IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

An Information-Theoretic Perspective of the Poisson Approximation via the Chen-Stein Method

Igal Sason

CCIT Report #817
October 2012

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

An Information-Theoretic Perspective of the Poisson Approximation via the Chen-Stein Method

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

Abstract

The first part of this work considers the entropy of the sum of (possibly dependent and non-identically distributed) Bernoulli random variables. Upper bounds on the error that follows from an approximation of this entropy by the entropy of a Poisson random variable with the same mean are derived via the Chen-Stein method. The second part of this work derives new lower bounds on the total variation distance and relative entropy between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution. The starting point of the derivation of the new bounds in the second part of this work is an introduction of a new lower bound on the total variation distance, whose derivation generalizes and refines the analysis by Barbour and Hall (1984), based on the Chen-Stein method for the Poisson approximation. A new lower bound on the relative entropy between these two distributions is introduced, and this lower bound is compared to a previously reported upper bound on the relative entropy by Kontoyiannis et al. (2005). The derivation of the new lower bound on the relative entropy follows from the new lower bound on the total variation distance, combined with a distribution-dependent refinement of Pinsker's inequality by Ordentlich and Weinberger (2005). Upper and lower bounds on the Bhattacharyya parameter, Chernoff information and Hellinger distance between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution with the same mean are derived as well via some relations between these quantities with the total variation distance and the relative entropy. The analysis in this work combines elements of information theory with the Chen-Stein method for the Poisson approximation. The resulting bounds are easy to compute, and their applicability is exemplified.

Index Terms

Chen-Stein method, Chernoff information, entropy, error bounds, error exponents, Poisson approximation, relative entropy, total variation distance.

AMS 2000 Subject Classification: Primary 60E07, 60E15, 60G50, 94A17.

I. Introduction

Convergence to the Poisson distribution, for the number of occurrences of possibly dependent events, naturally arises in various applications. Following the work of Poisson, there has been considerable interest in how well the Poisson distribution approximates the binomial distribution. This approximation was treated by a limit theorem in [15, Chapter 8], and later some non-asymptotic results have considered the accuracy of this approximation. Among these old and interesting results, Le Cam's inequality [35] provides an upper bound on the total variation distance between the distribution of the sum $S_n = \sum_{i=1}^n X_i$ of n independent Bernoulli random variables $\{X_i\}_{i=1}^n$, where $X_i \sim \mathrm{Bern}(p_i)$, and a Poisson distribution $\mathrm{Po}(\lambda)$ with mean $\lambda = \sum_{i=1}^n p_i$. This inequality states that

$$d_{\mathrm{TV}}\big(P_{S_n}, \mathrm{Po}(\lambda)\big) \triangleq \frac{1}{2} \sum_{k=0}^{\infty} \Big| \mathbb{P}(S_n = k) - \frac{e^{-\lambda}\lambda^k}{k!} \Big| \leq \sum_{i=1}^n p_i^2$$

so if, e.g., $X_i \sim \mathrm{Bern}\big(\frac{\lambda}{n}\big)$ for every $i \in \{1, \ldots, n\}$ (referring to the case that S_n is binomially distributed) then this upper bound is equal to $\frac{\lambda^2}{n}$, thus decaying to zero as n tends to infinity. This upper bound was later improved, e.g., by Barbour and Hall (see [4, Theorem 1]), replacing the above upper bound by $\Big(\frac{1-e^{-\lambda}}{\lambda}\Big) \sum_{i=1}^n p_i^2$ and therefore improving it by a factor of $\frac{1}{\lambda}$ when λ is large. This improved upper bound was also proved by Barbour and Hall to be essentially tight (see [4, Theorem 2]) with the following lower bound on the total variation distance:

$$d_{\mathrm{TV}}\big(P_{S_n}, \mathrm{Po}(\lambda)\big) \geq \frac{1}{32} \min\Big\{1, \frac{1}{\lambda}\Big\} \sum_{i=1}^n p_i^2$$

so the upper and lower bounds on the total variation distance differ by a factor of at most 32, irrespectively of the value of λ (it is noted that in [5, Remark 3.2.2], the factor $\frac{1}{32}$ in the lower bound was claimed to be improvable to $\frac{1}{14}$ with no explicit proof). The Poisson approximation and later also the compound Poisson approximation have been extensively treated in the literature (see, e.g., the reference list in [5] and this paper).

Among modern methods, the Chen-Stein method forms a powerful probabilistic tool that is used to calculate error bounds when the Poisson approximation serves to assess the distribution of a sum of (possibly dependent) Bernoulli random variables [10]. This method is based on the simple property of the Poisson distribution where $Z \sim \mathrm{Po}(\lambda)$ with $\lambda \in (0, \infty)$ if and only if $\lambda \, \mathbb{E}[f(Z+1)] - \mathbb{E}[Z \, f(Z)] = 0$ for all bounded functions f that are defined on $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$. This method provides a rigorous analytical treatment, via error bounds, to the case where W has approximately the Poisson distribution $\mathrm{Po}(\lambda)$ where it can be expected that $\lambda \, \mathbb{E}[f(W+1)] - \mathbb{E}[W \, f(W)] \approx 0$ for an arbitrary bounded function f that is defined on $\mathbb{N}_0$. The interested reader is referred to several comprehensive surveys on the Chen-Stein method in [3], [5], [6, Chapter 2], [9], [42, Chapter 2] and [43].

During the last decade, information-theoretic methods were exploited to establish convergence to Poisson and compound Poisson limits in suitable paradigms. An information-theoretic study of the convergence rate of the binomial-to-Poisson distribution, in terms of the relative entropy between the binomial and Poisson distributions, was provided in [19], and maximum entropy results for the binomial, Poisson and compound Poisson distributions were studied in [18], [28], [32], [46], [50], [51] and [52]. The law of small numbers refers to the phenomenon that, for random variables $\{X_i\}_{i=1}^n$ defined on $\mathbb{N}_0$, the sum $\sum_{i=1}^n X_i$ is approximately Poisson distributed with mean $\lambda = \sum_{i=1}^n p_i$ if (qualitatively) the following conditions hold: $\mathbb{P}(X_i = 0)$ is close to 1, $\mathbb{P}(X_i = 1)$ is uniformly small, $\mathbb{P}(X_i > 1)$ is negligible as compared to $\mathbb{P}(X_i = 1)$, and $\{X_i\}_{i=1}^n$ are weakly dependent (see [17], [44] and [45]). An information-theoretic study of the law of small numbers was provided in [33] via the derivation of upper bounds on the relative entropy between the distribution of the sum of possibly dependent Bernoulli random variables and the Poisson distribution with the same mean. An extension of the law of small numbers to a thinning limit theorem for convolutions of discrete distributions that are defined on $\mathbb{N}_0$ was introduced in [22], followed by an analysis of the convergence rate and some non-asymptotic results. Further work in this direction was studied in [30], and the work in [7] provides an information-theoretic study for the problem of compound Poisson approximation, which parallels the earlier study for the Poisson approximation in [33]. A recent follow-up to the works in [7] and [33] is provided in [36] and [37], considering connections between Stein characterizations and Fisher information functionals. Nice surveys on the line of work on information-theoretic aspects of the Poisson approximation are introduced in [28, Chapter 7] and [34]. Furthermore, [13, Chapter 2] surveys some commonly-used metrics between probability measures with some pointers to the Poisson approximation.

This paper provides an information-theoretic study of the Poisson approximation via the Chen-Stein method. The novelty of this paper is considered to be in the following aspects:

- Consider the entropy of a sum of (possibly dependent and non-identically distributed) Bernoulli random variables. Upper bounds on the error that follows from an approximation of this entropy by the entropy of a Poisson random variable with the same mean are derived via the Chen-Stein method (see Theorem 5 and its related results in Section II). The use of these new bounds is exemplified for some interesting applications of the Chen-Stein method in [2] and [3].
- Improved lower bounds on the relative entropy between the distribution of a sum of independent Bernoulli random variables and the Poisson distribution with the same mean are derived (see Theorem 7 in Section III). These new bounds are obtained by combining a derivation of some sharpened lower bounds on the total variation distance (see Theorem 6 and some related results in Section III) that improve the original lower bound in [4, Theorem 2], and a probability-dependent refinement of Pinsker's inequality [38]. The new lower bounds are compared with existing upper bounds.
- New upper and lower bounds on the Chernoff information and Bhattacharyya parameter are also derived in Section III via the introduction of new bounds on the Hellinger distance and relative entropy. The use of the new lower bounds on the relative entropy and Chernoff information is exemplified in the context of binary hypothesis testing. The impact of the improvements of these new bounds is studied as well.

To the best of our knowledge, among the publications of the *IEEE Trans. on Information Theory*, the Chen-Stein method for Poisson approximation was used so far only in two occasions. In [49], this probabilistic method was used by A. J. Wyner to analyze the redundancy and the distribution of the phrase lengths in one of the versions of the Lempel-Ziv data compression algorithm. In the second occasion, this method was applied in [16] in the context

of random networks. In [16], the authors relied on existing upper bounds on the total variation distance, applying them to analyze the asymptotic distribution of the number of isolated nodes in a random grid network where nodes are always active. The first part of this paper relies (as well) on some existing upper bounds on the total variation distance, with the purpose of obtaining error bounds on the Poisson approximation of the entropy for a sum of (possibly dependent) Bernoulli random variables, or more generally for a sum of non-negative, integer-valued and bounded random variables (this work relies on stronger versions of the upper bounds in [16, Theorems 2.2 and 2.4]).

The paper is structured as follows: Section II forms the first part of this work where the entropy of the sum of Bernoulli random variables is considered. Section III provides the second part of this work where new lower bounds on the total variation distance and relative entropy between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution are derived. The derivation of the new and improved lower bounds on the total variation distance relies on the Chen-Stein method for the Poisson approximation, and it generalizes and tightens the analysis that was used to derive the original lower bound on the total variation distance in [4]. The derivation of the new lower bound on the relative entropy follows from the new lower bounds on the total variation distance, combined with a distribution-dependent refinement of Pinsker's inequality in [38]. The new lower bound on the relative entropy is compared to a previously reported upper bound on the relative entropy from [33]. Upper and lower bounds on the Bhattacharyya parameter, Chernoff information and the Hellinger, local and Kolmogorov-Smirnov distances between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution with the same mean are also derived in Section III via some relations between these quantities with the total variation distance and the relative entropy. The analysis in this work combines elements of information theory with the Chen-Stein method for Poisson approximation. The use of these new bounds is exemplified in the two parts of this work, partially relying on some interesting applications of the Chen-Stein method for the Poisson approximation that were introduced in [2] and [3]. The bounds that are derived in this work are easy to compute, and their applicability is exemplified. Throughout the paper, the logarithms are expressed on the natural base (on base e).

II. Error Bounds on the Entropy of the Sum of Bernoulli Random Variables

This section considers the entropy of a sum of (possibly dependent and non-identically distributed) Bernoulli random variables. Section II-A provides a review of some reported results on the Poisson approximation, whose derivation relies on the Chen-Stein method, that are relevant to the analysis in this section. The original results of this section are introduced from Section II-B which provides an upper bound on the entropy difference between two discrete random variables in terms of their total variation distance. This bound is later in this section in the context of the Poisson approximation. Section II-C introduces some explicit upper bounds on the error that follows from the approximation of the entropy of a sum of Bernoulli random variables by the entropy of a Poisson random variable with the same mean. Some applications of the new bounds are exemplified in Section II-D, and these bounds are proved in Section II-E. Finally, a generalization of these bounds is introduced in Section II-F to address the case of the Poisson approximation for the entropy of a sum of non-negative, integer-valued and bounded random variables.

A. Review of Some Essential Results for the Analysis in Section II

Throughout the paper, we use the term 'distribution' to refer to the discrete probability mass function of an integer-valued random variable. In the following, we review briefly some known results that are used for the analysis later in this section.

Definition 1: Let P and Q be two probability measures defined on a set $\mathcal{X}$. Then, the total variation distance between P and Q is defined by

$$d_{\mathrm{TV}}(P,Q) \triangleq \sup_{\text{Borel } A \subseteq \mathcal{X}} |P(A) - Q(A)| \tag{1}$$

where the supermum is taken w.r.t. all the Borel subsets A of $\mathcal{X}$. If $\mathcal{X}$ is a countable set then (1) is simplified to

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| = \frac{||P-Q||_1}{2} \tag{2}$$

so the total variation distance is equal to one-half of the L_1-distance between the two probability distributions.

The following theorem combines [4, Theorems 1 and 2], and its proof relies on the Chen-Stein method:

Theorem 1: Let $W = \sum_{i=1}^{n} X_i$ be a sum of n independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$ for $i \in \{1, \ldots, n\}$, and $\mathbb{E}(W) = \lambda$. Then, the total variation distance between the probability distribution of W and the Poisson distribution with mean λ satisfies

$$\frac{1}{32}\left(1 \wedge \frac{1}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \leq d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \tag{3}$$

where $a \wedge b \triangleq \min\{a, b\}$ for every $a, b \in \mathbb{R}$.

Remark 1: The ratio between the upper and lower bounds in Theorem 1 is not larger than 32, irrespectively of the values of $\{p_i\}$. This shows that, for independent Bernoulli random variables, these bounds are essentially tight. The upper bound in (3) improves Le Cam's inequality (see [35], [47])) which states that $d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \sum_{i=1}^{n} p_i^2$ so the improvement, for large values of λ, is approximately by the factor $\frac{1}{\lambda}$.

Theorem 1 provides a non-asymptotic result for the Poisson approximation of sums of independent binary random variables via the use of the Chen-Stein method. In general, this method enables to analyze the Poisson approximation for sums of dependent random variables. To this end, the following notation was used in [2] and [3]:

Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0. \tag{4}$$

Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha \tag{5}$$

where it is assumed that $\lambda \in (0, \infty)$. For every $\alpha \in I$, let B_α be a subset of I that is chosen such that $\alpha \in B_\alpha$. This subset is interpreted in [2] as the neighborhood of dependence for α in the sense that X_α is independent or weakly dependent of all of the X_β for $\beta \notin B_\alpha$. Furthermore, the following coefficients were defined in [2, Section 2]:

$$b_1 \triangleq \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta \tag{6}$$

$$b_2 \triangleq \sum_{\alpha \in I} \sum_{\beta \in B_\alpha \setminus \{\alpha\}} p_{\alpha,\beta}, \quad p_{\alpha,\beta} \triangleq \mathbb{E}(X_\alpha X_\beta) \tag{7}$$

$$b_3 \triangleq \sum_{\alpha \in I} s_\alpha, \qquad s_\alpha \triangleq \mathbb{E}\big|\mathbb{E}(X_\alpha - p_\alpha \,|\, \sigma(\{X_\beta\})_{\beta \in I \setminus B_\alpha})\big| \tag{8}$$

where $\sigma(\cdot)$ in the conditioning of (8) denotes the σ-algebra that is generated by the random variables inside the parenthesis. In the following, we cite [2, Theorem 1] which essentially implies that when b_1, b_2 and b_3 are all small, then the total number W of events is approximately Poisson distributed.

Theorem 2: Let $W = \sum_{\alpha \in I} X_\alpha$ be a sum of (possibly dependent and non-identically distributed) Bernoulli random variables $\{X_\alpha\}_{\alpha \in I}$. Then, with the notation in (4)–(8), the following upper bound on the total variation distance holds:

$$d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq (b_1 + b_2)\left(\frac{1 - e^{-\lambda}}{\lambda}\right) + b_3\Big(1 \wedge \frac{1.4}{\sqrt{\lambda}}\Big). \tag{9}$$

Remark 2: A comparison of the right-hand side of (9) with the bound in [2, Theorem 1] shows a difference in a factor of 2 between the two upper bounds. This follows from a difference in a factor of 2 between the two definitions of the total variation distance in [2, Section 2] and Definition 1 here. It is noted, however, that Definition 1 in this work is consistent, e.g., with [4] and [5].

Remark 3: Theorem 2 forms a generalization of the upper bound in Theorem 1 by choosing $B_\alpha = \{\alpha\}$ for $\alpha \in I \triangleq \{1, \ldots, n\}$ (note that, due to the independence assumption of the Bernoulli random variables in Theorem 1, the neighborhood of dependence of α is α itself). In this setting, under the independence assumption,

$$b_1 = \sum_{i=1}^{n} p_i^2, \quad b_2 = b_3 = 0$$

which therefore gives, from (9), the upper bound on the right-hand side of (3).

Before proceeding to this analysis, the following maximum entropy result of the Poisson distribution is introduced.

Theorem 3: The Poisson distribution $\mathrm{Po}(\lambda)$ has the maximal entropy among all probability distributions with mean λ that can be obtained as sums of independent Bernoulli RVs:

$$\begin{aligned} &H(\mathrm{Po}(\lambda)) = \sup_{S \in B_\infty(\lambda)} H(S) \\ &B_\infty(\lambda) \triangleq \bigcup_{n \in \mathbb{N}} B_n(\lambda) \\ &B_n(\lambda) \triangleq \left\{ S : \, S = \sum_{i=1}^n X_i, \; X_i \sim \mathrm{Bern}(p_i) \text{ independent}, \; \sum_{i=1}^n p_i = \lambda \right\}. \end{aligned} \tag{10}$$

Furthermore, since the supremum of the entropy over the set $B_n(\lambda)$ is monotonic increasing in n, then

$$H(\mathrm{Po}(\lambda)) = \lim_{n \to \infty} \sup_{S \in B_n(\lambda)} H(S).$$

For $n \in \mathbb{N}$, the maximum entropy distribution in the class $B_n(\lambda)$ is the Binomial distribution of the sum of n i.i.d. Bernoulli random variables $\mathrm{Ber}\Big(\frac{\lambda}{n}\Big)$, so

$$H(\mathrm{Po}(\lambda)) = \lim_{n \to \infty} H\Big(\mathrm{Binomial}\Big(n, \frac{\lambda}{n}\Big)\Big).$$

Remark 4: Theorem 3 partially appears in [32, Proposition 2.1] (see [32, Eq. (2.20)]). This theorem follows directly from [18, Theorems 7 and 8].

Remark 5: The maximum entropy result for the Poisson distribution in Theorem 3 was strengthened in [28] by showing that the supermum on the right-hand side of (10) can be extended to the larger set of ultra-log-concave probability mass functions (that includes the binomial distribution). This result for the Poisson distribution was generalized in [29] and [31] to maximum entropy results for discrete compound Poisson distributions.

Calculation of the entropy of a Poisson random variable: In the next sub-section, we consider the approximation of the entropy of a sum of Bernoulli random variables by the entropy of a Poisson random variable with the same mean. To this end, it is required to evaluate the entropy of $Z \sim \mathrm{Po}(\lambda)$. It is straightforward to verify that

$$H(Z) = \lambda \log\Big(\frac{e}{\lambda}\Big) + \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda} \log k!}{k!} \tag{11}$$

so the entropy of the Poisson distribution (in nats) is given in terms of an infinite series that has no closed-form expression. Sequences of simple upper and lower bounds on this entropy, which are asymptotically tight, were derived in [1]. In particular, from [1, Theorem 2],

$$-\frac{31}{24\lambda^2} - \frac{33}{20\lambda^3} - \frac{1}{20\lambda^4} \leq H(Z) - \frac{1}{2}\log(2\pi e\lambda) + \frac{1}{12\lambda} \leq \frac{5}{24\lambda^2} + \frac{1}{60\lambda^3} \tag{12}$$

which gives tight bounds on the entropy of $Z \sim \mathrm{Po}(\lambda)$ for large values of λ. For $\lambda \geq 20$, the entropy of Z is approximated by the average of its upper and lower bounds in (12), asserting that the relative error of this approximation is less than 0.1% (and it decreases like $\frac{1}{\lambda^2}$ while increasing the value of λ). For $\lambda \in (0, 20)$, a truncation of the infinite series on the right-hand side of (11) after its first $\lceil 10\lambda \rceil$ terms gives an accurate approximation.

B. A New Bound on the Entropy Difference of Two Discrete Random Variables

The following theorem provides a new upper bound on the entropy difference between two discrete random variables in terms of their total variation distance. This theorem relies on the bound of Ho and Yeung in [23, Theorem 6] that forms an improvement over the previously reported bound in [11, Theorem 17.3.3] or [12, Lemma 2.7]. The following new bound is later used in this section in the context of the Poisson approximation.

Theorem 4: Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countable infinite set. Let X and Y be two discrete random variables where X takes values from a finite set $\mathcal{X} = \{a_1, \ldots, a_m\}$, for some $m \in \mathbb{N}$, and Y takes values from the entire set $\mathcal{A}$. Assume that

$$d_{\mathrm{TV}}(X, Y) \leq \eta \tag{13}$$

for some $\eta \in [0, 1)$, and let

$$M \triangleq \max\left\{m + 1, \frac{1}{1-\eta}\right\}. \tag{14}$$

Furthermore, let $\mu > 0$ be set such that

$$-\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) \leq \mu \tag{15}$$

then

$$|H(X) - H(Y)| \leq \eta \log(M-1) + h(\eta) + \mu \tag{16}$$

where h denote the binary entropy function.

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

$$P_{\widetilde{Y}}(a_i) = \begin{cases} P_Y(a_i) & \text{if } i \in \{1, \ldots, M-1\} \\ \sum_{j=M}^{\infty} P_Y(a_j) & \text{if } i = M. \end{cases} \tag{17}$$

Since $P_X(a_i) = 0$ for every $i > m$ and $M \geq m + 1$, then it follows from (17) that

$$\begin{aligned} & d_{\mathrm{TV}}(X, \widetilde{Y}) \\ &= \frac{1}{2}\sum_{i=1}^{m} |P_X(a_i) - P_{\widetilde{Y}}(a_i)| + \frac{1}{2}\sum_{i=m+1}^{M-1} P_{\widetilde{Y}}(a_i) + \frac{1}{2} P_{\widetilde{Y}}(a_M) \\ &= \frac{1}{2}\sum_{i=1}^{m} |P_X(a_i) - P_Y(a_i)| + \frac{1}{2}\sum_{i=m+1}^{\infty} P_Y(a_i) \\ &= d_{\mathrm{TV}}(X, Y). \end{aligned} \tag{18}$$

Hence, X and $\widetilde{Y}$ are two discrete random variables that take values from the set $\{a_1, \ldots, a_M\}$ (note that it includes the set $\mathcal{X}$) and $d_{\mathrm{TV}}(X, \widetilde{Y}) \leq \eta$ (see (13) and (18)). The bound in [23, Theorem 6] therefore implies that if $\eta \leq 1 - \frac{1}{M}$ (which is indeed the case, due to the way M is defined in (14)), then

$$|H(X) - H(\widetilde{Y})| \leq \eta \log(M-1) + h(\eta). \tag{19}$$

Since $\widetilde{Y}$ is a deterministic function of Y then $H(Y) = H(Y, \widetilde{Y}) \geq H(\widetilde{Y})$, and therefore (15) and (17) imply that

$$\begin{aligned} & |H(\widetilde{Y}) - H(Y)| \\ &= H(Y) - H(\widetilde{Y}) \\ &= -\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) + \left(\sum_{i=M}^{\infty} P_Y(a_i)\right) \log\left(\sum_{i=M}^{\infty} P_Y(a_i)\right) \\ &\leq -\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) \\ &\leq \mu. \end{aligned} \tag{20}$$

Finally, the bound in (16) follows from (19), (20) and the triangle inequality. ∎

C. New Error Bounds on the Entropy of Sums of Bernoulli Random Variables

The new bounds on the entropy of sums of Bernoulli random variables are introduced in the following. Their use is exemplified in Section II-D, and their proofs appear in Section II-E.

Theorem 5: Let I be an arbitrary finite index set with $|I| \triangleq n$. Under the assumptions of Theorem 2 and the notation used in Eqs. (4)–(8), let

$$\eta \triangleq (b_1 + b_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b_3\Big(1 \wedge \frac{1.4}{\sqrt{\lambda}}\Big) \tag{21}$$

$$M \triangleq \max\left\{n+2, \frac{1}{1-\eta}\right\} \tag{22}$$

$$\mu \triangleq \left[\Big(\lambda \log\Big(\frac{e}{\lambda}\Big)\Big)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right] \exp\left\{-\left[\lambda + (M-2)\,\log\left(\frac{M-2}{\lambda e}\right)\right]\right\} \tag{23}$$

where, in (23), $(x)_+ \triangleq \max\{x, 0\}$ for every $x \in \mathbb{R}$. Let $Z \sim \mathrm{Po}(\lambda)$ be a Poisson random variable with mean λ. If $\eta < 1$, then the difference between the entropies of Z and W satisfies the following inequality:

$$|H(Z) - H(W)| \leq \eta\,\log(M-1) + h(\eta) + \mu. \tag{24}$$

The following corollary refers to the entropy of a sum of independent Bernoulli random variables:

Corollary 1: Consider the setting in Theorem 5, and assume that the Bernoulli random variables $\{X_\alpha\}_{\alpha\in I}$ are also independent. Then, the following inequality holds:

$$0 \leq H(Z) - H(W) \leq \eta\,\log(M-1) + h(\eta) + \mu \tag{25}$$

where η in (21) is specialized to

$$\eta \triangleq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{\alpha\in I} p_\alpha^2. \tag{26}$$

The following bound forms a possible improvement of the result in Corollary 1.

Proposition 1: Assume that the conditions in Corollary 1 are satisfied. Then, inequality (25) holds with the new parameter

$$\eta \triangleq \theta\,\min\left\{1-e^{-\lambda},\ \frac{3}{4e(1-\sqrt{\theta})^{3/2}}\right\} \tag{27}$$

where

$$\lambda \triangleq \sum_{\alpha\in I} p_\alpha \tag{28}$$

$$\theta \triangleq \frac{1}{\lambda}\sum_{\alpha\in I} p_\alpha^2. \tag{29}$$

Remark 6: From (28) and (29), it follows that $0 \leq \theta \leq \max_{\alpha\in I} p_\alpha \triangleq p_{\max}$. The condition that $\eta < 1$ is mild since it is a meaningful upper bound on the total variation distance (which is bounded by 1).

Remark 7: Proposition 1 improves the bound in Corollary 1 only if θ is below a certain value that depends on λ. The maximal improvement that is obtained by Proposition 1, as compared to Corollary 1, is in the case where $\theta \to 0$ and $\lambda \to \infty$, and the corresponding improvement in the value of η is by a factor of $\frac{3}{4e} \approx 0.276$.

D. Applications of the New Error Bounds on the Entropy

In the following, the use of Theorem 5 is exemplified for the estimation of the entropy of sums of (possibly dependent) Bernoulli random variables. It starts with a simple example where the summands are independent binary random variables, and some interesting examples from [2, Section 3] and [3, Section 4] are considered next. These examples are related to sums of dependent Bernoulli random variables, where the use of Theorem 5 is exemplified for the calculation of error bounds on the entropy via the Chen-Stein method.

Example 1 (sums of independent Bernoulli random variables): Let $W = \sum_{i=1}^{n} X_i$ be a sum of n independent Bernoulli random variables where $X_i \sim \text{Bern}(p_i)$ for $i = 1, \ldots, n$. The calculation of the entropy of W involves the numerical computation of the probabilities

$$\big(P_W(0), P_W(1), \ldots, P_W(n)\big) = (1-p_1, p_1) * (1-p_2, p_2) * \ldots (1-p_n, p_n)$$

whose computational complexity is high for very large values of n, especially if the probabilities $p_1, \ldots, p_n$ are not the same. The bounds in Corollary 1 and Proposition 1 provide rigorous upper bounds on the accuracy of the Poisson approximation for $H(W)$. Lets exemplify this in the case where

$$p_i = 2ai, \quad \forall\, i \in \{1, \ldots, n\},\ a = 10^{-10},\ n = 10^8$$

then

$$\lambda = \sum_{i=1}^{n} p_i = an(n+1) = 1,000,000.01 \approx 10^6$$

and from (29)

$$\theta = \frac{1}{\lambda} \sum_{i=1}^{n} p_i^2 = \frac{2a(2n+1)}{3} = 0.0133.$$

The entropy of the Poisson random variable $Z \sim \text{Po}(\lambda)$ is evaluated via the bounds in (12) (since $\lambda \gg 1$, these bounds are tight), and they imply that $H(Z) = 8.327$ nats. From Corollary 1 (see Eq. (25) where $I = \{1, \ldots, n\}$), it follows that $0 \leq H(Z) - H(W) \leq 0.316$ nats, and Proposition 1 improves it to $0 \leq H(Z) - H(W) \leq 0.110$ nats. Hence, $H(W) \approx 8.272$ nats with a relative error of at most 0.7%.

Example 2 (random graphs): This problem, which appears in [2, Example 1], is described as follows: On the cube $\{0,1\}^n$, assume that each of the $n2^{n-1}$ edges is assigned a random direction by tossing a fair coin. Let $k \in \{0, 1, \ldots, n\}$ be fixed, and denote by $W \triangleq W(k, n)$ the random variable that is equal to the number of vertices at which exactly k edges point outward (so $k = 0$ corresponds to the event where all n edges, from a certain vertex, point inward). Let I be the set of all 2^n vertices, and X_α be the indicator that vertex $\alpha \in I$ has exactly k of its edges directed outward. Then $W = \sum_{\alpha \in I} X_\alpha$ with

$$X_\alpha \sim \text{Bern}(p), \quad p = 2^{-n} \binom{n}{k}, \quad \forall \alpha \in I.$$

This implies that $\lambda = \binom{n}{k}$ (since $|I| = 2^n$). Clearly, the neighborhood of dependence of a vertex $\alpha \in I$, denoted by B_α, is the set of vertices that are directly connected to α (including α itself since Theorem 2 requires that $\alpha \in B_\alpha$). It is noted, however, that B_α in [2, Example 1] was given by $B_\alpha = \{\beta : |\beta - \alpha| = 1\}$ so it excluded the vertex α. From (6), this difference implies that b_1 in their example should be modified to

$$\begin{aligned} b_1 &= |I|\,|B_\alpha| \left(2^{-n} \binom{n}{k}\right)^2 \\ &= 2^{-n}(n+1) \binom{n}{k}^2 \end{aligned} \tag{30}$$

so b_1 is larger than its value in [2, p. 14] by a factor of $1 + \frac{1}{n}$ which has a negligible effect if $n \gg 1$. As is noted in [2, p. 14], if α and β are two vertices that are connected by an edge, then a conditioning on the direction of this edge gives that

$$p_{\alpha,\beta} \triangleq \mathbb{E}(X_\alpha X_\beta) = 2^{2-2n} \binom{n-1}{k} \binom{n-1}{k-1}, \quad \forall\, \alpha \in I,\ \ \beta \in B_\alpha \setminus \{\alpha\}$$

and therefore, from (7),

$$b_2 = n\, 2^{2-n} \binom{n-1}{k} \binom{n-1}{k-1}.$$

Finally, as is noted in [2, Example 1], $b_3 = 0$ (this is because the conditional expectation of X_α given $(X_\beta)_{\beta \in I \setminus B_\alpha}$ is, similarly to the un-conditional expectation, equal to p_α; i.e., the directions of the edges outside the neighborhood of dependence of α are irrelevant to the directions of the edges connecting the vertex α).

In the following, Theorem 5 is applied to get a rigorous error bound on the Poisson approximation of the entropy $H(W)$. Table I presents numerical results for the approximated value of $H(W)$, and the maximal relative error that is associated with this approximation. Note that, by symmetry, the cases with $W(k,n)$ and $W(n-k,n)$ are equivalent, so $H\big(W(k,n)\big) = H\big(W(n-k,n)\big)$.

TABLE I
NUMERICAL RESULTS FOR THE POISSON APPROXIMATIONS OF THE ENTROPY $H(W)$ ($W = W(k,n)$) BY THE ENTROPY $H(Z)$ WHERE $Z \sim \mathrm{Po}(\lambda)$, JOINTLY WITH THE ASSOCIATED ERROR BOUNDS OF THESE APPROXIMATIONS. THESE ERROR BOUNDS ARE CALCULATED FROM THEOREM 5 FOR THE RANDOM GRAPH PROBLEM IN EXAMPLE 2.

n	k (or $n-k$)	$\lambda = \binom{n}{k}$	Approximation of $H(W)$	Maximal relative error
30	27	$4.060 \cdot 10^3$	5.573 nats	0.1%
30	26	$2.741 \cdot 10^4$	6.528 nats	0.5%
30	25	$1.425 \cdot 10^5$	7.353 nats	2.3%
50	48	$1.225 \cdot 10^3$	4.974 nats	$7.6 \cdot 10^{-10}$
50	46	$2.303 \cdot 10^5$	7.593 nats	$9.5 \cdot 10^{-8}$
50	44	$1.589 \cdot 10^7$	9.710 nats	$5.2 \cdot 10^{-6}$
50	42	$5.369 \cdot 10^8$	11.470 nats	$1.5 \cdot 10^{-4}$
50	40	$1.027 \cdot 10^{10}$	12.945 nats	$2.5 \cdot 10^{-3}$
100	95	$7.529 \cdot 10^7$	10.487 nats	$7.9 \cdot 10^{-20}$
100	90	$1.731 \cdot 10^{13}$	16.660 nats	$1.2 \cdot 10^{-14}$
100	85	$2.533 \cdot 10^{17}$	21.456 nats	$1.3 \cdot 10^{-10}$
100	80	$5.360 \cdot 10^{20}$	25.284 nats	$2.4 \cdot 10^{-7}$
100	75	$2.425 \cdot 10^{23}$	28.342 nats	$9.6 \cdot 10^{-5}$
100	70	$2.937 \cdot 10^{25}$	30.740 nats	1.1%

Example 3 (maxima of dependent Gaussian random variables): Consider a finite sequence of possibly dependent Gaussian random variables. The Chen-Stein method was used in [3, Section 4.4] and [26] to derive explicit upper bounds on the total variation distance between the distribution of the number of times (W) where this sequence exceeds a given level and the Poisson distribution with the same mean. The following example relies on the analysis in [3, Section 4.4], and it aims to provide a rigorous estimate of the entropy of the random variable that counts the number of times that the sequence of Gaussian random variables exceeds a given level. This estimation is done as an application of Theorem 5. In order to sharpen the error bound on the entropy, we derive a tightened upper bound on the coefficient b_2 in (7) for the studied example; this bound on b_2 improves the upper bound in [3, Eq. (21)], and it therefore also improves the error bound on the entropy of W. Note that the random variable W can be expressed as a sum of dependent Bernoulli random variables where each of these binary random variables is an indicator function that the corresponding Gaussian random variable in the sequence exceeds the fixed level. The probability that a Gaussian random variable with zero mean and a unit variance exceeds a certain high level is small, and the law of small numbers indicates that the Poisson approximation for W is good if the required level of crossings is high.

By referring to the setting in [3, Section 4.4], let $\{Z_i\}$ be a sequence of independent and standard Gaussian random variables (having a zero mean and a unit variance). Consider a 1-dependent moving average of Gaussian

random variables $\{Y_i\}$ that are defined, for some $\theta \in \mathbb{R}$, by

$$Y_i \triangleq \frac{Z_i + \theta Z_{i+1}}{\sqrt{1+\theta^2}}, \quad \forall\, i \geq 1. \tag{31}$$

This implies that $\mathbb{E}(Y_i) = 0$, $\mathbb{E}(Y_i^2) = 1$, and the lag-1 auto-correlation is equal to

$$\rho \triangleq \mathbb{E}(Y_i\, Y_{i+1}) = \frac{\theta}{1+\theta^2}\,. \tag{32}$$

Let $t > 0$ be a fixed level, $n \in \mathbb{N}$, and W be the number of elements in the sequence $\{Y_1, \ldots, Y_n\}$ that exceed the level t. Then, $W = \sum_{i=1}^n X_i$ is the sum of dependent Bernoulli random variables where $X_i \triangleq 1_{\{Y_i > t\}}$ for $i \in \{1, \ldots, n\}$ (note that W is a sum of independent Bernoulli random variables only if $\theta = 0$). The expected value of W is

$$\mathbb{E}(W) = n\mathbb{P}(Y_1 > t) = n\,\big(1 - \Phi(t)\big) \triangleq \lambda_n(t) \tag{33}$$

where

$$\Phi(t) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}}\, \mathrm{d}t\,, \quad \forall\, t \in \mathbb{R} \tag{34}$$

is the Gaussian cumulative distribution function. Considering the sequence of Bernoulli random variables $\{X_\alpha\}_{\alpha \in I}$ where $I = \{1, \ldots, n\}$ then, it follows from (4) that

$$p_\alpha = \mathbb{P}(Y_\alpha > t) = 1 - \Phi(t), \quad \forall\, \alpha \in I. \tag{35}$$

The neighborhood of dependence of an arbitrary $\alpha \in I$ is

$$B_\alpha \triangleq \{\alpha - 1,\, \alpha,\, \alpha + 1\} \cap I$$

since Y_α only depends in $Y_{\alpha-1}, Y_\alpha, Y_{\alpha+1}$. From (6), (33) and (35), and also because $|B_\alpha| \leq 3$ for every $\alpha \in I$, then the following upper bound on b_1 (see (6)) holds (see [3, Eq. (21)])

$$b_1 \leq |I| \max_{\alpha \in I} \{|B_\alpha|\, p_\alpha^2\} = \frac{3\lambda_n^2(t)}{n} \tag{36}$$

In the following, a tightened upper bound on b_2 (as is defined in (7)) is derived, which improves the bound in [3, Eq. (21)]. Since, by definition $X_\alpha = 1_{\{Y_\alpha > t\}}$, $X_\beta = 1_{\{Y_\beta > t\}}$, and (from (7)) $p_{\alpha,\beta} \triangleq \mathbb{E}(X_\alpha X_\beta)$, then

$$p_{\alpha,\beta} = \mathbb{P}\big(\min\{Y_\alpha, Y_\beta\} > t\big), \quad \forall\, \alpha \in I,\ \beta \in B_\alpha \setminus \{\alpha\}. \tag{37}$$

Note that for every $\alpha \in I$ and $\beta \in B_\alpha \setminus \{\alpha\}$, necessarily $\beta = \alpha \pm 1$ so Y_α and Y_β are jointly standard Gaussian random variables with the correlation ρ in (32) (it therefore follows that $\rho \in \big[-\frac{1}{2}, \frac{1}{2}\big]$, achieving these two extreme values at $\theta = \pm 1$). From [3, Eq. (23) in Lemma 1], it follows that

$$p_{\alpha,\beta} < \sqrt{\frac{2(1+\rho)}{\pi(1-\rho)}} \cdot \big[\varphi(u) - u\,(1 - \Phi(u))\big] \tag{38}$$

where

$$u \triangleq t\sqrt{\frac{2}{1+\rho}}\,, \tag{39}$$

$$\varphi(u) \triangleq \frac{1}{\sqrt{2\pi}}\, \exp\big(-\frac{u^2}{2}\big) = \Phi'(u). \tag{40}$$

Finally, since $|I| = n$ and $|B_\alpha| \leq 3$ for every $\alpha \in I$, then (7) and (38) lead to the following upper bound:

$$\begin{aligned} b_2 &\leq |I| \max_{\alpha \in I,\, \beta \in B_\alpha \setminus \{\alpha\}} \big\{\big(|B_\alpha| - 1\big) p_{\alpha,\beta}\big\} \\ &\leq 2n \sqrt{\frac{2(1+\rho)}{\pi(1-\rho)}} \cdot \big[\varphi(u) - u\,(1 - \Phi(u))\big] \end{aligned} \tag{41}$$

where Φ, ρ, u and φ are introduced, respectively, in Eqs. (32), (34), (39) and (40). This improves the upper bound on b_2 in [3, Eq. (21)] where the reason for this improvement is related to the weakening of an inequality in the

transition from [3, Eq. (23)] to [3, Eq. (24)]. As is noted in [3, Eq. (21)], since Y_α is independent of $(Y_\beta)_{\beta \notin B_\alpha}$, then it follows from (8) that $b_3 = 0$.

Having upper bounds on b_1 and b_2 (see (36) and (41)) and the exact value of b_3, we are ready to use Theorem 5 to get error bounds for the approximation of $H(W)$ by the entropy of a Poisson random variable Z with the same mean (i.e., $Z \sim \mathrm{Po}(\lambda_n(t))$ where $\lambda_n(t)$ is introduced in (33)). Table II presents numerical results for the Poisson approximation of the entropy, and the associated error bounds. It also shows the improvement in the error bound due to the tightening of the upper bound on b_2 in (41) (as compared to its original bound in [3, Eq. (21)]).

TABLE II

NUMERICAL RESULTS FOR THE POISSON APPROXIMATIONS OF THE ENTROPY $H(W)$ IN EXAMPLE 3. IT IS APPROXIMATED BY $H(Z)$ WHERE $Z \sim \mathrm{Po}(\lambda_n(t))$ IN (33), AND THE ASSOCIATED ERROR BOUNDS ARE COMPUTED FROM THEOREM 5. THE INFLUENCE OF THE TIGHTENED BOUND IN (41) IS EXAMINED BY A COMPARISON WITH THE LOOSENED UPPER BOUND ON b_2 IN [3, EQ. (21)].

n	θ (Eq. (31))	t (a fixed level)	$\mathbb{E}(W) = \lambda_n(t)$ (Eq. (33))	Poisson Approximation of $H(W)$	Maximal relative error with tightened and loosened bounds
10^4	$+1$	5	$2.87 \cdot 10^{-3}$	0.020 nats	1.9% (2.3%)
10^6	$+1$	5	0.287	0.672 nats	4.9% (6.0%)
10^8	$+1$	5	28.7	3.094 nats	4.9% (6.0%)
10^{10}	$+1$	5	$2.87 \cdot 10^3$	5.399 nats	3.3% (4.1%)
10^{12}	$+1$	5	$2.87 \cdot 10^5$	7.702 nats	2.7% (3.3%)
10^4	-1	5	$2.87 \cdot 10^{-3}$	0.020 nats	$3.8 \cdot 10^{-6}$
10^6	-1	5	0.287	0.672 nats	$9.6 \cdot 10^{-6}$
10^8	-1	5	28.7	3.094 nats	$9.3 \cdot 10^{-6}$
10^{10}	-1	5	$2.87 \cdot 10^3$	5.399 nats	$6.1 \cdot 10^{-6}$
10^{12}	-1	5	$2.87 \cdot 10^5$	7.702 nats	$4.8 \cdot 10^{-6}$
10^4	$+1$	6	$9.87 \cdot 10^{-6}$	$1.24 \cdot 10^{-4}$ nats	0.2% (0.2%)
10^6	$+1$	6	$9.87 \cdot 10^{-4}$	0.008 nats	0.3% (0.4%)
10^8	$+1$	6	$9.87 \cdot 10^{-2}$	0.327 nats	0.7% (0.8%)
10^{10}	$+1$	6	9.87	2.555 nats	1.0% (1.2%)
10^{12}	$+1$	6	$9.87 \cdot 10^2$	4.866 nats	0.6% (0.7%)

Table II supports the following observations, which are first listed and then explained:

- For fixed values of n and θ, the Poisson approximation is improved by increasing the level t.
- For fixed values of n and t, the error bounds for the Poisson approximation of the entropy improve when the value of θ is modified in a way that decreases the lag-1 auto-correlation ρ in (32).
- For fixed values of n and t, the effect of the tightened upper bound of b_2 (see (41)) on the error bound of the entropy $H(W)$ is more enhanced when ρ is increased (via a change in the value of θ).
- For fixed values of θ and t, the error bounds for the Poisson approximation are weakly dependent on n.

The explanation of these observations is, respectively, as follows:

- For fixed values of n and θ, by increasing the value of the positive level t, the probability that a standard Gaussian random variable Y_i (for $i \in \{1, \ldots, n\}$) exceeds the value t is decreased. The law of small numbers indicates on the enhancement of the accuracy of the Poisson approximation for W in this case.
- For fixed values of n and t, the expected value of W (i.e., $\lambda_n(t)$ in (33)) is kept fixed, and so is the upper bound on b_1 in (36). However, if the correlation ρ in (32) is decreased (by a proper change in the value of θ) then the value of u in (39) is increased, and the upper bound on b_2 (see (41)) is decreased. Since the upper bounds on b_1 and b_3 are not affected by a change in the value of θ and the upper bound on b_2 is decreased, then the upper bound on the total variation distance in Theorem 2 is decreased as well. This also decreases the error bound that refers to the Poisson approximation of the entropy in Theorem 5. Note that Table II compares the situation for $\theta = \pm 1$, which corresponds respectively to $\rho = \pm \frac{1}{2}$ (these are the two extreme values of ρ).

- When n and t are fixed, the balance between the upper bounds on b_1 and b_2 changes significantly while changing the value of θ. To exemplify this numerically, let $n = 10^8$ and $t = 5$ be the length of the sequence of Gaussian random variables and the considered level, respectively. If $\theta = 1$, the upper bounds on b_1 and b_2 in (36) and (41) are, respectively, equal to $2.47 \cdot 10^{-5}$ and 0.176 (the loosened bound on b_2 is equal to 0.218). In this case, b_2 dominates b_1 and therefore an improvement in the value of b_2 (or its upper bound) also improves the error bound for the Poisson approximation of the entropy $H(W)$ in Theorem 5. Consider now the case where $\theta = -1$ (while n, t are kept fixed); this changes the lag-1 autocorrelation ρ in (32) from its maximal value $(+\frac{1}{2})$ to its minimal value $(-\frac{1}{2})$. In this case, the upper bound on b_1 does not change, but the new bound on b_2 is decreased from 0.218 to $6.88 \cdot 10^{-17}$ (and the loosened bound on b_2, for $\theta = -1$, is equal to $9.90 \cdot 10^{-16}$). In the latter case, the situation w.r.t. the balance between the coefficients b_1 and b_2 is reversed, i.e., the bound on b_1 dominates the bound on b_2. Hence, the upper bound on the total variation distance and the error bound that follows from the Poisson approximation of the entropy $H(W)$ are reduced considerably when θ changes from $+1$ to -1. This is because, from Theorem 2, the upper bound on the total variation distance depends linearly on the sum $b_1 + b_2$ when $b_3 = 0$). A similar conclusion also holds w.r.t. the error bound on the entropy (see Theorem 5). In light of this comparison, the tightened bound on b_2 affects the error bound for the Poisson approximation of $H(W)$ when $\theta = 1$, in contrast to the case when $\theta = -1$.
- The numerical results in Table II show that the accuracy of the Poisson approximation is weakly dependent on the length n of the sequence $\{Y_i\}_{i=1}^n$. This is attributed to the fact that the probabilities p_i, for $i \in \{1, \ldots, n\}$, are not affected by n but they are only affected by choice of the level t. Hence, the law of small numbers does not necessarily indicate on an enhanced accuracy of the Poisson approximation for $H(W)$ when the length of the sequence n is increased.

E. Proofs of the New Bounds in Section II-C

1) Proof of Theorem 5: The random variable $W = \sum_{\alpha \in I} X_\alpha$ is a sum of Bernoulli random variables where $|I| = n < \infty$, then W gets values from the set $\{0, 1, \ldots, n\}$, and Z gets non-negative integer values. Theorem 4 therefore implies that

$$\big|H(W) - H(Z)\big| \leq \eta \log(M-1) + h(\eta) + \mu \tag{42}$$

and we need in the following to calculate proper constants η, μ and M for the Poisson approximation. The cardinality of the set of possible values of W is $m = n + 1$, so it follows from (14) that M is given by (22). The parameter η, which serves as an upper bound on the total variation distance $d_{\mathrm{TV}}(W, Z)$, is given in (21) due to the result in Theorem 2. The last thing that is now required is the calculation of μ. Let

$$\Pi_\lambda(k) \triangleq \frac{e^{-\lambda}\, \lambda^k}{k!}, \quad \forall\, k \in \{0, 1, \ldots\}$$

designate the probability distribution of $Z \sim \mathrm{Po}(\lambda)$, so μ is an upper bound on $\sum_{k=M}^\infty \big\{-\Pi_\lambda(k)\, \log \Pi_\lambda(k)\big\}$, which is an infinite sum that only depends on the Poisson distribution. Straightforward calculation gives that

$$\begin{aligned}
&\sum_{k=M}^\infty \big\{-\Pi_\lambda(k)\, \log \Pi_\lambda(k)\big\} \\
&= -\lambda\, \log \lambda \sum_{k=M-1}^\infty \Pi_\lambda(k) + \lambda \sum_{k=M}^\infty \Pi_\lambda(k) + \sum_{k=M}^\infty \Pi_\lambda(k)\, \log(k!)\,.
\end{aligned} \tag{43}$$

From Stirling's formula, for every $k \in \mathbb{N}$, the equality $k! = \sqrt{2\pi k}\, \big(\frac{k}{e}\big)^k\, e^{\eta_k}$ holds for some $\eta_k \in \big(\frac{1}{12k+1}, \frac{1}{12k}\big)$. This therefore implies that the third infinite sum on the right-hand side of (43) satisfies

$$\begin{aligned}
&\sum_{k=M}^\infty \Pi_\lambda(k)\, \log(k!) \\
&\leq \sum_{k=M}^\infty \Pi_\lambda(k)\, \log\left(\sqrt{2\pi k}\, \left(\frac{k}{e}\right)^k\, e^{\frac{1}{12k}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\log(2\pi)}{2} \sum_{k=M}^{\infty} \Pi_\lambda(k) + \sum_{k=M}^{\infty} \Pi_\lambda(k) \left[\Big(k+\frac{1}{2}\Big) \log(k) - k \right] + \frac{1}{12} \sum_{k=M}^{\infty} \frac{\Pi_\lambda(k)}{k} \\
&\leq \frac{\log(2\pi)}{2} \sum_{k=M}^{\infty} \Pi_\lambda(k) + \sum_{k=M}^{\infty} \big\{ k(k-1)\, \Pi_\lambda(k) \big\} + \frac{1}{12} \sum_{k=M}^{\infty} \Pi_\lambda(k) \\
&\overset{\text{(a)}}{=} \frac{\log(2\pi)}{2} \sum_{k=M}^{\infty} \Pi_\lambda(k) + \lambda^2 \sum_{k=M-2}^{\infty} \Pi_\lambda(k) + \frac{1}{12} \sum_{k=M}^{\infty} \Pi_\lambda(k) \\
&\leq \left(\frac{6\log(2\pi)+1}{12} + \lambda^2 \right) \sum_{k=M-2}^{\infty} \Pi_\lambda(k)
\end{aligned} \tag{44}
$$

where the equality in (a) follows from the identity $k(k-1)\,\Pi_\lambda(k) = \lambda^2\,\Pi_\lambda(k-2)$ for every $k \geq 2$. By combining (43) and (44), it follows that

$$
\begin{aligned}
&\sum_{k=M}^{\infty} -\Pi_\lambda(k)\, \log \Pi_\lambda(k) \\
&\leq \Big(\lambda\, \log\Big(\frac{e}{\lambda}\Big) \Big)_+ \sum_{k=M-1}^{\infty} \Pi_\lambda(k) + \left(\frac{6\log(2\pi)+1}{12} + \lambda^2 \right) \sum_{k=M-2}^{\infty} \Pi_\lambda(k) \\
&\leq \left[\Big(\lambda\, \log\Big(\frac{e}{\lambda}\Big) \Big)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12} \right] \sum_{k=M-2}^{\infty} \Pi_\lambda(k).
\end{aligned} \tag{45}
$$

Based on Chernoff's bound, since $Z \sim \mathrm{Po}(\lambda)$,

$$
\begin{aligned}
&\sum_{k=M-2}^{\infty} \Pi_\lambda(k) \\
&= \mathbb{P}(Z \geq M-2) \\
&\leq \inf_{\theta \geq 0} \Big\{ e^{-\theta(M-2)}\, \mathbb{E}\big[e^{\theta Z}\big] \Big\} \\
&= \inf_{\theta \geq 0} \Big\{ e^{-\theta(M-2)}\, e^{\lambda(e^\theta - 1)} \Big\} \\
&= \exp\left\{ -\left[\lambda + (M-2) \log\Big(\frac{M-2}{\lambda e}\Big) \right] \right\}
\end{aligned} \tag{46}
$$

where the last equality follows by substituting the optimized value $\theta = \log\big(\frac{M-2}{\lambda}\big)$ in the exponent (note that $\lambda \leq n = m-1 \leq M-2$, so optimized value of θ is indeed non-negative). Hence, by combining (45) and (46), it follows that

$$
\sum_{k=M}^{\infty} \big\{ -\Pi_\lambda(k)\, \log \Pi_\lambda(k) \big\} \leq \mu \tag{47}
$$

where the parameter μ is introduced in (23). This completes the proof of Theorem 5.

2) Proof of Corollary 1: For proving the right-hand side of (25), which holds under the assumption that the Bernoulli random variables $\{X_\alpha\}_{\alpha \in I}$ are independent, one chooses (similarly to Remark 3) the set $B_\alpha \triangleq \{\alpha\}$ as the neighborhood of dependence for every $\alpha \in I$. Note that this choice of B_α is taken because $\sigma\big(X_\beta\big)_{\beta \in I \setminus \{\alpha\}}\big)$ is independent of X_α. From (6)–(8), this choice gives that $b_1 = \sum_{\alpha \in I} p_\alpha^2$ and $b_2 = b_3 = 0$ which therefore implies the right-hand side of (25) as a special case of Theorem 5. Furthermore, due to the maximum entropy result of the Poisson distribution (see Theorem 3), then $H(Z) - H(W) \geq 0$. This completes the proof of Corollary 1.

3) Proof of Proposition 1: Under the assumption that the Bernoulli random variables $\{X_\alpha\}_{\alpha \in I}$ are independent, we rely here on two possible upper bounds on the total variation distance between the distributions of W and $Z \sim \mathrm{Po}(\lambda)$. The first bound is the one in [4, Theorem 1], used earlier in Corollary 1. This bound gets the form

$$
d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \left(\frac{1-e^{-\lambda}}{\lambda} \right) \sum_{\alpha \in I} p_\alpha^2 = \big(1 - e^{-\lambda}\big)\theta \tag{48}
$$

where θ is introduced in (29). The second bound appears in [8, Eq. (30)], and it improves the bound in [40, Eq. (10)] (see also [41, Eq. (4)]). This bound gets the form

$$d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \frac{3\theta}{4e\big(1-\sqrt{\theta}\big)^{3/2}}\,. \tag{49}$$

It therefore follows that

$$d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \eta \tag{50}$$

where η is defined in (27) to be the minimum of the upper bounds on the total variation distance in (48) and (49). The continuation of the proof of this proposition is similar to the proof of Corollary 1.

F. Generalization: Bounds on the Entropy for a Sum of Non-Negative, Integer-Valued and Bounded Random Variables

We introduce in the following a generalization of the bounds in Section II-C to consider the accuracy of the Poisson approximation for the entropy of a sum of non-negative, integer-valued and bounded random variables. The generalized version of Theorem 5 is first introduced, and it is then justified by relying on the proof of this theorem for sums of Bernoulli random variables with the approach of Serfling in [45, Section 7]. This approach enables to derive an explicit upper bound on the total variation distance between a sum of non-negative and integer-valued random variables and a Poisson distribution with the same mean. The requirement that the summands are bounded random variables is used to obtain an upper bound on the accuracy of the Poisson approximation for the entropy of a sum of non-negative, integer-valued and bounded random variables. The following proposition forms a generalized version of Theorem 5.

Proposition 2: Let I be an arbitrary finite index set, and let $|I| \triangleq n$. Let $\{X_\alpha\}_{\alpha\in I}$ be non-negative, integer-valued random variables, and assume that there exists some $A \in \mathbb{N}$ such that $X_\alpha \in \{0, 1, \ldots, A\}$ a.s. for every $\alpha \in I$. Let

$$W \triangleq \sum_{\alpha\in I} X_\alpha\,, \quad p_\alpha \triangleq \mathbb{P}(X_\alpha = 1)\,, \quad q_\alpha \triangleq \mathbb{P}(X_\alpha \geq 2)\,, \quad \lambda \triangleq \sum_{\alpha\in I} p_\alpha\,, \quad q \triangleq \sum_{\alpha\in I} q_\alpha \tag{51}$$

where $\lambda > 0$ and $q \geq 0$. Furthermore, for every $\alpha \in I$, let X'_α be a Bernoulli random variable that is equal to 1 if $X_\alpha = 1$, and let it be equal otherwise to zero. Referring to these Bernoulli random variables, let

$$b'_1 \triangleq \sum_{\alpha\in I} \sum_{\beta\in B_\alpha} p_\alpha p_\beta \tag{52}$$

$$b'_2 \triangleq \sum_{\alpha\in I} \sum_{\alpha\neq\beta\in B_\alpha} p'_{\alpha,\beta}, \quad p'_{\alpha,\beta} \triangleq \mathbb{E}(X'_\alpha X'_\beta) \tag{53}$$

$$b'_3 \triangleq \sum_{\alpha\in I} s'_\alpha, \qquad s'_\alpha \triangleq \mathbb{E}\big|\mathbb{E}(X'_\alpha - p_\alpha \,|\, \sigma(\{X_\beta\})_{\beta\in I\setminus B_\alpha})\big| \tag{54}$$

where, for every $\alpha \in I$, the subset $B_\alpha \subseteq I$ is determined arbitrarily such that it includes the element α. Furthermore, let

$$\eta_A \triangleq 2(b'_1 + b'_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b'_3\Big(1 \wedge \frac{1.4}{\sqrt{\lambda}}\Big) + q \tag{55}$$

$$M_A \triangleq \max\left\{nA+2, \frac{1}{1-\eta_A}\right\} \tag{56}$$

$$\mu_A \triangleq \left[\Big(\lambda\log\big(\frac{e}{\lambda}\big)\Big)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right]\exp\left\{-\left[\lambda + (M_A-2)\log\left(\frac{M_A-2}{\lambda e}\right)\right]\right\} \tag{57}$$

provided that $\eta_A < 1$. Then, the difference between the entropies (to base e) of W and $Z \sim \mathrm{Po}(\lambda)$ satisfies

$$|H(Z) - H(W)| \leq \eta_A\,\log(M_A - 1) + h(\eta_A) + \mu_A. \tag{58}$$

Proof: Following the approach in [45, Section 7], let $X'_\alpha \triangleq 1_{\{X_\alpha=1\}}$ be a Bernoulli random variable that is equal to the indicator function of the event $X_\alpha = 1$ and $\mathbb{P}(X'_\alpha = 1) = p_\alpha$ for every $\alpha \in I$. Let $W' \triangleq \sum_{\alpha\in I} X'_\alpha$ be the sum of the induced Bernoulli random variables. From the Chen-Stein method (see Theorem 2)

$$d_{\mathrm{TV}}\big(P_{W'}, \mathrm{Po}(\lambda)\big) \leq (b'_1 + b'_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b'_3\Big(1 \wedge \frac{1.4}{\sqrt{\lambda}}\Big) \tag{59}$$

with the constants b'_1, b'_2 and b'_3 as defined in (52)–(54). Furthermore, from [45, Eq. (7.2)], it follows that

$$\begin{aligned} &d_{\mathrm{TV}}\big(P_W,\, P_{W'}\big) \\ &\leq \mathbb{P}(W' \neq W) \\ &\leq \sum_{\alpha\in I} \mathbb{P}\big(X'_\alpha \neq X_\alpha\big) \\ &= \sum_{\alpha\in I} \mathbb{P}\big(X_\alpha \geq 2\big) \\ &= \sum_{\alpha\in I} q_\alpha = q. \end{aligned} \tag{60}$$

It therefore follows from (55), (59) and (60) that

$$d_{\mathrm{TV}}\big(P_W, \mathrm{Po}(\lambda)\big) \leq d_{\mathrm{TV}}\big(P_W,\, P_{W'}\big) + d_{\mathrm{TV}}\big(P_{W'}, \mathrm{Po}(\lambda)\big) \leq \eta_A.$$

The rest of this proof follows closely the proof of Theorem 5 (note that $P_W(k) = 0$ for $k > nA$, so W gets $m \triangleq nA + 1$ possible values). This completes the proof of Proposition 2. ■

III. Improved Lower Bounds on the Total Variation Distance, Relative Entropy and Some Related Quantities for Sums of Independent Bernoulli Random Variables

This section forms the second part of this work. As in the previous section, the presentation starts in Section III-A with a brief review of some reported results that are relevant to the analysis in this section. Improved lower bounds on the total variation distance between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution with the same mean are introduced in Section III-B. These improvements are obtained via the Chen-Stein method, by a non-trivial refinement of the analysis that was used for the derivation of the original lower bound by Barbour and Hall (see [4, Theorem 2]). Furthermore, the improved tightness of the new lower bounds and their connection to the original lower bound are further considered. Section III-C introduces an improved lower bound on the relative entropy between the above two distributions. The analysis that is used for the derivation of the lower bound on the relative entropy is based on the lower bounds on the total variation distance in Section III-B, combined with the use of the distribution-dependent refinement of Pinsker's inequality by Ordentlich and Weinberger [38] (where the latter is specialized to the Poisson distribution). The lower bound on the relative entropy is compared to some previously reported upper bounds on the relative entropy by Kontoyiannis et al. [33] in the context of the Poisson approximation. Upper and lower bounds on the Bhattacharyya parameter, Chernoff information and Hellinger distance between the distribution of the sum of independent Bernoulli random variables and the Poisson distribution are next derived in Section III-D. The discussion proceeds in Section III-E by exemplifying the use of some of the new bounds that are derived in this section in the context of the classical binary hypothesis testing. Finally, Section III-F proves the new results that are introduced in Sections III-C and III-D. It is emphasized that, in contrast to the setting in Section II where the Bernoulli random variables may be dependent summands, the analysis in this section depends on the assumption that the Bernoulli random variables are independent. This difference stems from the derivation of the improved lower bound on the total variation distance in Section III-B, which forms the starting point for the derivation of all the subsequent results that are introduced in this section, assuming an independence of the summands.

A. Review of Some Essential Results for the Analysis in Section III

The following definitions of probability metrics are particularized and simplified to the case of our interest where the probability mass functions are defined on $\mathbb{N}_0$.

Definition 2: Let P and Q be two probability mass functions that are defined on a same countable set $\mathcal{X}$. The Hellinger distance and the Bhattacharyya parameter between P and Q are, respectively, given by

$$d_{\mathrm{H}}(P,Q) \triangleq \left(\frac{1}{2}\sum_{x\in\mathcal{X}}\Big(\sqrt{P(x)}-\sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}} \tag{61}$$

$$\mathrm{BC}(P,Q) \triangleq \sum_{x\in\mathcal{X}}\sqrt{P(x)\,Q(x)} \tag{62}$$

so, these two probability metrics (including the total variation distance in Definition 1) are bounded between 0 and 1.

Remark 8: In general, these probability metrics are defined in the setting where $(\mathcal{X}, d)$ is a separable metric space. The interest in this work is in the specific case where $\mathcal{X}=\mathbb{N}_0$ and $d=|\cdot|$. In this case, the expressions of these probability metrics are simplified as above. For further study of probability metrics and their properties, the interested reader is referred to, e.g., [5, Appendix A.1], [13, Chapter 2] and [39, Section 3.3].

Remark 9: The Hellinger distance is related to the Bhattacharyya parameter via the equality

$$d_{\mathrm{H}}(P,Q)=\sqrt{1-\mathrm{BC}(P,Q)}. \tag{63}$$

Definition 3: The Chernoff information and relative entropy (a.k.a. divergence or Kullback-Leibler distance) between two probability mass functions P and Q that are defined on a countable set $\mathcal{X}$ are, respectively, given by

$$C(P,Q) \triangleq -\min_{\theta\in[0,1]}\log\left(\sum_{x\in\mathcal{X}}P^{\theta}(x)Q^{1-\theta}(x)\right) \tag{64}$$

$$D(P||Q) \triangleq \sum_{x\in\mathcal{X}}P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{65}$$

so $C(P,Q), D(P||Q)\in[0,\infty]$. Throughout this paper, the logarithms are on base e.

Proposition 3: For two probability mass functions P and Q that are defined on the same set $\mathcal{X}$

$$d_{\mathrm{TV}}(P,Q)\le\sqrt{2}\,d_{\mathrm{H}}(P,Q)\le\sqrt{D(P||Q)}. \tag{66}$$

The left-hand side of (66) is proved in [39, p. 99], and the right-hand side is proved in [39, p. 328].

Remark 10: It is noted that the Hellinger distance in the middle of (66) is not multiplied by the square-root of 2 in [39], due to a small difference in the definition of this distance where the factor of one-half on the right-hand side of (61) does not appear in the definition of the Hellinger distance in [39, p. 98]. However, this is just a matter of normalization of this distance (as otherwise, according to [39], the Hellinger distance varies between 0 and $\sqrt{2}$ instead of the interval $[0,1]$). The definition of this distance in (61) is consistent, e.g., with [5]. It makes the range of this distance to be between 0 and 1, similarly to the total variation, local and Kolmogorov-Smirnov distances and also the Bhattacharyya parameter that are considered in this paper.

The Chernoff information, $C(P,Q)$, is the best achievable exponent in the Bayesian probability of error for binary hypothesis testing (see, e.g., [11, Theorem 11.9.1]). Furthermore, if $X_1, X_2, \ldots, X_N$ are i.i.d. random variables, having distribution P with prior probability π_1 and distribution Q with prior probability π_2, the following upper bound holds for the best achievable overall probability of error:

$$P_{\mathrm{e}}^{(N)}\le\exp\big(-N\,C(P,Q)\big). \tag{67}$$

A distribution-dependent refinement of Pinsker's inequality [38]: : Pinsker's inequality provides a lower bound on the relative entropy in terms of the total variation distance between two probability measures that are defined on the same set. It states that

$$D(P||Q) \geq 2\Big(d_{\mathrm{TV}}(P,Q)\Big)^2. \tag{68}$$

In [38], a distribution-dependent refinement of Pinsker's inequality was introduced for an arbitrary pair of probability distributions P and Q that are defined on $\mathbb{N}_0$. It is of the form

$$D(P||Q) \geq \varphi(\pi_Q)\,\Big(d_{\mathrm{TV}}(P,Q)\Big)^2 \tag{69}$$

where

$$\pi_Q \triangleq \sup_{A\subseteq\mathbb{N}_0} \min\big\{Q(A), 1-Q(A)\big\} \tag{70}$$

and

$$\varphi(p) \triangleq \begin{cases} \frac{1}{1-2p}\,\log\left(\frac{1-p}{p}\right) & \text{if } 0<p<\frac{1}{2} \\ 2 & \text{if } p=\frac{1}{2} \end{cases} \tag{71}$$

so φ is monotonic decreasing in the interval $(0,\frac{1}{2}]$,

$$\lim_{p\to 0^+}\varphi(p) = +\infty, \quad \lim_{p\to\frac{1}{2}^-}\varphi(p) = 2$$

where the latter limit implies that φ is left-continuous at one-half. Note that it follows from (70) that $\pi_Q \in [0,\frac{1}{2}]$.

In Section III-C, we rely on this refinement of Pinsker's inequality and combine it with the new lower bound on the total variation distance between the distribution of a sum of independent Bernoulli random variables and the Poisson distribution with the same mean that is introduced in Section III-B. The combination of these two bounds provides a new lower bound on the relative entropy between these two distributions.

B. Improved Lower Bounds on the Total Variation Distance

In Theorem 1, we introduced the upper and lower bounds on the total variation distance in [6, Theorem 1 and 2] (see also [5, Theorem 2.M and Corollary 3.D.1]). This shows that these upper and lower bounds are essentially tight, where the lower bound is about $\frac{1}{32}$ of the upper bound. Furthermore, it was claimed in [5, Remark 3.2.2] (with no explicit proof) that the constant $\frac{1}{32}$ in the lower bound on the left-hand side of (3) can be improved to $\frac{1}{14}$. In this section, we obtain further improvements of this lower bound where, e.g., the ratio of the upper and new lower bounds on the total variation distance tends to 1.69 in the limit where $\lambda \to 0$, and this ratio tends to 10.54 in the limit where $\lambda \to \infty$. As will be demonstrated in the continuation of Section III, the effect of these improvements is enhanced considerably when considering improved lower bounds on the relative entropy and some other related information-theoretic measures. We further study later in this section the implications of the improvement in lower bounding the total variation distance, originating in this sub-section, and exemplify these improvements in the context of information theory and statistics.

Similarly to the proof of [4, Theorem 2], the derivation of the improved lower bound is also based on the Chen-Stein method, but it follows from a significant modification of the analysis that served to derive the original lower bound in [4, Theorem 2]. The following upper bound on the total variation distance is taken (as is) from [4, Theorem 1] (this bound also appears in Theorem 1 here). The motivation for improving the lower bound on the total variation distance is to take advantage of it to improve the lower bound on the relative entropy (via Pinsker's inequality or a refinement of it) and some other related quantities, and then to examine the benefit of this improvement in an information-theoretic context.

Theorem 6: Let $W = \sum_{i=1}^n X_i$ be a sum of n independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$ for $i \in \{1,\ldots,n\}$, and $\mathbb{E}(W) = \lambda$. Then, the total variation distance between the probability distribution of W and the Poisson distribution with mean λ satisfies

$$K_1(\lambda)\sum_{i=1}^n p_i^2 \leq d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_{i=1}^n p_i^2 \tag{72}$$

where K_1 is given by

$$K_1(\lambda) \triangleq \sup_{\substack{\alpha_1,\alpha_2 \in \mathbb{R},\\ \alpha_2 \leq \lambda + \frac{3}{2},\\ \theta > 0}} \left(\frac{1 - h_\lambda(\alpha_1,\alpha_2,\theta)}{2\, g_\lambda(\alpha_1,\alpha_2,\theta)} \right) \tag{73}$$

and

$$h_\lambda(\alpha_1,\alpha_2,\theta) \triangleq \frac{3\lambda + (2-\alpha_2+\lambda)^3 - (1-\alpha_2+\lambda)^3 + |\alpha_1-\alpha_2|\left(2\lambda + |3-2\alpha_2|\right) \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right)}{\theta\lambda} \tag{74}$$

$$x_+ \triangleq \max\{x, 0\}, \quad x_+^2 \triangleq \big(x_+\big)^2, \quad \forall\, x \in \mathbb{R} \tag{75}$$

$$g_\lambda(\alpha_1,\alpha_2,\theta) \triangleq \max\left\{ \left| \left(1 + \sqrt{\frac{2}{\theta\lambda e}} \cdot |\alpha_1 - \alpha_2| \right) \lambda + \max\big\{x(u_i)\big\} \right|, \right.$$
$$\left. \left| \left(2e^{-\frac{3}{2}} + \sqrt{\frac{2}{\theta\lambda e}} \cdot |\alpha_1 - \alpha_2| \right) \lambda - \min\big\{x(u_i)\big\} \right| \right\} \tag{76}$$

$$x(u) \triangleq (c_0 + c_1 u + c_2 u^2)\, \exp(-u^2), \quad \forall\, u \in \mathbb{R} \tag{77}$$

$$\{u_i\} \triangleq \Big\{ u \in \mathbb{R} :\ 2c_2u^3 + 2c_1u^2 - 2(c_2 - c_0)u - c_1 = 0 \Big\} \tag{78}$$

$$c_0 \triangleq (\alpha_2 - \alpha_1)(\lambda - \alpha_2) \tag{79}$$

$$c_1 \triangleq \sqrt{\theta\lambda}\,(\lambda + \alpha_1 - 2\alpha_2) \tag{80}$$

$$c_2 \triangleq -\theta\lambda. \tag{81}$$

Remark 11: The upper and lower bounds on the total variation distance in (72) scale like $\sum_{i=1}^n p_i^2$, similarly to the known bounds in Theorem 1. The ratio of the upper and lower bounds in Theorem 1 tends to 32.00 when either λ tends to zero or infinity. It was obtained numerically that the ratio of the upper and lower bounds in Theorem 6 improves by a factor of 18.96 when $\lambda \to 0$, a factor of 3.04 when $\lambda \to \infty$, and at least by a factor of 2.48 for all $\lambda \in (0, \infty)$. Alternatively, since the upper bound on the total variation distance in Theorems 1 and 6 is common, it follows that the ratio of the upper bound and new lower bound on the total variation distance is reduced to 1.69 when $\lambda \to 0$, it is 10.54 when $\lambda \to \infty$, and it is at most 12.91 for all $\lambda \in (0, \infty)$.

Remark 12: [14, Theorem 1.2] provides an asymptotic result for the total variation distance between the distribution of the sum W of n independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$ and the Poisson distribution with mean $\lambda = \sum_{i=1}^n p_i$. It shows that when $\sum_{i=1}^n p_i \to \infty$ and $\max_{1 \leq i \leq n} p_i \to 0$ as $n \to \infty$ then

$$d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \sim \frac{1}{\sqrt{2\pi e}\,\lambda} \sum_{i=1}^n p_i^2\,. \tag{82}$$

This implies that the ratio of the upper bound on the total variation distance in [4, Theorem 1] (see Theorems 1 here) and this asymptotic expression is equal to $\sqrt{2\pi e} \approx 4.133$. Therefore, in light of the previous remark (see Remark 11), it follows that the ratio between the exact asymptotic value in (82) and the new lower bound in (72) is equal to $\frac{10.54}{\sqrt{2\pi e}} \approx 2.55$. It therefore follows from Remark 11 that in the limit where $\lambda \to 0$, the new lower bound on the total variation in (72) is smaller than the exact value by no more than 1.69, and for $\lambda \gg 1$, it is smaller than the exact asymptotic result by a factor of 2.55.

Remark 13: Since $\{u_i\}$ in (78) are zeros of a cubic polynomial equation with real coefficients, then the size of the set $\{u_i\}$ is either 1 or 3. But since one of the values of u_i is a point where the global maximum of x is attained, and another value of u_i is the point where its global minimum is attained (note that $\lim_{u \to \pm\infty} x(u) = 0$ and x is differentiable, so the global maxima and minima of x are attained at finite values where the derivative of x is equal to zero), then the size of the set $\{u_i\}$ cannot be 1, which implies that it should be equal to 3.

Remark 14: The optimization that is required for the computation of K_1 in (73) w.r.t. the three parameters $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\theta \in \mathbb{R}^+$ is performed numerically. The numerical procedure for the computation of K_1 will be discussed later (after introducing the following corollary).

In the following, we introduce a closed-form lower bound on the total variation distance that is looser than the lower bound in Theorem 6, but which already improves the lower bound in [4, Theorem 2]. This lower bound follows from Theorem 6 by the special choice of $\alpha_1 = \alpha_2 = \lambda$ that is included in the optimization set for K_1 on the right-hand side of (73). Following this sub-optimal choice, the lower bound in the next corollary follows by a derivation of a closed-form expression for the third free parameter $\theta \in \mathbb{R}^+$. In fact, this was our first step towards the derivation of an improved lower bound on the total variation distance. After introducing the following corollary, we discuss it shortly, and suggest an optimization procedure for the computing K_1 on the left-hand side of (72).

Corollary 2: Under the assumptions in Theorem 6, then

$$\widetilde{K}_1(\lambda) \sum_{i=1}^n p_i^2 \leq d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 \tag{83}$$

where

$$\widetilde{K}_1(\lambda) \triangleq \frac{e}{2\lambda} \, \frac{1 - \frac{1}{\theta}\left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}} \tag{84}$$

$$\theta \triangleq 3 + \frac{7}{\lambda} + \frac{1}{\lambda} \cdot \sqrt{(3\lambda+7)\big[(3+2e^{-1/2})\lambda + 7\big]}. \tag{85}$$

Furthermore, the ratio of the upper and lower bounds on the total variation distance in (83) tends to $\frac{56}{e} \approx 20.601$ as $\lambda \to 0$, it tends to 10.539 as $\lambda \to \infty$, and this ratio is monotonic decreasing as a function of $\lambda \in (0, \infty)$ (see the upper plot in Figure 1, and the calculation of the two limits in Section III-F3).

Remark 15: The lower bound on the total variation distance on the left-hand side of (83) improves uniformly the lower bound in [4, Theorem 2] (i.e., the left-hand side of Eq. (3) here). The improvement is by factors of 1.55 and 3.03 for $\lambda \to 0$ and $\lambda \to \infty$, respectively. Note that this improvement is already remarkable since the ratio of the upper and lower bounds in [4, Theorems 1 and 2] (Theorem 1 here) is equal to 32 in these two extreme cases, and it is also uniformly upper bounded by 32 for all values of $\lambda \in (0, \infty)$. Furthermore, in light of Remark 11, the improvement of the lower bound on the total variation distance in Theorem 6 over its loosened version in Corollary 2 is especially significant for small values of λ, but it is marginal for large values of λ; this improvement is by a factor of 11.88 in the limit where $\lambda \to 0$, but asymptotically there is no improvement if $\lambda \to \infty$ where it even holds for $\lambda \geq 20$ (see Figure 1 where all the curves in this plot merge approximately for $\lambda \geq 20$). Note, however, that even if $\lambda \to \infty$, the lower bounds in Theorem 6 and Corollary 2 improve the original lower bound in Theorem 1 by a factor that is slightly above 3.

Remark 16: In light of Corollary 2, a simplified algorithm is suggested in the following for the computation of K_1 in (73). In general, what we compute numerically is a lower bound on K_1; but this is fine since K_1 is the coefficient of the lower bound on the left-hand side of (73), so its replacement by a lower bound still gives a valid lower bound on the total variation distance. The advantage of the suggested algorithm is its reduced complexity, as compared to a brute force search over the infinite three-dimensional region for $(\alpha_1, \alpha_2, \theta)$; the numerical computation that is involved with this algorithm takes less than a second on a standard PC. The algorithm proceeds as follows:

- It chooses the initial values $\alpha_1 = \alpha_2 = \lambda$, and θ as is determined on the right-hand side of (85). The corresponding lower bound on the total variation distance from Theorem 6, for this sub-optimal selection of the three free parameters $\alpha_1, \alpha_2, \theta$, is equal to the closed-form lower bound in Corollary 2.
- At this point, the algorithm performs several iterations where at each iteration, it defines a certain three-dimensional grid around the optimized point from the previous iteration (the zeroth iteration refers to the initial choice of parameters from the previous item, and to the closed-form lower bound in Corollary 2). At each iteration, the algorithm searches for the optimized point on the new grid (i.e., it computes the maximum of the expression inside the supremum on the right-hand side of (73) among all the points of the grid, and it also updates the new location of this point $(\alpha_1, \alpha_2, \theta)$ for the search that is made in the next iteration. Note that, from (73), the grid should exclude points $(\alpha_1, \alpha_2, \theta)$ when either $\theta < 0$ or $\alpha_2 > \lambda + \frac{3}{2}$.

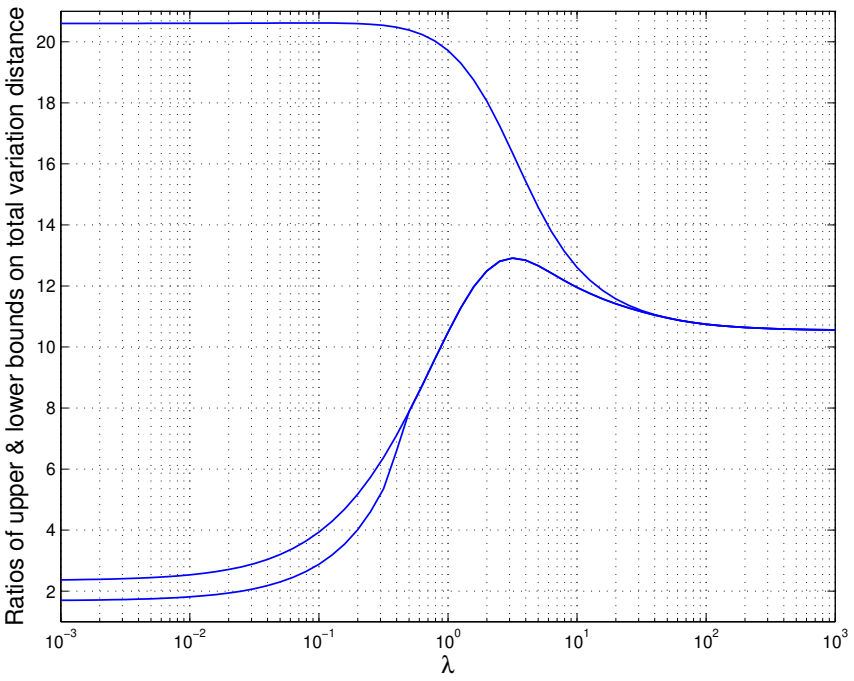

Fig. 1. The figure presents curves that correspond to ratios of upper and lower bounds on the total variation distance between the sum of independent Bernoulli random variables and the Poisson distribution with the same mean λ. The upper bound on the total variation distance for all these three curves is the bound by Barbour and Hall (see [4, Theorem 1] or Theorem 1 here). The lower bounds that the three curves refer to them are the following: the curve at the bottom (i.e., the one which provides the lowest ratio for a fixed λ) is the improved lower bound on the total variation distance that is introduced in Theorem 6. The curve slightly above it for small values of λ corresponds to looser lower bound when α_1 and α_2 in (73) are set to be equal (i.e., $\alpha_1 = \alpha_2 \triangleq \alpha$ is their common value), so that the optimization of K_1 for this curve is reduced to be a two-parameter maximization of K_1 over the two free parameters $\alpha \in \mathbb{R}$ and $\theta \in \mathbb{R}^+$. Finally, the curve at the top of this figure corresponds to the further loosening of this lower bound where α is set to be equal to λ; this leads to a single-parameter maximization of K_1 (over the parameter $\theta \in \mathbb{R}^+$) whose optimization leads to the closed-form expression of the lower bound in Corollary 5. For comparison, in order to assess the enhanced tightness of the new lower bounds, note that the ratio of the upper and lower bounds on the total variation distance from [4, Theorems 1 and 2] (or Theorem 1 here) is roughly equal to 32 for all values of λ.

- At the beginning of this recursive procedure, the algorithm take a very large neighborhood around the point that was selected at the previous iteration (or the initial selection of the point from the first item). The size of this neighborhood at each subsequent iteration shrinks, but the grid also becomes more dense around the new selected point from the previous iteration.

It is noted that numerically, the resulting lower bound on K_1 seems to be the exact value in (73) and not just a lower bound. However, the reduction in the computational complexity of (a lower bound on) K_1 provides a very fast algorithm. The conclusions of the last two remarks (i.e., Remarks 15 and 16 are supported by Figure 1.

C. Improved Lower Bounds on the Relative Entropy

The following theorem relies on the new lower bound on the total variation distance in Theorem 6, and the distribution-dependent refinement of Pinsker's inequality in [38]. Their combination serves to derive a new lower bound on the relative entropy between the distribution of a sum of independent Bernoulli random variables and a Poisson distribution with the same mean. The following upper bound on the relative entropy was introduced in [33, Theorem 1]. Together with the new lower bound on the relative entropy, it leads to the following statement:

Theorem 7: In the setting of Theorem 6, the relative entropy between the probability distribution of W and the Poisson distribution with mean $\lambda = \mathbb{E}(W)$ satisfies the following inequality:

$$K_2(\lambda) \left(\sum_{i=1}^{n} p_i^2 \right)^2 \leq D\big(P_W || \mathrm{Po}(\lambda)\big) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i} \tag{86}$$

where

$$K_2(\lambda) \triangleq m(\lambda) \big(K_1(\lambda)\big)^2 \tag{87}$$

with K_1 from (73), and

$$m(\lambda) \triangleq \begin{cases} \left(\frac{1}{2e^{-\lambda}-1}\right) \log\left(\frac{1}{e^{\lambda}-1}\right) & \text{if } \lambda \in (0, \log 2) \\ 2 & \text{if } \lambda \geq \log 2. \end{cases} \tag{88}$$

Remark 17: For the sake of simplicity, in order to have a bound in closed-form (that is not subject to numerical optimization), the lower bound on the relative entropy on the left-hand side of (86) can be loosened by replacing $K_1(\lambda)$ on the right-hand side of (87) with $\widetilde{K}_1(\lambda)$ in (84) and (85). In light of Remark 15, this possible loosening of the lower bound on the relative entropy has no effect if $\lambda \geq 30$.

Remark 18: The distribution-dependent refinement of Pinsker's inequality from [38] yields that, when applied to a Poisson distribution with mean λ, the coefficient $m(\lambda)$ in (87) is larger than 2 for $\lambda \in (0, \log 2)$, and it is approximately equal to $\log\big(\frac{1}{\lambda}\big)$ for $\lambda \approx 0$. Hence, for $\lambda \approx 0$, the refinement of Pinsker's inequality in [38] leads to a remarkable improvement in the lower bound that appears in (86)–(88), which is by approximately a factor of $\frac{1}{2}\log\big(\frac{1}{\lambda}\big)$. If, however, $\lambda \geq \log 2$ then there is no refinement of Pinsker's inequality (since $m(\lambda) = 2$ in (88)).

Remark 19: The combination of the original lower bound on the total variation distance from [4, Theorem 2] (see (3)) with Pinsker's inequality (see (68)) gives the following lower bound on the relative entropy:

$$D\big(P_W || \operatorname{Po}(\lambda)\big) \geq \frac{1}{512}\left(1 \wedge \frac{1}{\lambda^2}\right)\left(\sum_{i=1}^n p_i^2\right)^2. \tag{89}$$

In light of Remarks 11 and 18, it is possible to quantify the improvement that is obtained by the new lower bound of Theorem 7 in comparison to the looser lower bound in (89). The improvement of the new lower bound on the relative entropy is by a factor of $179.7\log\big(\frac{1}{\lambda}\big)$ for $\lambda \approx 0$, a factor of 9.22 for $\lambda \to \infty$, and at least by a factor of 6.14 for all $\lambda \in (0, \infty)$. The conclusions in the last two remarks (i.e., Remark 18 and 19) are supported by Figure 2 that refers to the special case of the relative entropy between the binomial and Poisson distributions.

Remark 20: In [20, Example 6], it is shown that if $\mathbb{E}(X) \leq \lambda$ then $D\big(P_X \,||\, \operatorname{Po}(\lambda)\big) \geq \frac{1}{2\lambda}\big(\mathbb{E}(X) - \lambda\big)^2$. Since $\mathbb{E}(S_n) = \lambda$ then this lower bound on the relative entropy is not informative for the relative entropy $D\big(P_{S_n} \,||\, \operatorname{Po}(\lambda)\big)$. Theorem 7 and the loosened bound in (89) are, however, informative in the studied case.

The author was notified in [21] about the existence of another recently derived lower bound on the relative entropy $D\big(P_X \,||\, \operatorname{Po}(\lambda)\big)$ in terms of the variance of a random variable X with values in $\mathbb{N}_0$ (this lower bound appears in a currently un-published work). The two bounds were derived independently, based on different approaches. In the setting where $X = \sum_{i=1}^n X_i$ is a sum of independent Bernoulli random variables $\{X_i\}_{i=1}^n$ with $\mathbb{E}(X_i) = p_i$ and $\lambda = \mathbb{E}(X) = \sum_{i=1}^n p_i$, the two lower bounds on the relative entropy scale like $\big(\sum_{i=1}^n p_i^2\big)^2$ but with a different scaling factor.

D. Bounds on Related Quantities

1) Bounds on the Hellinger Distance and Bhattacharyya Parameter: The following proposition introduces a sharpened version of Proposition 3.

Proposition 4: Let P and Q be two probability mass functions that are defined on a same set $\mathcal{X}$. Then, the following inequality suggests a sharpened version of the inequality in (66)

$$\sqrt{1 - \sqrt{1 - \big(d_{\mathrm{TV}}(P,Q)\big)^2}} \leq d_{\mathrm{H}}(P,Q) \leq \sqrt{1 - \exp\left(-\frac{D(P||Q)}{2}\right)} \tag{90}$$

and

$$\exp\left(-\frac{D(P||Q)}{2}\right) \leq \mathrm{BC}(P,Q) \leq \sqrt{1 - \big(d_{\mathrm{TV}}(P,Q)\big)^2}. \tag{91}$$

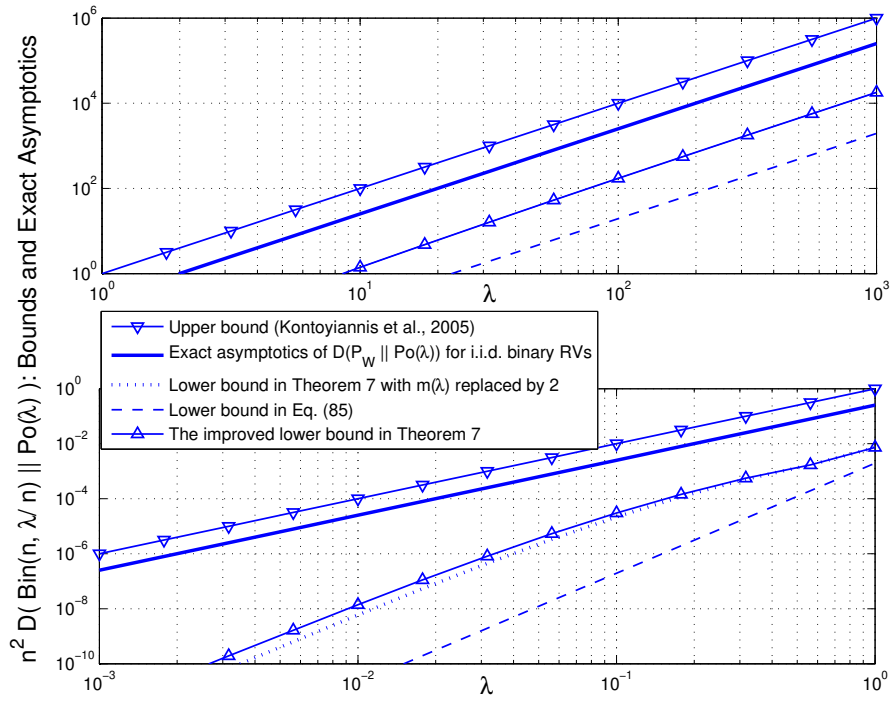

Fig. 2. This figure refers to the relative entropy between the binomial and Poisson distributions with the same mean λ. The horizontal axis refers to λ, and the vertical axis refers to a scaled relative entropy $n^2 D(\mathrm{Bin}(n, \frac{\lambda}{n})||\mathrm{Po}(\lambda))$ ($\sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \frac{\lambda}{n})$ when $X_i \sim \mathrm{Bern}(p_i)$ with $p_i \triangleq \frac{\lambda}{n}$ is fixed for all $i \in \{1, \ldots, n\}$). This scaling of the relative entropy is supported by the upper bound on the relative entropy by Kontoyiannis et al. (see [33, Theorem 1]) that is equal to $\frac{1}{\lambda}\sum_{i=1}^n \frac{p_i^3}{1-p_i} = \frac{\lambda^2}{n^2} + O(\frac{1}{n^3})$. It is also supported by the new lower bounds in Theorems 7 and Eq. (89) since the common term in these lower bounds is equal to $\left(\sum_{i=1}^n p_i^2\right)^2 = \frac{\lambda^4}{n^2}$, so a multiplication of these lower bounds on the relative entropy by n^2 gives an expression that only depends on λ. It follows from [19, Theorem 1] (see also [1, p. 2302]) that $D(\mathrm{Bin}(n, \frac{\lambda}{n})||\mathrm{Po}(\lambda)) = \frac{\lambda^2}{4n^2} + O(\frac{1}{n^3})$ (so, the exact value is asymptotically equal to one-quarter of the upper bound). This figure shows the upper and lower bounds, as well as the exact asymptotic result, in order to study the tightness of the existing upper bound and the new lower bounds. By comparing the dotted and dashed lines, this figure also shows the significant impact of the refinement of the lower bound on the total variation distance by Barbour and Hall (see [4, Theorem 2]) on the improved lower bound on the relative entropy (the former improvement is squared via Pinsker's inequality or its refinement). Furthermore, by comparing the dotted and solid lines of this figure, it shows that the probability-dependent refinement of Pinsker's inequality, applied to the Poisson distribution, affects the lower bound for $\lambda < \log(2)$.

Remark 21: A comparison of the upper and lower bounds on the Hellinger distance in (90) or the Bhattacharyya parameter in (91) gives the following lower bound on the relative entropy in terms of the total variation distance:

$$D(P||Q) \geq \log\left(\frac{1}{1-\big(d_{\mathrm{TV}}(P,Q)\big)^2}\right). \tag{92}$$

It is noted that (92) also follows from the combination of the last two inequalities in [25, p. 741]. It is tighter than Pinsker's inequality (see (68) if $d_{\mathrm{TV}}(P,Q) \geq 0.893$, having also the advantage of giving the right bound for the relative entropy (∞) when the total variation distance approaches to 1. However, (92) is a slightly looser bound on the relative entropy in comparison to Vajda's lower bound [48] that reads:

$$D(P||Q) \geq \log\left(\frac{1+d_{\mathrm{TV}}(P,Q)}{1-d_{\mathrm{TV}}(P,Q)}\right) - \frac{2d_{\mathrm{TV}}(P,Q)}{1+d_{\mathrm{TV}}(P,Q)}. \tag{93}$$

Corollary 3: Under the assumptions in Theorem 6, the Hellinger distance and Bhattacharyya parameter satisfy the following upper and lower bounds:

$$\sqrt{1-\sqrt{1-\big(K_1(\lambda)\big)^2\left(\sum_{i=1}^n p_i^2\right)^2}} \leq d_{\mathrm{H}}(P_W, \mathrm{Po}(\lambda)) \leq \sqrt{1-\exp\left(-\frac{1}{2\lambda}\sum_{i=1}^n \frac{p_i^3}{1-p_i}\right)} \tag{94}$$

and

$$\exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n}\frac{p_i^3}{1-p_i}\right) \leq \mathrm{BC}(P_W, \mathrm{Po}(\lambda)) \leq \sqrt{1-\big(K_1(\lambda)\big)^2\left(\sum_{i=1}^{n} p_i^2\right)^2} \tag{95}$$

where K_1 on the left-hand side of (94) and the right-hand side of (95) is introduced in (73).

Corollary 4: Let $\{S_n\}_{n=1}^{\infty}$ be a sequence of random variables where $S_n \triangleq \sum_{i=1}^{n} X_i^{(n)}$ is a sum of n independent Bernoulli random variables $\{X_i^{(n)}\}_{i=1}^{n}$ with $\mathbb{P}(X_i^{(n)}=1) = p_i^{(n)}$ (note that, for $n \neq m$, the binary random variables $X_i^{(n)}$ and $X_j^{(m)}$ may be dependent). Assume that $\mathbb{E}(S_n) = \sum_{i=1}^{n} p_i^{(n)} = \lambda$ for some $\lambda \in (0,\infty)$ and every $n \in \mathbb{N}$, and that there exist some fixed constants $c_1, c_2 > 0$ such that

$$\frac{c_1\lambda}{n} \leq p_i^{(n)} \leq \frac{c_2\lambda}{n}, \quad \forall\, i \in \{1,\ldots,n\}$$

(which implies that $c_1 \leq 1$ and $c_2 \geq 1$, and $c_1 = c_2 = 1$ if and only if the binary random variables $\{X_i^{(n)}\}_{i=1}^{n}$ are i.i.d.). Then, the following asymptotic results hold:

$$D\big(P_{S_n}||\mathrm{Po}(\lambda)\big) = O\Big(\frac{1}{n^2}\Big) \tag{96}$$

$$d_{\mathrm{TV}}\big(P_{S_n}, \mathrm{Po}(\lambda)\big) = O\Big(\frac{1}{n}\Big) \tag{97}$$

$$d_{\mathrm{H}}\big(P_{S_n}, \mathrm{Po}(\lambda)\big) = O\Big(\frac{1}{n}\Big) \tag{98}$$

$$\mathrm{BC}\big(P_{S_n}, \mathrm{Po}(\lambda)\big) = 1 - O\Big(\frac{1}{n^2}\Big) \tag{99}$$

so, the relative entropy between the distribution of S_n and the Poisson distribution with mean λ scales like $\frac{1}{n^2}$, the total variation and Hellinger distances scale like $\frac{1}{n}$, and the gap of the Bhattacharyya parameter to 1 scales like $\frac{1}{n^2}$.

2) Bounds on the Chernoff Information:

Proposition 5: Let P and Q be two probability mass functions that are defined on a same set $\mathcal{X}$. Then, the Chernoff information between P and Q is lower bounded in terms of the total variation distance as follows:

$$C(P,Q) \geq -\frac{1}{2}\,\log\Big(1-\big(d_{\mathrm{TV}}(P,Q)\big)^2\Big). \tag{100}$$

Corollary 5: Under the assumptions in Theorem 6, the Chernoff information satisfies the following lower bound:

$$C(P_W, \mathrm{Po}(\lambda)) \geq -\frac{1}{2}\,\log\left(1-\big(K_1(\lambda)\big)^2\left(\sum_{i=1}^{n} p_i^2\right)^2\right) \tag{101}$$

where K_1 is introduced in (73).

Remark 22: Remark 17 also applies to Corollaries 3 and 5.

Remark 23: The combination of Proposition 5 with the lower bound on the total variation distance in [4, Theorem 2] (see Theorem 1 here) gives the following looser lower bound on the Chernoff information:

$$C(P_W, \mathrm{Po}(\lambda)) \geq -\frac{1}{2}\,\log\left(1-\frac{1}{1024}\,\Big(1\wedge\frac{1}{\lambda^2}\Big)\left(\sum_{i=1}^{n} p_i^2\right)^2\right). \tag{102}$$

The impact of the tightened lower bound in (101), as compared to the bound in (102) is exemplified in Section III-E in the context of the Bayesian approach for binary hypothesis testing.

E. Applications of the New Bounds in Section III

In the following, we consider the use of the new bounds in Section III for binary hypothesis testing.

Example 4 (Application of the Chernoff-Stein lemma and lower bounds on the relative entropy): The Chernoff-Stein lemma considers the asymptotic error exponent in binary hypothesis testing when one of the probabilities of error is held fixed, and the other one has to be made as small as possible (see, e.g., [11, Theorem 11.8.3]).

Let $\{Y_j\}_{j=1}^N$ be a sequence of non-negative, integer-valued i.i.d. random variables with $\mathbb{E}(Y_1) = \lambda$ for some $\lambda \in (0, \infty)$. Let $Y_1 \sim Q$ where we consider the following two hypothesis:

- H_1: $Q = P_1$ where Y_j, for $j \in \{1, \ldots, N\}$, is a sum of n binary random variables $\{X_{i,j}\}_{i=1}^n$ with $\mathbb{E}(X_{i,j}) = p_i$ and $\sum_{i=1}^n p_i = \lambda$. It is assumed that the elements of the sequence $\{X_{i,j}\}$ are independent, and $n \in \mathbb{N}$ is fixed.
- H_2: $Q = P_2$ is the Poisson distribution with mean λ (i.e., $Y_1 \sim \mathrm{Po}(\lambda)$).

Note that in this case, if one of the Y_j exceeds the value n then H_1 is rejected automatically, so one may assume that $n \gg \max\{\lambda, 1\}$. More explicitly, if $Y_j \sim \mathrm{Po}(\lambda)$ for $j \in \{1, \ldots, N\}$, the probability of this event to happen is upper bounded (via the union and Chernoff bounds) by

$$\mathbb{P}(\exists\, j \in \{1, \ldots, N\} : Y_j \geq n+1) \leq N \exp\left\{-\left[\lambda + (n+1)\log\left(\frac{n+1}{\lambda e}\right)\right]\right\} \tag{103}$$

so, if $n \geq 10 \max\{\lambda, 1\}$, this probability is typically very small.

For an arbitrary $N \in \mathbb{N}$, let A_N be an acceptance region for hypothesis 1. Using standard notation, let

$$\alpha_N \triangleq P_1^N(A_N^{\mathrm{c}}), \quad \beta_N \triangleq P_2^N(A_N) \tag{104}$$

be the two types of error probabilities. Following [11, Theorem 11.8.3], for an arbitrary $\varepsilon \in \left(0, \frac{1}{2}\right)$, let

$$\beta_N^{\varepsilon} \triangleq \min_{A_N \subseteq \mathcal{Y}^N : \alpha_N < \varepsilon} \beta_N$$

where $\mathcal{Y} \triangleq \{0, 1, \ldots, n\}$ is the alphabet that is associated with hypothesis H_1. Then, the best asymptotic exponent of β_N^{ε} in the limit where $\varepsilon \to 0$ is

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{1}{N} \log \beta_N^{\varepsilon} = -D(P_1||P_2).$$

From [11, Eqs. (11.206), (11.207) and (11.227)], for the relative entropy typical set that is defined by

$$A_N^{(\varepsilon)}(P_1||P_2) \triangleq \left\{\underline{y} \in \mathcal{Y}^N : \left|\frac{1}{N}\log\left(\frac{P_1^N(\underline{y})}{P_2^N(\underline{y})}\right) - D(P_1||P_2)\right| \leq \varepsilon\right\} \tag{105}$$

then, it follows from the AEP for relative entropy that $\alpha_N < \varepsilon$ for N large enough (see, e.g., [11, Theorem 11.8.1]). Furthermore, for every N (see, e.g, [11, Theorem 11.8.2]),

$$\beta_N < \exp\Big(-N\big(D(P_1||P_2) - \varepsilon\big)\Big). \tag{106}$$

The error probability of the second type β_N is treated here separately from α_N. In this case, a lower bound on the relative entropy $D(P_1||P_2)$ gives an exponential upper bound on β_N. Let $\varepsilon \to 0$ (more explicitly, let ε be chosen to be small enough as compared to a lower bound on $D(P_1||P_2)$). In the following two simple examples, we calculate the improved lower bound in Theorem 7, and compare it to the lower bound in (89). More importantly, we study the impact of Theorem 7 on the reduction of the number of samples N that are required for achieving $\beta_N < \varepsilon$. The following two cases are used to exemplify this issue:

1) Let the probabilities $\{p_i\}_{i=1}^n$ (that correspond to hypothesis 1) be given by

$$p_i = \frac{i\, p_n}{n}, \quad \forall\, i \in \{1, \ldots, n\}.$$

For $\lambda \in (0, \infty)$, in order to satisfy the equality $\sum_{i=1}^n p_i = \lambda$ then $p_n = \frac{2\lambda}{n+1}$, and $\sum_{i=1}^n p_i^2 = \frac{2\lambda^2}{3}\,\frac{2n+1}{n(n+1)}$. From Theorem 7, the improved lower bound on the relative entropy reads

$$D(P_1||P_2) \geq K_2(\lambda)\left(\frac{2\lambda^2}{3}\,\frac{2n+1}{n(n+1)}\right)^2 \tag{107}$$

where K_2 is introduced in (87), and the weaker lower bound in (89) gets the form

$$D(P_1||P_2) \geq \left(\frac{\lambda^4}{1152}\right) \min\left\{1, \frac{1}{\lambda^2}\right\} \left(\frac{2n+1}{n(n+1)}\right)^2. \tag{108}$$

Lets examine the two bounds on the relative entropy for $\lambda = 10$ and $n = 100$ to find accordingly a proper value of N such that $\beta_N < 10^{-10}$, and choose $\varepsilon = 10^{-10}$. Note that the probability of the event that one of the N Poisson random variables $\{Y_j\}_{j=1}^N$, under hypothesis H_2, exceeds the value n is upper bounded in (103) by $1.22N \cdot 10^{-62}$, so it is neglected for all reasonable amounts of samples N. In this setting, the two lower bounds on the relative entropy in (107) and (108), respectively, are equal to $2.47 \cdot 10^{-4}$ and $3.44 \cdot 10^{-5}$ nats. For these two lower bounds, the exponential upper bound in (106) ensures that $\beta_N < 10^{-10}$ for $N \geq 9.32 \cdot 10^4$ and $N \geq 6.70 \cdot 10^5$, respectively. Hence, the improved lower bound on the relative entropy in Theorem 7 implies here a reduction in the required number of samples by a factor of 7.17.

2) In the second case, assume that the probabilities $\{p_i\}_{i=1}^n$ scale exponentially in i (instead of the linear scaling in the previous case). Let $\alpha \in (0, 1)$ and consider the case where

$$p_i = p_1 \alpha^{i-1}, \quad \forall\, i \in \{1, \ldots, n\}.$$

For $\lambda \in (0, \infty)$, in order to hold the equality $\sum_{i=1}^n p_i = \lambda$ then $p_1 = \frac{\lambda(1-\alpha)}{1-\alpha^n}$, and $\sum_{i=1}^n p_i^2 = \frac{\lambda^2(1-\alpha)}{1+\alpha} \frac{1+\alpha^n}{1-\alpha^n}$. Hence, the improved lower bound in Theorem 7 and the other bound in (89) imply respectively that

$$D(P_1||P_2) \geq \lambda^4\, K_2(\lambda) \left(\frac{1-\alpha}{1+\alpha}\, \frac{1+\alpha^n}{1-\alpha^n}\right)^2 \tag{109}$$

and

$$D(P_1||P_2) \geq \left(\frac{\lambda^4}{512}\right) \min\left\{1, \frac{1}{\lambda^2}\right\} \left(\frac{1-\alpha}{1+\alpha}\, \frac{1+\alpha^n}{1-\alpha^n}\right)^2. \tag{110}$$

The choice $\alpha = 0.05$, $\lambda = 0.1$ and $n = 100$, implies that the two lower bounds on the relative entropy in (109) and (110) are respectively equal to $2.48 \cdot 10^{-5}$ and $1.60 \cdot 10^{-7}$. The exponential upper bound in (106) therefore ensures that $\beta_N < 10^{-10}$ for $N \geq 9.26 \cdot 10^5$ and $N \geq 1.44 \cdot 10^8$, respectively. Hence, the improvement in Theorem 7 leads in this case to the conclusion that one can achieve the target error probability of the second type while reducing the number of samples $\{Y_j\}_{j=1}^N$ by a factor of 155.

Example 5 (Application of the lower bounds on the Chernoff information to binary hypothesis testing): We turn to consider binary hypothesis testing with the Bayesian approach (see, e.g., [11, Section 11.9]). In this setting, one wishes to minimize the overall probability of error while we refer to the two hypotheses in Example 4. The best asymptotic exponent in the Bayesian approach is the Chernoff information (see (64)), and the overall error probability satisfies the following exponential upper bound:

$$P_{\mathrm{e}}^{(N)} \leq \exp\bigl(-N\, C(P_1, P_2)\bigr) \tag{111}$$

so, a lower bound on the Chernoff information provides an upper bound on the overall error probability. In the following, the two lower bounds on the Chernoff information in (101) and (102), and the advantage of the former lower bound is studied in the two cases of Example 4 in order to examine the impact of its improved tightness on the reduction of the number of samples N that are required to achieve an overall error probability below $\varepsilon = 10^{-10}$. We refer, respectively, to cases 1 and 2 of Example 4.

1) In case 1 of Example 4, the two lower bounds on the Chernoff information in Corollary 5 and Remark 23 (following the calculation of $\sum_{i=1}^n p_i^2$ for these two cases) are

$$C(P_1, P_2) \geq \begin{cases} -\frac{1}{2} \log\left(1 - \bigl(K_1(\lambda)\bigr)^2 \left(\frac{2\lambda^2}{3}\, \frac{2n+1}{n(n+1)}\right)^2\right) & \text{From Eq. (101) (Corollary 5)} \\ -\frac{1}{2} \log\left(1 - \frac{\lambda^4}{2304}\, \min\left\{1, \frac{1}{\lambda^2}\right\} \left(\frac{2n+1}{n(n+1)}\right)^2\right) & \text{From Eq. (102) (Remark 23).} \end{cases}$$

As in the first case of Example 4, let $\lambda = 10$ and $n = 100$. The lower bounds on the Chernoff information are therefore equal to

$$C(P_1, P_2) \geq \begin{cases} 6.16 \cdot 10^{-5} & \text{From Eq. (101)} \\ 8.59 \cdot 10^{-6} & \text{From Eq. (102).} \end{cases} \tag{112}$$

Hence, in order to achieve the target $P_{\mathrm{e}}^{(N)} \leq 10^{-10}$ for the overall error probability, the lower bounds on the Chernoff information in (112) and the exponential upper bound on the overall error probability in (111) imply that

$$N \geq \begin{cases} 3.74 \cdot 10^5 & \text{From Eqs. (101) and (111)} \\ 2.68 \cdot 10^6 & \text{From Eqs. (102) and (111)} \end{cases} \tag{113}$$

so, the number of required samples is approximately reduced by a factor of 7.

2) For the second case in Example 4, the lower bounds on the Chernoff information in Eqs. (101) and (102) read

$$C(P_1, P_2) \geq \begin{cases} -\frac{1}{2} \log\left(1 - \lambda^4 \left(K_1(\lambda)\right)^2 \left(\frac{1-\alpha}{1+\alpha} \frac{1+\alpha^n}{1-\alpha^n}\right)^2\right) & \text{From Eq. (101)} \\ -\frac{1}{2} \log\left(1 - \frac{\lambda^4}{1024} \min\left\{1, \frac{1}{\lambda^2}\right\} \left(\frac{1-\alpha}{1+\alpha} \frac{1+\alpha^n}{1-\alpha^n}\right)^2\right) & \text{From Eq. (102)} \end{cases}$$

so, the same choice of parameters $\alpha = 0.05$, $\lambda = 0.1$ and $n = 100$ as in Example 4 implies that

$$C(P_1, P_2) \geq \begin{cases} 4.93 \cdot 10^{-6} & \text{From Eq. (101)} \\ 4.00 \cdot 10^{-8} & \text{From Eq. (102).} \end{cases} \tag{114}$$

For obtaining the target $P_{\mathrm{e}}^{(N)} \leq 10^{-10}$ for the overall error probability, the lower bounds on the Chernoff information in (114) and the exponential upper bound on the overall error probability in (111) imply that

$$N \geq \begin{cases} 4.68 \cdot 10^6 & \text{From Eqs. (101) and (111)} \\ 5.76 \cdot 10^8 & \text{From Eqs. (102) and (111)} \end{cases} \tag{115}$$

so, the improved lower bound on the Chernoff information implies in this case a reduction in the required number of samples N by a factor of 123.

F. Proofs of the New Results in Section III

1) Proof of Theorem 6: The proof of Theorem 6 starts similarly to the proof of [4, Theorem 2]. However, it significantly deviates from the original analysis in order to derive an improved lower bound on the total variation distance. In the following, we introduce the proof of Theorem 6.

Let $\{X_i\}_{i=1}^n$ be independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$. Let $W \triangleq \sum_{i=1}^n X_i$, $V_i \triangleq \sum_{j \neq i} X_j$ for every $i \in \{1, \ldots, n\}$, and $Z \sim \mathrm{Po}(\lambda)$ with mean $\lambda \triangleq \sum_{i=1}^n p_i$. From the basic equation of the Chen-Stein method, the equality

$$\mathbb{E}[\lambda f(Z+1) - Z f(Z)] = 0. \tag{116}$$

holds for an arbitrary bounded function $f : \mathbb{N}_0 \to \mathbb{R}$. Furthermore

$$\begin{aligned} &\mathbb{E}\big[\lambda f(W+1) - W f(W)\big] \\ &= \sum_{j=1}^n p_j \, \mathbb{E}\big[f(W+1)\big] - \sum_{j=1}^n \mathbb{E}\big[X_j f(W)\big] \\ &= \sum_{j=1}^n p_j \, \mathbb{E}\big[f(W+1)\big] - \sum_{j=1}^n p_j \, \mathbb{E}\big[f(V_j+1) \,|\, X_j = 1\big] \\ &\overset{\text{(a)}}{=} \sum_{j=1}^n p_j \, \mathbb{E}\big[f(W+1) - f(V_j+1)\big] \\ &= \sum_{j=1}^n p_j^2 \, \mathbb{E}\big[f(W+1) - f(V_j+1) \,|\, X_j = 1\big] \end{aligned}$$

$$= \sum_{j=1}^{n} p_j^2 \, \mathbb{E}\big[f(V_j+2) - f(V_j+1) \,|\, X_j = 1\big]$$
$$\overset{\text{(b)}}{=} \sum_{j=1}^{n} p_j^2 \, \mathbb{E}\big[f(V_j+2) - f(V_j+1)\big] \tag{117}$$

where equalities (a) and (b) hold since X_j and V_j are independent random variables for every $j \in \{1, \ldots, n\}$. By subtracting (116) from (117), it follows that for an arbitrary bounded function $f : \mathbb{N}_0 \to \mathbb{R}$

$$\mathbb{E}\big[\lambda f(W+1) - W f(W)\big] - \mathbb{E}\big[\lambda f(Z+1) - Z f(Z)\big] = \sum_{j=1}^{n} p_j^2 \, \mathbb{E}\big[f(V_j+2) - f(V_j+1)\big]. \tag{118}$$

In the following, an upper bound on the left-hand side of (118) is derived, based on total variation distance between the two distributions of W and Z.

$$\begin{aligned}
&\mathbb{E}\big[\lambda f(W+1) - W f(W)\big] - \mathbb{E}\big[\lambda f(Z+1) - Z f(Z)\big] \\
&= \sum_{k=0}^{\infty} \big(\lambda f(k+1) - k f(k)\big) \big(\mathbb{P}(W=k) - \mathbb{P}(Z=k)\big) \\
&\leq \sum_{k=0}^{\infty} \big|\lambda f(k+1) - k f(k)\big| \, \big|\mathbb{P}(W=k) - \mathbb{P}(Z=k)\big| \qquad (119) \\
&\leq \sup_{k \in \mathbb{N}_0} \big|\lambda f(k+1) - k f(k)\big| \sum_{k=0}^{\infty} \big|\mathbb{P}(W=k) - \mathbb{P}(Z=k)\big| \\
&= 2 d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \sup_{k \in \mathbb{N}_0} \big|\lambda f(k+1) - k f(k)\big| \qquad (120)
\end{aligned}$$

where the last equality follows from (2). Hence, the combination of (118) and (120) gives the following lower bound on the total variation distance:

$$d_{\mathrm{TV}}(P_W, \mathrm{Po}(\lambda)) \geq \frac{\sum_{j=1}^{n} \Big\{ p_j^2 \, \mathbb{E}\big[f(V_j+2) - f(V_j+1)\big] \Big\}}{2 \, \sup_{k \in \mathbb{N}_0} \big|\lambda f(k+1) - k f(k)\big|} \tag{121}$$

which holds, in general, for an arbitrary bounded function $f : \mathbb{N}_0 \to \mathbb{R}$.

At this point, we deviate from the proof of [4, Theorem 2] by generalizing and refining (in a non-trivial way) the original analysis. The general problem with the current lower bound in (121) is that it is not calculable in closed form for a given f, so one needs to choose a proper function f and derive a closed-form expression for a lower bound on the right-hand side of (121). To this end, let

$$f(k) \triangleq (k - \alpha_1) \, \exp\left(-\frac{(k-\alpha_2)^2}{\theta \lambda}\right), \quad \forall \, k \in \mathbb{N}_0 \tag{122}$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\theta \in \mathbb{R}^+$ are fixed constants (note that θ in (122) needs to be positive for f to be a bounded function). In order to derive a lower bound on the total variation distance, we calculate a lower bound on the numerator and an upper bound on the denominator of the right-hand side of (121) for the function f in (122).

Referring to the numerator of the right-hand side of (121) with f in (122), for every $j \in \{1, \ldots, n\}$,

$$\begin{aligned}
&f(V_j + 2) - f(V_j + 1) \\
&= \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} \frac{\mathrm{d}}{\mathrm{d}u} \left((u + \alpha_2 - \alpha_1) \exp\Big(-\frac{u^2}{\theta\lambda}\Big) \right) \mathrm{d}u \\
&= \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} \left(1 - \frac{2u(u + \alpha_2 - \alpha_1)}{\theta\lambda} \right) \exp\Big(-\frac{u^2}{\theta\lambda}\Big) \mathrm{d}u \\
&= \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} \left(1 - \frac{2u^2}{\theta\lambda} \right) \exp\Big(-\frac{u^2}{\theta\lambda}\Big) \mathrm{d}u - \frac{2(\alpha_2 - \alpha_1)}{\theta\lambda} \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} u \exp\Big(-\frac{u^2}{\theta\lambda}\Big) \mathrm{d}u \\
&= \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} \left(1 - \frac{2u^2}{\theta\lambda} \right) \exp\Big(-\frac{u^2}{\theta\lambda}\Big) \mathrm{d}u \\
&\quad -(\alpha_2 - \alpha_1) \left[\exp\left(-\frac{(V_j + 2 - \alpha_2)^2}{\theta\lambda} \right) - \exp\left(-\frac{(V_j + 1 - \alpha_2)^2}{\theta\lambda} \right) \right]. \qquad (123)
\end{aligned}$$

We rely in the following on the inequality

$$(1 - 2x)\, e^{-x} \geq 1 - 3x, \quad \forall\, x \geq 0.$$

Applying it to the integral on the right-hand side of (123) gives that

$$\begin{aligned}
&f(V_j + 2) - f(V_j + 1) \\
&\geq \int_{V_j+1-\alpha_2}^{V_j+2-\alpha_2} \left(1 - \frac{3u^2}{\theta\lambda} \right) \mathrm{d}u - (\alpha_2 - \alpha_1) \left[\exp\left(-\frac{(V_j + 2 - \alpha_2)^2}{\theta\lambda} \right) - \exp\left(-\frac{(V_j + 1 - \alpha_2)^2}{\theta\lambda} \right) \right] \\
&\geq 1 - \frac{\big(V_j + 2 - \alpha_2\big)^3 - \big(V_j + 1 - \alpha_2\big)^3}{\theta\lambda} \\
&\quad -\big|\alpha_2 - \alpha_1\big| \cdot \left| \exp\left(-\frac{(V_j + 2 - \alpha_2)^2}{\theta\lambda} \right) - \exp\left(-\frac{(V_j + 1 - \alpha_2)^2}{\theta\lambda} \right) \right|. \qquad (124)
\end{aligned}$$

In order to proceed, note that if $x_1, x_2 \geq 0$ then (based on the mean-value theorem of calculus)

$$\begin{aligned}
&|e^{-x_2} - e^{-x_1}| \\
&= \big|e^{-c}\,(x_1 - x_2)\big| \quad \text{for some} \;\; c \in [x_1, x_2] \\
&\leq e^{-\min\{x_1, x_2\}}\, |x_1 - x_2|
\end{aligned}$$

which, by applying it to the second term on the right-hand side of (124), gives that for every $j \in \{1, \ldots, n\}$

$$\begin{aligned}
&\left| \exp\left(-\frac{(V_j + 2 - \alpha_2)^2}{\theta\lambda} \right) - \exp\left(-\frac{(V_j + 1 - \alpha_2)^2}{\theta\lambda} \right) \right| \\
&\leq \exp\left(-\frac{\min\Big\{(V_j + 2 - \alpha_2)^2,\, (V_j + 1 - \alpha_2)^2\Big\}}{\theta\lambda} \right) \cdot \left(\frac{(V_j + 2 - \alpha_2)^2 - (V_j + 1 - \alpha_2)^2}{\theta\lambda} \right). \qquad (125)
\end{aligned}$$

Since $V_j = \sum_{i \neq j} X_i \geq 0$ then

$$\begin{aligned}
&\min\Big\{(V_j + 2 - \alpha_2)^2,\, (V_j + 1 - \alpha_2)^2\Big\} \\
&\geq \begin{cases} 0 & \text{if } \alpha_2 \geq 1 \\ (1 - \alpha_2)^2 & \text{if } \alpha_2 < 1 \end{cases} \\
&= \big(1 - \alpha_2\big)_+^2 \qquad (126)
\end{aligned}$$

where

$$x_+ \triangleq \max\{x, 0\}, \quad x_+^2 \triangleq \big(x_+\big)^2, \quad \forall\, x \in \mathbb{R}.$$

Hence, the combination of the two inequalities in (125)–(126) gives that

$$\begin{aligned}
&\left|\exp\left(-\frac{(V_j+2-\alpha_2)^2}{\theta\lambda}\right) - \exp\left(-\frac{(V_j+1-\alpha_2)^2}{\theta\lambda}\right)\right| \\
&\leq \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \left(\frac{\big|(V_j+2-\alpha_2)^2 - (V_j+1-\alpha_2)^2\big|}{\theta\lambda}\right) \\
&= \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \frac{|2V_j+3-2\alpha_2|}{\theta\lambda} \\
&\leq \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \frac{2V_j+|3-2\alpha_2|}{\theta\lambda}
\end{aligned} \tag{127}$$

and therefore, a combination of the inequalities in (124) and (127) gives that

$$\begin{aligned}
&f(V_j+2) - f(V_j+1) \\
&\geq 1 - \frac{\big(V_j+2-\alpha_2\big)^3 - \big(V_j+1-\alpha_2\big)^3}{\theta\lambda} \\
&\quad -\big|\alpha_2-\alpha_1\big| \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \frac{2V_j+|3-2\alpha_2|}{\theta\lambda}\,.
\end{aligned} \tag{128}$$

Let $U_j \triangleq V_j - \lambda$, then

$$\begin{aligned}
&f(V_j+2) - f(V_j+1) \\
&\geq 1 - \frac{\big(U_j+\lambda+2-\alpha_2\big)^3 - \big(U_j+\lambda+1-\alpha_2\big)^3}{\theta\lambda} \\
&\quad -\big|\alpha_2-\alpha_1\big| \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \frac{2U_j+2\lambda+|3-2\alpha_2|}{\theta\lambda} \\
&= 1 - \frac{3U_j^2 + 3\big(3-2\alpha_2+2\lambda\big)U_j + (2-\alpha_2+\lambda)^3 - (1-\alpha_2+\lambda)^3}{\theta\lambda} \\
&\quad -\big|\alpha_2-\alpha_1\big| \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \frac{2U_j+2\lambda+|3-2\alpha_2|}{\theta\lambda}\,.
\end{aligned} \tag{129}$$

In order to derive a lower bound on the numerator of the right-hand side of (121), for the function f in (122), we need to calculate the expected value of the right-hand side of (129). To this end, the first and second moments of U_j are calculated as follows:

$$\begin{aligned}
&\mathbb{E}(U_j) \\
&= \mathbb{E}(V_j) - \lambda \\
&= \sum_{i\neq j} p_i - \sum_{i=1}^n p_i \\
&= -p_j
\end{aligned} \tag{130}$$

and

$$\begin{aligned}
&\mathbb{E}(U_j^2) \\
&= \mathbb{E}\big((V_j-\lambda)^2\big) \\
&= \mathbb{E}\left[\left(\sum_{i\neq j}(X_i-p_i) - p_j\right)^2\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(a)}}{=} \sum_{i\neq j} \mathbb{E}\big[(X_i - p_i)^2\big] + p_j^2 \\
&\overset{\text{(b)}}{=} \sum_{i\neq j} p_i(1-p_i) + p_j^2 \\
&= \sum_{i\neq j} p_i - \sum_{i\neq j} p_i^2 + p_j^2 \\
&= \lambda - p_j - \sum_{i\neq j} p_i^2 + p_j^2. \qquad (131)
\end{aligned}$$

where equalities (a) and (b) hold since, by assumption, the binary random variables $\{X_i\}_{i=1}^n$ are independent and $\mathbb{E}(X_i) = p_i$, $\mathrm{Var}(X_i) = p_i(1-p_i)$. By taking expectations on both sides of (129), one obtains from (130) and (131) that

$$\begin{aligned}
&\mathbb{E}\big[f(V_j+2) - f(V_j+1)\big] \\
&\geq 1 - \frac{3\Big(\lambda - p_j - \sum_{i\neq j} p_i^2 + p_j^2\Big) + 3\big(3 - 2\alpha_2 + 2\lambda\big)\big(-p_j\big) + (2 - \alpha_2 + \lambda)^3 - (1 - \alpha_2 + \lambda)^3}{\theta\lambda} \\
&\quad - \big|\alpha_2 - \alpha_1\big| \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \cdot \left(\frac{-2p_j + 2\lambda + |3 - 2\alpha_2|}{\theta\lambda}\right) \\
&= 1 - \frac{3\lambda + (2 - \alpha_2 + \lambda)^3 - (1 - \alpha_2 + \lambda)^3 - \Big[3p_j(1-p_j) + 3\sum_{i\neq j} p_i^2 + 3\big(3 - 2\alpha_2 + 2\lambda\big)p_j\Big]}{\theta\lambda} \\
&\quad - \left(\frac{\big|\alpha_2 - \alpha_1\big|\,\big(2\lambda - 2p_j + |3 - 2\alpha_2|\big)}{\theta\lambda}\right) \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) \\
&\geq 1 - \frac{3\lambda + (2 - \alpha_2 + \lambda)^3 - (1 - \alpha_2 + \lambda)^3 - \big(9 - 6\alpha_2 + 6\lambda\big)p_j}{\theta\lambda} \\
&\quad - \left(\frac{\big|\alpha_2 - \alpha_1\big|\,\big(2\lambda + |3 - 2\alpha_2|\big)}{\theta\lambda}\right) \cdot \exp\left(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\right) . \qquad (132)
\end{aligned}$$

Therefore, from (132), the following lower bound on the right-hand side of (121) holds

$$\begin{aligned}
&\sum_{j=1}^n \Big\{p_j^2\,\mathbb{E}\big[f(V_j+2) - f(V_j+1)\big]\Big\} \geq \left(\frac{3\big(3 - 2\alpha_2 + 2\lambda\big)}{\theta\lambda}\right) \sum_{j=1}^n p_j^3 \\
&\quad + \left(1 - \frac{3\lambda + (2 - \alpha_2 + \lambda)^3 - (1 - \alpha_2 + \lambda)^3 + |\alpha_1 - \alpha_2|\big(2\lambda + |3 - 2\alpha_2|\big)\exp\Big(-\frac{(1-\alpha_2)_+^2}{\theta\lambda}\Big)}{\theta\lambda}\right) \sum_{j=1}^n p_j^2 . \qquad (133)
\end{aligned}$$

Note that if $\alpha_2 \leq \lambda + \frac{3}{2}$, which is a condition that is involved in the maximization of (73), then the first term on the right-hand side of (133) can be removed, and the resulting lower bound on the numerator of the right-hand side of (121) gets the form

$$\sum_{j=1}^n \Big\{p_j^2\,\mathbb{E}\big[f(V_j+2) - f(V_j+1)\big]\Big\} \geq \Big(1 - h_\lambda(\alpha_1, \alpha_2, \theta)\Big) \sum_{j=1}^n p_j^2 \qquad (134)$$

where the function h_λ is introduced in (74).

We turn now to derive an upper bound on the denominator of the right-hand side of (121). Therefore, we need to derive a closed-form upper bound on $\sup_{k\in\mathbb{N}_0} \big|\lambda\, f(k+1) - k\, f(k)\big|$ with the function f in (122). For every $k \in \mathbb{N}_0$

$$\lambda\, f(k+1) - k\, f(k) = \lambda\,\big[f(k+1) - f(k)\big] + (\lambda - k)\, f(k). \qquad (135)$$

In the following, we derive bounds on each of the two terms on the right-hand side of (135), and we start with the first term. Let

$$t(u) \triangleq (u+\alpha_2-\alpha_1)\,\exp\left(-\frac{u^2}{\theta\lambda}\right), \quad \forall\, u \in \mathbb{R}$$

then $f(k) = t(k-\alpha_2)$ for every $k \in \mathbb{N}_0$, and by the mean value of calculus

$$\begin{aligned}
&f(k+1) - f(k) \\
&= t(k+1-\alpha_2) - t(k-\alpha_2) \\
&= t'(c_k) \quad \text{for some } c_k \in [k-\alpha_2, k+1-\alpha_2] \\
&= \left(1 - \frac{2c_k^2}{\theta\lambda}\right) \exp\left(-\frac{c_k^2}{\theta\lambda}\right) + \left(\frac{2(\alpha_1-\alpha_2)c_k}{\theta\lambda}\right) \exp\left(-\frac{c_k^2}{\theta\lambda}\right). \qquad (136)
\end{aligned}$$

By referring to the first term on the right-hand side of (136), let

$$p(u) \triangleq (1-2u)e^{-u}, \quad \forall\, u \geq 0$$

then the global maximum and minimum of p over the non-negative real line are obtained at $u=0$ and $u=\frac{3}{2}$, respectively, and therefore

$$-2e^{-\frac{3}{2}} \leq p(u) \leq 1, \quad \forall\, u \geq 0.$$

Let $u = \frac{c_k^2}{\theta\lambda}$, then it follows that the first term on the right-hand side of (136) satisfies the inequality

$$-2e^{-\frac{3}{2}} \leq \left(1 - \frac{2c_k^2}{\theta\lambda}\right) \exp\left(-\frac{c_k^2}{\theta\lambda}\right) \leq 1. \qquad (137)$$

Furthermore, by referring to the second term on the right-hand side of (136), let

$$q(u) \triangleq ue^{-u^2}, \quad \forall\, u \in \mathbb{R}$$

then the global maximum and minimum of q over the real line are obtained at $u = +\frac{\sqrt{2}}{2}$ and $u = -\frac{\sqrt{2}}{2}$, respectively, and therefore

$$-\frac{1}{2}\sqrt{\frac{2}{e}} \leq q(u) \leq +\frac{1}{2}\sqrt{\frac{2}{e}}\,, \quad \forall\, u \in \mathbb{R}.$$

Let this time $u = \sqrt{\frac{c_k}{\theta\lambda}}$, then it follows that the second term on the right-hand side of (136) satisfies

$$\left|\left(\frac{2(\alpha_1-\alpha_2)c_k}{\theta\lambda}\right) \cdot \exp\left(-\frac{c_k^2}{\theta\lambda}\right)\right| \leq \sqrt{\frac{2}{\theta\lambda e}} \cdot |\alpha_1-\alpha_2|. \qquad (138)$$

Hence, by combining the equality in (136) with the two inequalities in (137) and (138), it follows that the first term on the right-hand side of (135) satisfies

$$-\left(2\lambda e^{-\frac{3}{2}} + \sqrt{\frac{2\lambda}{\theta e}} \cdot |\alpha_1-\alpha_2|\right) \leq \lambda\big[f(k+1)-f(k)\big] \leq \lambda + \sqrt{\frac{2\lambda}{\theta e}} \cdot |\alpha_1-\alpha_2|\,, \quad \forall\, k \in \mathbb{N}_0. \qquad (139)$$

We continue the analysis by a derivation of bounds on the second term of the right-hand side of (135). For the function f in (122), it is equal to

$$\begin{aligned}
&(\lambda-k)\,f(k) \\
&= (\lambda-k)(k-\alpha_1)\exp\left(-\frac{(k-\alpha_2)^2}{\theta\lambda}\right) \\
&= \big[(\lambda-\alpha_2)+(\alpha_2-k)\big]\,\big[(k-\alpha_2)+(\alpha_2-\alpha_1)\big]\,\exp\left(-\frac{(k-\alpha_2)^2}{\theta\lambda}\right) \\
&= \Big[(\lambda-\alpha_2)(k-\alpha_2)+(\alpha_2-\alpha_1)(\lambda-\alpha_2)-(k-\alpha_2)^2+(\alpha_1-\alpha_2)(k-\alpha_2)\Big]\,\exp\left(-\frac{(k-\alpha_2)^2}{\theta\lambda}\right) \\
&= \big[\sqrt{\theta\lambda}\,(\lambda-\alpha_2)\,v_k - \theta\lambda\,v_k^2 - \sqrt{\theta\lambda}\,(\alpha_2-\alpha_1)\,v_k + (\alpha_2-\alpha_1)(\lambda-\alpha_2)\big]\,e^{-v_k^2}\,, \quad v_k \triangleq \frac{k-\alpha_2}{\sqrt{\theta\lambda}} \;\; \forall\, k \in \mathbb{N}_0 \\
&= (c_0 + c_1 v_k + c_2 v_k^2)\,e^{-v_k^2} \qquad (140)
\end{aligned}$$

where the coefficients c_0, c_1 and c_2 are introduced in Eqs. (79)–(81), respectively. In order to derive bounds on the left-hand side of (140), lets find the global maximum and minimum of the function x in (77):

$$x(u) \triangleq (c_0 + c_1 u + c_2 u^2) e^{-u^2} \quad \forall\, u \in \mathbb{R}.$$

Note that $\lim_{u \to \pm\infty} x(u) = 0$ and x is differentiable over the real line, so the global maximum and minimum of x are attained at finite points and their corresponding values are finite. By setting the derivative of x to zero, the candidates for the global maximum and minimum of x over the real line are the real zeros $\{u_i\}$ of the cubic polynomial equation in (78). Note that by their definition in (78), the values of $\{u_i\}$ are *independent* of the value of $k \in \mathbb{N}_0$, and also the size of the set $\{u_i\}$ is equal to 3 (see Remark 13). Hence, it follows from (140) that

$$\min_{i \in \{1,2,3\}} \{x(u_i)\} \leq (\lambda - k)\, f(k) \leq \max_{i \in \{1,2,3\}} \{x(u_i)\}, \quad \forall\, k \in \mathbb{N}_0 \tag{141}$$

where these bounds on the second term on the right-hand side of (135) are independent of the value of $k \in \mathbb{N}_0$.

In order to get bounds on the left-hand side of (135), note that from the bounds on the first and second terms on the right-hand side of (135) (see (139) and (141), respectively) then for every $k \in \mathbb{N}_0$

$$\begin{aligned} &\min_{i \in \{1,2,3\}} \{x(u_i)\} - \left(2\lambda e^{-\frac{3}{2}} + \sqrt{\frac{2\lambda}{\theta e}} \cdot |\alpha_1 - \alpha_2| \right) \\ &\leq \lambda\, f(k+1) - k\, f(k) \\ &\leq \max_{i \in \{1,2,3\}} \{x(u_i)\} + \lambda + \sqrt{\frac{2\lambda}{\theta e}} \cdot |\alpha_1 - \alpha_2| \end{aligned} \tag{142}$$

which yields that the following inequality is satisfied:

$$\sup_{k \in \mathbb{N}_0} |\lambda\, f(k+1) - k\, f(k)| \leq g_\lambda(\alpha_1, \alpha_2, \theta) \tag{143}$$

where the function g_λ is introduced in (76). Finally, by combining the inequalities in Eqs. (121), (134) and (143), the lower bound on the total variation distance in (72) follows. The existing upper bound on the total variation distance in (72) was derived in [4, Theorem 1] (see Theorem 1 here). This completes the proof of Theorem 6.

2) Proof of Corollary 2: Corollary 2 follows as a special case of Theorem 6 when the proposed function f in (122) is chosen such that two of its three free parameters (i.e., α_1 and α_2) are determined sub-optimally, and its third parameter (θ) is determined optimally in terms of the sub-optimal selection of the two other parameters. More explicitly, let α_1 and α_2 in (122) be set to be equal to λ (i.e., $\alpha_1 = \alpha_2 = \lambda$). From (79)–(81), this setting implies that $c_0 = c_1 = 0$ and $c_2 = -\theta\lambda < 0$ (since $\theta, \lambda > 0$). The cubic polynomial equation in (78), which corresponds to this (possibly sub-optimal) setting of α_1 and α_2, is

$$2c_2 u^3 - 2c_2 u = 0$$

whose zeros are $u = 0, \pm 1$. The function x in (77) therefore gets the form

$$x(u) = c_2 u^2 e^{-u^2} \quad \forall\, u \in \mathbb{R}$$

so $x(0) = 0$ and $x(\pm 1) = \frac{c_2}{e} < 0$. It implies that

$$\min_{i \in \{1,2,3\}} x(u_i) = \frac{c_2}{e}, \quad \max_{i \in \{1,2,3\}} x(u_i) = 0,$$

and therefore h_λ and g_λ in (74) and (76), respectively, are simplified to

$$h_\lambda(\lambda, \lambda, \theta) = \frac{3\lambda + 7}{\theta\lambda}, \tag{144}$$

$$g_\lambda(\lambda, \lambda, \theta) = \lambda\, \max\left\{1, 2e^{-\frac{3}{2}} + \theta e^{-1}\right\}. \tag{145}$$

This sub-optimal setting of α_1 and α_2 in (122) implies that the coefficient K_1 in (73) is replaced with a loosened version

$$K_1'(\lambda) \triangleq \sup_{\theta > 0} \left(\frac{1 - h_\lambda(\lambda, \lambda, \theta)}{2 g_\lambda(\lambda, \lambda, \theta)} \right). \tag{146}$$

Let $\theta \geq e - \frac{2}{\sqrt{e}}$, then (145) is simplified to $g_\lambda(\lambda, \lambda, \theta) = \lambda \left(2e^{-\frac{3}{2}} + \theta e^{-1}\right)$. It therefore follows from (72), (73) and (144)–(146) that

$$d_{\mathrm{TV}}\big(P_W, \mathrm{Po}(\lambda)\big) \geq \widetilde{K}_1(\lambda) \sum_{i=1}^n p_i^2 \tag{147}$$

where

$$\widetilde{K}_1(\lambda) = \sup_{\theta \geq e - \frac{2}{\sqrt{e}}} \left(\frac{1 - \frac{3\lambda+7}{\theta\lambda}}{2\lambda\big(2e^{-\frac{3}{2}} + \theta e^{-1}\big)} \right) \tag{148}$$

and, in general, $K_1'(\lambda) \geq \widetilde{K}_1(\lambda)$ due to the above restricted constraint on θ (see (146) versus (148)). Differentiation of the function inside the supremum w.r.t. θ and by setting its derivative to zero, one gets the following quadratic equation in θ:

$$\lambda\,\theta^2 - 2(3\lambda + 7)\,\theta - 2(3\lambda + 7)e^{-1} = 0$$

whose positive solution is the optimized value of θ in (85). Furthermore, it is clear that this value of θ in (85) is larger than, e.g., 3, so it satisfies the constraint in (148). This completes the proof of Corollary 2.

3) Discussion on the Connections of Theorem 6 and Corollary 2 to [4, Theorem 2]: As was demonstrated in the previous sub-section, Theorem 6 implies the satisfiability of the lower bound on the total variation distance in Corollary 2. In the following, it is proved that Corollary 2 implies the lower bound on the total variation distance in [4, Theorem 2] (see also Theorem 1 here), and the improvement in the tightness of the lower bound in Corollary 2 is explicitly quantified in the two extreme cases where $\lambda \to 0$ and $\lambda \to \infty$. The observation that Corollary 2 provides a tightened lower bound, as compared to [4, Theorem 2], is justified by the fact that the lower bound in (147) with the coefficient $\widetilde{K}_1(\lambda)$ in (148) was loosened in the proof of [4, Theorem 2] by a sub-optimal selection of the parameter θ which leads to a lower bound on $\widetilde{K}_1(\lambda)$ (the sub-optimal selection of θ in the proof of [4, Theorem 2] is $\theta = 21 \max\{1, \frac{1}{\lambda}\}$). On the other hand, the optimized value of θ that is used in (85) provides an exact closed-form expression for $\widetilde{K}_1(\lambda)$ in (148), and it leads to the derivation of the bound in Corollary 2. This therefore justifies the observation that the lower bound on the total variation distance in Corollary 2 implies the original lower bound in [4, Theorem 2].

From [4, Theorems 1 and 2], the ratio between the upper and lower bounds on the total variation distance (these bounds also appear in (3)) is equal to 32 in the two extreme cases where $\lambda \to 0$ or $\lambda \to \infty$. In order to quantify the improvement that is obtained by Corollary 2 (that follows by the optimal selection of the parameter θ), we calculate in the following the ratio of the same upper bound and the new lower bound in this corollary at these two extreme cases. In the limit where one lets λ tend to infinity, this ratio tends to

$$\begin{aligned}
&\lim_{\lambda\to\infty} \frac{\left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2}{\left(\frac{1-\frac{3\lambda+7}{\lambda\theta}}{2\lambda\left(2e^{-3/2}+\theta\, e^{-1}\right)}\right) \sum_{i=1}^n p_i^2} \qquad (\theta = \theta(\lambda) \text{ is given in Eq. (85)}) \\
&= 2 \lim_{\lambda\to\infty} \frac{2e^{-3/2} + \theta\, e^{-1}}{1 - \frac{3\lambda+7}{\lambda\theta}} \\
&= \frac{2}{e} \lim_{\lambda\to\infty} \frac{\theta\big(2e^{-1/2} + \theta\big)}{\theta - \big(3 + \frac{7}{\lambda}\big)} \\
&\overset{\text{(a)}}{=} \frac{2\left(3 + \sqrt{3(3+2e^{-1/2})}\right)\left(3 + 2e^{-1/2} + \sqrt{3(3+2e^{-1/2})}\right)}{e\,\sqrt{3(3+2e^{-1/2})}} \\
&= \frac{2}{e}\left(3 + \sqrt{3(3+2e^{-1/2})}\right)\left(1 + \sqrt{1 + \frac{2}{3}\cdot e^{-1/2}}\right) \\
&= \frac{6}{e}\left(1 + \sqrt{1 + \frac{2}{3}\cdot e^{-1/2}}\right)^2 \approx 10.539
\end{aligned} \tag{149}$$

where equality (a) holds since, from (85), $\lim_{\lambda\to\infty}\theta = 3+\sqrt{3(3+2e^{-1/2})}$. Furthermore, the limit of this ratio when λ tends to zero is equal to

$$\begin{aligned}
&2\lim_{\lambda\to 0}\left(\frac{1-e^{-\lambda}}{\lambda}\right)\lim_{\lambda\to 0}\left(\frac{\lambda\big(2e^{-3/2}+\theta\,e^{-1}\big)}{1-\frac{3\lambda+7}{\lambda\theta}}\right)\\
&=2\lim_{\lambda\to 0}\left(\frac{\lambda\theta\,(2e^{-3/2}+\theta\,e^{-1})}{\theta-\big(3+\frac{7}{\lambda}\big)}\right)\\
&\overset{\text{(a)}}{=}\frac{28}{e}\lim_{\lambda\to 0}\left(\frac{2e^{-1/2}+\theta)}{\theta-\big(3+\frac{7}{\lambda}\big)}\right)\\
&\overset{\text{(b)}}{=}\frac{56}{e}\approx 20.601
\end{aligned}\tag{150}$$

where equalities (a) and (b) hold since, from (85), it follows that $\lim_{\lambda\to 0}(\lambda\theta)=14$. Note that the two limits in (149) and (150) are indeed consistent with the limits of the upper curve in Figure. 1 (see p. 20). This implies that Corollary 2 improves the original lower bound on the total variation distance in [4, Theorem 2] by a factor of $\frac{32}{10.539}\approx 3.037$ in the limit where $\lambda\to\infty$, and it improves it by a factor of $\frac{32}{20.601}\approx 1.553$ in the other extreme case where $\lambda\to 0$ while still having a closed-form expression lower bound in Corollary 2 where the only reason for this improvement that is related to the optimal choice of the free parameter θ, versus its sub-optimal choice in the proof of [4, Theorem 2], shows a sensitivity of the resulting lower bound to the selection of θ. This observation in fact motivated us to further improve the lower bound on the total variation distance in Theorem 6 by introducing the two additional parameters $\alpha_1,\alpha_2\in\mathbb{R}$ of the proposed function f in (122) (which, according to the proof in the previous sub-section, are set to be both equal to λ). The further improvement in the lower bound at the expense of a feasible increase in computational complexity is shown in the plot of Figure. 1 (by comparing the upper and lower curves of this plot which correspond to the ratio of the upper bound in [4, Theorem 1] and new lower bounds in Corollary 2 and Theorem 6, respectively. It is interesting to note that no improvement is obtained however in Theorem 6, as compared to Corollary 2, for $\lambda\geq 20$, as is shown in Figure 1 (since the the upper and lower curves in this plot merge for $\lambda\geq 20$, and their common limit in the extreme case where $\lambda\to\infty$ is given in (149); this therefore implies that the two new lower bounds in Theorem 6 and Corollary 2 coincide for these values of λ; however, for this range of values of λ, the lower bound on the total variation distance in Corollary 2 has the advantage of being expressed in closed form (i.e., there is no need for a numerical optimization of this bound). Due to the above discussion, another important reasoning for our motivation to improve the lower bound on the total variation distance in Theorem 6 and Corollary 2 is that the factors of improvements that are obtained by these lower bounds (as compared to the original bound) are squared, according to Pinsker's inequality, when one wishes to derive lower bounds on the relative entropy, and this improvement becomes significant in many inequalities in information theory and statistics where the relative entropy appears in the error exponent (as is exemplified in Section III-E). Finally, it is noted that the reason for introducing this type of discussion, which partially motivates our paper, in a sub-section that refers to proofs (of the second half of this work) is because this kind of discussion follows directly from the proofs of Theorem 6 and Corollary 2, and therefore it was introduced here.

4) Proof of Theorem 7: In the following we prove Theorem 7 by obtaining a lower bound on the relative entropy between the distribution P_W of a sum of independent Bernoulli random variables $\{X_i\}_{i=1}^n$ with $X_i\sim\text{Bern}(p_i)$ and the Poisson distribution $\text{Po}(\lambda)$ with mean $\lambda\triangleq\sum_{i=1}^n p_i$. A first lower bound on the relative entropy follows from a combination of Pinsker's inequality (see Eq. (68)) with the lower bound on the total variation distance between these distributions (see Theorem 6). The combination of the two gives that

$$D\big(P_W\,||\,\text{Po}(\lambda)\big)\geq 2\big(K_1(\lambda)\big)^2\left(\sum_{i=1}^n p_i^2\right)^2.\tag{151}$$

This lower bound can be tightened via the distribution-dependent refinement of Pinsker's inequality in [38], which is introduced shortly in Section III-A. Following the technique of this refinement, let $Q\triangleq\Pi_\lambda$ be the probability mass function that corresponds to the Poisson distribution $\text{Po}(\lambda)$, i.e.,

$$Q(k)=\frac{e^{-\lambda}\,\lambda^k}{k!}\quad\forall\,k\in\mathbb{N}_0.$$

If $\lambda \leq \log 2$ then $Q(0) = e^{-\lambda} \geq \frac{1}{2}$. Hence, from (70), the maximization of $\min\big\{Q(A),\, 1 - Q(A)\big\}$ over all the subsets $A \subseteq \mathbb{N}_0$ is obtained for $A = \{0\}$ (or, symmetrically, for $A = \mathbb{N}_0 \setminus \{0\} = \mathbb{N}$) which implies that, if $\lambda \leq \log 2$, one gets from Eqs. (69), (70) and (71) that

$$D\big(P_W \,||\, \mathrm{Po}(\lambda)\big) \geq \varphi(\Pi_Q)\big(K_1(\lambda)\big)^2 \left(\sum_{i=1}^{n} p_i^2\right)^2 . \tag{152}$$

where

$$\Pi_Q = \min\big\{e^{-\lambda}, 1 - e^{-\lambda}\big\} = 1 - e^{-\lambda} \tag{153}$$

and, since $\Pi_Q < \frac{1}{2}$ then

$$\begin{aligned} \varphi(\Pi_Q) &= \frac{1}{1 - 2\Pi_Q} \cdot \log\left(\frac{1 - \Pi_Q}{\Pi_Q}\right) \\ &= \left(\frac{1}{2e^{-\lambda} - 1}\right) \log\left(\frac{1}{e^{\lambda} - 1}\right) . \end{aligned} \tag{154}$$

Hence, the combination of (151), (152), and (154) gives the lower bound on the relative entropy in Theorem 7 (see Eqs. (86), (87) and (88) in this theorem). The upper bound on the considered relative entropy is a known result (see [33, Theorem 1]), which is cited here in order to have both upper and lower bounds in the same inequality (see Eq. (86)). This completes the proof of Theorem 7.

5) Proof of Proposition 4: We start by proving the tightened upper and lower bounds on the Hellinger distance in terms of the total variation distance and relative entropy between the two considered distributions. These refined bounds in (90) improve the original bounds in (66). It is noted that the left-hand side of (66) is proved in [39, p. 99], and the right-hand side is proved in [39, p. 328]. The following is the proof of the refined bounds on the Hellinger distance in (90).

Lets start with the proof of the left-hand side of (90). To this end, let P and Q be two probability mass functions that are defined on a same set $\mathcal{X}$. From (2), (61) and the Cauchy-Schwartz inequality

$$\begin{aligned} & d_{\mathrm{TV}}(P, Q) \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} \Big|\sqrt{P(x)} - \sqrt{Q(x)}\Big| \, \Big(\sqrt{P(x)} + \sqrt{Q(x)}\Big) \\ &\leq \frac{1}{2} \left(\sum_{x \in \mathcal{X}} \Big(\sqrt{P(x)} - \sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}} \left(\sum_{x \in \mathcal{X}} \Big(\sqrt{P(x)} + \sqrt{Q(x)}\Big)^2\right)^{\frac{1}{2}} \\ &= d_{\mathrm{H}}(P, Q) \cdot \left(1 + \sum_{x \in \mathcal{X}} \sqrt{P(x)\, Q(x)}\right)^{\frac{1}{2}} \\ &= d_{\mathrm{H}}(P, Q) \left(2 - \big(d_{\mathrm{H}}(P, Q)\big)^2\right)^{\frac{1}{2}} . \end{aligned} \tag{155}$$

Let $c \triangleq \big(d_{\mathrm{TV}}(P, Q)\big)^2$ and $x \triangleq \big(d_{\mathrm{H}}(P, Q)\big)^2$, then it follows by squaring both sides of (155) that $x(2 - x) \geq c$, which therefore implies that

$$1 - \sqrt{1 - c} \leq x \leq 1 + \sqrt{1 - c} . \tag{156}$$

The right-hand side of (156) is satisfied automatically since $0 \leq d_{\mathrm{H}}(P, Q) \leq 1$ implies that $x \leq 1$. The left-hand side of (156) gives the lower bound on the left-hand side of (90). Next, we prove the upper bound on the right-hand

side of (90). By Jensen's inequality

$$\begin{aligned}
&\Big(d_{\mathrm{H}}(P,Q)\Big)^2 \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}\left\{\Big(\sqrt{P(x)}-\sqrt{Q(x)}\Big)^2\right\} \\
&= 1-\sum_{x\in\mathcal{X}}\sqrt{P(x)\,Q(x)} \\
&= 1-\sum_{x\in\mathcal{X}}P(x)\sqrt{\frac{Q(x)}{P(x)}} \\
&= 1-\sum_{x\in\mathcal{X}}P(x)\,e^{\frac{1}{2}\log\left(\frac{Q(x)}{P(x)}\right)} \\
&\leq 1-e^{\frac{1}{2}\sum_{x\in\mathcal{X}}P(x)\log\left(\frac{Q(x)}{P(x)}\right)} \\
&= 1-e^{-\frac{1}{2}D(P||Q)}
\end{aligned} \tag{157}$$

which completes the proof of (90). The other bound on the Bhattacharyya parameter in (91) follows from (90) and the simple relation in (63) between the Bhattacharyya parameter and Hellinger distance. This completes the proof of Proposition 4.

Remark 24: The weaker bounds in (66), proved in [39], follow from their refined version in (90) by using the simple inequalities

$$\sqrt{1-x}\leq 1-\frac{x}{2}\,,\quad \forall\, x\in[0,1]$$

and

$$e^{-x}\geq 1-x,\quad \forall\, x\geq 0.$$

6) Proof of Corollary 3: This corollary is a direct consequence of Theorems 6 and 7, and Proposition 4.

7) Proof of Corollary 4: Under the conditions in Corollary 4, the asymptotic scaling of the total variation distance, relative entropy, Hellinger distance and Bhattacharyya parameter follow from their (upper and lower) bounds in Theorems 6 and 7 and Eqs. (94) and (95), respectively. This completes the proof of Corollary 4.

8) Proof of Proposition 5: Let P and Q be two arbitrary probability mass functions that are defined on a same set $\mathcal{X}$. We derive in the following the lower bound on the Chernoff information in terms of the total variation distance between P and Q, as is stated in (100).

$$\begin{aligned}
C(P,Q) &\overset{\text{(a)}}{\geq} -\log\left(\sum_{x\in\mathcal{X}}\sqrt{P(x)\,Q(x)}\right) \\
&\overset{\text{(b)}}{=} -\log\,\mathbf{BC}(P,Q) \\
&\overset{\text{(c)}}{=} -\log\,\Big(1-\big(d_{\mathrm{H}}(P,Q)\big)^2\Big) \\
&\overset{\text{(d)}}{\geq} -\frac{1}{2}\log\Big(1-\big(d_{\mathrm{TV}}(P,Q)\big)^2\Big)
\end{aligned}$$

where inequality (a) follows by selecting the possibly sub-optimal choice $\theta=\frac{1}{2}$ in (64), equality (b) holds by definition of the Bhattacharyya parameter (see (62)), equality (c) follows from the equality in (63) that relates the Hellinger distance and Bhattacharyya parameter, and inequality (d) follows from the lower bound on the Hellinger distance in terms of the total variation distance (see (90)). This completes the proof of Proposition 5.

9) Proof of Corollary 5: This corollary is a direct consequence of the lower bound on the total variation distance in Theorem 6, and the lower bound on the Chernoff information in terms of the total variation distance in Proposition 5.

ACKNOWLEDGMENT

I thank Ioannis Kontoyiannis for inviting me to present part of this work at the 2012 Information Theory Workshop (ITW 2012) in Lausanne, Switzerland, September 2012. I also thank Louis H. Y. Chen for expressing his interest in this work during the 2012 International Workshop on Applied Probability that took place in June 2012 in Jerusalem, Israel. These two occasions were stimulating for the writing of this paper. I am thankful to Peter Harremoës for personal communications during the 2012 International Symposium on Information Theory (ISIT 2012) at MIT, and to Abraham J. Wyner for notifying me (during ISIT 2012 as well) about his related work to the Chen-Stein method and Poisson approximation in the context of pattern recognition and the Lempel-Ziv algorithm [49].

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] R. Arratia, L. Goldstein and L. Gordon, "Poisson approximation and the Chen-Stein method," *Statistical Science*, vol. 5, no. 4, pp. 403–424, November 1990.

[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 and L. H. Y. Chen, *An Introduction to Stein's Method*, Lecture Notes Series, Institute for Mathematical Sciences, Singapore University Press and World Scientific, 2005.

[7] A. D. Barbour, O. Johnson, I. Kontoyiannis and M. Madiman, "Compound Poisson approximation via information functionals," *Electronic Journal of Probability*, vol. 15, paper no. 42, pp. 1344–1369, August 2010.

[8] V. Čekanavičius and B. Roos, "An expansion in the exponent for compound binomial approximations," *Lithuanian Mathematical Journal*, vol. 46, no. 1, pp. 54–91, 2006.

[9] S. Chatterjee, P. Diaconis and E. Meckes, "Exchangeable pairs and Poisson approximation," *Probability Surveys*, vol. 2, pp. 64–106, 2005.

[10] L. H. Y. Chen, "Poisson approximation for dependent trials," *Annals of Probability*, vol. 3, no. 3, pp. 534–545, June 1975.

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

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

[13] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer Texts in Statistics, 2008.

[14] P. Deheuvels and D. Pfeifer, "A semigroup approach to Poisson approximation," *Annals of Probability*, vol. 14, no. 2, pp. 663–676, April 1986.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, volume 1, third edition, John Wiley & Sons, New York, 1968.

[16] M. Franceschetti and R. Meester, "Critical node lifetimes in random networks via the Chen-Stein method," *IEEE Trans. on Information Theory*, vol. 52, no. 6, pp. 2831–2837, June 2006.

[17] D. Freedman, "The Poisson approximation for dependent events," *Annals of Probability*, vol. 2, no. 2, pp. 256–269, April 1974.

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

[19] P. Harremoës and P. S. Ruzankin, "Rate of convergence to Poisson law in terms of information divergence," *IEEE Trans. on Information Theory*, vol. 50, no. 9, pp. 2145–2149, September 2004.

[20] P. Harremoës and C. Vignat, "Lower bounds on information divergence," February 2011. Online available at http://arxiv.org/pdf/1102.2536.pdf.

[21] P. Harremoës, *personal communications*, July 2012.

[22] P. Harremoës, O. Johnson and I. Kontoyiannis, "Thinning, entropy and the law of thin numbers," *IEEE Trans. on Information Theory*, vol. 56, no. 9, pp. 4228–4244, September 2010.

[23] 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.

[24] J. L. Hodges and L. Le Cam, "The Poisson approximation to the Poisson binomial distribution," *Annals of Mathematical Statistics*, vol. 31, no. 3, pp. 737–740, September 1960.

[25] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions," *Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 700–718, September 1958.

[26] L. Holst and S. Janson, "Poisson approximation using the Stein-Chen method and coupling: Number of exceedances of Gaussian random variables," *Annals of Probability*, vol. 18, no. 2, pp. 713–723, April 1990.
[27] O. Johnson, *Information Theory and the Central Limit Theorem*, Imperial College Press, 2004.
[28] O. Johnson, "Log-concavity and maximum entropy property of the Poisson distribution," *Stochastic Processes and their Applications*, vol. 117, no. 6, pp. 791–802, November 2006.
[29] O. Johnson, I. Kontoyiannis and M. Madiman, "A criterion for the compound Poisson distribution to be maximum entropy," *Proceedings 2009 IEEE International Symposium on Information Theory*, pp. 1899–1903, Seoul, South Korea, July 2009.
[30] O. Johnson and Y. Yu, "Monotonicity, thinning and discrete versions of the entropy power inequality," *IEEE Trans. on Information Theory*, vol. 56, no. 11, pp. 5387–5395, November 2010.
[31] O. Johnson, I. Kontoyiannis and M. Madiman, "Log-concavity, ultra-log concavity, and a maximum entropy property of discrete compound Poisson measures," to appear in *Discrete Applied Mathematics*, 2012. Online available from http://arxiv.org/pdf/0912.0581v2.pdf.
[32] 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.
[33] 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.
[34] I. Kontoyiannis, P. Harremoës, O. Johnson and M. Madiman, "Information-theoretic ideas in Poisson approximation and concentration," slides of a short course (available from the homepage of the first co-author), September 2006.
[35] L. Le Cam, "An approximation theorem for the Poisson binomial distribution," *Pacific Journal of Mathematics*, vol. 10, no. 4, pp. 1181–1197, Spring 1960.
[36] C. Ley and Y. Swan, "On a connection between Stein characterizations and Fisher information," preprint, November 2011. Online available at http://arxiv.org/pdf/1111.2368v1.pdf.
[37] C. Ley and Y. Swan, "Discrete Stein characterizations and discrete information distances," preprint, December 2011. Online available at http://arxiv.org/pdf/1201.0143v1.pdf.
[38] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.
[39] R. D. Reiss, *Approximate Distributions of Order Statistics with Applications to Non-Parametric Statistics*, Springer Series in Statistics, Springer-Verlag, 1989.
[40] B. Roos, "Sharp constants in the Poisson approximation," *Statistics and Probability Letters*, vol. 52, no. 2, pp. 155–168, April 2001.
[41] B. Roos, "Kerstan's method for compound Poisson approximation," *Annals of Probability*, vol. 31, no. 4, pp. 1754–1771, October 2003.
[42] S. M. Ross and E. A. Peköz, *A Second Course in Probability*, Probability Bookstore, 2007.
[43] N. Ross, "Fundamentals of Stein's Method," *Probability Surveys*, vol. 8, pp. 210–293, 2011.
[44] R. J. Serfling, "A general Poisson approximation theorem," *Annals of Probability*, vol. 3, no. 4, pp. 726–731, August 1975.
[45] R. J. Serfling, "Some elementary results on Poisson approximation in a sequence of Bernoulli trials," *Siam Review*, vol. 20, no. 3, pp. 567–579, July 1978.
[46] 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.
[47] J. M. Steele, "Le Cam's inequality and Poisson approximation," *The American Mathematical Monthly*, vol. 101, pp. 48–54, 1994.
[48] I. Vajda, "Note on discrimination information and variation," *IEEE Trans. on Information Theory*, vol. 16, no. 6, pp. 771–773, November 1970.
[49] A. J. Wyner, "The redundancy and distribution of the phrase lengths of the fixed-database Lempel-Ziv algorithm," *IEEE Trans. on Information Theory*, vol. 43, no. 5, pp. 1452–1464, September 1997.
[50] Y. Yu, "On the maximum entropy properties of the binomial distribution," *IEEE Trans. on Information Theory*, vol. 54, no. 7, pp. 3351–3353, July 2008.
[51] Y. Yu, "On the entropy of compound distributions on non-negative integers," *IEEE Trans. on Information Theory*, vol. 55, no. 8, pp. 3645–3650, August 2009.
[52] Y. Yu, "Monotonic convergence in an information-theoretic law of small numbers," *IEEE Trans. on Information Theory*, vol. 55, no. 12, pp. 5412–5422, December 2009.