A} \tag{6}$$

$$P_V(v) = \frac{P_X(v) - \min\{P_X(v), P_Y(v)\}}{1 - p}, \quad \forall\, v \in \mathcal{A} \tag{7}$$

$$P_W(w) = \frac{P_Y(w) - \min\{P_X(w), P_Y(w)\}}{1 - p}, \quad \forall\, w \in \mathcal{A}. \tag{8}$$

If $J = 1$, let $\hat{X} = \hat{Y} = U$, and if $J = 0$ let $\hat{X} = V$ and $\hat{Y} = W$. For every $x, y \in \mathcal{A}$

$$\begin{aligned}
&P_{\hat{X}}(x) \\
&= p\, \mathbb{P}(\hat{X} = x \,|\, J = 1) + (1 - p)\, \mathbb{P}(\hat{X} = x \,|\, J = 0) \\
&= p\, P_U(x) + (1 - p)\, P_V(x) \\
&= P_X(x)
\end{aligned}$$

and similarly $P_{\hat{Y}}(y) = P_Y(y)$, so $(\hat{X}, \hat{Y})$ is indeed a coupling of (X, Y). Furthermore,

$$\mathbb{P}(\hat{X} = \hat{Y}) \geq \mathbb{P}(J = 1) = p \tag{9}$$

so, from (4) and (9), it follows that the proposed construction for $(\hat{X}, \hat{Y})$ forms a maximal coupling of (X, Y) and also $\mathbb{P}(\hat{X} = \hat{Y}) = p$. ■

The following result is a simple consequence of Theorem 1 (see, e.g., [31, Chapter 2]), and it is also used for the derivation of the new bounds on the entropy in Section III.

Theorem 2: Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$. If $(\hat{X}, \hat{Y})$ is a maximal coupling of (X, Y) then

$$\mathbb{P}(\hat{X} \neq \hat{Y}) = d_{\text{TV}}(X, Y). \tag{10}$$

Proof: This follows from (2) and (3), and the equality $\min\{a, b\} = \frac{a+b-|a-b|}{2}$ for all $a, b \in \mathbb{R}$. ■

This work refines bounds on the difference of the entropies of two discrete random variables via the use of maximal couplings, leading to sharpened bounds that depend on both the local and total variation distances. The reader is also referred to a recent work in [21] that derived bounds for information measures by relying on the notion of the *minimum entropy coupling*.

The main observation of this work is that if the local distance between two probability distributions on a finite alphabet is smaller than the total variation distance, then the bounds on the entropy difference can be significantly strengthened. The second observation made in this work is that there is an extension of the new bound to countably infinite alphabets, where just knowing the total variation distance between two distributions does not imply anything about the difference of the respective entropies. The new bound that follows from the second observation is applied in this work to obtain refined bounds on the entropy of sums of independent (possibly non-identically distributed) Bernoulli random variables that arise in numerous applications. The application of the new bounds to the Poisson approximation is facilitated by using bounds on the total variation and local distances which follow from Stein's method, and the improvement that is obtained by these bounds is exemplified in this work. For comparison, a looser version of the new bounds was earlier applied in [33] to get bounds on the entropy of sums of dependent and non-identically distributed Bernoulli random variables.

The continuation of this paper is structured as follows: Section II introduces a known bound, due to Zhang [38], on the difference of the entropies of two discrete random variables in terms of the total variation distance. A shortened proof that is based on maximal coupling serves to motivate the derivation of some refined bounds. These new bounds, proved in Section III via maximal coupling, depend on both the local and total variation distances. Section IV exemplifies the use of the new bounds with a link to Stein's method, and it also compares them with some previously known bounds. Finally, the paper is concluded in Section V. Throughout this paper, the logarithms and the entropies are to the base e.

II. A Proof of a Known Bound on the Entropy of Discrete Random Variables via Coupling

The following theorem relies on a bound that first appeared in [38, Eq. (4)] and proved by coupling. It was later introduced in [17, Theorem 6] by re-proving the inequality in a different way (without coupling), and it was also strengthened there by showing an explicit case where the following bound is tight. As is proved in [38, Section 3], the bound on the entropy difference that is introduced in the following theorem improves the bound in [8, Theorem 17.3.3] or [9, Lemma 2.7].

Theorem 3: Let X and Y be two discrete random variables that take values in a finite set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X, Y) \, \log(M-1) + h\big(d_{\text{TV}}(X, Y)\big) \tag{11}$$

where h denotes the binary entropy function. Furthermore, there is a case where the bound is tight.

The following proof of Theorem 3 exemplifies the use of maximal coupling in proving an information-theoretic result.

Proof: Let $(\hat{X}, \hat{Y})$ be a maximal coupling of (X, Y). Since $H(X) = H(\hat{X})$ and $H(Y) = H(\hat{Y})$ (note that the marginal probability mass functions of (X, Y) and $(\hat{X}, \hat{Y})$ are the same), it follows from Fano's inequality and Theorem 2 (see (10)) that

$$\begin{aligned}
&\big|H(X) - H(Y)\big| \\
&= \big|H(\hat{X}) - H(\hat{Y})\big| \\
&= \big|H(\hat{X}|\hat{Y}) - H(\hat{Y}|\hat{X})\big| \\
&\leq \max\Big\{H(\hat{X}|\hat{Y}), H(\hat{Y}|\hat{X})\Big\} \\
&\leq \mathbb{P}(\hat{X} \neq \hat{Y}) \, \log(M-1) + h\big(\mathbb{P}(\hat{X} \neq \hat{Y})\big) \\
&= d_{\text{TV}}(X, Y) \, \log(M-1) + h\big(d_{\text{TV}}(X, Y)\big).
\end{aligned}$$

This proves the bound in (11) (see [38, Eq. (4)]). If $d_{\text{TV}}(X, Y) \leq \varepsilon$ for some $\varepsilon \in \big[0, 1 - \frac{1}{M}\big]$, the replacement of $d_{\text{TV}}(X, Y)$ in the last bound by ε is valid; this holds since the function $f(x) \triangleq x \log(M-1) + h(x)$ is monotonic increasing over the interval $[0, 1 - \frac{1}{M}]$ (since $f'(x) = \log(M-1) + \log\big(\frac{1-x}{x}\big) > 0$ for $0 < x < 1 - \frac{1}{M}$). Otherwise, if $\varepsilon > 1 - \frac{1}{M}$,

$$\big|H(X) - H(Y)\big| \leq \max\Big\{H(X), H(Y)\Big\} \leq \log(M).$$

Cases where the bound is tight [17]: If $\varepsilon \in [0, 1 - \frac{1}{M}]$, the bound is tight when

$$\begin{aligned}
&X \sim P_X = \left(1 - \varepsilon, \frac{\varepsilon}{M-1}, \ldots, \frac{\varepsilon}{M-1}\right) \\
&Y \sim P_Y = (1, 0, \ldots, 0)
\end{aligned}$$

which implies that

$$\begin{aligned}
&d_{\text{TV}}(X, Y) = \varepsilon, \\
&|H(X) - H(Y)| = H(X) = h(\varepsilon) + \varepsilon \log(M-1).
\end{aligned}$$

If $\varepsilon \in (1 - \frac{1}{M}, 1]$ then the bound is tight when

$$X \sim \Big(\frac{1}{M}, \ldots, \frac{1}{M}\Big), \quad Y \sim (1, 0, \ldots, 0)$$

so, $d_{\text{TV}}(X, Y) = 1 - \frac{1}{M} < \varepsilon$ and $|H(X) - H(Y)| = \log(M)$. ■

III. New Bounds on the Entropy of Discrete Random Variables via Coupling

In the cases where the known bound in Theorem 3 was shown to be tight in [17] (see the last part of the proof in Section II), it is easy to verify that the local distance is equal to the total variation distance. However, as is shown in the following, if it is not the case (i.e., the local distance is smaller than the total variation distance), then the bound in Theorem 3 is *necessarily* not tight. Furthermore, this section provides new bounds that depend on both the total variation and local distances. If these two distances are equal then the new bound is particularized to the bound in Theorem 3 but otherwise, the new bound improves the bound in Theorem 3. The general approach for proving the following new inequalities relies on the construction of the

maximal coupling that is introduced in the proof of Theorem 1. The new results are stated and proved in the following.

Theorem 4: Let X and Y be two discrete random variables that take values in a finite set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X,Y) \, \log(M\alpha - 1) + h\big(d_{\text{TV}}(X,Y)\big) \tag{12}$$

where

$$\alpha \triangleq \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \tag{13}$$

denotes the ratio of the local and total variation distances (so, $\alpha \in [\frac{2}{M}, 1]$), and h denotes the binary entropy function. Furthermore, if the probability mass functions of X and Y satisfy the condition that $\frac{1}{2} \leq \frac{P_X}{P_Y} \leq 2$ whenever $P_X, P_Y > 0$, then the bound in (12) is tightened to

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X,Y) \, \log\left(\frac{M\alpha - 1}{4}\right) + h\big(d_{\text{TV}}(X,Y)\big). \tag{14}$$

Remark 1: Since, in general, $\alpha \leq 1$ then the case where $\alpha = 1$ is the worst case for the bound in (12). In the latter case, it is particularized to the bound in Theorem 3 (see [17, Theorem 6] or [38, Eq. (4)]).

Remark 2: If $\alpha \leq \frac{1}{N}$ for some integer N (since $\alpha \in \left[\frac{2}{M}, 1\right]$ then $N \in \{1, \ldots, \lfloor \frac{M}{2} \rfloor\}$), the bound in (12) implies that

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X,Y) \, \log\left(\frac{M - N}{N}\right) + h\big(d_{\text{TV}}(X,Y)\big). \tag{15}$$

The bounds in (15) and [17, Theorem 7] are similar *but they hold under different conditions*. The bound in [17, Theorem 7] requires that $P_X, P_Y \leq \frac{1}{N}$ everywhere, whereas the bound in (15) holds under the requirement that the ratio α of the local and total variation distances satisfies $\alpha \leq \frac{1}{N}$. None of these conditions implies the other.

We prove in the following Theorem 4.

Proof: Assume without loss of generality (w.l.o.g.) that $H(X) - H(Y) \geq 0$ (note that the terms $|H(X) - H(Y)|$, $d_{\text{loc}}(X,Y)$ and $d_{\text{TV}}(X,Y)$ are invariant under a switch of X and Y). Let $(\hat{X}, \hat{Y})$ be the maximal coupling of (X,Y) according to the construction in the proof of Theorem 1. Then,

$$\begin{aligned}
&|H(X) - H(Y)| \\
&= H(X) - H(Y) \\
&= H(\hat{X}) - H(\hat{Y}) \\
&= H(\hat{X}|J) - H(\hat{Y}|J) + I(\hat{X};J) - I(\hat{Y};J).
\end{aligned} \tag{16}$$

The conditional entropy $H(\hat{X}|J)$ satisfies

$$\begin{aligned}
&H(\hat{X}|J) \\
&= \mathbb{P}(J=0) \, H(\hat{X}|J=0) + \mathbb{P}(J=1) \, H(\hat{X}|J=1) \\
&\overset{\text{(a)}}{=} d_{\text{TV}}(X,Y) \, H(V|J=0) + \big(1 - d_{\text{TV}}(X,Y)\big) \, H(U|J=1) \\
&\overset{\text{(b)}}{=} d_{\text{TV}}(X,Y) \, H(V) + \big(1 - d_{\text{TV}}(X,Y)\big) \, H(U)
\end{aligned} \tag{17}$$

where equality (a) holds since $J \sim \text{Bernoulli}(p)$ with

$$p = \mathbb{P}(J=1) = \mathbb{P}(\hat{X} = \hat{Y}) = 1 - d_{\text{TV}}(X,Y)$$

(see the proof of Theorem 1 and the result in Theorem 2), and because $\hat{X}$ is equal to V or U when J gets that values zero or one, respectively. Furthermore, equality (b) holds since U, V, W, J are independent random variables (due to the construction shown in the proof of Theorem 1). Similarly,

$$H(\hat{Y}|J) = d_{\text{TV}}(X,Y)\, H(W) + \big(1 - d_{\text{TV}}(X,Y)\big)\, H(U). \tag{18}$$

Combining (16)–(18) yields that

$$|H(X) - H(Y)| = d_{\text{TV}}(X,Y)\,(H(V) - H(W)) + I(\hat{X};J) - I(\hat{Y};J). \tag{19}$$

From (7) and (8), it follows that

$$P_V(a)\, P_W(a) = 0, \quad \forall\, a \in \mathcal{A} \tag{20}$$

and also, for every $a \in \mathcal{A}$,

$$\begin{aligned} & P_V(a) + P_W(a) \\ &= \frac{P_X(a) + P_Y(a) - 2\min\{P_X(a), P_Y(a)\}}{d_{\text{TV}}(X,Y)} \\ &= \frac{|P_X(a) - P_Y(a)|}{d_{\text{TV}}(X,Y)} \\ &\leq \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \triangleq \alpha. \end{aligned} \tag{21}$$

In the following, we derive upper bounds on $H(V) - H(W)$ and $I(\hat{X};J) - I(\hat{Y};J)$, and rely on (19) to get an upper bound on $|H(X) - H(Y)|$. Let $\mathcal{A} \triangleq \{a_1, \ldots, a_M\}$, and

$$s_i \triangleq P_V(a_i), \;\; t_i \triangleq P_W(a_i), \quad \forall\, i \in \{1, \ldots, M\}.$$

From (20) and (21),

$$s_i t_i = 0, \quad s_i + t_i \leq \alpha, \quad \forall\, i \in \{1, \ldots, M\}$$

and $H(V) - H(W) = -\sum_{i=1}^{M} s_i \log(s_i) + \sum_{i=1}^{M} t_i \log(t_i)$. Hence, for fixed α and M (since $|\mathcal{A}| = M$, then $\alpha \in [\frac{2}{M}, 1]$),

$$H(V) - H(W) \leq g(\alpha) \tag{22}$$

where $g(\alpha)$ is the solution of the optimization problem

$$\begin{aligned} & \text{maximize} \left(-\sum_{i=1}^{M} s_i \log(s_i) + \sum_{i=1}^{M} t_i \log(t_i) \right) \\ & \text{subject to} \\ & \begin{cases} s_i, t_i \geq 0, \;\; s_i + t_i \leq \alpha \\ s_i t_i = 0, \;\; \forall\, i \in \{1, \ldots, M\} \\ \displaystyle\sum_{i=1}^{M} s_i = \sum_{i=1}^{M} t_i = 1 \end{cases} \end{aligned} \tag{23}$$

with the $2M$ variables $s_1, t_1, \ldots s_M, t_M$. Fortunately, this *non-convex* optimization problem admits a closed-form solution.

Lemma 1: The solution of the non-convex optimization problem in (23), denoted by $g(\alpha)$, is the following:

$$g(\alpha) = \log\Big(M - \left\lceil \frac{1}{\alpha} \right\rceil\Big) + \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \log \alpha + \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \log \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \tag{24}$$

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

Proof: Let's first show that the solution on the right-hand side of (24) forms an upper bound on $g(\alpha)$, and then show that this upper bound is tight.

For the derivation of the upper bound, note that due to the above constraints,

$$\begin{aligned} 1 = \sum_{i=1}^{M} t_i &\overset{\text{(a)}}{\leq} \alpha \left|\{i \in \{1, \ldots, M\} : t_i > 0\}\right| \\ &\Rightarrow \left|\{i \in \{1, \ldots, M\} : t_i > 0\}\right| \geq \frac{1}{\alpha} \\ &\overset{\text{(b)}}{\Rightarrow} \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \leq M - \frac{1}{\alpha} \\ &\overset{\text{(c)}}{\Rightarrow} \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \leq M - \left\lceil \frac{1}{\alpha} \right\rceil \end{aligned} \tag{25}$$

where inequality (a) holds since $s_i + t_i \leq \alpha$ and $s_i, t_i \geq 0$ for every $i \in \{1, \ldots, M\}$, (b) follows from the constraint that $s_i \, t_i = 0$ for every i, and (c) holds since the cardinality of the support of $\{s_i\}$ is an integer, and $\left\lfloor M - \frac{1}{\alpha} \right\rfloor = M - \left\lceil \frac{1}{\alpha} \right\rceil$. Hence,

$$-\sum_{i=1}^{M} s_i \, \log(s_i) \leq \log\Big(M - \left\lceil \frac{1}{\alpha} \right\rceil\Big)$$

and the solution of the optimization problem in (23) satisfies

$$g(\alpha) \leq \log\Big(M - \left\lceil \frac{1}{\alpha} \right\rceil\Big) + f(\alpha) \tag{26}$$

where $f(\alpha)$ solves the optimization problem

$$\begin{aligned} &\text{maximize} \ \sum_{i=1}^{M} t_i \log(t_i) \\ &\text{subject to} \\ &\begin{cases} 0 \leq t_i \leq \alpha, \quad \forall \, i \in \{1, \ldots, M\} \\ \displaystyle\sum_{i=1}^{M} t_i = 1 \end{cases} \end{aligned} \tag{27}$$

with the M optimization variables $t_1, \ldots, t_M$. Note that the objective function in (27) is convex, and the feasible set is a bounded polyhedron. Furthermore, the maximum of a convex function over a bounded polyhedron is attained at one of its vertices (see, e.g., [30, Corollary 32.3.3]; this property follows from the convex-hull description of a bounded polyhedron and Jensen's inequality). Since the objective function and the feasible set in (27) are invariant to a permutation of the variables $t_1, \ldots, t_M$, then an optimal point is given by

$$\begin{aligned} &t_1 = \ldots = t_l = \alpha, \quad l \triangleq \left\lfloor \frac{1}{\alpha} \right\rfloor \\ &t_{l+1} = 1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor, \quad t_{l+2} = \ldots = t_M = 0 \end{aligned}$$

where $l \leq \frac{M}{2}$ (since $\alpha \in [\frac{2}{M}, 1]$); as requested, $t_i \in [0, \alpha]$ for $i \in \{1, \ldots, M\}$. This implies that the solution of the optimization problem in (27) is given by

$$f(\alpha) = \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \log \alpha + \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \log \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right). \tag{28}$$

From (26) and (28), it follows that the right-hand side of (24) forms an upper bound on $g(\alpha)$. It remains to show that this bound is tight. To this end, we separate into the following two cases:

Case 1: Suppose that $N \triangleq \frac{1}{\alpha}$ is an integer. In this case, the upper bound on $g(\alpha)$ (see (26) and (28)) gets the simplified form

$$g(\alpha) \leq \log\Big(M - \frac{1}{\alpha}\Big) + \log \alpha = \log(M\alpha - 1).$$

This upper bound on $g(\alpha)$ is achieved by the point $(s_1, t_1, \ldots, s_M, t_M)$ where

$$\begin{aligned} &t_1 = \ldots = t_N = \alpha, \quad t_{N+1} = \ldots = t_M = 0 \\ &s_1 = \ldots = s_N = 0, \quad s_{N+1} = \ldots = s_M = \frac{1}{M - N}. \end{aligned}$$

Note that this point is included in the feasible set of the optimization problem in (23) since $\frac{1}{M-N} = \frac{\alpha}{M\alpha - 1} \leq \alpha$ where the last inequality holds because $\alpha \in [\frac{2}{M}, 1]$. The value of the objective function in (23) at this point is equal to

$$\begin{aligned} &-\sum_{i=1}^{M} s_i \log(s_i) + \sum_{i=1}^{M} t_i \log(t_i) \\ &= \log\Big(M - \frac{1}{\alpha}\Big) + \log \alpha = \log(M\alpha - 1) \end{aligned}$$

so this upper bound on $g(\alpha)$ is tight if $\frac{1}{\alpha}$ is an integer.

Case 2: Suppose that $\frac{1}{\alpha}$ is not an integer. In this case, let $l \triangleq \lfloor \frac{1}{\alpha} \rfloor$ so $l + 1 = \left\lceil \frac{1}{\alpha} \right\rceil$, and consider the $(2M)$-dimensional vector $(s_1, t_1, \ldots, s_M, t_M)$ where

$$\begin{aligned} &t_1 = \ldots = t_l = \alpha, \quad t_{l+1} = 1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \\ &t_{l+2} = \ldots = t_M = 0 \\ &s_1 = \ldots = s_{l+1} = 0 \\ &s_{l+2} = \ldots = s_M = \frac{1}{M - l - 1} = \frac{1}{M - \left\lceil \frac{1}{\alpha} \right\rceil}. \end{aligned} \tag{29}$$

To verify that it is included in the feasible set of (23), note that due to the constraints of this optimization problem

$$\begin{aligned} &1 = \sum_{i=1}^{M} s_i \leq \alpha \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \\ &\Rightarrow \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \geq \frac{1}{\alpha} \\ &\Rightarrow \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \geq \left\lceil \frac{1}{\alpha} \right\rceil \end{aligned}$$

and, by combining it with (25), it follows that

$$\left\lceil \frac{1}{\alpha} \right\rceil \leq \left|\{i \in \{1, \ldots, M\} : s_i > 0\}\right| \leq M - \left\lceil \frac{1}{\alpha} \right\rceil$$

so $\lceil \frac{1}{\alpha} \rceil \leq \frac{M}{2}$. This implies that for $j \in \{l+2, \ldots, M\}$ (note also that $\alpha \in [\frac{2}{M}, 1]$)

$$\begin{aligned} s_j &= \frac{1}{M-l-1} \\ &= \frac{1}{M - \lceil \frac{1}{\alpha} \rceil} \\ &\leq \frac{2}{M} \\ &\leq \alpha \end{aligned}$$

and $t_{l+1} = 1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \leq \alpha$, so the vector is indeed included in the feasible set of (23). The value of the objective function in (23) at the selected point in (29) is equal to

$$\begin{aligned} &-\sum_{i=1}^{M} s_i \log(s_i) + \sum_{i=1}^{M} t_i \log(t_i) \\ &= \log\Big(M - \left\lceil \frac{1}{\alpha} \right\rceil\Big) + \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \log \alpha + \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \log\left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \\ &= g(\alpha) \end{aligned}$$

so the upper bound on $g(\alpha)$ from (26) and (28) is tight, and this completes the proof of Lemma 1. ■

Corollary 1: The solution of the non-convex optimization problem in (23) satisfies the inequality

$$g(\alpha) \leq \log(M\alpha - 1)$$

and this bound is tight if and only if $\frac{1}{\alpha}$ is an integer.

Proof: From Lemma 1 (see Eq. (24)), it follows that

$$\begin{aligned} &g(\alpha) \\ &\leq \log\Big(M - \frac{1}{\alpha}\Big) + \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \log \alpha + \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \log\left(1 - \alpha\Big(\frac{1}{\alpha} - 1\Big)\right) \\ &= \log\Big(M - \frac{1}{\alpha}\Big) + \log(\alpha) \\ &= \log(M\alpha - 1) \end{aligned}$$

and the above inequality turns to be an equality if and only if $\frac{1}{\alpha}$ is an integer. ■

By combining (22) and Corollary 1, it follows that

$$H(V) - H(W) \leq \log(M\alpha - 1)$$

and therefore from (19)

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X, Y) \log(M\alpha - 1) + I(J; \hat{X}) - I(J; \hat{Y}). \tag{30}$$

Finally, the bound in (12) follows from the inequality

$$I(J; \hat{X}) - I(J; \hat{Y}) \leq H(J) = h\big(d_{\text{TV}}(X, Y)\big). \tag{31}$$

We move to derive a refinement of the bound in (12) when $\frac{1}{2} \leq \frac{P_X}{P_Y} \leq 2$. In this case, the starting point is the inequality in (30) where it is aimed to improve the upper bound in (31). To

this end,

$$\begin{aligned}
&I(J;\hat{X}) - I(J;\hat{Y}) \\
&= H(J|\hat{Y}) - H(J|\hat{X}) \\
&\leq H(J) - H(J|\hat{X}) \\
&= h\big(d_{\text{TV}}(X,Y)\big) - H(J|\hat{X})
\end{aligned} \tag{32}$$

and, from [18, Theorem 11],

$$H(J|\hat{X}) \geq 2\log 2\, \mathbb{P}\big(J \neq J_{\text{MAP}}(\hat{X})\big) \tag{33}$$

where $J_{\text{MAP}}(\hat{X})$ is the maximum a-posteriori (MAP) estimator of J based on $\hat{X}$ (note that the minimum on the left-hand side of [18, Eq. (110)] is achieved by the MAP estimator). In the following, the estimator $J_{\text{MAP}}(\hat{X})$ on the right-hand side of (33) is calculated.

1) If $\hat{X} \notin \text{supp}(P_V)$ then a.s. $J = 1$ (otherwise, $J = 0$ and $\hat{X} = V$, so $\hat{X} \in \text{supp}(P_V)$ a.s.). Hence,

$$\hat{X} \notin \text{supp}(P_V) \;\Rightarrow\; J_{\text{MAP}}(\hat{X}) = 1.$$

From (7), it follows that $\hat{X} \notin \text{supp}(P_V)$ if and only if $P_X(\hat{X}) \leq P_Y(\hat{X})$.

2) If $\hat{X} \in \text{supp}(P_V)$ then, from (7), $P_X(\hat{X}) > P_Y(\hat{X})$. Hence, from (6) and (7) with $p = 1 - d_{\text{TV}}(X,Y)$,

$$\begin{aligned}
P_U(\hat{X}) &= \frac{P_Y(\hat{X})}{1 - d_{\text{TV}}(X,Y)} \\
P_V(\hat{X}) &= \frac{P_X(\hat{X}) - P_Y(\hat{X})}{d_{\text{TV}}(X,Y)}.
\end{aligned}$$

Since U, V, J are independent, then from (5)

$$\begin{aligned}
\mathbb{P}(J = 1, \hat{X}) &= \mathbb{P}(J = 1)\, P_U(\hat{X}) = P_Y(\hat{X}) \\
\mathbb{P}(J = 0, \hat{X}) &= \mathbb{P}(J = 0)\, P_V(\hat{X}) = P_X(\hat{X}) - P_Y(\hat{X})
\end{aligned}$$

so, if $\hat{X} \in \text{supp}(P_V)$, then

$$J_{\text{MAP}}(\hat{X}) = \begin{cases} 1 & \text{if } \frac{P_X(\hat{X})}{2} \leq P_Y(\hat{X}) < P_X(\hat{X}) \\ 0 & \text{if } P_Y(\hat{X}) < \frac{P_X(\hat{X})}{2}. \end{cases}$$

To conclude, the MAP estimator of J that is based on the observation $\hat{X}$ is given by

$$J_{\text{MAP}}(\hat{X}) = \begin{cases} 1 & \text{if } \frac{P_X(\hat{X})}{2} \leq P_Y(\hat{X}) \\ 0 & \text{if } P_Y(\hat{X}) < \frac{P_X(\hat{X})}{2}. \end{cases}$$

It therefore implies that if $\frac{P_Y}{P_X} \geq \frac{1}{2}$ whenever $P_X > 0$, then $J_{\text{MAP}}(\hat{X}) = 1$ independently of $\hat{X}$, so in this case

$$\mathbb{P}\big(J \neq J_{\text{MAP}}(\hat{X})\big) = \mathbb{P}(J = 0) = d_{\text{TV}}(X,Y).$$

Hence, from (32), (33) and the last equality, if $\frac{P_Y}{P_X} \geq \frac{1}{2}$ whenever $P_X > 0$ then

$$I(J;\hat{X}) - I(J;\hat{Y}) \leq h\big(d_{\text{TV}}(X,Y)\big) - 2\log 2 \cdot d_{\text{TV}}(X,Y).$$

A combination of the last inequality with (30) finally gives the refined bound in (14). Since it was assumed at the beginning of the proof that $H(X) \geq H(Y)$ while it is not necessarily known

in advance which entropy is larger, the requirement on $\frac{P_Y}{P_X}$ can be symmetrized by requiring that $\frac{1}{2} \leq \frac{P_X}{P_Y} \leq 2$ whenever $P_X, P_Y > 0$. This completes the proof of Theorem 4. ■

Corollary 2: Let X and Y be two discrete random variables that take values in a finite set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Assume that for some positive constants $\varepsilon_1, \varepsilon_2$

$$d_{\text{TV}}(X,Y) \leq \varepsilon_1 \leq 1 - \frac{1}{M\varepsilon_2}\,, \tag{34}$$

$$\frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \leq \varepsilon_2 \leq 1. \tag{35}$$

Then,

$$|H(X) - H(Y)| \leq \varepsilon_1 \, \log(M\varepsilon_2 - 1) + h(\varepsilon_1). \tag{36}$$

Proof: From (12), (13), (35), and since $\alpha \leq \varepsilon_2$

$$|H(X) - H(Y)| \leq d_{\text{TV}}(X,Y) \, \log(M\varepsilon_2 - 1) + h\big(d_{\text{TV}}(X,Y)\big).$$

The function $q(\varepsilon) \triangleq \varepsilon c + h(\varepsilon)$ is monotonic increasing over the interval $\left[0, \frac{e^c}{1+e^c}\right]$ ($q'(\varepsilon) = c + \log\left(\frac{1-\varepsilon}{\varepsilon}\right) > 0$ if and only if $0 < \varepsilon < \frac{e^c}{1+e^c}$). Referring to the right-hand side of the above inequality, let $c \triangleq \log(M\varepsilon_2 - 1)$, so $\frac{e^c}{1+e^c} = 1 - \frac{1}{M\varepsilon_2}$. Hence, if the conditions in (34) and (35) are satisfied then the inequality in (36) holds. ■

Remark 3: By considering the pair of probability mass functions $P_{X,Y}$ and $P_X \times P_Y$ (without abuse of notation, let $H(P_X) \triangleq H(X)$), then

$$\begin{aligned} &H(P_X \times P_Y) - H(P_{X,Y}) \\ &= H(X) + H(Y) - H(X,Y) \\ &= I(X;Y). \end{aligned}$$

Hence, Theorem 4 and Corollary 2 provide bounds on the mutual information between two discrete random variables of finite support, where these bounds are expressed in terms of the local and total variation distances between the joint distribution of (X,Y) and the product of its marginal distributions. The specialization of Theorem 4 to this setting tightens the bound in [38, Theorem 1], and the former bound is particularized to the latter known bound in the case where the local and total variation distances are equal (which is the extreme case).

Remark 4: The bound in [38, Theorem 1] was improved in [27, Proposition 1] without any further assumptions. It is noted that by introducing the additional requirement where there exists some constant $\varepsilon_2 \in [0,1]$ such that for every $y \in \mathcal{Y}$

$$\frac{d_{\text{loc}}(P_X, P_{X|Y=y})}{d_{\text{TV}}(P_X, P_{X|Y=y})} \leq \varepsilon_2$$

then it enables to refine the bound in [27, Proposition 1]. This follows by combining the proof of [27, Proposition 1] with (36) (see Corollary 2) where Eq. (36) replaces the use of [38, Eq. (4)] in [27, Eq. (35)]. The same thing also applies to [28, Proposition 2], referring to its proof in [28, p. 305].

For countably infinite alphabets, just knowing the total variation distance between two distributions does not imply anything about the difference of entropies (i.e., one has discontinuity of the entropy). The following theorem shows that is one of the distributions is finitely supported, and some knowledge of the tail behavior of the other distribution is available, then having bounds

on the local and total variation distances allows one to bound the difference of the entropies even in this case.

Theorem 5: Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countable infinite set. Let X and Y be discrete random variables where X takes values in the set $\mathcal{X} = \{a_1, \ldots, a_m\}$ for some $m \in \mathbb{N}$, and Y takes values in the set $\mathcal{A}$. Assume that for some $\eta_1, \eta_2, \eta_3 > 0$, the local and total variation distances between X and Y satisfy

$$\eta_2 \leq d_{\text{TV}}(X, Y) \leq \eta_1, \quad d_{\text{loc}}(X, Y) \leq \eta_3 \tag{37}$$

where $\eta_3 \leq \eta_2 \leq \eta_1 < 1$. Let M be an integer such that

$$\sum_{i=M}^{\infty} P_Y(a_i) \leq \eta_3, \quad M \geq \max\left\{m+1, \frac{\eta_2}{(1-\eta_1)\eta_3}\right\} \tag{38}$$

and let $\eta_4 > 0$ satisfy

$$-\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) \leq \eta_4. \tag{39}$$

Then, the following inequality holds:

$$|H(X) - H(Y)| \leq \eta_1 \log\left(\frac{M\eta_3}{\eta_2} - 1\right) + h(\eta_1) + \eta_4. \tag{40}$$

Remark 5: The inequality $\eta_3 \leq \eta_2 \leq \eta_1 < 1$ (after (37)) is easily satisfied since $d_{\text{loc}}(X, Y) \leq d_{\text{TV}}(X, Y) \leq 1$; so, if $d_{\text{TV}}(X, Y) < 1$ then it is possible to choose η_1, η_2 and η_3 that satisfy this inequality.

Proof: Let $\widetilde{Y}$ be a random variable that is defined to be equal to Y if $Y \in \{a_1, \ldots, a_{M-1}\}$, and it is set to be equal to a_M if $Y = a_i$ for some $i \geq M$. Hence, the probability mass function of $\widetilde{Y}$ is related to that of Y as follows:

$$P_{\widetilde{Y}}(a_i) = \begin{cases} P_Y(a_i) & \text{if } i \in \{1, \ldots, M-1\} \\ \sum_{j=M}^{\infty} P_Y(a_j) & \text{if } i = M. \end{cases} \tag{41}$$

Since $P_X(a_i) = 0$ for every $i > m$ and also $M \geq m+1$ (see the second inequality in (38)), then it follows from (41) that

$$\begin{aligned} & d_{\text{TV}}(X, \widetilde{Y}) \\ &= \frac{1}{2}\sum_{i=1}^{m} |P_X(a_i) - P_{\widetilde{Y}}(a_i)| + \frac{1}{2}\sum_{i=m+1}^{M} P_{\widetilde{Y}}(a_i) \\ &= \frac{1}{2}\sum_{i=1}^{m} |P_X(a_i) - P_Y(a_i)| + \frac{1}{2}\sum_{i=m+1}^{\infty} P_Y(a_i) \\ &= d_{\text{TV}}(X, Y). \end{aligned} \tag{42}$$

Hence, X and $\widetilde{Y}$ are discrete random variables that take values in the set $\{a_1, \ldots, a_M\}$ (note that it includes the set $\mathcal{X}$), and from (37) and (42)

$$0 < \eta_2 \leq d_{\text{TV}}(X, \widetilde{Y}) \leq \eta_1. \tag{43}$$

Furthermore, the local distance between X and $\widetilde{Y}$ satisfies

$$\begin{aligned}
&d_{\text{loc}}(X,\widetilde{Y}) \\
&= \max_{i\in\{1,\ldots,M\}} |P_X(a_i) - P_{\widetilde{Y}}(a_i)| \\
&\overset{\text{(a)}}{=} \max\left\{\max_{i\in\{1,\ldots,M-1\}} |P_X(a_i) - P_Y(a_i)|\,, \sum_{i=M}^{\infty} P_Y(a_i)\right\} \\
&\overset{\text{(b)}}{\leq} \max\{d_{\text{loc}}(X,Y), \eta_3\} \\
&\overset{\text{(c)}}{=} \eta_3
\end{aligned} \tag{44}$$

where (a), (b) and (c) above follow from the equality in (41) (note also that $m \leq M-1$), the first inequality in (38) and the second inequality in (37), respectively. From (43) and (44)

$$d_{\text{TV}}(X,\widetilde{Y}) \leq \eta_1 \triangleq \varepsilon_1 \tag{45}$$

$$\frac{d_{\text{loc}}(X,\widetilde{Y})}{d_{\text{TV}}(X,\widetilde{Y})} \leq \frac{\eta_3}{\eta_2} \triangleq \varepsilon_2 \tag{46}$$

where $0 < \varepsilon_1 < 1$ and $0 < \varepsilon_2 \leq 1$ (since, by assumption, $0 < \eta_3 \leq \eta_2 \leq \eta_1 < 1$). The integer M is set to satisfy the inequality $M \geq \frac{\eta_2}{\eta_3(1-\eta_1)}$ (see (38)), so from (45) and (46)

$$\varepsilon_1 \leq 1 - \frac{1}{M\varepsilon_2}.$$

Hence, it follows from Theorem 4 that

$$|H(X) - H(\widetilde{Y})| \leq \eta_1 \log\Big(\frac{M\eta_3}{\eta_2} - 1\Big) + h(\eta_1). \tag{47}$$

Since $\widetilde{Y}$ is a deterministic function of Y then $H(Y) \geq H(\widetilde{Y})$, and from (41)

$$\begin{aligned}
&|H(\widetilde{Y}) - H(Y)| \\
&= H(Y) - H(\widetilde{Y}) \\
&= -\sum_{i=M}^{\infty} P_Y(a_i)\,\log P_Y(a_i) + \left(\sum_{i=M}^{\infty} P_Y(a_i)\right)\log\left(\sum_{i=M}^{\infty} P_Y(a_i)\right) \\
&\leq -\sum_{i=M}^{\infty} P_Y(a_i)\,\log P_Y(a_i) \leq \eta_4.
\end{aligned} \tag{48}$$

Finally, the bound in (40) follows from (47), (48) and the triangle inequality. ■

Corollary 3: In the setting of X and Y in Theorem 5, assume that $d_{\text{TV}}(X,Y) \leq \eta$ for some $\eta \in (0,1)$. Let $M \triangleq \max\Big\{m+1, \frac{1}{1-\eta}\Big\}$, and assume that for some $\mu > 0$

$$-\sum_{i=M}^{\infty} P_Y(a_i)\,\log P_Y(a_i) \leq \mu$$

then $|H(X) - H(Y)| \leq \eta \log(M-1) + h(\eta) + \mu$.

Proof: This follows from Theorem 5 by setting $\eta_2 = \eta_3 = d_{\text{loc}}(X,Y)$ (note that $d_{\text{loc}}(X,Y) \leq d_{\text{TV}}(X,Y)$), and then η_1 and η_4 are replaced by η and μ, respectively. ■

Remark 6: The result in Corollary 3 can be obtained by a simplification of the proof of Theorem 5 where (47) is replaced by the bound in [38, Eq. (4)] (see (11)), without the refinement which takes the local distance into consideration.

IV. EXAMPLES

In the following, we exemplify the use of the new bounds in Section III, and also compare them with some previously known bounds.

Example 1: Let X be a discrete random variable that gets values in the set $\mathcal{A} = \{a_1, \ldots, a_M\}$. Let's express its arbitrary probability mass function in the form

$$P_X(a_i) = \frac{1 + u_i \xi_i}{M} \quad \forall\, i \in \{1, \ldots, M\} \tag{49}$$

where

$$\begin{aligned} &u_i \in \{-1, 1\},\ \xi_i \geq 0, \\ &0 \leq 1 + u_i \xi_i \leq M, \quad \forall\, i \in \{1, \ldots, M\} \end{aligned}$$

and

$$\sum_{i=1}^{M} u_i \xi_i = 0$$

where the latter equality is equivalent to $\sum_{i=1}^{M} P_X(a_i) = 1$.

In the following, we derive a lower bound on the entropy $H(X)$. Let Y be a random variable that takes the values from $\mathcal{A}$ with equal probability, so $H(Y) = \log M$. The local and total variation distances between X and Y are equal to

$$d_{\text{TV}}(X, Y) = \frac{1}{2M} \sum_{i=1}^{M} \xi_i = \frac{\xi_{\text{avg}}^{(M)}}{2}$$

$$d_{\text{loc}}(X, Y) = \frac{1}{M} \max_{1 \leq i \leq M} \xi_i = \frac{\xi_{\max}^{(M)}}{M}$$

where $\xi_{\text{avg}}^{(M)}$ and $\xi_{\max}^{(M)}$ denote the average and maximal values of $\{\xi_i\}_{i=1}^{M}$, respectively. From (13)

$$\alpha_M \triangleq \frac{d_{\text{loc}}(X, Y)}{d_{\text{TV}}(X, Y)} = \frac{2\xi_{\max}^{(M)}}{M \xi_{\text{avg}}^{(M)}}$$

so

$$\alpha_M = \frac{2K_M}{M}$$

where

$$K_M \triangleq \frac{\xi_{\max}^{(M)}}{\xi_{\text{avg}}^{(M)}}. \tag{50}$$

From (12) (where also $H(Y) = \log M \geq H(X)$), it follows that

$$\log M - \frac{\xi_{\text{avg}}^{(M)}\ \log(2K_M - 1)}{2} - h\left(\frac{\xi_{\text{avg}}^{(M)}}{2}\right) \leq H(X) \leq \log M$$

and, since the binary entropy function is bounded between 0 and $\log 2$, the above inequality can be loosened to

$$1 - \frac{\xi_{\text{avg}}^{(M)}}{2} \frac{\log(2K_M - 1)}{\log M} - \frac{\log 2}{\log M} \leq \frac{H(X)}{\log M} \leq 1 \tag{51}$$

which implies (since $K_M \geq 1$) that

$$\lim_{M \to \infty} \frac{\xi_{\text{avg}}^{(M)}\ \log K_M}{\log M} = 0 \ \Rightarrow\ \lim_{M \to \infty} \frac{H(X)}{\log M} = 1. \tag{52}$$

For comparison, the bound in Theorem 3 gives that

$$1 - \frac{\xi_{\text{avg}}^{(M)}}{2} \cdot \frac{\log(M-1)}{\log M} - \frac{1}{\log M} \cdot h\left(\frac{\xi_{\text{avg}}^{(M)}}{2}\right) \leq \frac{H(X)}{\log M} \leq 1 \tag{53}$$

which implies that

$$\lim_{M\to\infty} \xi_{\text{avg}}^{(M)} = 0 \;\Rightarrow\; \lim_{M\to\infty} \frac{H(X)}{\log M} = 1. \tag{54}$$

The latter condition in (54) is stronger than (52). To see this, note that $1 \leq K_M \leq \frac{M}{2}$ (since $\frac{2}{M} \leq \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \leq 1$). On the other hand, as a concrete example for the case where the condition in (52) holds whereas the condition in (54) does not hold, let M be an arbitrary even number, and

$$u_i = (-1)^i,\; \xi_i = \beta \in (0,1], \quad i \in \{1,\ldots,M\}$$

where, indeed, $\sum_{i=1}^{M} u_i \xi_i = \beta \sum_{i=1}^{M} (-1)^i = 0$. In this case, $P_X(a_i) = \frac{1-\beta}{M}$ for odd numbers $i \in \{1,\ldots,M\}$, and $P_X(a_i) = \frac{1+\beta}{M}$ for even numbers i. Furthermore, in this case $K_M = 1$ for every even M, so the condition in (52) holds by letting the even number M tend to infinity. On the other hand, the condition in (54) is not satisfied since $\lim_{M\to\infty} \xi_{\text{avg}}^{(M)} = \beta > 0$. The upper and lower bounds in (53) tend to 1 and $1 - \frac{\beta}{2}$, respectively, so the gap between these asymptotic bounds is increased linearly with β. Therefore, Theorem 4 gives a simple lower bound on the entropy $H(X)$ in terms of the average and maximal values of $\{\xi_i\}_{i=1}^{M}$, which improves the lower bound on the entropy that follows from the known bound in Theorem 3 (see (53)).

For comparison, the bound in [17, Theorem 7] is also applied to this example. In this case, since $P_X, P_Y \leq \frac{1+\xi_{\max}}{M}$ then P_X and P_Y are less than or equal to $\frac{1}{N_M}$ with $N_M \triangleq \left\lfloor \frac{M}{1+\xi_{\max}^{(M)}} \right\rfloor$. Similarly to the above analysis, it is easy to verify from [17, Theorem 7] that

$$\lim_{M\to\infty} \frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\max}^{(M)}\right)}{\log M} = 0 \;\Rightarrow\; \lim_{M\to\infty} \frac{H(X)}{\log M} = 1. \tag{55}$$

Since

$$\frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\max}^{(M)}\right)}{\log M} \geq \frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\text{avg}}^{(M)}\right)}{\log M} \geq -\frac{\log e}{e \log M}$$

where the right-hand side of this inequality holds since the function $f(x) = x \log x$ for $x > 0$ achieves its minimal value at $x = \frac{1}{e}$, it follows that if the limit on the left-hand side of (55) is zero then also

$$\lim_{M\to\infty} \frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\text{avg}}^{(M)}\right)}{\log M} = 0.$$

Therefore, the definition of K_M in (50) gives that

$$\begin{aligned}
&\lim_{M\to\infty} \frac{\xi_{\text{avg}}^{(M)} \log K_M}{\log M} \\
&= \lim_{M\to\infty} \frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\max}^{(M)}\right)}{\log M} - \lim_{M\to\infty} \frac{\xi_{\text{avg}}^{(M)} \log\left(\xi_{\text{avg}}^{(M)}\right)}{\log M} \\
&= 0.
\end{aligned}$$

This shows that the conclusion in (52) implies the one in (55).

A special case of (49) *with numerical results*: As a special case of the probability mass function in (49), let $M = 2^m$ for some $m \in \mathbb{N}$, let $u_i = (-1)^i$ for $i \in \{1, \ldots, M\}$, and $\xi_i = \beta$ for some $\beta \in [0, 1]$. In this special case,

$$P_X(a_i) = \begin{cases} 2^{-m}\,(1-\beta) & \text{if } i \in \{1, 3, \ldots, 2^m - 1\} \\ 2^{-m}\,(1+\beta) & \text{if } i \in \{2, 4, \ldots, 2^m\}. \end{cases}$$

Let Y be a random variable that gets all the values in the set $\{a_1, \ldots, a_M\}$ with equal probability (i.e., 2^{-m}). Then, the local and total variation distances between X and Y are

$$d_{\text{loc}}(X, Y) = \frac{\beta}{M}, \quad d_{\text{TV}}(X, Y) = \frac{\beta}{2}$$

so, from (13), $\alpha = \frac{2}{M}$. The entropies of X and Y are

$$H(X) = (m-1)\log 2 + h\Big(\frac{1-\beta}{2}\Big), \quad H(Y) = m\,\log 2$$

so, $H(Y) - H(X) = \log 2 - h\big(\frac{1-\beta}{2}\big)$ independently of m.

For comparison, the known bound in Theorem 3 that only depends on the total variation distance between X and Y (with no further knowledge about their probability mass functions) gives

$$H(Y) - H(X) \leq \frac{m\beta}{2} \cdot \log 2 + h\Big(\frac{\beta}{2}\Big) + \frac{\beta}{2} \cdot \log(1 - 2^{-m})$$

so this upper bound increases almost linearly with m, in contrast to the exact value that is independent of m. The new bound in (12), which depends on both the local and total variation distances between X and Y (but again, without any further information on their probability mass functions) gives

$$\begin{aligned} &H(Y) - H(X) \\ &\leq d_{\text{TV}}(X, Y)\,\log(M\alpha - 1) + h\big(d_{\text{TV}}(X, Y)\big) \\ &= h\Big(\frac{\beta}{2}\Big). \end{aligned} \tag{56}$$

Similarly to the exact value, but in contrast to the former bound, the latter bound is independent of m. Furthermore, if $\beta \to 0$ and $m\beta \to \infty$, then the exact value of $H(Y) - H(X)$ as well as the latter bound (that follows from Theorem 4) tend to zero, whereas the former bound that follows from Theorem 3 tends to infinity. This shows the difference in the two bounds, exemplifying the possible advantage of taking into account the local distance in addition to the total variation distance.

For $\beta \in [0, \frac{1}{2}]$, the condition $\frac{1}{2} \leq \frac{P_X}{P_Y} \leq 2$ is fulfilled, so the tightened bound in (14) gives that

$$0 \leq H(Y) - H(X) \leq h\Big(\frac{\beta}{2}\Big) - \beta\,\log 2. \tag{57}$$

If $\beta = \frac{1}{2}$, $H(Y) - H(X) = \log 2 - h\big(\frac{1}{2}\big) = 0.131$ nats, the upper bound in (56) is equal to 0.562 nats, and the tightened version of this bound in (57) is equal to 0.216 nats.

It is noted that since P_X is majorized by P_Y (see [18, Definition 1 on p. 5934]), then according to [18, Theorem 3]

$$H(Y) - H(X) \geq D(P_X || P_Y)$$

and since P_Y refers to a uniform distribution over a set of cardinality $M = 2^m$ then $H(Y) = m \log 2$, and

$$D(P_X || P_Y) = m \log 2 - H(P_X)$$

so, the above lower bound is achieved here with equality.

In Example 1, the probability mass function of the discrete random variable X was known explicitly. However, in many interesting applications, this is not necessarily the case. If the exact distribution of X is not available or is numerically hard to compute, a derivation of some good bounds on the local and total variation distances between X and another random variable Y with a known probability mass function can be valuable to get a rigorous bound on the difference $|H(X)-H(Y)|$ via Theorems 4 or 5. As a result of the calculation of such a bound on the entropy difference, it provides bounds on the entropy of X in terms of another entropy (the entropy of Y) which is assumed to be easily calculable. For example, assume that $X = \sum_{i=1}^n X_i$ is expressed as a sum of Bernoulli random variables that are either independent or weakly dependent, and may be also non-identically distributed. Let $X_i \sim \text{Bernoulli}(p_i)$, and assume that $\sum_{i=1}^n p_i = \lambda$ where all of the p_i's are much smaller than 1. In this case, the approximation of X by a Poisson distribution with mean λ (according to the law of small numbers [20]) raises the question: How close is $H(X)$ to the entropy of the Poisson distribution with mean λ ? (note that the latter entropy of the Poisson distribution is calculated efficiently in [2]). This question is especially interesting because the support of the Poisson distribution is the infinite countable set of non-negative integers, and the entropy is known not to be continuous when the support is not finite; hence, a small total variation distance does not yield in general a small difference of the two entropies. This question was addressed in [33] via the use of Corollary 3, combined with an upper bound on the total variation distance between X and Y; the latter bound is calculated via the use of the Chen-Stein method (see, e.g., [31, Chapter 2]).

Example 2 (Poisson approximation): In the following, we wish to tighten the bounds on the entropy of a sum of independent Bernoulli random variables that are not necessarily identically distributed. The bound provided in [33, Corollary 1] relies on an upper bound on the total variation distance between this sum and a Poisson random variable with the same mean (see [4, Theorem 1] or [5, Theorem 2.M]). In order to tighten the bound on the entropy in the considered setting, we further rely on a new lower bound on the total variation distance (see [34, Theorem 1 and Corollary 1]) and an upper bound on the local distance (see [5, Theorem 2.Q and Corollary 9.A.2]). The latter two bounds provide an upper bound on the ratio of the local and total variation distances, which enables to apply the bound in Theorem 5; it improves the bound in Corollary 3 which solely relies on an upper bound on the total variation distance. It is noted that the latter looser bound, which relies on Corollary 3 was used in [33] for estimating the entropy of a sum of Bernoulli random variables in the more general setting where the summands are possibly dependent.

Let $X = \sum_{i=1}^n X_i$ be a sum of independent Bernoulli random variables such that $X_i \sim \text{Bernoulli}(p_i)$ for every $i \in \{1, \ldots, n\}$. Let $Y \sim \text{Po}(\lambda)$, where $\lambda = \mathbb{E}[X] = \sum_{i=1}^n p_i$, be a Poisson random variable with mean λ. From [4, Theorem 1] (or [5, Theorem 2.M]), the following upper bound on the total variation distance holds:

$$d_{\text{TV}}(X, Y) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2. \tag{58}$$

Furthermore, from [34, Corollary 1], the following lower bound on the total variation distance holds:

$$d_{\text{TV}}(X, Y) \geq k \sum_{i=1}^n p_i^2 \tag{59}$$

where

$$k \triangleq \frac{e}{2\lambda} \, \frac{1 - \frac{1}{\theta}\left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}} \tag{60}$$

$$\theta \triangleq 3 + \frac{7}{\lambda} + \frac{1}{\lambda} \cdot \sqrt{(3\lambda + 7)\big[(3 + 2e^{-1/2})\lambda + 7\big]}. \tag{61}$$

An upper bound on the local distance between a sum of independent Bernoulli random variables and a Poisson distribution with the same mean λ follows as a special case of [5, Corollary 9.A.2] by setting $l = 1$ (so that the distribution Q_l in this corollary is specialized for $l = 1$ to the Poisson distribution $\mathrm{Po}(\lambda)$, according to [5, Eq. (1.12) on p. 177]). Since the upper bound on the right-hand side of the inequality in [5, Corollary 9.A.2] does not depend on the (time) index j, it follows that the same bound also holds while referring to

$$d_{\text{loc}}(X, Y) \triangleq \sup_{j \in \mathbb{N}_0} \big|\mathbb{P}(X = j) - \mathrm{Po}(\lambda)\{j\}\big|.$$

Based on the notation used in this corollary, it implies that if $\left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \leq \frac{1}{8}$ then the local distance between a sum of independent Bernoulli random variables $X_i \sim \mathrm{Bernoulli}(p_i)$ and a Poisson random variable with mean $\lambda = \sum_{i=1}^n p_i$ is upper bounded by

$$\begin{aligned} & d_{\text{loc}}(X, Y) \\ & \leq 4 \left(2 \max_{j \in \mathbb{N}_0} \mathbb{P}(Y = j)\right) \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \\ & \overset{(a)}{\leq} 4 \, \min\left\{\sqrt{\frac{2}{e\lambda}}, \, 2e^{-\lambda} \, I_0(\lambda)\right\} \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \end{aligned} \tag{62}$$

where inequality (a) holds due to [5, Proposition A.2.7 on pp. 262–263], and I_0 denotes the modified Bessel function of order zero. Since an upper bound on the total variation distance also forms an upper bound on the local distance, then a combination of (58) and (62) gives that

$$d_{\text{loc}}(X, Y) \leq \min\left\{1, 4\sqrt{\frac{2}{e\lambda}}, \, 8e^{-\lambda} \, I_0(\lambda)\right\} \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2. \tag{63}$$

We now apply Theorem 5 to get rigorous bounds on the entropy $H(X)$ by estimating how close it is to $H\big(\mathrm{Po}(\lambda)\big)$. Note that the improvement in the tightness of the bound in Theorem 5, in comparison to the looser bound in Corollary 3, is more significant when the ratio α of the local and total variation distances is close to zero. This happens to be the case if $\lambda \gg 1$ where due to the asymptotic expansion of I_0 (see [1, Eq. (9.7.1) on p. 377] or [12, Eq. (8.451.5) on p. 973])

$$I_0(\lambda) \approx \frac{e^{\lambda}}{\sqrt{2\pi\lambda}} \left(1 + \frac{1}{8\lambda} + \frac{9}{128\lambda^2} + \ldots\right), \quad \text{if } \lambda \gg 1$$

one gets from Eqs. (59)–(61) and (63), combined with the limit in [34, Eq. (47)], that

$$\begin{aligned}
\alpha &= \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \\
&\overset{(\text{if } \lambda \gg 1)}{\leq} \frac{4\sqrt{\frac{2}{\pi\lambda}}\,\left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_{i=1}^{n} p_i^2}{\frac{e}{6}\left(1+\sqrt{1+\frac{2}{3}\cdot e^{-1/2}}\right)^{-2}\,\left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_{i=1}^{n} p_i^2} \\
&= \frac{24}{e}\sqrt{\frac{2}{\pi}}\left(1+\sqrt{1+\frac{2}{3}\cdot e^{-1/2}}\right)^{2}\,\sqrt{\frac{1}{\lambda}} \\
&\approx \frac{33.634}{\sqrt{\lambda}}
\end{aligned} \tag{64}$$

so, for large values of λ, the upper bound on the parameter α in (13) decays to zero like the square-root of $\frac{1}{\lambda}$.

As a possible application, consider a noiseless binary-adder multiple-access channel (MAC) with n independent users where each user transmits binary symbols, and the channel output is the algebraic sum of the input symbols. The capacity region of this MAC channel is an n-dimensional polyhedron. One feature of this capacity region is the sum of the rates that is given by $R_{\text{SUM}} \triangleq \sum_{i=1}^{n} R_i$, and it is upper bounded by the joint mutual information between the input symbols $X_1, \ldots, X_n$ and the corresponding channel output $Y = \sum_{i=1}^{n} X_i$, i.e.,

$$R_{\text{SUM}} \leq \max_{P_{\mathbf{X}}: P_{\mathbf{X}} = P_{X_1} \ldots P_{X_n}} I(X_1, \ldots, X_n; Y)$$

where $H(Y|X_1, \ldots, X_n) = 0$ since the MAC is noiseless and the output symbol is the sum of the n input symbols, and therefore $I(X_1, \ldots, X_n; Y) = H(Y)$.[1] Hence, in the considered setting, the maximal sum rate is the maximal entropy of the sum of n independent binary random variables where $X_i \sim \text{Bernoulli}(p_i)$ for $i \in \{1, \ldots, n\}$. Under the constraint that $\sum_{i=1}^{n} \mathbb{E}[X_i] \leq \lambda$, it follows from the maximal entropy result in [14], [19] and [35] that the entropy of Y is maximized when the n independent inputs are i.i.d. with mean $p = \frac{\lambda}{n}$, and consequently the channel output Y is Binomially distributed with $Y \sim \text{Binom}\big(n, \frac{\lambda}{n}\big)$. For a very large number of users, the calculation of the entropy of the Binomial distribution is difficult, and it would be much easier to calculate the entropy $H\big(\text{Po}(\lambda)\big)$ for a Poisson distribution with mean λ (see [2]).

In the following, we make use of Theorem 5 to get an upper bound on the entropy difference

$$H\big(\text{Po}(\lambda)\big) - H\Big(\text{Binom}\big(n, \frac{\lambda}{n}\big)\Big) \tag{65}$$

where, due to the maximal entropy result for the Poisson distribution (see, e.g., [14], [19] or [35]), this difference is positive. Let $X \sim \text{Binom}\big(n, \frac{\lambda}{n}\big)$ be a sum of n i.i.d. Bernoulli random variables with probability of success $p = \frac{\lambda}{n}$, and let $Y \sim \text{Po}(\lambda)$. From (58), the total variation distance in this case is upper bounded by

$$d_{\text{TV}}(X,Y) \leq \frac{\lambda(1-e^{-\lambda})}{n} \triangleq \eta_1. \tag{66}$$

[1]The reader is referred to [7] for the consideration of the sum-rate for two noiseless multiple-access channels with some similarity to the binary adder channel, see footnote in [7, p. 43].

From (59) and (60), the following inequality holds:

$$d_{\mathrm{TV}}(X,Y) \geq \frac{e}{2}\,\frac{1-\frac{1}{\theta}\,\left(3+\frac{7}{\lambda}\right)}{\theta+2e^{-1/2}}\,\frac{\lambda}{n} \triangleq \eta_2 \tag{67}$$

where θ is given in (61). Furthermore, for using Theorem 5, one needs an upper bound on the local distance between the Poisson and Binomial distributions. Eq. (63) gives that

$$d_{\mathrm{loc}}(X,Y) \leq \min\left\{1, 4\sqrt{\frac{2}{\pi\lambda}},\, 8e^{-\lambda}\, I_0(\lambda)\right\}\frac{\lambda\big(1-e^{-\lambda}\big)}{n} \triangleq \eta_3. \tag{68}$$

Following the notation in Theorem 5, it follows that $m = n+1$. From (38), one needs to choose an integer M such that

$$M \geq \max\left\{n+2,\, \frac{\eta_2}{\eta_3(1-\eta_1)}\right\} \tag{69}$$

and

$$\sum_{j=M}^{\infty} \Pi_\lambda(j) \leq \eta_3 \tag{70}$$

where $\Pi_\lambda(j) \triangleq \frac{e^{-\lambda}\,\lambda^j}{j!}$ for $j \in \mathbb{N}_0$ designates the probability mass function of $\mathrm{Po}(\lambda)$. Based on Chernoff's bound,

$$\begin{aligned}
&\sum_{j=M}^{\infty} \Pi_\lambda(j)\\
&= \mathbb{P}(Y \geq M)\\
&\leq \inf_{\theta\geq 0}\left\{e^{-\theta M}\,\mathbb{E}\big[e^{\theta Y}\big]\right\}\\
&= \inf_{\theta\geq 0}\left\{e^{-\theta M}\,e^{\lambda(e^{\theta}-1)}\right\}\\
&= \exp\left\{-\left[\lambda + M\log\Big(\frac{M}{\lambda e}\Big)\right]\right\}.
\end{aligned} \tag{71}$$

Let $M \geq \lambda e^2$, then it follows from (70) and (71) that it is sufficient for M to satisfy the condition

$$\exp\big(-(\lambda+M)\big) \leq \eta_3.$$

Combining it with (69) leads to the following possible choice of M:

$$M \triangleq \max\left\{n+2,\, \frac{\eta_2}{\eta_3(1-\eta_1)},\, \lambda e^2,\, \log\left(\frac{1}{\eta_3}\right)-\lambda\right\} \tag{72}$$

where η_1, η_2 and η_3 are introduced in (66), (67), and (68) respectively. Finally, for the use of Theorem 5, one needs to choose $\eta_4 > 0$ such that $\sum_{j=M}^{\infty}\big\{-\Pi_\lambda(j)\ \log\big(\Pi_\lambda(j)\big)\big\} \leq \eta_4$. Straightforward calculation gives that

$$\begin{aligned}
&\sum_{j=M}^{\infty}\big\{-\Pi_\lambda(j)\,\log\Pi_\lambda(j)\big\}\\
&= -\lambda\,\log\lambda\sum_{j=M-1}^{\infty}\Pi_\lambda(j) + \lambda\sum_{j=M}^{\infty}\Pi_\lambda(j) + \sum_{j=M}^{\infty}\Pi_\lambda(j)\,\log(j!)\,.
\end{aligned} \tag{73}$$

From Stirling's formula, the equality $j! = \sqrt{2\pi j}\left(\frac{j}{e}\right)^j e^{\xi_j}$ holds for every $j \in \mathbb{N}$ and for some $\xi_j \in \left(\frac{1}{12j+1}, \frac{1}{12j}\right)$. This therefore implies that the third infinite sum on the right-hand side of (73) satisfies

$$\begin{aligned}
&\sum_{j=M}^{\infty} \Pi_\lambda(j)\, \log(j!) \\
&\leq \sum_{j=M}^{\infty} \Pi_\lambda(j)\, \log\left(\sqrt{2\pi j}\left(\frac{j}{e}\right)^j e^{\frac{1}{12j}}\right) \\
&= \frac{\log(2\pi)}{2}\sum_{j=M}^{\infty}\Pi_\lambda(j) + \sum_{j=M}^{\infty}\Pi_\lambda(j)\left[\left(j+\frac{1}{2}\right)\log(j) - j\right] + \frac{1}{12}\sum_{j=M}^{\infty}\frac{\Pi_\lambda(j)}{j} \\
&\leq \frac{\log(2\pi)}{2}\sum_{j=M}^{\infty}\Pi_\lambda(j) + \sum_{j=M}^{\infty}\left\{j(j-1)\,\Pi_\lambda(j)\right\} + \frac{1}{12}\sum_{j=M}^{\infty}\Pi_\lambda(j) \\
&\overset{\text{(a)}}{=} \frac{\log(2\pi)}{2}\sum_{j=M}^{\infty}\Pi_\lambda(j) + \lambda^2\sum_{j=M-2}^{\infty}\Pi_\lambda(j) + \frac{1}{12}\sum_{j=M}^{\infty}\Pi_\lambda(j) \\
&\leq \left(\frac{6\log(2\pi)+1}{12} + \lambda^2\right)\sum_{j=M-2}^{\infty}\Pi_\lambda(j)
\end{aligned} \tag{74}$$

where equality (a) follows from the identity

$$j(j-1)\,\Pi_\lambda(j) = \lambda^2\,\Pi_\lambda(j-2), \quad \forall\, j \geq 2.$$

By combining (73) and (74), it follows that

$$\begin{aligned}
&\sum_{j=M}^{\infty}\left\{-\Pi_\lambda(j)\,\log\Pi_\lambda(j)\right\} \\
&\leq \left(\lambda\,\log\left(\frac{e}{\lambda}\right)\right)_+ \sum_{j=M-1}^{\infty}\Pi_\lambda(j) + \left(\frac{6\log(2\pi)+1}{12} + \lambda^2\right)\sum_{j=M-2}^{\infty}\Pi_\lambda(j) \\
&\leq \left[\left(\lambda\,\log\left(\frac{e}{\lambda}\right)\right)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right]\sum_{j=M-2}^{\infty}\Pi_\lambda(j)
\end{aligned}$$

where M is introduced in (72), and $(x)_+ \triangleq \max\{x, 0\}$ for every $x \in \mathbb{R}$. From (71) and the last inequality, it follows that η_4 can be chosen to be

$$\eta_4 \triangleq \left[\left(\lambda\log\left(\frac{e}{\lambda}\right)\right)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right]\cdot\exp\left\{-\left[\lambda + (M-2)\,\log\left(\frac{M-2}{\lambda e}\right)\right]\right\} \tag{75}$$

At this stage, we are ready to apply Theorem 5 to derive a bound on the non-negative difference of the entropies in (65). From Theorem 5, it follows that

$$\begin{aligned}
0 &\leq H\big(\mathrm{Po}(\lambda)\big) - H\Big(\mathrm{Binom}\big(n, \frac{\lambda}{n}\big)\Big) \\
&\leq \eta_1\,\log\left(\frac{M\eta_3}{\eta_2} - 1\right) + h(\eta_1) + \eta_4.
\end{aligned} \tag{76}$$

For comparison, it follows from Corollary 3 that the upper bound on the right-hand side of (76) is replaced by

$$\eta_1\,\log(\tilde{M} - 1) + h(\eta_1) + \eta_4 \tag{77}$$

where

$$\tilde{M} \triangleq \max\Big\{n+2,\ \frac{1}{1-\eta_1}\Big\}. \tag{78}$$

Note that the bound in (76) improves the bound in (77) if $\eta_3 < \eta_2$ (i.e., if the upper bound on the local distance is smaller than the lower bound on the total variation distance). Furthermore, the latter bound does not take into account the parameters η_2 and η_3. As a numerical example, for $n = 10^6$ and $p = 0.1$, lets check the bound on the entropy difference in (65) for $\lambda = np$ (i.e., $\lambda = 10^5$). Eqs. (66)–(68), (72), (75) and (78) yield that

$$\begin{aligned}&\eta_1 = 10^{-1},\ \eta_2 = 9.5 \cdot 10^{-3},\ \eta_3 = 1.0 \cdot 10^{-3},\ \eta_4 \approx 0,\\ &M = \tilde{M} = 10^6 + 2\end{aligned}$$

and the two bounds in (76) and (77) are, respectively, equal to 1.483 and 1.707 nats, respectively. The value of $H\big(\mathrm{Po}(\lambda)\big)$ is 7.175 nats, so the entropy $H\big(\mathrm{Binom}(n, \frac{\lambda}{n})\big)$ ranges between 5.693 to 7.175 nats. Note that for $n = 10^6$ and $\lambda = 10^4$, where $p = \frac{\lambda}{n}$ is decreased from 10^{-1} to 10^{-2}, the upper bounds on (65) are decreased, respectively, to 0.183 and 0.194 nats, and $H\big(\mathrm{Po}(\lambda)\big) = 6.024$ nats. The Poisson approximation is more accurate in the latter case, consistently with the law of small numbers (see, e.g., [20]).

Remark 7: Example 2 considers the use of Theorem 5 for the estimation of the entropy of a sum of independent Bernoulli random variables. The more general case of the estimation of the entropy (via rigorous bounds) for a sum of possibly dependent Bernoulli random variables was considered in [33] by using the looser bound in Corollary 3 with an upper bound on the total variation distance that follows from the Chen-Stein method (see [3, Theorem 1]). It is noted that, in principle, also the sharper bound in Theorem 5 can be applied to obtain bounds on the entropy for a sum of possibly dependent Bernoulli random variables. To this end, in addition to the upper bound on the total variation distance in [3, Theorem 1], one needs to rely on a lower bound on the total variation distance (see [5, Chapter 3]) and an upper bound on the local distance (see [5, Theorem 2.Q on p. 42]). It is noted, however, that these distance bounds are much simplified in the setting of independent summands (see Example 2).

Remark 8: The Chen-Stein method for the Poisson approximation was adapted in [23] to the setting of the geometric distribution, and it yields a convenient method for assessing the accuracy of the geometric approximation to the distribution of the number of failures preceding the first success in dependent trials. A recent study of upper bounds on the total variation and local distances for the geometric approximation (respectively, denoted by d_1 and d_2 in [24]) enables to apply the entropy bounds in Theorem 5 and Corollary 3 in a conceptually similar way to Example 2. Furthermore, the entropy bound in Corollary 3 can be applied to compound geometric and negative binomial approximations, based on upper bounds on the total variation distance that were derived via Stein's method in [11] and [36], respectively.

V. Summary and Outlook

This paper is motivated by the fundamental question of quantifying the continuity (or lack of it) of entropy, with respect to natural topologies on discrete probability distributions. This question has been studied in the literature for the topology induced by the total variation distance, and there it is well known that the entropy is continuous when the alphabet is finite, but not when the alphabet is countably infinite (see, e.g., [17] and references therein). To set terminology, the local and total variation distances are introduced in Definition 3 (see Section I): the local distance between two discrete random variables is defined to be the l^∞ distance between their

probability mass functions, and the total variation distance is half the l^1 distance; it is easy to show that the local distance is less than or equal to the total variation distance.

A key tool in this paper is an explicit construction for maximal coupling, i.e., a coupling $(\hat{X}, \hat{Y})$ of the random variables X and Y that maximizes the probability $\mathbb{P}(\hat{X} = \hat{Y})$. The notion of *maximal coupling* is also known to be useful for the derivation of error bounds via Stein's method (see, e.g., [31, Chapter 2] and [32]). Stein's method also serves in this paper to exemplify the use of the new bounds in the context of the Poisson approximation [5]; this is done by using good upper and lower bounds on the total variation distance (see [4] and [34]) and a good upper bound on the local distance [5]. The link between Stein's method and information theory was pioneered in [6] in the context of the compound Poisson approximation, and it has been further recently studied in [22].

This paper starts by introducing preliminary material in Sections I and II; Theorems 1–3 are known results on maximal coupling, and a bound from [38] on the difference of the entropies of two discrete random variables in terms of the total variation distance. Note that the proofs of these known results are important for the analysis in this paper.

The new results in this paper are the following:

1) For two given distributions on a finite alphabet, if the local distance is strictly smaller than the total variation distance, then Theorem 4 provides a new bound which can be significantly better than the previously best known bound (Theorem 3) due to Zhang [38].
2) For countably infinite alphabets, a knowledge of the total variation distance between two distributions is not sufficient for establishing an informative bound on the difference of entropies (i.e., one has discontinuity of entropy). Theorem 5 demonstrates that if one of the distributions is finitely supported and some knowledge of the other distribution is available, then the knowledge of the local and total variation distances (or bounds on these distances) allows one to bound the difference of the entropies even in this case.
3) Refined bounds on the entropy of near-uniform random variables on large alphabets, as well as of sums of independent Bernoulli random variables (which arise in numerous applications, see [3] and references therein) are obtained in Section IV (see Examples 1 and 2). These refined bounds are compared with previously known bounds. One special case where the entropy can be explicitly evaluated and compared to various bounds is worked out, and it is shown that Theorem 4 improves significantly the known bound in Theorem 3.

A natural question that arises in the context of this paper is what if one only has bounds on the local distance ? A treatment of this problem (which does not exist in the literature) possibly gives further insight into why the local distance is useful in combination of the total variation distance. In the finite alphabet case, the two metrics are equivalent since

$$d_{\text{loc}}(X,Y) \leq d_{\text{TV}}(X,Y) \leq \frac{M}{2} \cdot d_{\text{loc}}(X,Y)$$

and, hence, generate the same topology; so the bare continuity of entropy is guaranteed for finite alphabets, and so is the discontinuity of the entropy for infinite alphabets. But are there tight bounds on the difference of entropies just based on the local distance for finite alphabets ?

The following proposition suggests a simple bound on the difference of entropies of two discrete random variables that are finitely supported, based only on their local distance:

Proposition 1: Let X and Y be discrete random variables that take values in a finite set $\mathcal{A}$, and let $|\mathcal{A}| = M$. If $d_{\text{loc}}(X,Y) \leq \frac{1}{e}$, then

$$|H(X) - H(Y)| \leq -M\, d_{\text{loc}}(X,Y)\, \log\bigl(d_{\text{loc}}(X,Y)\bigr) \tag{79}$$

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

Proof: The derivation of this bound forms a small modification of the proof of [8, Theorem 17.3.3]. Let P_X and P_Y denote the probability mass functions of X and Y, respectively, and let $r(u) \triangleq |P_X(u) - P_Y(u)|$ for every $u \in \mathcal{A}$. From [8, Eqs. (17.27)–(17.30)], if $r(u) \leq \frac{1}{2}$ for every $u \in \mathcal{A}$, then

$$|H(X) - H(Y)| \leq \sum_{u \in \mathcal{A}} -r(u) \, \log r(u).$$

The bound in (79) now follows from the simple inequality $r(u) \leq d_{\text{loc}}(X, Y)$ for every $u \in \mathcal{A}$ (by definition), and due to the fact that the function $f(x) = -x \log(x)$ is monotonic increasing over the interval $[0, \frac{1}{e}]$. ■

Remark 9: The bound in (79) does not necessarily hold if $d_{\text{loc}}(X, Y) > \frac{1}{e}$. As a counter example, let $\mathcal{A}$ be a set of 3 elements, and let

$$P_X = \Big(\frac{1}{2}, \frac{1}{2}, 0\Big), \quad P_Y = (0, 0, 1).$$

Then $d_{\text{loc}}(X, Y) = 1$ and $H(X) - H(Y) = \log 2$, so (79) is not satisfied due to the violation of the condition on the local distance in Proposition 1.

Remark 10: A slight loosening of the bound in (79) gives that if $d_{\text{loc}}(X, Y) \leq \frac{1}{e}$, then

$$|H(X) - H(Y)| \leq M \, h\big(d_{\text{loc}}(X, Y)\big)$$

where h is the binary entropy function. In the simple case where the probability mass functions of X and Y are equal to $P_X = (1-\varepsilon, \varepsilon)$ and $P_Y = (1, 0)$, respectively, we have $d_{\text{loc}}(X, Y) = \varepsilon$; if $0 < \varepsilon \leq \frac{1}{e}$, the bound on $|H(X) - H(Y)|$ is twice larger than its exact value that is equal to $h(\varepsilon)$. Even in this simple case, the bound on the difference of the entropies that only depends on the local distance is not tight. On one hand, it will be of interest to derive tighter bounds on the difference of entropies for finite alphabets that are just based on the local distance; on the other hand, even the simple bound in Proposition 1 provides some insight into why the local distance is useful in combination of the total variation distance for upper bounding the difference of entropies for finite alphabets (see Theorem 4).

Acknowledgment

An anonymous reviewer of this journal paper and the conference version at ISIT 2013 is gratefully acknowledged for suggestions that led to an improvement of the presentation, and for raising the question in Section V that led to the bound in Proposition 1. The Associate Editor, Ioannis Kontoyiannis, is acknowledged for handling the manuscript. This research work was supported by the Israeli Science Foundation (ISF), grant number 12/12.

REFERENCES

[1] M. Abramowitz and A. I. Stegun, *Handbook of Mathmetical Functions*, Dover Publications, New York, 1965.

[2] J. A. Adell, A. Lekouna and Y. Yu, "Sharp bounds on the entropy of the Poisson law and related quantities," *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2299–2306, May 2010.

[3] R. Arratia, L. Goldstein and L. Gordon, "Two moments suffice for Poisson approximations: The Chen-Stein method," *Annals of Probability*, vol. 17, no. 1, pp. 9–25, January 1989.

[4] A. D. Barbour and P. Hall, "On the rate of Poisson Convergence," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 95, no. 3, pp. 473–480, 1984.

[5] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford University Press, 1992.

[6] A. D. Barbour, O. Johnson, I. Kontoyiannis and M. Madiman, "Compound Poisson approximation via information functionals," *Electronic Journal of Probability*, vol. 15, pp. 1344–1368, 2010.

[7] S. C. Chung and J. Wolf, "On the T-user M-frequency noiseless multiple-access channel with and without intensity information," *IEEE Trans. on Information Theory*, vol. 27, no. 1, pp. 41–48, January 1981.

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

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[10] I. Csiszár, "Almost independence and secrecy capacity," *Problems of Information Transmission*, vol. 32, no. 1, pp. 40–47, March 1996.

[11] F. Daly, "Stein's method for compound geometric approximation," *Journal of Applied Probability*, vol. 47, no. 1, pp. 146–156, March 2010.

[12] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, (edited by A. Jeffrey), fifth edition, Academic Press, 1994.

[13] R. M. Gray, *Entropy and Information Theory*, second edition, Springer, 2010.

[14] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. on Information Theory*, vol. 47, no. 5, pp. 2039–2041, July 2001.

[15] P. Harremoës, "Information topologies with applications," *Entropy, Search, Complexity*, vol. 16, pp. 113–150, 2007.

[16] S. W. Ho and R. W. Yeung, "On the discontinuity of the Shannon information measures," *IEEE Trans. on Information Theory*, vol. 55, no. 12, pp. 5362–5374, December 2009.

[17] 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.

[18] S. W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 5930–5942, December 2010.

[19] S. Karlin and Y. Rinott, "Entropy inequalities for classes of probability distributions I: the univariate case," *Advances in Applied Probability*, vol. 13, no. 1, pp. 93–112, March 1981.

[20] I. Kontoyiannis, P. Harremoës and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Information Theory*, vol. 51, no. 2, pp. 466–472, February 2005.

[21] M. Kovačević, I. Stanojević and V. Šenk, "The entropy of couplings," March 2013. [Online]. Available: http://arxiv.org/abs/1303.3235.

[22] C. Ley and Y. Swan, "Local Pinsker inequalities via Stein's discrete density approach," *IEEE Trans. on Information Theory*, vol. 59, no. 9, pp. 5584–5591, September 2013.

[23] E. Peköz, "Stein's method for geometric approximation," *Journal of Applied Probability*, vol. 33, no. 3, pp. 707–713, September 1996.

[24] E. Peköz, A. Röllin and N. Ross, "Total variation error bounds for geometric approximation," *Bernoulli*, vol. 19, no. 2, pp. 610–632, May 2013.

[25] M. S. Pinsker, "On estimation of information via variation," *Problems of Information Transmission*, vol. 41, no. 2, pp. 71–75, 2005.

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

[27] 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.

[28] V. V. Prelov, "Mutual information of several random variables and its estimation via variation," *Problems of Information Transmission*, vol. 45, no. 4, pp. 3–17, December 2009.

[29] V. V. Prelov, "On computation of information via variation and inequalities for the entropy function," *Problems of Information Transmission*, vol. 46, no. 2, pp. 24–29, June 2010.

[30] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

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

[32] N. Ross, "Fundamentals of Stein's Method," *Probability Surveys*, vol. 8, pp. 210–293, 2011.

[33] I. Sason, "On the entropy of sums of Bernoulli random variables via the Chen-Stein method," *Proceedings of the 2012 IEEE International Workshop on Information Theory*, pp. 542–546, Lausanne, Switzerland, September 2012.

[34] I. Sason, "Improved lower bounds on the total variation distance for the Poisson approximation," *Statistics and Probability Letters*, vol. 83, no. 10, pp. 2422–2431, October 2013.

[35] L. A. Shepp and I. Olkin, "Entropy of the sum of independent Bernoulli random variables and the multinomial distribution," *Contributions to Probability*, pp. 201–206, Academic Press, New York, 1981.

[36] K. Teerapabolarn and A. Boondirek, "Negative binomial approximation with Stein's method and Stein's identity," *International Mathematical Forum*, vol. 5, no. 51, pp. 2541–2551, 2010.

[37] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. on Information Theory*, vol. 46, no. 4, pp. 1602–1609, July 2000.

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