Entropy and the Law of Small Numbers

Ioannis Kontoyiannis, *Member, IEEE*, Peter Harremoës, *Member, IEEE*, and Oliver Johnson

***Abstract*—Two new information-theoretic methods are introduced for establishing Poisson approximation inequalities. First, using only elementary information-theoretic techniques it is shown that, when $S_n = \sum_{i=1}^n X_i$ is the sum of the (possibly dependent) binary random variables $X_1, X_2, \ldots, X_n$, with $E(X_i) = p_i$ and $E(S_n) = \lambda$, then**

$$D(P_{S_n} \| \mathrm{Po}(\lambda)) \le \sum_{i=1}^n p_i^2 + \left[\sum_{i=1}^n H(X_i) - H(X_1, X_2, \ldots, X_n) \right]$$

where $D(P_{S_n} \| \mathrm{Po}(\lambda))$ is the relative entropy between the distribution of S_n and the Poisson(λ) distribution. The first term in this bound measures the individual smallness of the X_i and the second term measures their dependence. A general method is outlined for obtaining corresponding bounds when approximating the distribution of a sum of general discrete random variables by an infinitely divisible distribution.

Second, in the particular case when the X_i are independent, the following sharper bound is established:

$$D(P_{S_n} \| \mathrm{Po}(\lambda)) \le \frac{1}{\lambda} \sum_{i=1}^n \frac{p_i^3}{1 - p_i}$$

and it is also generalized to the case when the X_i are general integer-valued random variables. Its proof is based on the derivation of a subadditivity property for a new discrete version of the Fisher information, and uses a recent logarithmic Sobolev inequality for the Poisson distribution.

***Index Terms*—Convergence in relative entropy, Fisher information, law of small numbers, logarithmic Sobolev inequality, Poisson approximation, subadditivity, total variation.**

I. Introduction

Let $X_1, X_2, \ldots, X_n$ be binary random variables. A classical result in probability states that, if the X_i are independent and identically distributed (i.i.d.) with common parameter $p_i = E(X_i) = \lambda/n$, then, when n is large, the distribution of their sum

$$S_n = X_1 + X_2 + \cdots + X_n$$

is close to $\mathrm{Po}(\lambda)$, the Poisson distribution with parameter λ. More generally, analogous results apply when the X_i are possibly dependent and not necessarily identically distributed. The distribution of S_n is close to $\mathrm{Po}(\lambda)$ as long as:

a) the sum $\sum p_i$ of the parameters p_i of the X_i is close to λ;
b) none of the X_i dominate the sum, i.e., all the p_i are small;
c) the variables X_i are not strongly dependent.

Such results are often referred to as "laws of small numbers" or "Poisson approximation results." See [1], [17, Sec. 2.6], [3] for details.

Our purpose here is to illustrate how techniques based on information-theoretic ideas can be used to establish general Poisson approximation inequalities. In Section II we prove the following.

Proposition 1. Poisson Approximation in Relative Entropy: If $S_n = \sum_{i=1}^n X_i$ is the sum of n (possibly dependent) binary random variables $X_1, X_2, \ldots, X_n$ with parameters $p_i = E(X_i)$ and with $E(S_n) = \sum_{i=1}^n p_i = \lambda$, then the distribution P_{S_n} of S_n satisfies

$$D(P_{S_n} \| \mathrm{Po}(\lambda)) \le \sum_{i=1}^n p_i^2 + \left[\sum_{i=1}^n H(X_i) - H(X_1, X_2, \ldots, X_n) \right]. \quad (1)$$

For two probability distributions P and Q on a discrete set S, the relative entropy between P and Q is defined as

$$D(P \| Q) = \sum_{x \in S} P(x) \log \frac{P(x)}{Q(x)}$$

and the entropy of a discrete random variable (or random vector) X with distribution P on S is

$$H(X) = H(P) = -\sum_{x \in S} P(x) \log P(x)$$

where $\log$ denotes the natural logarithm.

Whenever a), b) and c) hold, we expect the two terms in the right-hand side of (1) to be small, and hence the distribution of S_n to be close to $\mathrm{Po}(\lambda)$ in the relative entropy sense. Although $D(P\|Q)$ is not a proper metric, it is a natural measure of "dissimilarity" in the context of statistics [26], [11, Ch. 12], and it can be used to define a topology on probability measures [20]. Also, bounds in relative entropy can be translated into bounds in total variation via Pinsker's inequality [11]

$$\frac{1}{2} \|P - Q\|_{\mathrm{TV}}^2 \le D(P \| Q). \quad (2)$$

For example, if the X_i are independent, (1) reduces to

$$D(P_{S_n} \| \mathrm{Po}(\lambda)) \le \sum_{i=1}^n p_i^2. \quad (3)$$

Manuscript received November 1, 2002; revised June 11, 2004. The work of I. Kontoyiannis was supported in part by the National Science Foundation under Grants CCR-0073378 and DMS-9615444, and by a fellowship from the Sloan Foundation. The work of P. Harremoës was supported in part by a grant from the Cowi Foundation, and by a Postdoctoral Fellowship from the Villum Kann Rasmussen Foundation.

I. Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Providence, RI 02912 USA (e-mail:yiannis@dam.brown.edu).

P. Harremoës is with the Department of Mathematics, University of Copenhagen, DK-2100 København Ø, Denmark (e-mail: moes@math.ku.dk).

O. Johnson is with the Statistical Laboratory, Centre for Mathematical Sciences, Cambridge CB3 0WB, U.K. (e-mail: otj1000@cam.ac.uk).

Communicated by P. Narayan, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.840861

Although this is reminiscent of the simple total variation bound due to Le Cam [27]

$$\|P_{S_n} - \text{Po}(\lambda)\|_{\text{TV}} \leq \sum_{i=1}^{n} p_i^2$$

(which, incidentally, only holds when the X_i are independent), applying Pinsker's inequality (2) to (3) leads to the suboptimal bound

$$\|P_{S_n} - \text{Po}(\lambda)\|_{\text{TV}} \leq \Big[2\sum_{i=1}^{n} p_i^2\Big]^{1/2}. \quad (4)$$

The proof of Proposition 1 uses only elementary information-theoretic facts that are established using little more than Jensen's inequality. To get sharper bounds for the case of independent random variables X_i, in Section III we employ a new discrete version of the Fisher information which we call scaled Fisher information, and we prove the following.

Theorem 1. Poisson Approximation for Independent Variables: If $S_n = \sum_{i=1}^n X_i$ is the sum of n independent binary random variables $X_1, X_2, \ldots, X_n$, with $E(S_n) = \sum_{i=1}^n p_i = \lambda$, then

$$D(P_{S_n}\|\text{Po}(\lambda)) \leq \frac{1}{\lambda}\sum_{i=1}^{n} \frac{p_i^3}{1-p_i}. \quad (5)$$

The proof of Theorem 1 combines a natural discrete analog of Stam's subbativity of the Fisher information [35], [7] and a recent logarithmic Sobolev inequality of Bobkov and Ledoux [8]. As we discuss extensively in Section III, Theorem 1 is a significant improvement over Proposition 1, and in certain cases it leads to total variation bounds that are asymptotically optimal up to multiplicative constants in the convergence rate. Moreover, (5) is a nontrivial improvement over existing results, as it gives a bound for the relative entropy and not just the total variation distance.

For an information-theoretic interpretation, consider a triangular array of binary random variables

$$\{(X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)}),\ n \geq 1\}$$

such that the right-hand side of (1) goes to zero as $n \to \infty$ (as, for example, when the $X_i^{(n)}$ are i.i.d. Bernoulli (λ/n)). Then the distribution of S_n converges to $\text{Po}(\lambda)$, i.e., P_{S_n} comes closer and closer to the "most random" distribution among all those that can be obtained by summing a finite number of Bernoulli random variables. Let $\mathcal{P}(\lambda)$ denote the set of all distributions of sums S_n of n *independent* binary random variables with $E(S_n) = \lambda$, for any finite n. Then [19]

$$H(\text{Po}(\lambda)) = \sup\{H(P) : P \in \mathcal{P}(\lambda)\}.$$

So, roughly and somewhat incorrectly speaking, the entropy of S_n "increases" to the maximum entropy $H(\text{Po}(\lambda))$ as n grows. This invites a tempting analogy with the second law of thermodynamics, stating that the uncertainty of a physical system increases with time, until the system reaches equilibrium in its maximum entropy state.

Corresponding information-theoretic interpretations and proofs have been given for numerous classical results of proability theory, including the central limit theorem [28], [9], [4], [21], the convergence of Markov chains [31], [24], [6], many large-deviations results [12], [16], [13], the martingale convergence theorem [5], [6], and the Hewitt–Savage 0–1 law [29]. See also the powerful comments in [18, pp. 211, 215]. Finally, we mention that Johnstone and MacGibbon considered the problem of Poisson convergence from the information-theory angle in [22]. Their approach is different from ours, and parallels that in [9], [4] for the central limit theorem.

II. General Bounds in Relative Entropy

Before giving the proof of Proposition 1 we introduce some notation and briefly recall two elementary, well-known facts. The first one formalizes the intuitive idea that we cannot do better in a hypothesis test by simply preprocessing the data. Suppose X and Y are random variables with distributions P and Q, respectively, let f be an arbitrary function, and write P', Q' for the distribution of $f(X)$ and $f(Y)$, respectively. The following "data processing" inequality is an easy consequence of Jensen's inequality [14, Lemma 1.3.11]:

$$D(P'\|Q') \leq D(P\|Q).$$

Next, given X and Y with joint distribution $P_{X,Y}$ and marginals P_X and P_Y, let $I(X;Y) = H(X) - H(X \mid Y)$ denote their mutual information. The "chain rule" is the simple expansion

$$D(P_{X,Y}\|Q_X \times Q_Y) = D(P_X\|Q_X) + D(P_Y\|Q_Y) + I(X;Y)$$

for any two probability distributions Q_X and Q_Y.

Proof of Proposition 1: If we define $S_n' = \sum_{i=1}^n Z_i$, where Z_i are independent Poisson (p_i) random variables, then the distribution $P_{S_n'}$ of S_n' is $\text{Po}(\lambda)$ and

$$\begin{aligned} D(P_{S_n}\|\text{Po}(\lambda)) &= D(P_{S_n}\|P_{S_n'}) \\ &\overset{\text{(a)}}{\leq} D(P_{X_1,\ldots,X_n}\|P_{Z_1,\ldots,Z_n}) \\ &\overset{\text{(b)}}{=} \sum_{i=1}^{n} D(P_{X_i}\|\text{Po}(p_i)) \\ &\quad + \sum_{i=1}^{n-1} I(X_i; (X_{i+1},\ldots,X_n)) \end{aligned} \quad (6)$$

where (a) follows from the data processing inequality, and (b) follows by applying the chain rule $(n-1)$ times. Using simple calculus we obtain the bound

$$D(\text{Bern}(p)\|\text{Po}(p)) = (1-p)\log\frac{(1-p)}{e^{-p}} + p\log\frac{p}{pe^{-p}} \leq p^2$$

which, applied to each term in the first sum in (6), gives

$$\begin{aligned} D(P_{S_n}\|\text{Po}(\lambda)) &\leq \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n-1} I(X_i; (X_{i+1},\ldots,X_n)) \\ &= \sum_{i=1}^{n} p_i^2 + \Big[\sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n)\Big] \end{aligned} \quad (7)$$

where in the last step we expanded the definition of the mutual informations. □

The first term in the preceding bound makes precise what we mean by the requirement that "all the p_i be small" whereas the second term quantifies their degree of dependence. It is worth noting that this difference between the sum of the entropies of the X_i and their joint entropy can also be written as the relative entropy $D(P_{X_1^n} \| P_{X_1} \times \cdots \times P_{X_n})$ between their joint distribution and the product of their marginals. This expression also admits a natural interpretation as a measure of how far the X_i are from being independent.

As indicated in the Introduction, although the result of Proposition 1 is generally good enough to prove convergence to the Poisson distribution, for finite n it often gives a suboptimal convergence rate. This is also illustrated in the following two examples.

A Markov Chain: Let

$$\{(X_1^{(n)}, X_2^{(n)}, \dots, X_n^{(n)}),\ \ n \geq 1\}$$

be a triangular array of binary random variables such that each row $(X_1^{(n)}, \dots, X_n^{(n)})$ is a Markov chain with transition matrix

$$\begin{pmatrix} \frac{n}{n+1} & \frac{1}{n+1} \\ \frac{n-1}{n+1} & \frac{2}{n+1} \end{pmatrix}$$

and with each $X_i^{(n)}$ having (the stationary) Bernoulli $(\frac{1}{n})$ distribution. The convergence of the distribution of $S_n = \sum_{i=1}^n X_i^{(n)}$ to $\text{Po}\,(1)$ is a well-studied problem; see, e.g., [10] and the references therein. Applying Proposition 1 (or, equivalently, inequality (7)) in this case translates to

$$\begin{aligned} D(P_{S_n} \| \text{Po}(1)) &\leq \sum_{i=1}^n \frac{1}{n^2} + \sum_{i=1}^{n-1} I(X_i^{(n)}; X_{i+1}^{(n)}) \\ &= \frac{1}{n} + (n-1) I(X_1^{(n)}; X_2^{(n)}) \end{aligned}$$

since

$$I(X_i^{(n)}; (X_{i+1}^{(n)}, \dots, X_n^{(n)})) = I(X_i^{(n)}; X_{i+1}^{(n)})$$

by the Markov property, and stationarity implies that

$$I(X_i^{(n)}; X_{i+1}^{(n)}) = I(X_1^{(n)}; X_2^{(n)}).$$

A straightforward calculation yields that

$$\begin{aligned} (n-1) I(X_1^{(n)}; X_2^{(n)}) = (n-1) &\left[h\left(\frac{1}{n}\right) - h\left(\frac{1}{n+1}\right) \right] \\ &+ \frac{n-1}{n} h\left(\frac{1}{n+1}\right) - \frac{n-1}{n} h\left(\frac{2}{n+1}\right) \end{aligned}$$

where $h(p)$ denotes the binary entropy function

$$h(p) = -p \log p - (1-p) \log(1-p)$$

and simple calculus shows that all three terms above converge to zero as $n \to \infty$. In fact, this expression can be bounded above by

$$h\left(\frac{1}{n+1}\right) + \frac{\log n}{n} \leq 3 \frac{\log n}{n}$$

where the last inequality holds for all $n \geq 3$, so putting it all together

$$D(P_{S_n} \| \text{Po}(1)) \leq 3 \frac{\log n}{n} + \frac{1}{n}.$$

[A corresponding bound can similarly be derived if instead of stationarity we assume that $X_1^{(n)}$ has $p_1^{(n)} = E(X_1^{(n)}) < 1/n$.] As mentioned earlier, although this bound is sufficient to prove that P_{S_n} converges to the Poisson distribution, it leads to a convergence rate in total variation of order $\sqrt{(\log n)/n}$, compared to the $O(1/n)$ bound derived in [3], [33], [34].

A Compound Poisson Approximation Example: Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameters $p_i = E(X_i)$, write

$$\lambda = \sum_{i=1}^n p_i,$$

and let $\alpha_1, \alpha_2, \dots, \alpha_n$ be i.i.d., independent of the X_i, with distribution

$$\alpha_i = \begin{cases} 1, & \text{with prob. } 1/2 \\ 2, & \text{with prob. } 1/2. \end{cases}$$

We will show that the distribution of the sum

$$S_n = \sum_{i=1}^n \alpha_i X_i$$

is close to the compound Poisson distribution with parameters $(\lambda/2, \lambda/2)$, which we denote by $\text{Po}(\lambda/2, \lambda/2)$. Recall that if Z_1 and Z_2 are i.i.d. Poisson $(\lambda/2)$ random variables, then $Z = (Z_1 + 2Z_2)$ has $\text{Po}(\lambda/2, \lambda/2)$ distribution. Alternatively, we can write $Z = \sum_{i=1}^n Y_i$ where the Y_i are independent $\text{Po}(p_i/2, p_i/2)$ random variables. Arguing as before, the data processing inequality and the chain rule imply that

$$\begin{aligned} D(P_{S_n} \| \text{Po}(\lambda/2, \lambda/2)) &\leq D(P_{\alpha_1 X_1, \dots, \alpha_n X_n} \| P_{Y_1, \dots, Y_n}) \\ &= \sum_{i=1}^n D(P_{\alpha_i X_i} \| P_{Y_i}) \end{aligned}$$

and it is straightforward to calculate

$$\begin{aligned} D(P_{\alpha_i X_i} \| P_{Y_i}) \leq p_i^2 + (1-p_i)[p_i + \log(1-p_i)] \\ - \frac{p_i}{2} \log(1 + p_i/4) \leq p_i^2 \end{aligned}$$

so that

$$D(P_{S_n} \| \text{Po}(\lambda/2, \lambda/2)) \leq \sum_{i=1}^n p_i^2.$$

A General Method: Finally, we outline a simple general strategy for approximating the distribution P_{S_n} of the sum of n nonnegative integer-valued random variables $X_1, X_2, \dots, X_n$ by the distribution of some infinitely divisible discrete random variable Z with $E(S_n) = E(Z)$.

First, use the infinitely divisibility of P_Z to represent Z as $Z = \sum_{i=1}^n Y_i$ where the Y_i are independent and have the same distribution as Z but with different parameters. Then apply the data processing inequality and the chain rule as before to obtain

$$D(P_{S_n}\|P_Z) \le \sum_{i=1}^n D(P_{X_i}\|P_{Y_i}) + \Big[\sum_{i=1}^n H(X_i) - H(X_1,\dots X_n)\Big]$$

and estimate the last two terms in above inequality. The first term should be small if the X_i are individually small and well-approximated by the corresponding Y_i, and the second term should be small if the X_i are sufficiently weakly dependent.

III. Tighter Bounds for Independent Random Variables

Next we take a different point of view that yields tighter bounds than Proposition 1. Recall that in [22], [30], [23], the Fisher information of a random variable X with distribution P on $\mathbb{Z}_+ = \{0,1,2,\dots\}$, is defined in a way analogous to that for continuous random variables, via

$$J(X) = E\Big[\Big(\frac{P(X-1)-P(X)}{P(X)}\Big)^2\Big]$$

with the convention that $P(-1)=0$. However, as Kagan [23] acknowledges, this definition is really only useful if X is supported on the entire $\mathbb{Z}_+$: If X has bounded support then for some n, $P(n)>0$ but $P(n+1)=0$, which implies that $J(X)=\infty$.

Partly in order to avoid this difficulty, we proceed along a different route. Recalling that the Poisson distribution is characterized by the recurrence $\lambda P(x) = (x+1)P(x+1)$ for all x, we let the *scaled score function of a random variable* X with mean λ and distribution P on $\mathbb{Z}_+$ be

$$\rho_X(x) = \frac{(x+1)P(x+1)}{\lambda P(x)} - 1, \qquad x\in\mathbb{Z}_+$$

and we define the *scaled Fisher information of* X as

$$K(X) = \lambda E[\rho_X(X)^2].$$

From this we easily see that

$$K(X) \ge 0$$

with equality iff $\rho_X(X)=0$ with probability 1, i.e., iff X has a Poisson (λ) distribution. Moreover, as we show next, the smaller the value of $K(X)$, the closer P is to the Poisson (λ) distribution. The proof of Proposition 2, given in Section III-B, is an easy consequence of a recent logarithmic Sobolev inequality of Bobkov and Ledoux [8].

Proposition 2. Relative Entropy and $K(X)$: If X is a random variable with distribution P on $\mathbb{Z}_+$ and with $E(X)=\lambda$, then

$$D(P\|\mathrm{Po}(\lambda)) \le K(X) \tag{8}$$

as long as either P has full support (i.e., $P(k)>0$ for all k), or finite support (i.e., there exists $N\in\mathbb{Z}_+$ such that $P(k)=0$ for all $k>N$).

Note that from (8) and Pinsker's inequality (2) we have that

$$\|P - \mathrm{Po}(\lambda)\|_{\mathrm{TV}} \le \sqrt{2K(X)}. \tag{9}$$

We also give a direct proof of (9) in Section III-B, based on a simple Poincaré inequality for the Poisson measure.

A. Results

The main step in the proof of Theorem 1 will be to establish a form of subadditivity for the scaled Fisher information. It is worth noting that in the Gaussian case the Fisher information is also subadditive [35], [7], but, in contrast to the present setting, subadditivity alone does not suffice to prove the central limit theorem [4]. Proposition 3 is proved in Section III-B.

Proposition 3. Subadditivity of Scaled Fisher Information: If $S_n = \sum_{i=1}^n X_i$ is the sum of n independent integer-valued random variables $X_1, X_2, \dots, X_n$, with means $E(X_i)=p_i$ and $E(S_n)=\sum_{i=1}^n p_i = \lambda$, then

$$K(S_n) \le \sum_{i=1}^n \frac{p_i}{\lambda} K(X_i).$$

Proof of Theorem 1: If the X_i are independent Bernoulli (p_i) random variables with $\sum_{i=1}^n p_i = \lambda$, then $K(X_i) = p_i^2/(1-p_i)$ and Proposition 3 gives

$$K(S_n) \le \frac{1}{\lambda}\sum_{i=1}^n \frac{p_i^3}{1-p_i}.$$

Combining this with $X=S_n$ in Proposition 2 yields inequality (5). □

Example 1: If the X_i are i.i.d. Bernoulli (λ/n) random variables, from Theorem 1 combined with Pinsker's inequality (2) we obtain that for any $\epsilon>0$

$$\|P_{S_n} - \mathrm{Po}(\lambda)\|_{\mathrm{TV}} \le (2+\epsilon)\frac{\lambda}{n}, \qquad \text{for } n \ge \lambda/\epsilon.$$

This is a definite improvement over the earlier $2\lambda/\sqrt{n}$ bound from (4), and, except for the constant factor, it is asymptotically of the right order; see [3], [15] for details.

Example 2: If the X_i are i.i.d. Bernoulli $(\mu/\sqrt{n})$ random variables, Theorem 1 together with Pinsker's inequality (2) yield

$$\big\|P_{S_n} - \mathrm{Po}(\mu\sqrt{n})\big\|_{\mathrm{TV}} \le \frac{\mu}{\sqrt{n}}\sqrt{\frac{2}{1-\mu/\sqrt{n}}} \approx \frac{\mu}{\sqrt{n}}\sqrt{2}$$

which is of the same order as the optimal asymptotic rate, as $n\to\infty$

$$\big\|P_{S_n} - \mathrm{Po}(\mu\sqrt{n})\big\|_{\mathrm{TV}} \sim \frac{\mu}{\sqrt{n}}\sqrt{1/(2\pi e)}$$

derived in [15].

Example 3: If the X_i are geometric random variables with respective distributions $P_i(x) = (1-q_i)^x q_i,\ x\ge 0$, then $K(X_i) = (1-q_i)^2/q_i$. Letting $S_n = \sum_{i=1}^n X_i$ and assuming that

$$E(S_n) = \sum_{i=1}^n \frac{1-q_i}{q_i} = \lambda$$

combining Proposition 3 and the bound (9) yields

$$\|P_{S_n} - \mathrm{Po}(\lambda)\|_{\mathrm{TV}} \leq \sqrt{\frac{2}{\lambda}\sum_{i=1}^{n}\frac{(1-q_i)3}{q_i^2}}.$$

In particular, taking all the $q_i = n/(n+\lambda)$ gives the elegant estimate

$$\|P_{S_n} - \mathrm{Po}(\lambda)\|_{\mathrm{TV}} \leq \frac{\sqrt{2}\lambda}{\sqrt{n(n+\lambda)}} \leq \sqrt{2}\frac{\lambda}{n}.$$

To see how tight the result of Proposition 3 is in general, note that the following lower bound of Cramér–Rao type holds: Since for all a and any random variable S with mean λ and variance σ^2

$$0 \leq \lambda E(\rho_S(S) - a(S-\lambda))^2 = K(S) + \lambda\left(a^2\sigma^2 - 2a\left(\frac{\sigma^2-\lambda}{\lambda}\right)\right) \quad (10)$$

choosing $a = (\sigma^2 - \lambda)/(\sigma^2\lambda)$, we obtain that

$$K(S) \geq \left(\sigma^2 - \lambda\right)^2/(\sigma^2\lambda).$$

In Example 1, where $S = S_n = \sum_{i=1}^n X_i$ is the sum of n i.i.d. Bernoulli (λ/n) random variables, the lower bound (10) coincides with the upper bound given in Proposition 3. Similarly, in Example 3 with all the $q_i = n/(n+\lambda)$, the upper bound from Proposition 3 holds with equality. Therefore, any remaining slackness in our bounds comes from either Proposition 2 or Pinsker's inequality.

Finally, in Proposition 4 below, we establish a formal connection between relative entropy and the probability distribution $(x+1)P(x+1)/\lambda$ implicitly used in our definition of the scaled Fisher information. It is proved in the next section.

Proposition 4: Let X be an integer-valued random variable with distribution P and mean λ. If X is the sum of independent Bernoulli random variables, then

$$D(P\|\mathrm{Po}(\lambda)) = \int_0^\infty D(P_t\|\widetilde{P}_t)dt \quad (11)$$

where $P_t(r) = \Pr(X_t = r)$ is the distribution of $X_t = X + \mathrm{Po}(t)$ where $\mathrm{Po}(t)$ is an independent Poisson (t) random variable, and

$$\widetilde{P}_t(r) = (r+1)\Pr(X_t = r+1)/(\lambda+t).$$

More generally, the same result holds for any random variable X that has $K(X) < \infty$ and satisfies the logarithmic Sobolev inequality of Proposition 2.

This result is reminiscent of the well-known de Bruijn identity, which states that the (differential) relative entropy between a random variable X and a Gaussian with the same variance can be written as a weighted integral of (continuous) Fisher informations of convex combinations of X and an independent $N(0,t)$ random variable; see [11], [4]. In a similar vain, if we formally expand the logarithm in the integrand in (11) as a Taylor series, then the first term in the expansion (the quadratic term) turns out to be equal to $K(X_t)/2(\lambda+t)$. Therefore,

$$D(P\|\mathrm{Po}(\lambda)) \approx \int_0^\infty \frac{K(X+\mathrm{Po}(t))}{2(\lambda+t)}dt$$

giving an alternative formula to Proposition 2, also relating scaled Fisher information and relative entropy.

B. Proofs

Although subsequently in several places we formally divide by a quantity which may be zero, this is taken care of by the usual conventions, $0\log(0/a) = 0$, $0\log(0/0) = 0$, and $0\log(a/0) = \infty$, for any $a > 0$.

Proof of Proposition 2: Let $\mathrm{Po}_\lambda(k)$ denote the $\mathrm{Po}(\lambda)$ probabilities. In the case when P has full support, the result follows immediately from Corollary 4 of [8], upon considering the function $f(k) = P(k)/\mathrm{Po}_\lambda(k),\ k \geq 0$.

In the case of finite support, for $\epsilon > 0$ let X^ϵ have the mixture distribution

$$P^\epsilon = \epsilon\,\mathrm{Po}_\lambda + (1-\epsilon)P.$$

Then $E(X^\epsilon) = \lambda$ and P^ϵ has full support, so by the previous part

$$D(P^\epsilon\|\mathrm{Po}(\lambda)) \leq K(X^\epsilon). \quad (12)$$

But since $P(k) = 0$ for $k \geq N+1$, then $P^\epsilon(k)/\mathrm{Po}_\lambda(k) = \epsilon$ for those k, and letting $\epsilon \downarrow 0$ in the left-hand side of (12) we get

$$D(P^\epsilon\|\mathrm{Po}(\lambda)) = \sum_{k=0}^{N} P^\epsilon(k)\log\left[\frac{P^\epsilon(k)}{\mathrm{Po}_\lambda(k)}\right] + \Pr\{\mathrm{Po}(\lambda) > N\}\epsilon\log\epsilon \to D(P\|\mathrm{Po}(\lambda)).$$

Moreover

$$\frac{(k+1)P^\epsilon(k+1)}{\lambda P^\epsilon(k)} = 1, \qquad k \geq N+1$$

so

$$K(X^\epsilon) = \sum_{k=0}^{N} P^\epsilon(k)\Big[\frac{(k+1)P^\epsilon(k+1)}{\lambda P^\epsilon(k)} - 1\Big]^2 \to K(X)$$

as $\epsilon \downarrow 0$, and this completes the proof. □

Next we prove the bound in (9) using a classical Poincaré inequality for the Poisson distribution. We actually establish the following (apparently stronger) bound for the Hellinger distance $\|P - \mathrm{Po}(\lambda)\|_H$ between P and $\mathrm{Po}(\lambda)$:

$$\|P - \mathrm{Po}(\lambda)\|_{\mathrm{TV}}^2 \leq \|P - \mathrm{Po}(\lambda)\|_H^2 \leq 2K(X).$$

Proof of (9): For any function $f : \mathbb{Z}_+ \to \mathbb{R}$, define $\Delta f(x) = f(x+1) - f(x)$. It is well known that, writing $\mathrm{Po}_\lambda(x)$ for the $\mathrm{Poisson}(\lambda)$ probabilities, then for all functions g in $L^2(q)$

$$\sum_x \mathrm{Po}_\lambda(x)(g(x)-\mu)^2 \leq \lambda\sum_x \mathrm{Po}_\lambda(x)(\Delta g(x))^2 \quad (13)$$

where $\mu = \sum_x g(x)\mathrm{Po}_\lambda(x)$ is the mean of g under $\mathrm{Po}(\lambda)$; see, for example, Klaassen [25].

Using the simple fact that

$$(\sqrt{u}-1)^2 \le (\sqrt{u}-1)^2(\sqrt{u}+1)^2 = (u-1)^2, \qquad \text{for all } u \ge 0$$

we get that

$$K(X) = \lambda \sum_x P(x)\left(\frac{P(x+1)\mathrm{Po}_\lambda(x)}{\mathrm{Po}_\lambda(x+1)P(x)} - 1\right)^2 \ge \lambda \sum_x P(x)\left(\sqrt{\frac{P(x+1)\mathrm{Po}_\lambda(x)}{\mathrm{Po}_\lambda(x+1)P(x)}} - 1\right)^2$$

and applying (13) to the function $g(x) = \sqrt{P(x)/\mathrm{Po}_\lambda(x)}$ we obtain

$$K(X) \ge \lambda \sum_x \mathrm{Po}_\lambda(x)\left(\sqrt{\frac{P(x+1)}{\mathrm{Po}_\lambda(x+1)}} - \sqrt{\frac{P(x)}{\mathrm{Po}_\lambda(x)}}\right)^2 \ge \sum_x \mathrm{Po}_\lambda(x)\left(\sqrt{\frac{P(x)}{\mathrm{Po}_\lambda(x)}} - \mu\right)^2 = 1-\mu^2$$

where $\mu = \sum_x \sqrt{P(x)\mathrm{Po}_\lambda(x)}$. Therefore, the Hellinger distance $\|P - \mathrm{Po}(\lambda)\|_H$ satisfies

$$\|P - \mathrm{Po}(\lambda)\|_H^2 = (2-2\mu) \le 2(1-\mu^2) \le 2K(X),$$

and since

$$\|P - \mathrm{Po}(\lambda)\|_{\mathrm{TV}} \le \sqrt{\|P - \mathrm{Po}(\lambda)\|_H}$$

(see, e.g., [32, p. 360]) the result follows. □

For the proof of Proposition 3, as in the case of normal convergence in Fisher information, we exploit the theory of L^2 spaces and the fact that scaled score functions of sums are conditional expectations (projections) of the original scaled score functions.

Lemma. Convolution: If X and Y are nonnegative integer-valued random variables with probability distributions P and Q and means p and q, respectively, then,

$$\rho_{X+Y}(z) = E[\alpha_X \rho_X(X) + \alpha_Y \rho_Y(Y) \mid X+Y=z],$$

where $\alpha_X = p/(p+q)$, $\alpha_Y = q/(p+q)$.

Proof: Writing $F(z+1) = \sum_x P(x)Q(z-x+1)$ for the distribution of $X+Y$, we get (see the first equation at the bottom of the page) as required, where (a) follows by moving x to $(x+1)$ in the first sum. □

Proof of Proposition 3: It suffices to prove the case $n=2$. By the Lemma

$$0 \le E\left[\frac{p_1}{\lambda}\rho_{X_1}(X_1) + \frac{p_2}{\lambda}\rho_{X_2}(X_2) - \rho_{X_1+X_2}(X_1+X_2)\right]^2 = E\left[\frac{p_1}{\lambda}\rho_{X_1}(X_1) + \frac{p_2}{\lambda}\rho_{X_2}(X_2)\right]^2 - E[\rho_{X_1+X_2}(X_1+X_2)]^2$$

therefore, noting that $E[\rho_X(X)]=0$ for any random variable X

$$\begin{aligned} K(X_1+X_2) &= (p_1+p_2)E[\rho_{X_1+X_2}(X_1+X_2)]^2 \\ &\le \lambda E\left[\frac{p_1}{\lambda}\rho_{X_1}(X_1) + \frac{p_2}{\lambda}\rho_{X_2}(X_2)\right]^2 \\ &= \frac{p_1}{\lambda}\left(p_1 E[\rho_{X_1}(X_1)]^2\right) + \frac{p_2}{\lambda}\left(p_2 E[\rho_{X_2}(X_2)]^2\right) \\ &= \frac{p_1}{\lambda}K(X_1) + \frac{p_2}{\lambda}K(X_2) \end{aligned}$$

as claimed. □

Proof of Proposition 4: Assume for the moment that the relative entropy between P_t and $\mathrm{Po}(\lambda+t)$ tends to zero as $t \to \infty$ (this will be established below). Then we can write $D(P\|\mathrm{Po}(\lambda))$ as the integral in the second equation at the bottom of the page. Since the probabilities P_t satisfy a differential-difference equation $\frac{\partial P_t}{\partial t}(x) = P_t(x-1) - P_t(x)$, we have

$$\frac{\partial}{\partial t}E[\log(X_t)!] = \sum_r \frac{\partial P_t}{\partial t}(r)\log r! = \sum_r (P_t(r-1) - P_t(r))\log r! = E\log(X_t+1)$$

and similarly

$$\frac{\partial}{\partial t}H(X_t) = -\sum_r \frac{\partial P_t}{\partial t}(r)\log P_t(r) = \sum_r P_t(r)\log\left(\frac{P_t(r)}{P_t(r+1)}\right).$$

Substituting these two expressions in the expansion of $D(P\|\mathrm{Po}(\lambda))$ the result follows.

$$\begin{aligned} \rho_{X+Y}(z) &= \sum_x \frac{(z+1)P(x)Q(z-x+1)}{(p+q)F(z)} - 1 \\ &= \sum_x \left[\frac{xP(x)Q(z-x+1)}{(p+q)F(z)} + \frac{(z-x+1)P(x)Q(z-x+1)}{(p+q)F(z)}\right] - 1 \\ &= \alpha_X\left[\sum_x \frac{xP(x)}{pP(x-1)}\frac{P(x-1)Q(z-x+1)}{F(z)} - 1\right] + \alpha_Y\left[\sum_x \frac{(z-x+1)Q(z-x+1)}{qQ(z-x)}\frac{P(x)Q(z-x)}{F(z)} - 1\right] \\ &\overset{(a)}{=} \sum_x \frac{P(x)Q(z-x)}{F(z)}\left[\alpha_X\rho_X(x) + \alpha_Y\rho_Y(z-x)\right] \end{aligned}$$

$$\begin{aligned} D(P\|\mathrm{Po}(\lambda)) &= -\int_0^\infty \frac{\partial}{\partial t}D(P_t\|\mathrm{Po}(\lambda+t))dt \\ &= -\int_0^\infty \frac{\partial}{\partial t}\Big((\lambda+t) - E[X_t\log(\lambda+t)] + E[\log(X_t!)] - H(X_t)\Big)dt \\ &= \int_0^\infty \left(\log(\lambda+t) - \frac{\partial}{\partial t}E[\log(X_t!)] + \frac{\partial}{\partial t}H(X_t)\right)dt. \end{aligned}$$

Finally, it remains to establish our initial assumption. If X is the sum of independent Bernoulli random variables then it has finite support and Proposition 2 holds; moreover, $K(X)$ is easily seen to be finite by Proposition 3. More generally, using Propositions 2 and 3 we have

$$D(P_t \| \mathrm{Po}(\lambda + t)) \leq K(X + \mathrm{Po}(t)) \leq \frac{\lambda}{\lambda + t} K(X) \to 0$$

as $t \to \infty$, as required.

□

ACKNOWLEDGMENT

Interesting conversations with Persi Diaconis, Stu Geman, and Matt Harrison are greatfully acknowledged. We also thank the Associate Editor and the two referees for their useful comments on an earlier version of the manuscript.

REFERENCES

[1] R. Arratia, L. Goldstein, and L. Gordon, "Two moments suffice for Poisson approximations: The Chen-Stein method," *Ann. Probab.*, vol. 17, pp. 9–25, 1989.
[2] R. R. Bahadur, *Some Limit Theorems in Statistics*. Philadelphia, PA: SIAM, 1971.
[3] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*. New York: Clarendon/Oxford Univ. Press, 1992.
[4] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.
[5] ——, "Information theory and martingales," in *Proc. IEEE Int. Symp. Information Theory*, Budapest, Hungary, 1991.
[6] ——, "Limits of information, Markov chains, and projection," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000.
[7] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.
[8] S. G. Bobkov and M. Ledoux, "On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures," *J. Funct. Anal.*, vol. 156, pp. 347–365, 1998.
[9] L. D. Brown, "A proof of the central limit theorem motivated by the Cramér-Rao inequality," in *Statistics and Probability: Essays in Honor of C. R. Rao*. Amsterdam, The Netherlands: North-Holland, 1982, pp. 141–148.
[10] V. Čekanavičius, "On the convergence of Markov binomial to Poisson distribution," *Statist. Probab. Lett.*, vol. 58, no. 1, pp. 83–91, 2002.
[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[12] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 1984.
[13] I. Csiszár, T. M. Cover, and B. S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 6, pp. 788–801, Nov. 1987.
[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[15] P. Deheuvels and D. Pfeifer, "A semigroup approach to Poisson approximation," *Ann. Probab.*, vol. 14, no. 2, pp. 663–676, 1986.
[16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.
[17] R. Durrett, *Probability: Theory and Examples*, 2nd ed. Belmont, CA: Duxbury, 1996.
[18] B. V. Gnedenko and V. Y. Korolev, *Random Summation: Limit Theorems and Applications*. Boca Raton, FL: CRC, 1996.
[19] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 2039–2041, Jul. 2001.
[20] ——, "The information topology," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 431.
[21] O. Johnson, "Entropy inequalities and the central limit theorem," *Stochastic Process. Appl.*, vol. 88, pp. 291–304, 2000.
[22] I. M. Johnstone and B. MacGibbon, "Une mesure d'information caractérisant la loi de Poisson," in *Séminaire de Probabilités, XXI*. Berlin, Germany: Springer-Verlag, 1987, pp. 563–573.
[23] A. Kagan, "Letter to the editor: 'A discrete version of the Stam inequality and a characterization of the Poisson distribution'," *J. Statist. Plann. Inference*, vol. 92, no. 1–2, pp. 7–12, 2001.
[24] D. G. Kendall, "Information theory and the limit-theorem for Markov chains and processes with a countable infinity of states," *Ann. Inst. Statist. Math.*, vol. 15, pp. 137–143, 1963.
[25] C. A. J. Klaassen, "On an inequality of Chernoff," *Ann. Probab.*, vol. 13, no. 3, pp. 966–974, 1985.
[26] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1997.
[27] L. Le Cam, "An approximation theorem for the Poisson binomial distribution," *Pacific J. Math.*, vol. 10, pp. 1181–1197, 1960.
[28] J. V. Linnik, "An information-theoretic proof of the central limit theorem with Lindeberg conditions," *Theor. Probab. Appl.*, vol. 4, pp. 288–299, 1959.
[29] N. O'Connell. (2000) Information-Theoretic Proof of the Hewitt-Savage Zero-One Law. Tech. Rep,. Hewlett-Packard Labs., Bristol, U.K.. [Online]. Available: http://www.hpl.hp.com/techreports/2000/HPL-BRIMS-2000-18.html
[30] V. Papathanasiou, "Some characteristic properties of the Fisher information matrix via Cacoullos-type inequalities," *J. Multivariate Anal.*, vol. 44, no. 2, pp. 256–265, 1993.
[31] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. I, Berkeley, CA, 1961, pp. 547–561.
[32] L. Saloff-Coste, "Lectures on finite Markov chains," in *Lectures on Probability Theory and Statistics (Saint-Flour, 1996) (Lecture Notes in Mathematics)*. Berlin, Germany : Springer-Verlag, 1997, vol. 1665, pp. 301–413.
[33] R. F. Serfozo, "Compound Poisson approximations for sums of random variables," *Ann. Probab.*, vol. 14, pp. 1391–1398, 1986.
[34] ——, "Correction to: 'Compound Poisson approximations for sums of random variables'," *Ann. Probab.*, vol. 16, pp. 429–439, 1988.
[35] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, 1959.