Entropy Bounds for Discrete Random Variables via Coupling

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

***Abstract*—This work provides new bounds on the difference between 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 couplings, and the bounds 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 entropy bounds is exemplified for the Poisson approximation, where bounds on the local and total variation distances follow from Stein's method. The full paper version for this work is available at http://arxiv.org/abs/1209.5259.**

***Index Terms*—Entropy, local distance, maximal coupling, Poisson approximation, Stein's method, total variation distance.**

I. Introduction

The question of quantifying the continuity (or lack of it) of entropy, with respect to the topology on discrete probability distributions induced by the total variation distance, is of basic interest. The interplay between the entropy difference of two discrete random variables and their total variation distance has been extensively studied (see, e.g., [7], [8], [9, Lemma 1], [12], [13], [16], [18], [22] and [26]).

The new bounds on the entropy difference of discrete random variables, as introduced in this work [23], 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., [19, Chapter 2] and [20]). The link between Stein's method and information theory was pioneered in [5] in the context of the compound Poisson distribution, and a recent work [17] 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 work [23].

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 $\mathbb{P}(\hat{X} = \hat{Y})$ is as large as possible 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, the total variation distance is half the l^1 distance, and it can be verified 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, the term 'distribution' refers to the probability mass function of a discrete random variable defined over a finite or countably infinite alphabet.

This work refines bounds on the entropy difference 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 [16] 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 (see Section III). 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 (i.e., one has discontinuity of entropy). It is shown in this work that if one of the distributions is finitely supported, then knowing also something about the local distance and the tail behavior of the other distribution allows to bound the difference of entropies in this case. This 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 (see Section IV). 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 [21, Section II] and [22] to get bounds on the entropy of sums of dependent and non-identically distributed Bernoulli random variables.

II. A Known Bound on the Entropy of Discrete Random Variables

The following theorem relies on a bound that first appeared in [26, Eq. (4)] and proved by coupling. It was later introduced in [13, 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 [26, Section 3], the bound on the entropy difference that is introduced in the following theorem improves the bound in [7, Theorem 17.3.3] or [8, Lemma 2.7].

Theorem 1: Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. If $d_{\text{TV}}(X,Y) \leq \varepsilon$, then

$$|H(X) - H(Y)| \leq \begin{cases} \varepsilon \log(M-1) + h(\varepsilon) & \text{if } \varepsilon \in \left[0, 1 - \frac{1}{M}\right] \\ \log(M) & \text{if } \varepsilon > 1 - \frac{1}{M} \end{cases}$$

where h denotes the binary entropy function.

A shortened proof of this theorem appears in [23, Section II] via the use of maximal coupling.

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

In the cases where the known bound in Theorem 1 is tight, it can be verified that the local distance is equal to the total variation distance [23]. 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 1 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 1 but otherwise, it improves the bound in Theorem 1. The general approach for proving the following new inequalities relies on maximal coupling (see [23]). The new bound is the following:

Theorem 2: Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$\begin{aligned} &|H(X) - H(Y)| \\ &\leq d_{\text{TV}}(X,Y) \, \log(M\alpha - 1) + h\big(d_{\text{TV}}(X,Y)\big) \end{aligned} \tag{3}$$

where

$$\alpha \triangleq \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \tag{4}$$

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 (3) is tightened to

$$\begin{aligned} &|H(X) - H(Y)| \\ &\leq d_{\text{TV}}(X,Y) \, \log\left(\frac{M\alpha - 1}{4}\right) + h\big(d_{\text{TV}}(X,Y)\big). \end{aligned} \tag{5}$$

Proof: See [23, Section III]. ■

Remark 1: Since, in general, $\alpha \leq 1$ then the case where $\alpha = 1$ is the worst case for the bound in (3). In the latter case, it is particularized to the bound in Theorem 1 (see [13, Theorem 6] or [26, Eq. (4)]).

Remark 2: If $\alpha \leq \frac{1}{N}$ for some integer N (since $\alpha \in \left[\frac{2}{M}, 1\right]$ then it yields that $N \in \{1, \ldots, \lfloor \frac{M}{2} \rfloor\}$), the bound in (3) implies that

$$\begin{aligned} &|H(X) - H(Y)| \\ &\leq d_{\text{TV}}(X,Y) \, \log\left(\frac{M-N}{N}\right) + h\big(d_{\text{TV}}(X,Y)\big). \end{aligned} \tag{6}$$

The bounds in (6) and [13, Theorem 7] are similar *but they hold under different conditions*. The bound in [13, Theorem 7] requires that $P_X, P_Y \leq \frac{1}{N}$ everywhere, whereas the bound in (6) 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.

Corollary 1: Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Assume that for some positive constants $\varepsilon_1, \varepsilon_2$

$$d_{\text{TV}}(X,Y) \leq \varepsilon_1 \leq 1 - \frac{1}{M\varepsilon_2}\,, \tag{7}$$

$$\frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \leq \varepsilon_2 \leq 1. \tag{8}$$

Then,

$$|H(X) - H(Y)| \leq \varepsilon_1 \, \log(M\varepsilon_2 - 1) + h(\varepsilon_1). \tag{9}$$

Proof: See [23, Section III]. ■

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

$$H(P_X \times P_Y) - H(P_{X,Y}) = I(X;Y).$$

Hence, Theorem 2 and Corollary 1 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 2 to this setting tightens the bound in [26, 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).

We proceed to consider the entropy difference of discrete random variables in the case of countably infinite alphabets.

Theorem 3: Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countably 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

$$\eta_2 \leq d_{\text{TV}}(X,Y) \leq \eta_1, \quad d_{\text{loc}}(X,Y) \leq \eta_3 \tag{10}$$

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

$$\sum_{i=M}^{\infty} P_Y(a_i) \leq \eta_3, \quad M \geq \max\left\{m+1, \frac{\eta_2}{(1-\eta_1)\eta_3}\right\} \tag{11}$$

and let $\eta_4 > 0$ satisfy

$$-\sum_{i=M}^{\infty} P_Y(a_i) \, \log P_Y(a_i) \leq \eta_4. \tag{12}$$

Then, the following inequality holds:

$$|H(X) - H(Y)| \le \eta_1 \log\!\left(\frac{M\eta_3}{\eta_2} - 1\right) + h(\eta_1) + \eta_4. \quad (13)$$

Proof: See [23]. ■

Corollary 2: In the setting of X and Y in Theorem 3, assume that $d_{\text{TV}}(X,Y) \le \eta$ for some $\eta \in (0,1)$. Let $M \triangleq \max\left\{m+1, \frac{1}{1-\eta}\right\}$, and assume that for some $\mu > 0$

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

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

Proof: This corollary follows from Theorem 3 by setting $\eta_2 = \eta_3 = d_{\text{loc}}(X,Y)$ (note that $d_{\text{loc}}(X,Y) \le d_{\text{TV}}(X,Y)$), and then η_1 and η_4 are replaced by η and μ, respectively. ■

Remark 4: The result in Corollary 2 coincides with [21, Theorem 4], which gives a bound on the entropy difference in terms of the total variation distance by relying on the bound in [26, Eq. (4)] or [13, Theorem 6].

IV. An Example: The Poisson Approximation

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

In many interesting applications, the exact distribution of X is not available or is numerically hard to compute. In such cases, 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 2 or 3. 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 [15]) 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 [1]). This question is especially interesting because the support of the Poisson distribution is the countably infinite set of non-negative integers, so a small total variation distance does not necessarily yield a small difference between the two entropies. This question was addressed in [21, Section 2] and [22] via the use of Corollary 2 (which coincides with [21, Theorem 4]), combined with an upper bound on the total variation distance between X and Y where the latter bound is calculated via the use of the Chen-Stein method (see, e.g., [19, Chapter 2]).

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 [21, Proposition 1] relies on an upper bound on the total variation distance between this sum and a Poisson random variable with the same mean (see [3, Theorem 1] or [4, Theorem 2.M]). In order to tighten the bound on the entropy in the considered setting, we further rely on a lower bound on the total variation distance (see [24, Theorem 1 and Corollary 1]) and an upper bound on the local distance (see [4, 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 3; it improves the bound in Corollary 2 which solely relies on an upper bound on the total variation distance. It is noted that the latter looser bound, which relies on Corollary 2 was used in [22] 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 with $X_i \sim \text{Bernoulli}(p_i)$ for $i \in \{1, \ldots, n\}$. Let $\sum_{i=1}^n p_i = \lambda$, and let $Y \sim \text{Po}(\lambda)$ be a Poisson random variable with mean λ. From [3, Theorem 1] (or [4, Theorem 2.M]), the following upper bound on the total variation distance holds:

$$d_{\text{TV}}(X,Y) \le \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2. \quad (14)$$

Furthermore, from [24, Corollary 1], the following lower bound on the total variation distance holds:

$$d_{\text{TV}}(X,Y) \ge k \sum_{i=1}^n p_i^2 \quad (15)$$

where

$$k \triangleq \frac{e}{2\lambda} \, \frac{1 - \frac{1}{\theta}\left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}} \quad (16)$$

$$\theta \triangleq 3 + \frac{7}{\lambda} + \frac{1}{\lambda} \cdot \sqrt{(3\lambda+7)\big[(3+2e^{-1/2})\lambda + 7\big]}. \quad (17)$$

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 [4, 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 $\text{Po}(\lambda)$, according to [4, Eq. (1.12) on p. 177]). Since the upper bound on the right-hand side of the inequality in [4, Corollary 9.A.2] does not depend on the (time) index j, it follows that the same bound also holds while referring to $d_{\text{loc}}(X,Y) \triangleq \sup_{j \in \mathbb{N}_0} \big|\mathbb{P}(X = j) - \text{Po}(\lambda)\{j\}\big|$. Based on the notation used in this corollary, it implies that if $\left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \le \frac{1}{8}$ then the local distance between a sum of independent Bernoulli random variables $X_i \sim \text{Bernoulli}(p_i)$ and a Poisson random variable with mean $\lambda = \sum_{i=1}^n p_i$ is upper bounded by

$$d_{\text{loc}}(X,Y) \le 4 \min\left\{\sqrt{\frac{2}{e\lambda}},\, 2e^{-\lambda} I_0(\lambda)\right\} \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \quad (18)$$

where this inequality holds due to [4, 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 (14) and (18) gives that

$$d_{\text{loc}}(X,Y) \le \min\left\{1, 4\sqrt{\frac{2}{e\lambda}},\, 8e^{-\lambda}\, I_0(\lambda)\right\} \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2. \quad (19)$$

We now apply Theorem 3 to get rigorous bounds on the entropy $H(X)$ by estimating how close it is to $H\bigl(\text{Po}(\lambda)\bigr)$. Note that the improvement in the tightness of the bound in Theorem 3, in comparison to the looser bound in Corollary 2, is more remarkable 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, e.g., [10, Eq. (8.451.5) on p. 973])

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

one gets from Eqs. (15)–(17) and (19), combined with the limit in [24, Eq. (48)], that

$$\alpha = \frac{d_{\text{loc}}(X,Y)}{d_{\text{TV}}(X,Y)} \overset{(\text{if } \lambda \gg 1)}{\le} \frac{33.634}{\sqrt{\lambda}} \quad (20)$$

so, for large values of λ, the upper bound on the parameter α in (4) 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 $X_1, \dots, X_n$ and the corresponding channel output $Y = \sum_{i=1}^{n} X_i$, i.e.,

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

where, since the MAC is noiseless and the output symbol is the sum of the n input symbols then $H(Y|X_1, \dots, X_n) = 0$, and therefore $I(X_1, \dots, X_n; Y) = H(Y)$.[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 $i \in \{1, \dots, n\}$. Under the constraint that $\sum_{i=1}^{n} \mathbb{E}[X_i] \le \lambda$, it follows from the maximal entropy result in [11], [14] and [25] 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 $Y \sim \text{Binom}\bigl(n, \frac{\lambda}{n}\bigr)$. 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 $H\bigl(\text{Po}(\lambda)\bigr)$ for a Poisson distribution with mean λ (see [1]).

In the following, we make use of Theorem 3 to get an upper bound on the entropy difference

$$H\bigl(\text{Po}(\lambda)\bigr) - H\Bigl(\text{Binom}\bigl(n, \frac{\lambda}{n}\bigr)\Bigr) \quad (21)$$

where, due to the maximal entropy result for the Poisson distribution (see, e.g., [11], [14] or [25]), this difference is positive. Let $X \sim \text{Binom}\bigl(n, \frac{\lambda}{n}\bigr)$ 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 (14), the total variation distance in this case is upper bounded by

$$d_{\text{TV}}(X,Y) \le \frac{\lambda(1-e^{-\lambda})}{n} \triangleq \eta_1. \quad (22)$$

From (15) and (16), the following inequality holds:

$$d_{\text{TV}}(X,Y) \ge \frac{e}{2}\, \frac{1 - \frac{1}{\theta}\left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}}\, \frac{\lambda}{n} \triangleq \eta_2 \quad (23)$$

where θ is given in (17). Furthermore, for using Theorem 3, one needs an upper bound on the local distance between the Poisson and Binomial distributions. Eq. (19) gives that

$$d_{\text{loc}}(X,Y) \le \min\left\{1, 4\sqrt{\frac{2}{\pi\lambda}},\, 8e^{-\lambda}\, I_0(\lambda)\right\} \frac{\lambda\bigl(1-e^{-\lambda}\bigr)}{n} \triangleq \eta_3. \quad (24)$$

Following the notation in Theorem 3, it follows that $m = n+1$. From (11), one needs to choose an integer M such that

$$M \ge \max\left\{n+2,\, \frac{\eta_2}{\eta_3(1-\eta_1)}\right\} \quad (25)$$

and

$$\sum_{j=M}^{\infty} \Pi_{\lambda}(j) \le \eta_3 \quad (26)$$

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

$$\sum_{j=M}^{\infty} \Pi_{\lambda}(j) = \mathbb{P}(Y \ge M) \le \exp\left\{-\left[\lambda + M \ln\Bigl(\frac{M}{\lambda e}\Bigr)\right]\right\}.$$

Let $M \ge \lambda e^2$, then it follows from (26) and (27) that it is sufficient for M to satisfy the condition $\exp\bigl(-(\lambda+M)\bigr) \le \eta_3$. Combining it with (25) leads to the following possible choice:

$$M \triangleq \max\left\{n+2,\, \frac{\eta_2}{\eta_3(1-\eta_1)},\, \lambda e^2,\, \ln\Bigl(\frac{1}{\eta_3}\Bigr) - \lambda\right\} \quad (27)$$

where η_1, η_2 and η_3 are introduced in (22), (23), and (24) respectively. Finally, for the use of Theorem 3, one needs to choose $\eta_4 > 0$ such that $\sum_{j=M}^{\infty}\bigl\{-\Pi_{\lambda}(j)\, \log\bigl(\Pi_{\lambda}(j)\bigr)\bigr\} \le \eta_4$. From the analysis in [21, Eqs. (43)–(47)], it follows from the last inequality and [21, Eq. (47)] that η_4 here is equal to μ in [21, Eq. (23)], i.e.,

$$\eta_4 \triangleq \left[\Bigl(\lambda \log\Bigl(\frac{e}{\lambda}\Bigr)\Bigr)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right] \exp\left\{-\left[\lambda + (M-2)\,\log\left(\frac{M-2}{\lambda e}\right)\right]\right\} \quad (28)$$

[1] The reader is referred to [6] for the consideration of the sum-rate for two noiseless multiple-access channels with some similarity to the binary adder channel, see footnote in [6, p. 43].

where M is introduced in (27), and $(x)_+ \triangleq \max\{x, 0\}$ for every $x \in \mathbb{R}$. At this stage, we are ready to apply Theorem 3 to derive a bound on the non-negative difference between the entropies in (21). From Theorem 3, it follows that

$$\begin{aligned} 0 &\le H\big(\text{Po}(\lambda)\big) - H\Big(\text{Binom}\big(n, \frac{\lambda}{n}\big)\Big) \\ &\le \eta_1 \log\left(\frac{M\eta_3}{\eta_2} - 1\right) + h(\eta_1) + \eta_4. \end{aligned} \tag{29}$$

For comparison, it follows from Corollary 2 that the upper bound on the right-hand side of (29) is replaced by

$$\eta_1 \log(\tilde{M} - 1) + h(\eta_1) + \eta_4 \tag{30}$$

where

$$\tilde{M} \triangleq \max\Big\{n + 2, \frac{1}{1 - \eta_1}\Big\}. \tag{31}$$

Note that the bound in (29) improves the bound in (30) 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 $p = 0.1$, let's check the bound on the entropy difference in (21) for $\lambda = np$ (i.e., $\lambda = 10^5$). Eqs. (22)–(24), (27), (28) and (31) yield that $\eta_1 = 10^{-1}$, $\eta_2 = 9.5 \cdot 10^{-3}$, $\eta_3 = 1.0 \cdot 10^{-3}$, $\eta_4 \approx 0$, and $M = \tilde{M} = 10^6 + 2$; the two bounds in (29) and (30) are, respectively, equal to 1.483 and 1.707 nats, respectively. The value of $H\big(\text{Po}(\lambda)\big)$ is 7.175 nats, so the entropy $H\big(\text{Binom}(n, \frac{\lambda}{n})\big)$ 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 (21) are decreased, respectively, to 0.183 and 0.194 nats, and $H\big(\text{Po}(\lambda)\big) = 6.024$ nats. The Poisson approximation is more accurate in the latter case, consistently with the law of small numbers (see, e.g., [15]).

Remark 5: The above example considers the use of Theorem 3 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 [22] by using the looser bound in Corollary 2 with an upper bound on the total variation distance that follows from the Chen-Stein method (see [2, Theorem 1]). It is noted that, in principle, also the sharper bound in Theorem 3 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 [2, Theorem 1], one needs to rely on a lower bound on the total variation distance (see [4, Chapter 3]) and an upper bound on the local distance (see [4, Theorem 2.Q on p. 42]). It is noted, however, that these distance bounds are much simplified in the setting of independent summands.

Acknowledgment

This research work was supported by the Israeli Science Foundation (ISF), grant number 12/12.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] C. Ley and Y. Swan, "Stein's density approach for discrete distributions and information inequalities," submitted to the *IEEE Trans. on Information Theory*, November 2012. [Online]. Available: http://arxiv.org/abs/1211.3668.

[18] 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.

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

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

[21] I. Sason, "An information-theoretic perspective of the Poisson approximation via the Chen-Stein method," unpublished, June 2012. [Online]. Available: http://arxiv.org/abs/1206.6811.

[22] 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.

[23] I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to the *IEEE Trans. on Information Theory*, September 2012. [Online]. Available: http://arxiv.org/abs/1209.5259.

[24] I. Sason, "Improved lower bounds on the total variation distance for the Poisson approximation," submitted to the *Statistics and Probability Letters*, April 2013. [Online]. Available: http://arxiv.org/abs/1301.7504.

[25] 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.

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