DOI: 10.1214/EJP.v15-799

Electronic Journal of Probability

Vol. 15 (2010), Paper no. 42, pages 1344–1368.

Journal URL
http://www.math.washington.edu/~ejpecp/

Compound Poisson Approximation via Information Functionals[*]

A.D. Barbour [†] O. Johnson [‡] I. Kontoyiannis [§] M. Madiman [¶]

Abstract

An information-theoretic development is given for the problem of compound Poisson approximation, which parallels earlier treatments for Gaussian and Poisson approximation. Nonasymptotic bounds are derived for the distance between the distribution of a sum of independent integer-valued random variables and an appropriately chosen compound Poisson law. In the case where all summands have the same conditional distribution given that they are non-zero, a bound on the relative entropy distance between their sum and the compound Poisson distribution is derived, based on the data-processing property of relative entropy and earlier Poisson approximation results. When the summands have arbitrary distributions, corresponding bounds are derived in terms of the total variation distance. The main technical ingredient is the introduction of two "information functionals," and the analysis of their properties. These information functionals play a role analogous to that of the classical Fisher information in normal approximation. Detailed comparisons are made between the resulting inequalities and related bounds.

Key words: Compound Poisson approximation, Fisher information, information theory, relative

[*] Portions of this paper are based on the Ph.D. dissertation [27] of M. Madiman, advised by I. Kontoyiannis, at the Division of Applied Mathematics, Brown University.

[†] Angewandte Mathematik, Universität Zürich–Irchel, Winterthurerstrasse 190, CH–8057 Zürich, Switzerland. Email: a.d.barbour@math.unizh.ch

[‡] Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK. Email: O.Johnson@bristol.ac.uk

[§] Department of Informatics, Athens University of Economics & Business, Patission 76, Athens 10434, Greece. Email: yiannis@aueb.gr . I.K. was supported in part by a Marie Curie International Outgoing Fellowship, PIOF-GA-2009-235837.

[¶] Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven CT 06511, USA. Email: mokshay.madiman@yale.edu

entropy, Stein’s method.

AMS 2000 Subject Classification: Primary 60E15,60E07,60F05,94A17.

Submitted to EJP on April 21, 2010, final version accepted August 12, 2010.

1 Introduction and main results

The study of the distribution of a sum $S_n = \sum_{i=1}^n Y_i$ of weakly dependent random variables Y_i is an important part of probability theory, with numerous classical and modern applications. This work provides an information-theoretic treatment of the problem of approximating the distribution of S_n by a compound Poisson law, when the Y_i are discrete, independent random variables. Before describing the present approach, some of the relevant background is briefly reviewed.

1.1 Normal approximation and entropy

When $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed (i.i.d.) random variables with mean zero and variance $\sigma^2 < \infty$, the central limit theorem (CLT) and its various refinements state that the distribution of $T_n := (1/\sqrt{n})\sum_{i=1}^n Y_i$ is close to the $N(0,\sigma^2)$ distribution for large n. In recent years the CLT has been examined from an information-theoretic point of view and, among various results, it has been shown that, if the Y_i have a density with respect to Lebesgue measure, then the density f_{T_n} of the normalized sum T_n converges *monotonically* to the normal density with mean zero and variance σ^2; that is, the entropy $h(f_{T_n}) := -\int f_{T_n} \log f_{T_n}$ of f_{T_n} *increases* to the $N(0,\sigma^2)$ entropy as $n \to \infty$, which is *maximal* among all random variables with fixed variance σ^2. [Throughout, 'log' denotes the natural logarithm.]

Apart from this intuitively appealing result, information-theoretic ideas and techniques have also provided nonasymptotic inequalities, for example giving accurate bounds on the relative entropy $D(f_{T_n}\|\phi) := \int f_{T_n} \log(f_{T_n}/\phi)$ between the density of T_n and the limiting normal density ϕ. Details can be found in [8; 19; 17; 3; 2; 35; 28] and the references in these works.

The gist of the information-theoretic approach is based on estimates of the Fisher information, which acts as a "local" version of the relative entropy. For a random variable Y with a differentiable density f and variance $\sigma^2 < \infty$, the *(standardized) Fisher information* is defined as,

$$J_N(Y) := E\left[\frac{\partial}{\partial y}\log f(Y) - \frac{\partial}{\partial y}\log\phi(Y)\right]^2,$$

where ϕ is the $N(0,\sigma^2)$ density. The functional J_N satisfies the following properties:

(A) $J_N(Y)$ is the variance of the (standardized) score function, $r_Y(y) := \frac{\partial}{\partial y}\log f(y) - \frac{\partial}{\partial y}\log\phi(y)$, $y \in \mathbb{R}$.

(B) $J_N(Y) = 0$ if and only if Y is Gaussian.

(C) J_N satisfies a subadditivity property for sums.

(D) If $J_N(Y)$ is small then the density f of Y is approximately normal and, in particular, $D(f\|\phi)$ is also appropriately small.

Roughly speaking, the information-theoretic approach to the CLT and associated normal approximation bounds consists of two steps; first a strong version of Property (C) is used to show that $J_N(T_n)$ is close to zero for large n, and then Property (D) is applied to obtain precise bounds on the relative entropy $D(f_{T_n}\|\phi)$.

1.2 Poisson approximation

More recently, an analogous program was carried out for Poisson approximation. The Poisson law was identified as having maximum entropy within a natural class of discrete distributions on $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ [16; 34; 18], and Poisson approximation bounds in terms of relative entropy were developed in [23]; see also [21] for earlier related results. The approach of [23] follows a similar outline to the one described above for normal approximation. Specifically, for a random variable Y with values in $\mathbb{Z}_+$ and distribution P, the *scaled Fisher information of Y* was defined as,

$$J_\pi(Y) := \lambda E[\rho_Y(Y)^2] = \lambda \mathrm{Var}(\rho_Y(Y)), \tag{1.1}$$

where λ is the mean of Y and the scaled score function ρ_Y is given by,

$$\rho_Y(y) := \frac{(y+1)P(y+1)}{\lambda P(y)} - 1, \quad y \geq 0. \tag{1.2}$$

[Throughout, we use the term 'distribution' to refer to the discrete probability mass function of an integer-valued random variable.]

As discussed briefly before the proof of Theorem 1.1 in Section 2 the functional $J_\pi(Y)$ was shown in [23] to satisfy Properties (A-D) exactly analogous to those of the Fisher information described above, with the Poisson law playing the role of the Gaussian distribution. These properties were employed to establish optimal or near-optimal Poisson approximation bounds for the distribution of sums of nonnegative integer-valued random variables [23]. Some additional relevant results in earlier work can be found in [36][31][29][10].

1.3 Compound Poisson approximation

This work provides a parallel treatment for the more general – and technically significantly more difficult – problem of approximating the distribution P_{S_n} of a sum $S_n = \sum_{i=1}^n Y_i$ of independent $\mathbb{Z}_+$-valued random variables by an appropriate compound Poisson law. This and related questions arise naturally in applications involving counting; see, e.g., [7; 1; 4; 14]. As we will see, in this setting the information-theoretic approach not only gives an elegant alternative route to the classical asymptotic results (as was the case in the first information-theoretic treatments of the CLT), but it actually yields fairly sharp finite-n inequalities that are competitive with some of the best existing bounds.

Given a distribution Q on $\mathbb{N} = \{1, 2, \ldots\}$ and a $\lambda > 0$, recall that the compound Poisson law $\mathrm{CPo}(\lambda, Q)$ is defined as the distribution of the random sum $\sum_{i=1}^Z X_i$, where $Z \sim \mathrm{Po}(\lambda)$ is Poisson distributed with parameter λ and the X_i are i.i.d. with distribution Q, independent of Z.

Relevant results that can be seen as the intellectual background to the information-theoretic approach for compound Poisson approximation were recently established in [20; 38], where it was shown that, like the Gaussian and the Poisson, the compound Poisson law has a maximum entropy property within a natural class of probability measures on $\mathbb{Z}_+$. Here we provide nonasymptotic, computable and accurate bounds for the distance between P_{S_n} and an appropriately chosen compound Poisson law, partly based on extensions of the information-theoretic techniques introduced in [23] and [21] for Poisson approximation.

In order to state our main results we need to introduce some more terminology. When considering the distribution of $S_n = \sum_{i=1}^n Y_i$, we find it convenient to write each Y_i as the product $B_i X_i$ of two independent random variables, where B_i takes values in $\{0,1\}$ and X_i takes values in $\mathbb{N}$. This is done uniquely and without loss of generality, by taking B_i to be $\mathrm{Bern}(p_i)$ with $p_i = \Pr\{Y_i \neq 0\}$, and X_i having distribution Q_i on $\mathbb{N}$, where $Q_i(k) = \Pr\{Y_i = k \,|\, Y_i \geq 1\} = \Pr\{Y_i = k\}/p_i$, for $k \geq 1$.

In the special case of a sum $S_n = \sum_{i=1}^n Y_i$ of random variables $Y_i = B_i X_i$ where all the X_i have the same distribution Q, it turns out that the problem of approximating P_{S_n} by a compound Poisson law can be reduced to a Poisson approximation inequality. This is achieved by an application of the so-called "data-processing" property of the relative entropy, which then facilitates the use of a Poisson approximation bound established in [23]. The result is stated in Theorem 1.1 below; its proof is given in Section 2.

Theorem 1.1. *Consider a sum $S_n = \sum_{i=1}^n Y_i$ of independent random variables $Y_i = B_i X_i$, where the X_i are i.i.d. $\sim Q$ and the B_i are independent* $\mathrm{Bern}(p_i)$*. Then the relative entropy between the distribution P_{S_n} of S_n and the* $\mathrm{CPo}(\lambda, Q)$ *distribution satisfies,*

$$D(P_{S_n} \| \mathrm{CPo}(\lambda, Q)) \leq \frac{1}{\lambda} \sum_{i=1}^n \frac{p_i^3}{1-p_i},$$

where $\lambda := \sum_{i=1}^n p_i$.

Recall that, for distributions P and Q on $\mathbb{Z}_+$, the *relative entropy*, or *Kullback-Leibler divergence*, $D(P\|Q)$, is defined by,

$$D(P\|Q) := \sum_{x \in \mathbb{Z}_+} P(x) \log \Big[\frac{P(x)}{Q(x)} \Big].$$

Although not a metric, relative entropy is an important measure of closeness between probability distributions [12][13] and it can be used to obtain total variation bounds via Pinsker's inequality [13],

$$d_{\mathrm{TV}}(P, Q)^2 \leq \tfrac{1}{2} D(P\|Q),$$

where, as usual, the total variation distance is

$$d_{\mathrm{TV}}(P, Q) := \tfrac{1}{2} \sum_{x \in \mathbb{Z}_+} \big| P(x) - Q(x) \big| = \max_{A \subset \mathbb{Z}_+} \big| P(A) - Q(A) \big|.$$

In the general case where the distributions Q_i corresponding to the X_i in the summands $Y_i = B_i X_i$ are not identical, the data-processing argument used in the proof of Theorem 1.1 can no longer be applied. Instead, the key idea in this work is the introduction of two "information functionals," or simply "informations," which, in the present context, play a role analogous to that of the Fisher information J_N and the scaled Fisher information J_π in Gaussian and Poisson approximation, respectively.

In Section 3 we will define two such information functionals, $J_{\mathbf{Q},1}$ and $J_{Q,2}$, and use them to derive compound Poisson approximation bounds. Both $J_{\mathbf{Q},1}$ and $J_{Q,2}$ will be seen to satisfy natural analogs of Properties (A-D) stated above, except that only a weaker version of Property (D) will be established: When either $J_{\mathbf{Q},1}(Y)$ or $J_{Q,2}(Y)$ is close to zero, the distribution of Y is close to a compound Poisson law in the sense of total variation rather than relative entropy. As in normal and Poisson approximation, combining the analogs of Properties (C) and (D) satisfied by the two new information functionals, yields new compound Poisson approximation bounds.

Theorem 1.2. *Consider a sum $S_n = \sum_{i=1}^n Y_i$ of independent random variables $Y_i = B_i X_i$, where each X_i has distribution Q_i on $\mathbb{N}$ with mean q_i, and each $B_i \sim \mathrm{Bern}(p_i)$. Let $\lambda = \sum_{i=1}^n p_i$ and $Q = \sum_{i=1}^n \frac{p_i}{\lambda} Q_i$. Then,*

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \le H(\lambda, Q) q \left\{ \left[\sum_{i=1}^n \frac{p_i^3}{1-p_i} \right]^{1/2} + D(\mathbf{Q}) \right\},$$

where P_{S_n} is the distribution of S_n, $q = \sum_{i=1}^n \frac{p_i}{\lambda} q_i$, $H(\lambda, Q)$ denotes the Stein factor defined in (1.4) *below, and $D(\mathbf{Q})$ is a measure of the dissimilarity of the distributions $\mathbf{Q} = (Q_i)$, which vanishes when the Q_i are identical:*

$$D(\mathbf{Q}) := \sum_{j=1}^{\infty} \sum_{i=1}^{n} \frac{j p_i}{q} \left| Q_i(j) - Q(j) \right|. \tag{1.3}$$

Theorem 1.2 is an immediate consequence of the subadditivity property of $J_{\mathbf{Q},1}$ established in Corollary 4.2, combined with the total variation bound in Proposition 5.3. The latter bound states that, when $J_{\mathbf{Q},1}(Y)$ is small, the total variation distance between the distribution of Y and a compound Poisson law is also appropriately small. As explained in Section 5, the proof of Proposition 5.3 uses a basic result that comes up in the proof of compound Poisson inequalities via Stein's method, namely, a bound on the sup-norm of the solution of the Stein equation. This explains the appearance of the Stein factor, defined next. But we emphasize that, apart from this point of contact, the overall methodology used in establishing the results in Theorems 1.2 and 1.4 is entirely different from that used in proving compound Poisson approximation bounds via Stein's method.

Definition 1.3. *Let Q be a distribution on $\mathbb{N}$. If $\{jQ(j)\}$ is a non-increasing sequence, set $\delta = [\lambda\{Q(1) - 2Q(2)\}]^{-1}$ and let,*

$$H_0(\lambda, Q) = \begin{cases} 1 & \text{if } \delta \ge 1 \\ \sqrt{\delta}(2 - \sqrt{\delta}) & \text{if } \delta < 1. \end{cases}$$

For general Q and any $\lambda > 0$, the Stein factor *$H(\lambda, Q)$ is defined as:*

$$H(\lambda, Q) = \begin{cases} H_0(\lambda, Q), & \text{if } \{jQ(j)\} \text{ is non-increasing} \\ e^{\lambda} \min\left\{1, \frac{1}{\lambda Q(1)}\right\}, & \text{otherwise.} \end{cases} \tag{1.4}$$

Note that in the case when all the Q_i are identical, Theorem 1.2 yields,

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q))^2 \le H(\lambda, Q)^2 q^2 \sum_{i=1}^n \frac{p_i^3}{1-p_i}, \tag{1.5}$$

where q is the common mean of the $Q_i = Q$, whereas Theorem 1.1 combined with Pinsker's inequality yields a similar, though not generally comparable, bound,

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q))^2 \le \frac{1}{2\lambda} \sum_{i=1}^n \frac{p_i^3}{1-p_i}. \tag{1.6}$$

See Section 6 for detailed comparisons in special cases.

The third and last main result, Theorem 1.4, gives an analogous bound to that of Theorem 1.2, with only a single term in the right-hand-side. It is obtained from the subadditivity property of the second information functional $J_{\mathbf{Q},2}$, Proposition 4.3, combined with the corresponding total variation bound in Proposition 5.1.

Theorem 1.4. *Consider a sum $S_n = \sum_{i=1}^n Y_i$ of independent random variables $Y_i = B_i X_i$, where each X_i has distribution Q_i on $\mathbb{N}$ with mean q_i, and each $B_i \sim \mathrm{Bern}(p_i)$. Assume all Q_i have have full support on $\mathbb{N}$, and let $\lambda = \sum_{i=1}^n p_i$, $Q = \sum_{i=1}^n \frac{p_i}{\lambda} Q_i$, and P_{S_n} denote the distribution of S_n. Then,*

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq H(\lambda, Q) \left\{ \sum_{i=1}^n \left[p_i^3 \sum_y Q_i(y) y^2 \Big(\frac{Q_i^{*2}(y)}{2Q_i(y)} - 1 \Big)^2 \right] \right\}^{1/2},$$

where Q_i^{*2} denotes the convolution $Q_i * Q_i$ and $H(\lambda, Q)$ denotes the Stein factor defined in (1.4) *above.*

The accuracy of the bounds in the three theorems above is examined in specific examples in Section 6, where the resulting estimates are compared with what are probably the sharpest known bounds for compound Poisson approximation. Although the main conclusion of these comparisons – namely, that in broad terms our bounds are competitive with some of the best existing bounds and, in certain cases, may even be the sharpest – is certainly encouraging, we wish to emphasize that the main objective of this work is the development of an elegant conceptual framework for compound Poisson limit theorems via information-theoretic ideas, akin to the remarkable information-theoretic framework that has emerged for the central limit theorem and Poisson approximation.

The rest of the paper is organized as follows. Section 2 contains basic facts, definitions and notation that will remain in effect throughout. It also contains a brief review of earlier Poisson approximation results in terms of relative entropy, and the proof of Theorem 1.1. Section 3 introduces the two new information functionals: The *size-biased information* $J_{\mathbf{Q},1}$, generalizing the scaled Fisher information of [23], and the *Katti-Panjer information* $J_{Q,2}$, generalizing a related functional introduced by Johnstone and MacGibbon in [21]. It is shown that, in each case, Properties (A) and (B) analogous to those stated in Section 1.1 for Fisher information hold for $J_{\mathbf{Q},1}$ and $J_{Q,2}$. In Section 4 we consider Property (C) and show that both $J_{\mathbf{Q},1}$ and $J_{Q,2}$ satisfy natural subadditivity properties on convolution. Section 5 contains bounds analogous to that Property (D) above, showing that both $J_{\mathbf{Q},1}(Y)$ and $J_{Q,2}(Y)$ dominate the total variation distance between the distribution of Y and a compound Poisson law.

2 Size-biasing, compounding and relative entropy

In this section we collect preliminary definitions and notation that will be used in subsequent sections, and we provide the proof of Theorem 1.1.

The compounding operation in the definition of the compound Poisson law in the Introduction can be more generally phrased as follows. [Throughout, the empty sum $\sum_{i=1}^0 [\ldots]$ is taken to be equal to zero].

Definition 2.1. *For any $\mathbb{Z}_+$-valued random variable $Y \sim R$ and any distribution Q on $\mathbb{N}$, the* compound distribution $C_Q R$ *is that of the sum,*

$$\sum_{i=1}^{Y} X_i,$$

where the X_i are i.i.d. with common distribution Q, independent of Y.

For example, the compound Poisson law $\mathrm{CPo}(\lambda, Q)$ is simply $C_Q\mathrm{Po}(\lambda)$, and the *compound binomial distribution* $C_Q\mathrm{Bin}(n, p)$ is that of the sum $S_n = \sum_{i=1}^{n} B_i X_i$ where the B_i are i.i.d. $\mathrm{Bern}(p)$ and the X_i are i.i.d. with distribution Q, independent of the B_i. More generally, if the B_i are Bernoulli with different parameters p_i, we say that S_n is a *compound Bernoulli sum* since the distribution of each summand $B_i X_i$ is $C_Q\mathrm{Bern}(p_i)$.

Next we recall the size-biasing operation, which is intimately related to the Poisson law. For any distribution P on $\mathbb{Z}_+$ with mean λ, the *(reduced) size-biased distribution $P^{\#}$* is,

$$P^{\#}(y) = \frac{(y+1)P(y+1)}{\lambda}, \quad y \geq 0.$$

Recalling that a distribution P on $\mathbb{Z}_+$ satisfies the recursion,

$$(k+1)P(k+1) := \lambda P(k), \quad k \in \mathbb{Z}_+, \tag{2.1}$$

if and only if $P = \mathrm{Po}(\lambda)$, it is immediate that $P = \mathrm{Po}(\lambda)$ if and only if $P = P^{\#}$. This also explains, in part, the definition (1.1) of the scaled Fisher information in [23]. Similarly, the *Katti-Panjer recursion* states that P is the $\mathrm{CPo}(\lambda, Q)$ law if and only if,

$$kP(k) = \lambda \sum_{j=1}^{k} jQ(j)P(k-j), \quad k \in \mathbb{Z}_+; \tag{2.2}$$

see the discussion in [20] for historical remarks on the origin of (2.2).

Before giving the proof of Theorem 1.1 we recall two results related to Poisson approximation bounds from [23]. First, for any random variable $X \sim P$ on $\mathbb{Z}_+$ with mean λ, a modified log-Sobolev inequality of [9] was used in [23, Proposition 2] to show that,

$$D(P \| \mathrm{Po}(\lambda)) \leq J_\pi(X), \tag{2.3}$$

as long as P has either full support or finite support. Combining this with the subadditivity property of J_π and elementary computations, yields [23, Theorem 1] that states: If T_n is the sum of n independent $B_i \sim \mathrm{Bern}(p_i)$ random variables, then,

$$D(P_{T_n} \| \mathrm{Po}(\lambda)) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i}, \tag{2.4}$$

where P_{T_n} denotes the distribution of T_n and $\lambda = \sum_{i=1}^{n} p_i$.

Proof of Theorem 1.1. Let $Z_n \sim \mathrm{Po}(\lambda)$ and $T_n = \sum_{i=1}^n B_i$. Then the distribution of S_n is also that of the sum $\sum_{i=1}^{T_n} X_i$; similarly, the $\mathrm{CPo}(\lambda, Q)$ law is the distribution of the sum $Z = \sum_{i=1}^{Z_n} X_i$. Thus, writing $\mathbf{X} = (X_i)$, we can express $S_n = f(\mathbf{X}, T_n)$ and $Z = f(\mathbf{X}, Z_n)$, where the function f is the same in both places. Applying the data-processing inequality and then the chain rule for relative entropy [13],

$$\begin{aligned} D(P_{S_n} \| \mathrm{CPo}(\lambda, Q)) &\leq D(P_{\mathbf{X},T_n} \| P_{\mathbf{X},Z_n}) \\ &= \Big[\sum_i D(P_{X_i} \| P_{X_i})\Big] + D(P_{T_n} \| P_{Z_n}) \\ &= D(P_{T_n} \| \mathrm{Po}(\lambda)), \end{aligned}$$

and the result follows from the Poisson approximation bound (2.4). □

3 Information functionals

This section contains the definitions of two new information functionals for discrete random variables, along with some of their basic properties.

3.1 Size-biased information

For the first information functional we consider, some knowledge of the summation structure of the random variables concerned is required.

Definition 3.1. *Consider the sum $S = \sum_{i=1}^n Y_i \sim P$ of n independent $\mathbb{Z}_+$-valued random variables $Y_i \sim P_i = C_{Q_i} R_i$, $i = 1, 2, \ldots, n$. For each j, let $Y_j' \sim C_{Q_j}(R_j^\#)$ be independent of the Y_i, and let $S^{(j)} \sim P^{(j)}$ be the same sum as S but with Y_j' in place of Y_j.*

Let q_i denote the mean of each Q_i, $p_i = E(Y_i)/q_i$ and $\lambda = \sum_i p_i$. Then the size-biased information of S relative to the sequence $\mathbf{Q} = (Q_i)$ *is,*

$$J_{\mathbf{Q},1}(S) := \lambda E[r_1(S; P, \mathbf{Q})^2],$$

where the score function r_1 is defined by,

$$r_1(s; P, \mathbf{Q}) := \frac{\sum_i p_i P^{(i)}(s)}{\lambda P(s)} - 1, \quad s \in \mathbb{Z}_+.$$

For simplicity, in the case of a single summand $S = Y_1 \sim P_1 = C_Q R$ we write $r_1(\cdot; P, Q)$ and $J_{Q,1}(Y)$ for the score and the size-biased information of S, respectively. [Note that the score function r_1 is only infinite at points x outside the support of P, which do not affect the definition of the size-biased information functional.]

Although at first sight the definition of $J_{\mathbf{Q},1}$ seems restricted to the case when all the summands Y_i have distributions of the form $C_{Q_i} R_i$, we note that this can always be achieved by taking $p_i = \Pr\{Y_i \geq 1\}$ and letting $R_i \sim \mathrm{Bern}(p_i)$ and $Q_i(k) = \Pr\{Y_i = k | Y_i \geq 1\}$, for $k \geq 1$, as before.

We collect below some of the basic properties of $J_{\mathbf{Q},1}$ that follow easily from the definition.

1. Since $E[r_1(S;P,\mathbf{Q})]=0$, the functional $J_{\mathbf{Q},1}(S)$ is in fact the variance of the score $r_1(S;P,\mathbf{Q})$.

2. In the case of a single summand $S=Y_1\sim C_QR$, if Q is the point mass at 1 then the score r_1 reduces to the score function ρ_Y in (1.2). Thus $J_{\mathbf{Q},1}$ can be seen as a generalization of the scaled Fisher information J_π of [23] defined in (1.1).

3. Again in the case of a single summand $S=Y_1\sim C_QR$, we have that $r_1(s;P,Q)\equiv 0$ if and only if $R^\#=R$, i.e., if and only if R is the Po(λ) distribution. Thus in this case $J_{Q,1}(S)=0$ if and only if $S\sim \text{CPo}(\lambda,Q)$ for some $\lambda>0$.

4. In general, writing $F^{(i)}$ for the distribution of the leave-one-out sum $\sum_{j\neq i}Y_i$,

$$r_1(\cdot;P,\mathbf{Q})\equiv 0 \iff \sum p_iF^{(i)}*(C_{Q_i}R_i-C_{Q_i}R_i^\#)\equiv 0.$$

   Hence within the class of ultra log-concave R_i (a class which includes compound Bernoulli sums), since the moments of R_i are no smaller than the moments of $R_i^\#$ with equality if and only if R_i is Poisson, the score $r_1(\cdot;P,\mathbf{Q})\equiv 0$ if and only if the R_i are all Poisson, i.e., if and only if P is compound Poisson.

3.2 Katti-Panjer information

Recall that the recursion (2.1) characterizing the Poisson distribution was used as part of the motivation for the definition of the scaled Fisher information J_π in (1.1) and (1.2). In an analogous manner, we employ the Katti-Panjer recursion (2.2) that characterizes the compound Poisson law to define another information functional.

Definition 3.2. *Given a $\mathbb{Z}_+$-valued random variable $Y\sim P$ and an arbitrary distribution Q on $\mathbb{N}$, the* Katti-Panjer information of Y relative to Q *is defined as,*

$$J_{Q,2}(Y):=E[r_2(Y;P,Q)^2],$$

where the score function r_2 is,

$$r_2(y;P,Q):=\frac{\lambda\sum_{j=1}^\infty jQ(j)P(y-j)}{P(y)}-y,\quad y\in\mathbb{Z}_+,$$

and where λ is the ratio of the mean of Y to the mean of Q.

From the definition of the score function r_2 it is immediate that,

$$\begin{aligned} E[r_2(Y;P,Q)] &= \sum_y P(y)r_2(y;P,Q)\\ &= \lambda\Bigg[\sum_{y:P(y)>0}\sum_j jQ(j)P(y-j)\Bigg]-E(Y)\\ &= \lambda\Bigg[\sum_j jQ(j)\Bigg]-E(Y)=0, \end{aligned}$$

therefore $J_{Q,2}(Y)$ is equal to the variance of $r_2(Y;P,Q)$. [This computation assumes that P has full support on $\mathbb{Z}_+$; see the last paragraph of this section for further discussion of this point.] Also, in view of the Katti-Panjer recursion (2.2) we have that $J_{Q,2}(Y) = 0$ if and only if $r_2(y; P, Q)$ vanishes for all y, which happens if and only if the distribution P of Y is CPo(λ, Q).

In the special case when Q is the unit mass at 1, the Katti-Panjer information of $Y \sim P$ reduces to,

$$J_{Q,2}(Y) = E\Big[\Big(\frac{\lambda P(Y-1)}{P(Y)} - Y\Big)^2\Big] = \lambda^2 I(Y) + (\sigma^2 - 2\lambda), \tag{3.1}$$

where λ, σ^2 are the mean and variance of Y, respectively, and $I(Y)$ denotes the functional,

$$I(Y) := E\Big[\Big(\frac{P(Y-1)}{P(Y)} - 1\Big)^2\Big], \tag{3.2}$$

proposed by Johnstone and MacGibbon [21] as a discrete version of the Fisher information (with the convention $P(-1) = 0$). Therefore, in view of (3.1) we can think of $J_{Q,2}(Y)$ as a generalization of the "Fisher information" functional $I(Y)$ of [21].

Finally note that, although the definition of $J_{Q,2}$ is more straightforward than that of $J_{\mathbf{Q},1}$, the Katti-Panjer information suffers the drawback that – like its simpler version $I(Y)$ in [21] – it is only finite for random variables Y with full support on $\mathbb{Z}_+$. As noted in [22] and [23], the definition of $I(Y)$ cannot simply be extended to all $\mathbb{Z}_+$-valued random variables by just ignoring the points outside the support of P, where the integrand in (3.2) becomes infinite. This was, partly, the motivation for the definition of the scaled scored function J_π in [23]. Similarly, in the present setting, the important properties of $J_{Q,2}$ established in the following sections *fail* unless P has full support, unlike for the size-biased information $J_{\mathbf{Q},1}$.

4 Subadditivity

The subadditivity property of Fisher information (Property (C) in the Introduction) plays a key role in the information-theoretic analysis of normal approximation bounds. The corresponding property for the scaled Fisher information (Proposition 3 of [23]) plays an analogous role in the case of Poisson approximation. Both of these results are based on a convolution identity for each of the two underlying score functions. In this section we develop natural analogs of the convolution identities and resulting subadditivity properties for the functionals $J_{\mathbf{Q},1}$ and $J_{Q,2}$.

4.1 Subadditivity of the size-biased information

The proposition below gives the natural analog of Property (C) in the the Introduction, for the information functional $J_{\mathbf{Q},1}$. It generalizes the convolution lemma and Proposition 3 of [23].

Proposition 4.1. *Consider the sum $S_n = \sum_{i=1}^n Y_i \sim P$ of n independent $\mathbb{Z}_+$-valued random variables $Y_i \sim P_i = C_{Q_i} R_i$, $i = 1, 2, \ldots, n$. For each i, let q_i denote the mean of Q_i, $p_i = E(Y_i)/q_i$ and $\lambda = \sum_i p_i$. Then,*

$$r_1(s; P, \mathbf{Q}) = E\left[\sum_{i=1}^n \frac{p_i}{\lambda} r_1(Y_i; P_i, Q_i) \middle| S_n = s\right], \tag{4.1}$$

and hence,

$$J_{\mathbf{Q},1}(S_n) \le \sum_{i=1}^{m} \frac{p_i}{\lambda} J_{Q_i,1}(Y_i). \tag{4.2}$$

Proof. In the notation of Definition 3.1 and the subsequent discussion, writing $F^{(i)} = P_1 * \ldots * P_{i-1} * P_{i+1} * \ldots * P_m$, so that $P^{(i)} = F^{(i)} * C_{Q_i} R_i^{\#}$, the right-hand side of the projection identity (4.1) equals,

$$\begin{array}{rl} & \displaystyle\sum_{i=1}^{n} \sum_{x} \frac{P_i(x) F^{(i)}(s-x)}{P(s)} \left(\frac{p_i}{\lambda} \left(\frac{C_{Q_i} R_i^{\#}(x)}{P_i(x)} - 1 \right) \right) \\ = & \displaystyle\frac{1}{\lambda P(s)} \left(\sum_{i=1}^{n} \sum_{x} p_i C_{Q_i} R_i^{\#}(x) F^{(i)}(s-x) \right) - 1 \\ = & \displaystyle\frac{1}{\lambda P(s)} \left(\sum_{i=1}^{n} p_i P^{(i)}(s) \right) - 1 \\ = & r_1(s; P, \mathbf{Q}), \end{array}$$

as required. The subadditivity result follows using the conditional Jensen inequality, exactly as in the proof of Proposition 3 of [23]. □

Corollary 4.2. *Under the assumptions of Proposition* 4.1*, if each $Y_i = B_i X_i$, where $B_i \sim$ Bern(p_i) and $X_i \sim Q_i$ where $p_i = \Pr\{Y_i \neq 0\}$ and $Q_i(k) = \Pr\{Y_i = k | Y_i \ge 1\}$, then,*

$$J_{\mathbf{Q},1}(S_n) \le \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i},$$

where $\lambda = \sum_i p_i$.

Proof. Consider $Y = BX$, where $B \sim R =$ Bern(p) and $X \sim Q$. Since $R^{\#} = \delta_0$ then $C_Q(R^{\#}) = \delta_0$. Further, Y takes the value 0 with probability $(1-p)$ and the value X with probability p. Thus,

$$\begin{array}{rcl} r_1(x; C_Q R, Q) & = & \displaystyle\frac{C_Q(R^{\#})(x)}{C_Q R(x)} - 1 \\ & = & \displaystyle\frac{\delta_0(x)}{(1-p)\delta_0(x) + pQ(x)} - 1 \\ & = & \begin{cases} \frac{p}{1-p} & \text{for } x = 0 \\ -1 & \text{for } x > 0. \end{cases} \end{array}$$

Consequently,

$$J_{Q,1}(Y) = \frac{p^2}{1-p}, \tag{4.3}$$

and using Proposition 4.1 yields,

$$J_{\mathbf{Q},1}(S_n) \le \sum_{i=1}^{n} \frac{p_i}{\lambda} J_{Q_i,1}(Y_i) = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i},$$

as claimed. □

4.2 Subadditivity of the Katti-Panjer information

When S_n is supported on the whole of $\mathbb{Z}_+$, the score r_2 satisfies a convolution identity and the functional $J_{Q,2}$ is subadditive. The following Proposition contains the analogs of (4.1) and (4.2) in Proposition 4.1 for the Katti-Panjer information $J_{Q,2}(Y)$. These can also be viewed as generalizations of the corresponding results for the Johnstone-MacGibbon functional $I(Y)$ established in [21].

Proposition 4.3. *Consider a sum $S_n = \sum_{i=1}^n Y_i$ of independent random variables $Y_i = B_i X_i$, where each X_i has distribution Q_i on $\mathbb{N}$ with mean q_i, and each $B_i \sim \text{Bern}(p_i)$. Let $\lambda = \sum_{i=1}^n p_i$ and $Q = \sum_{i=1}^n \frac{p_i}{\lambda} Q_i$. If each Y_i is supported on the whole of $\mathbb{Z}_+$, then,*

$$r_2(s;S_n,Q) = E\left[\sum_{i=1}^n r_2(Y_i;P_i,Q_i)\middle| S_n = s\right],$$

and hence,

$$J_{Q,2}(S_n) \le \sum_{i=1}^n J_{Q_i,2}(Y_i).$$

Proof. Write $r_{2,i}(\cdot)$ for $r_2(\cdot;P_i,Q_i)$, and note that $E(Y_i) = p_i q_i$, for each i. Therefore, $E(S_n) = \sum_i p_i q_i$ which equals λ times the mean of Q. As before, let $F^{(i)}$ denote the distribution of the leave-one-out sum $\sum_{j\neq i} Y_j$, and decompose the distribution P_{S_n} of S_n as $P_{S_n}(s) = \sum_x P_i(x)F^{(i)}(s-x)$. We have,

$$\begin{aligned}
r_2(s;S_n,Q) &= \frac{\lambda \sum_{j=1}^\infty jQ(j)P_{S_n}(s-j)}{P_{S_n}(s)} - s \\
&= \sum_{i=1}^n \frac{p_i \sum_{j=1}^\infty jQ_i(j)P_{S_n}(s-j)}{P_{S_n}(s)} - s \\
&= \sum_{i=1}^n \sum_x \frac{p_i \sum_{j=1}^\infty jQ_i(j)P_i(x-j)F^{(i)}(s-x)}{P_{S_n}(s)} - s \\
&= \sum_{i=1}^n \sum_x \frac{P_i(x)F^{(i)}(s-x)}{P_{S_n}(s)} \left[\frac{p_i \sum_{j=1}^\infty jQ_i(j)P_i(x-j)}{P_i(x)}\right] - s \\
&= E\left[\sum_{i=1}^n r_{2,i}(Y_i)\middle| S_n = s\right]
\end{aligned}$$

proving the projection identity. And using the conditional Jensen inequality, noting that the cross-terms vanish because $E[r_2(X;P,Q) = 0]$ for any $X \sim P$ with full support (cf. the discussion in Section 3.2), the subadditivity result follows, as claimed. □

5 Information functionals dominate total variation

In the case of Poisson approximation, the modified log-Sobolev inequality (2.3) directly relates the relative entropy to the scaled Fisher information J_π. However, the known (modified) log-Sobolev inequalities for compound Poisson distributions [37; 24], only relate the relative entropy to functionals different from $J_{\mathbf{Q},1}$ or $J_{Q,2}$. Instead of developing subadditivity results for those other functionals, we build, in part, on some of the ideas from Stein's method and prove relationships between the total variation distance and the information functionals $J_{\mathbf{Q},1}$ and $J_{Q,2}$. (Note, however, that Lemma 5.4 does offer a partial result showing that the relative entropy can be bounded in terms of $J_{\mathbf{Q},1}$.)

To illustrate the connection between these two information functionals and Stein's method, we find it simpler to first examine the Katti-Panjer information. Recall that, for an arbitrary function $h : \mathbb{Z}_+ \to \mathbb{R}$, a function $g : \mathbb{Z}_+ \to \mathbb{R}$ satisfies the *Stein equation* for the compound Poisson measure $\mathrm{CPo}(\lambda, Q)$ if,

$$\lambda \sum_{j=1}^{\infty} jQ(j)g(k+j) = kg(k) + \Big[h(k) - E[h(Z)]\Big], \quad g(0) = 0, \tag{5.1}$$

where $Z \sim \mathrm{CPo}(\lambda, Q)$. [See, e.g., [15] for details as well as a general review of Stein's method for Poisson and compound Poisson approximation.] Letting $h = \mathbb{I}_A$ for some $A \subset \mathbb{Z}_+$, writing g_A for the corresponding solution of the Stein equation, and taking expectations with respect to an arbitrary random variable $Y \sim P$ on $\mathbb{Z}_+$,

$$P(A) - \Pr\{Z \in A\} = E\left\{\lambda \sum_{j=1}^{\infty} jQ(j)g_A(Y+j) - Y g_A(Y)\right\}.$$

Then taking absolute values and maximizing over all $A \subset \mathbb{Z}_+$,

$$d_{\mathrm{TV}}(P, \mathrm{CPo}(\lambda, Q)) \le \sup_{A \subset \mathbb{Z}_+} \left| E\left\{\lambda \sum_{j=1}^{\infty} jQ(j)g_A(Y+j) - Y g_A(Y)\right\}\right|. \tag{5.2}$$

Noting that the expression in the expectation above is reminiscent of the Katti-Panjer recursion (2.2), it is perhaps not surprising that this bound relates directly to the Katti-Panjer information functional:

Proposition 5.1. *For any random variable $Y \sim P$ on $\mathbb{Z}_+$, any distribution Q on $\mathbb{N}$ and any $\lambda > 0$,*

$$d_{\mathrm{TV}}(P, \mathrm{CPo}(\lambda, Q)) \le H(\lambda, Q)\sqrt{J_{Q,2}(Y)},$$

where $H(\lambda, Q)$ is the Stein factor defined in (1.4).

Proof. We assume without loss of generality that Y is supported on the whole of $\mathbb{Z}_+$, since, otherwise, $J_{Q,2}(Y) = \infty$ and the result is trivial. Continuing from the inequality in (5.2),

$$\begin{aligned} d_{\mathrm{TV}}(P, \mathrm{CPo}(\lambda, Q)) &\le \Big(\sup_{A \subset \mathbb{Z}_+} \|g_A\|_\infty\Big) \sum_{y=0}^{\infty} \left|\lambda \sum_{j=1}^{\infty} jQ(j)P(y-j) - yP(y)\right| \\ &\le H(\lambda, Q) \sum_{y=0}^{\infty} P(y)|r_2(y; P, Q)| \\ &\le H(\lambda, Q)\sqrt{\sum_{y=0}^{\infty} P(y) r_2(y; P, Q)^2}, \end{aligned}$$

where the first inequality follows from rearranging the first sum, the second inequality follows from Lemma 5.2 below, and the last step is simply the Cauchy-Schwarz inequality. □

The following uniform bound on the sup-norm of the solution to the Stein equation (5.1) is the only auxiliary result we require from Stein's method. See [5] or [15] for a proof.

Lemma 5.2. *If g_A is the solution to the Stein equation* (5.1) *for $g = \mathbb{I}_A$, with $A \subset \mathbb{Z}_+$, then $\|g_A\|_\infty \leq H(\lambda, Q)$, where $H(\lambda, Q)$ is the Stein factor defined in* (1.4).

5.1 Size-biased information dominates total variation

Next we establish an analogous bound to that of Proposition 5.1 for the size-biased information $J_{\mathbf{Q},1}$. As this functional is not as directly related to the Katti-Panjer recursion (2.2) and the Stein equation (5.2), the proof is technically more involved.

Proposition 5.3. *Consider a sum $S = \sum_{i=1}^n Y_i \sim P$ of independent random variables $Y_i = B_i X_i$, where each X_i has distribution Q_i on $\mathbb{N}$ with mean q_i, and each $B_i \sim$* Bern(p_i)*. Let $\lambda = \sum_{i=1}^n p_i$ and $Q = \sum_{i=1}^n \frac{p_i}{\lambda} Q_i$. Then,*

$$d_{\mathrm{TV}}(P, \mathrm{CPo}(\lambda, Q)) \leq H(\lambda, Q) q \left(\sqrt{\lambda J_{\mathbf{Q},1}(S)} + D(\mathbf{Q}) \right),$$

where $H(\lambda, Q)$ is the Stein factor defined defined in (1.4)*, $q = (1/\lambda) \sum_i p_i q_i$ is the mean of Q, and $D(\mathbf{Q})$ is the measure of the dissimilarity between the distributions $\mathbf{Q} = (Q_i)$, defined in* (1.3).

Proof. For each i, let $T^{(i)} \sim F^{(i)}$ denote the leave-one-out sum $\sum_{j \neq i} Y_i$, and note that, as in the proof of Corollary 4.2, the distribution $F^{(i)}$ is the same as the distribution $P^{(i)}$ of the modified sum $S^{(i)}$ in Definition 3.1. Since Y_i is nonzero with probability p_i, we have, for each i,

$$\begin{aligned} E[Y_i g_A(S)] &= E[Y_i g_A(Y_i + T^{(i)})] \\ &= \sum_{j=1}^{\infty} \sum_{s=0}^{\infty} p_i Q_i(j) F^{(i)}(s) j g_A(j+s) \\ &= \sum_{j=1}^{\infty} \sum_{s=0}^{\infty} p_i j Q_i(j) P^{(i)}(s) g_A(s+j), \end{aligned}$$

where, for $A \subset \mathbb{Z}_+$ arbitrary, g_A denotes the solution of the Stein equation (5.1) with $h = \mathbb{I}_A$. Hence, summing over i and substituting in the expression in the right-hand-side of equation (5.2) with S in place of Y, yields,

$$\begin{aligned}
&E\Big\{\lambda\sum_{j=1}^{\infty} jQ(j)g_A(S+j)\Big\} - E[Sg_A(S)] \\
&\qquad = \sum_{s=0}^{\infty}\sum_{j=1}^{\infty} g_A(s+j)\left(\lambda jQ(j)P(s) - \sum_i p_i jQ_i(j)P^{(i)}(s)\right) \\
&\qquad = \sum_{s=0}^{\infty}\sum_{j=1}^{\infty} g_A(s+j)jQ(j)\left(\sum_i p_i(P(s) - P^{(i)}(s))\right) \\
&\qquad\qquad + \sum_{s=0}^{\infty}\sum_{j=1}^{\infty} g_A(s+j)\left(\sum_i p_i j(Q(j) - Q_i(j))P^{(i)}(s)\right) \\
&\qquad = -\sum_{s=0}^{\infty}\sum_{j=1}^{\infty} g_A(s+j)jQ(j)\lambda P(s)\left(\frac{\sum_i p_i P^{(i)}(s)}{\lambda P(s)} - 1\right) \\
&\qquad\qquad + \sum_{s=0}^{\infty}\sum_{j=1}^{\infty} g_A(s+j)\left(\sum_i p_i j(Q(j) - Q_i(j))P^{(i)}(s)\right). \qquad (5.3)
\end{aligned}$$

By the Cauchy-Schwarz inequality, the first term in (5.3) is bounded in absolute value by,

$$\sqrt{\lambda\sum_{j,s} g_A(s+j)^2 jQ(j)P(s)}\sqrt{\lambda\sum_{j,s} jQ(j)P(s)\left(\frac{\sum_i p_i P^{(i)}(s)}{\lambda P(s)} - 1\right)^2},$$

and for the second term, simply bound $\|g_A\|_\infty$ by $H(\lambda, Q)$ using Lemma 5.2, deducing a bound in absolute value of

$$H(\lambda, Q)\sum_{i,j} p_i j|Q(j) - Q_i(j)|.$$

Combining these two bounds with the expression in (5.3) and the original total-variation inequality (5.2) completes the proof, upon substituting the uniform sup-norm bound given in Lemma 5.2. □

Finally, recall from the discussion in the beginning of this section that the scaled Fisher information J_π satisfies a modified log-Sobolev inequality (2.3), which gives a bound for the relative entropy in terms of the functional J_π. For the information functionals $J_{\mathbf{Q},1}$ and $J_{Q,2}$ considered in this work, we instead established analogous bounds in terms of total variation. However, the following partial result holds for $J_{\mathbf{Q},1}$:

Lemma 5.4. *Let $Y = BX \sim P$, where $B \sim \text{Bern}(p)$ and $X \sim Q$ on $\mathbb{N}$. Then:*

$$D(P\|\text{CPo}(p, Q)) \leq J_{Q,1}(Y).$$

Proof. Recall from (4.3) that $J_{Q,1}(Y) = \frac{p^2}{1-p}$. Further, since the CPo$(p, Q)$ mass function at s is at least $e^{-p}pQ(s)$ for $s \geq 1$, we have, $D(C_Q\text{Bern}(p)\|\text{CPo}(p, Q)) \leq (1-p)\log(1-p) + p$, which yields the result. □

6 Comparison with existing bounds

In this section, we compare the bounds obtained in our three main results, Theorems 1.1, 1.2 and 1.4, with inequalities derived by other methods. Throughout, $S_n = \sum_{i=1}^n Y_i = \sum_{i=1}^n B_i X_i$, where the B_i and the Y_i are independent sequences of independent random variables, with $B_i \sim \mathrm{Bern}(p_i)$ for some $p_i \in (0,1)$, and with $X_i \sim Q_i$ on $\mathbb{N}$; we write $\lambda = \sum_{i=1}^n p_i$.

There is a large body of literature developing bounds on the distance between the distribution P_{S_n} of S_n and compound Poisson distributions; see, e.g., [15] and the references therein, or [33, Section 2] for a concise review.

We begin with the case in which all the $Q_i = Q$ are identical, when, in view of a remark of Le Cam [26, bottom of p.187] and Michel [30], bounds computed for the case $X_i = 1$ a.s. for all i are also valid for any Q. One of the earliest results is the following inequality of Le Cam [25], building on earlier results by Khintchine and Doeblin,

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \sum_{i=1}^n p_i^2. \tag{6.1}$$

Barbour and Hall (1984) used Stein's method to improve the bound to

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i^2. \tag{6.2}$$

Roos [32] gives the asymptotically sharper bound

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \left(\frac{3}{4e} + \frac{7\sqrt{\theta}(3 - 2\sqrt{\theta})}{6(1 - \sqrt{\theta})^2} \right) \theta, \tag{6.3}$$

where $\theta = \lambda^{-1} \sum_{i=1}^n p_i^2$, which was strengthened and simplified in form in Equation (30) of Čekanavičius and Roos [11] to give

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \frac{3\theta}{4e(1 - \sqrt{\theta})^{3/2}}. \tag{6.4}$$

In this setting, the bound (1.6) that was derived from Theorem 1.1 yields

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \left(\frac{1}{2\lambda} \sum_{i=1}^n \frac{p_i^3}{1 - p_i} \right)^{1/2}. \tag{6.5}$$

The bounds (6.2) – (6.5) are all derived using the observation made by Le Cam and Michel, taking Q to be degenerate at 1. For the application of Theorem 1.4, however, the distribution Q must have support the whole of $\mathbb{N}$, so Q cannot be replaced by the point mass at 1 in the formula; the bound that results from Theorem 1.4 can be expressed as

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq H(\lambda, Q) \left(K(Q) \sum_{i=1}^n p_i^3 \right)^{1/2},$$

$$\text{with } K(Q) = \sum_y Q(y) y^2 \left(\frac{Q^{*2}(y)}{2Q(y)} - 1 \right)^2. \tag{6.6}$$

Illustration of the effectiveness of these bounds with geometric Q and equal p_i is given in Section 6.2.

For non-equal Q_i, the bounds are more complicated. We compare those given in Theorems 1.2 and 1.4 with three other bounds. The first is Le Cam's bound (6.1) that still remains valid as stated in the case of non-equal Q_i. The second, from Stein's method, has the form

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq G(\lambda, Q) \sum_{i=1}^{n} q_i^2 p_i^2, \tag{6.7}$$

see Barbour and Chryssaphinou [6, eq. (2.24)], where q_i is the mean of Q_i and $G(\lambda, Q)$ is a Stein factor: if $jQ(j)$ is non-increasing, then

$$G(\lambda, Q) = \min\left\{1,\ \delta\left[\frac{\delta}{4} + \log^{+}\left(\frac{2}{\delta}\right)\right]\right\},$$

where $\delta = [\lambda\{Q(1) - 2Q(2)\}]^{-1} \geq 0$. The third is that of Roos [33], Theorem 2, which is in detail very complicated, but correspondingly accurate. A simplified version, valid if $jQ(j)$ is decreasing, gives

$$d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q)) \leq \frac{\alpha_2}{(1 - 2e\alpha_2)_+}, \tag{6.8}$$

where

$$\alpha_2 = \sum_{i=1}^{n} g(2p_i) p_i^2 \min\left(\frac{q_i^2}{e\lambda}, \frac{\nu_i}{2^{3/2}\lambda}, 1\right),$$

$\nu_i = \sum_{y \geq 1} Q_i(y)^2 / Q(y)$ and $g(z) = 2z^{-2} e^{z}(e^{-z} - 1 + z)$. We illustrate the effectiveness of these bounds in Section 6.3; in our examples, Roos's bounds are much the best.

6.1 Broad comparisons

Because of their apparent complexity and different forms, general comparisons between the bounds are not straightforward, so we consider two particular cases below in Sections 6.2 and 6.3. However, the following simple observation on approximating compound binomials by a compound Poisson gives a first indication of the strength of one of our bounds.

Proposition 6.1. *For equal p_i and equal Q_i:*

1. *If $n > (\sqrt{2}p(1-p))^{-1}$, then the bound of Theorem* 1.1 *is stronger than Le Cam's bound* (6.1)*;*
2. *If $p < 1/2$, then the bound of Theorem* 1.1 *is stronger than the bound* (6.2)*;*
3. *If* $0.012 < p < 1/2$ *and $n > (\sqrt{2}p(1-p))^{-1}$ are satisfied, then the bound of Theorem* 1.1 *is stronger than all three bounds in* (6.1)*,* (6.2) *and* (6.3)*.*

Proof. The first two observations follow by simple algebra, upon noting that the bound of Theorem 1.1 in this case reduces to $\frac{p}{\sqrt{2(1-p)}}$; the third is shown numerically, noting that here $\theta = p$. □

Although of no real practical interest, the bound of Theorem 1.1 is also better than (6.4) for $0.27 < p < 1/2$.

One can also examine the rate of convergence of the total variation distance between the distribution P_{S_n} and the corresponding compound Poisson distribution, under simple asymptotic schemes. We think of situations in which the p_i and Q_i are not necessarily equal, but are all in some reasonable sense comparable with one another; we shall also suppose that $jQ(j)$ is more or less a fixed and decreasing sequence. Two ways in which p varies with n are considered:

Regime I. $p = \lambda/n$ for fixed λ, and $n \to \infty$;

Regime II. $p = \sqrt{\frac{\mu}{n}}$, so that $\lambda = \sqrt{\mu n} \to \infty$ as $n \to \infty$.

Under these conditions, the Stein factors $H(\lambda, Q)$ are of the same order as $1/\sqrt{np}$. Table 1 compares the asymptotic performance of the various bounds above. The poor behaviour of the bound in Theorem 1.2 shown in Table 1 occurs because, for large values of λ, the quantity $D(\mathbf{Q})$ behaves much like λ, unless the Q_i are identical or near-identical.

Bound	$d_{\text{TV}}(P_{S_n}, \text{CPo}(\lambda, Q))$ to leading order	**I**	**II**
Le Cam (6.1)	np^2	n^{-1}	1
Roos (6.8)	$np^2 \min(1, 1/(np))$	n^{-1}	$n^{-1/2}$
Stein's method (6.7)	$np^2 \min(1, \log(np)/np)$	n^{-1}	$n^{-1/2} \log n$
Theorem 1.2	p	1	$n^{1/4}$
Theorem 1.4 (6.6)	p	n^{-1}	$n^{-1/2}$

Table 1: Comparison of the first-order asymptotic performance of the bounds in (6.1), (6.7) and (6.8), with those of Theorems 1.2 and 1.4 for comparable but non-equal Q_i, in the two limiting regimes $p \asymp 1/n$ and $p \asymp 1/\sqrt{n}$.

6.2 Example. Compound binomial with equal geometrics

We now examine the finite-n behavior of the approximation bounds (6.1) – (6.3) in the particular case of equal p_i and equal Q_i, when Q_i is geometric with parameter $\alpha > 0$, $Q(j) = (1-\alpha)\alpha^{j-1}$, $j \geq 1$.

If $\alpha < \frac{1}{2}$, then $\{jQ(j)\}$ is decreasing and, with $\delta = [\lambda(1 - 3\alpha + 2\alpha^2)]^{-1}$, the Stein factor in (6.6) becomes

$$H(\lambda, Q) = \min\{1, \sqrt{\delta}(2 - \sqrt{\delta})\}.$$

The resulting bounds are plotted in Figures 1 – 3.

6.3 Example. Sums with unequal geometrics

Here, we consider finite-n behavior of the approximation bounds (6.1), (6.7) and (6.8) in the particular case when the distributions Q_i are geometric with parameters $\alpha_i > 0$. The resulting bounds are plotted in Figures 4 and 5.

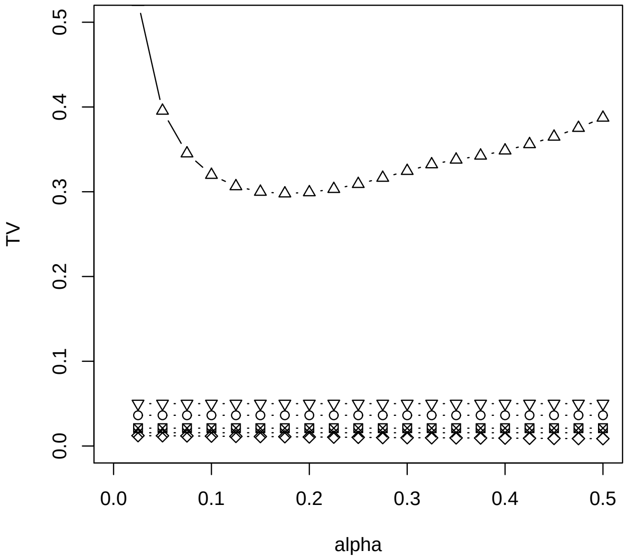

Figure 1: Bounds on the total variation distance $d_{\mathrm{TV}}(C_Q\mathrm{Bin}(p,Q),\mathrm{CPo}(\lambda,Q))$ for $Q \sim \mathrm{Geom}(\alpha)$, plotted against the parameter α, with $n = 100$ and $\lambda = 5$ fixed. The values of the bound in (6.5) are plotted as $\circ$; those in (6.6) as $\triangle$; those of the Stein's method bound in (6.2) as $\bigtriangledown$; Čekanavičius and Roos's bounds in (6.4) as $\times$ and Roos' bounds in (6.8) as $\boxtimes$. The true total variation distances, computed numerically in each case, are plotted as $\diamond$.

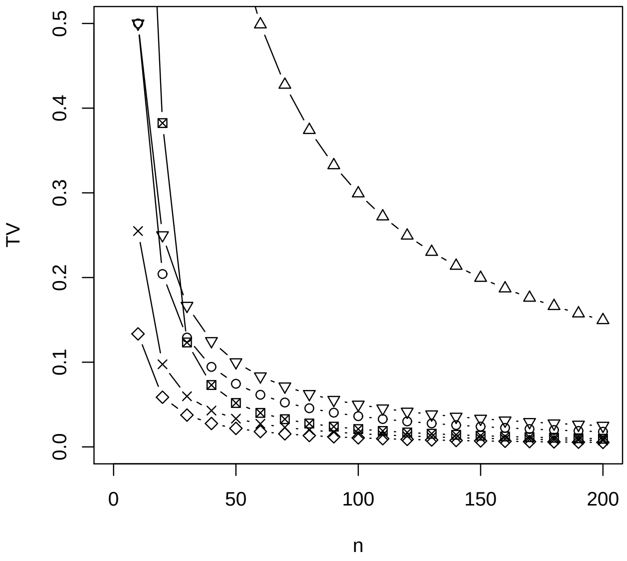

Figure 2: Bounds on the total variation distance $d_{\mathrm{TV}}(C_Q\mathrm{Bin}(p,Q),\mathrm{CPo}(\lambda,Q))$ for $Q\sim\mathrm{Geom}(\alpha)$ as in Figure 1, here plotted against the parameter n, with $\alpha=0.2$ and $\lambda=5$ fixed.

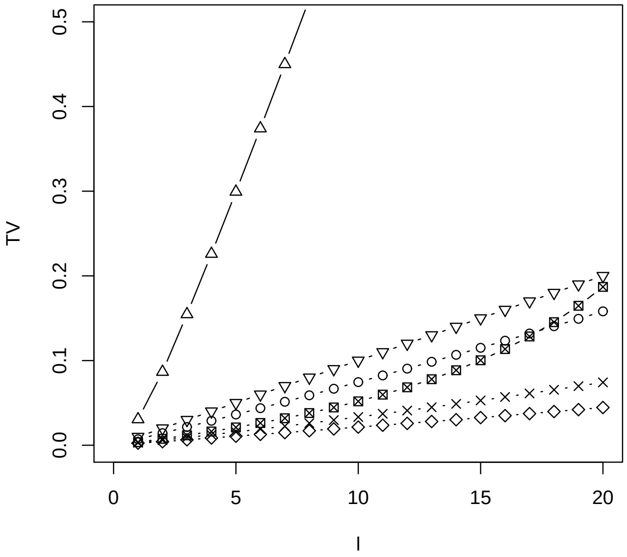

Figure 3: Bounds on the total variation distance $d_{\mathrm{TV}}(C_Q\mathrm{Bin}(p,Q),\mathrm{CPo}(\lambda,Q))$ for $Q\sim\mathrm{Geom}(\alpha)$ as in Figure 1, here plotted against the parameter λ, with $\alpha=0.2$ and $n=100$ fixed.

In this case, it is clear that the best bounds by a considerable margin are those of Roos [33] given in (6.8).

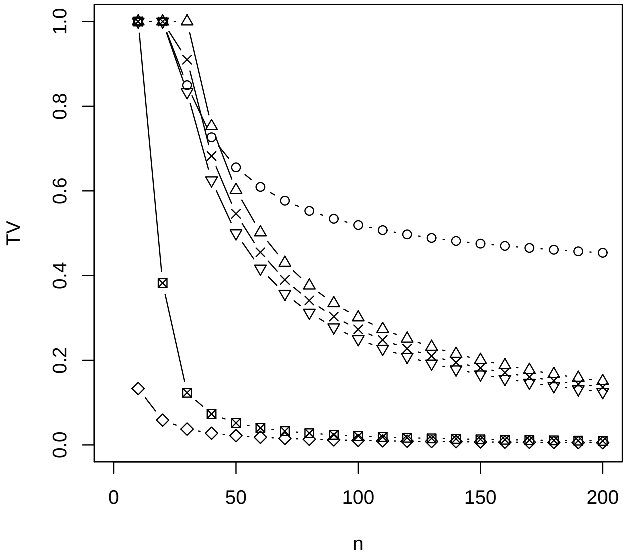

Figure 4: Bounds on the total variation distance $d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q))$ for $Q_i \sim \mathrm{Geom}(\alpha_i)$, where α_i are uniformly spread between 0.15 and 0.25, n varies, and p is as in regime I, $p = 5/n$. Again, bounds based on $J_{\mathbf{Q},1}$ are plotted as $\circ$; those based on $J_{Q,2}$ as $\triangle$; Le Cam's bound in (6.1) as $\bigtriangledown$; the Stein's method bound in (6.7) as $\times$, and Roos' bound from Theorem 2 of [33] as $\boxtimes$. The true total variation distances, computed numerically in each case, are plotted as $\diamond$.

References

[1] D. Aldous *Probability approximations via the Poisson clumping heuristic*. Springer-Verlag, New York, 1989. MR0969362

[2] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. On the rate of convergence in the entropic central limit theorem. *Probab. Theory Related Fields*, 129(3):381–390, 2004. MR2128238

[3] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.*, 17(4):975–982 (electronic), 2004. MR2083473

[4] A. D. Barbour and L. H. Y. Chen. *Stein's method and applications.* Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore, **5**, Published jointly by Singapore University Press, Singapore, 2005. MR2201882

[5] A. D. Barbour, L. H. Y. Chen, and W.-L. Loh. Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.*, 20(4):1843–1866, 1992. MR1188044

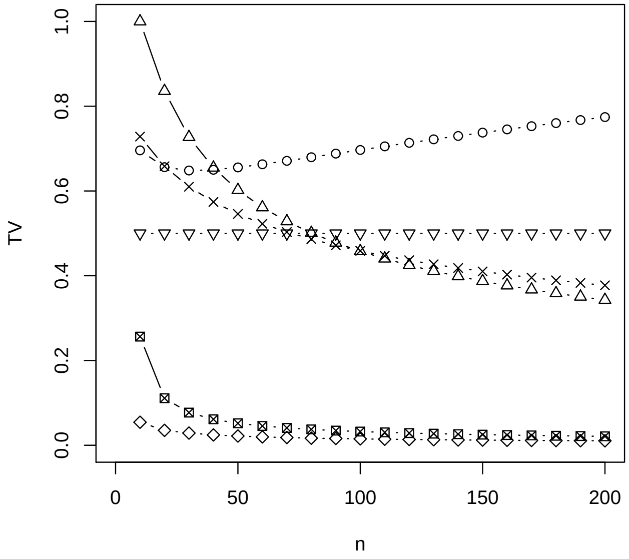

Figure 5: Bounds on the total variation distance $d_{\mathrm{TV}}(P_{S_n}, \mathrm{CPo}(\lambda, Q))$ for $Q_i \sim \mathrm{Geom}(\alpha_i)$ as in Figure 4, where α_i are uniformly spread between 0.15 and 0.25, n varies, and p is as in Regime II, $p = \sqrt{0.5/n}$.

[6] A. D. Barbour and O. Chryssaphinou. Compound Poisson approximation: a user's guide. *Ann. Appl. Probab.*, 11(3):964–1002, 2001. MR1865030

[7] A. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. The Clarendon Press Oxford University Press, New York, 1992. MR1163825

[8] A. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14:336–342, 1986. MR0815975

[9] S. Bobkov and M. Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, 156(2):347–365, 1998. MR1636948

[10] I.S. Borisov, and I.S. Vorozheĭkin. Accuracy of approximation in the Poisson theorem in terms of χ^2 distance. *Sibirsk. Mat. Zh.*, 49(1):8–22, 2008. MR2400567

[11] V. Čekanavičius and B. Roos. An expansion in the exponent for compound binomial approximations. *Liet. Mat. Rink.*, 46(1):67–110, 2006. MR2251442

[12] T. Cover and J. Thomas. *Elements of Information Theory*. J. Wiley, New York, 1991. MR1122806

[13] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981. MR0666545

[14] P. Diaconis and S. Holmes. *Stein's method: expository lectures and applications*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 46. Beachwood, OH, 2004. MR2118599

[15] T. Erhardsson. Stein's method for Poisson and compound Poisson approximation. In A. D. Barbour and L. H. Y. Chen, editors, *An Introduction to Stein's Method*, volume 4 of *IMS Lecture Note Series*, pages 59–111. Singapore University Press, 2005. MR2235449

[16] P. Harremoës. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Trans. Inform. Theory*, 47(5):2039–2041, 2001. MR1842536

[17] O. Johnson. *Information theory and the central limit theorem*. Imperial College Press, London, 2004. MR2109042

[18] O. Johnson. Log-concavity and the maximum entropy property of the Poisson distribution. *Stochastic Processes and Their Applications*, 117(6):791–802, 2007. MR2327839

[19] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields*, 129(3):391–409, 2004. MR2128239

[20] O. Johnson, I. Kontoyiannis, and M. Madiman, Log-concavity, ultra-log-concavity and a maximum entropy property of discrete compound Poisson measures. *Preprint*, October 2009. Earlier version online at `arXiv:0805.4112v1`, May 2008.

[21] I. Johnstone and B. MacGibbon. Une mesure d'information caractérisant la loi de Poisson. In *Séminaire de Probabilités, XXI*, pages 563–573. Springer, Berlin, 1987. MR0942005

[22] A. Kagan. A discrete version of the Stam inequality and a characterization of the Poisson distribution. *J. Statist. Plann. Inference*, 92(1-2):7–12, 2001. MR1809692

[23] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. Inform. Theory*, 51(2):466–472, February 2005. MR2236061

[24] I. Kontoyiannis and M. Madiman. Measure concentration for Compound Poisson distributions. *Elect. Comm. Probab.*, 11:45–57, 2006. MR2219345

[25] L. Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.*, 10:1181–1197, 1960. MR0142174

[26] L. Le Cam. On the distribution of sums of independent random variables. In *Proc. Internat. Res. Sem., Statist. Lab., Univ. California, Berkeley, Calif.*, pages 179–202. Springer-Verlag, New York, 1965. MR0199871

[27] M. Madiman. *Topics in Information Theory, Probability and Statistics*. PhD thesis, Brown University, Providence RI, August 2005. MR2624419

[28] M. Madiman and A. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Trans. Inform. Theory*, 53(7), 2317–2329, July 2007. MR2319376

[29] T. Matsunawa. Some strong ε-equivalence of random variables. *Ann. Inst. Statist. Math.*, 34(2):209–224, 1982. MR0666413

[30] R. Michel. An improved error bound for the compound Poisson approximation of a nearly homogeneous portfolio. *ASTIN Bull.*, 17:165–169, 1987.

[31] M. Romanowska. A note on the upper bound for the distrance in total variation between the binomial and the Poisson distribution. *Statistica Neerlandica*, 31(3):127–130, 1977. MR0467889

[32] B. Roos. Sharp constants in the Poisson approximation. *Statist. Probab. Lett.*, 52:155–168, 2001. MR1841404

[33] B. Roos. Kerstan's method for compound Poisson approximation. *Ann. Probab.*, 31(4):1754–1771, 2003. MR2016599

[34] F. Topsøe. Maximum entropy versus minimum risk and applications to some classical discrete distributions. *IEEE Trans. Inform. Theory*, 48(8):2368–2376, 2002. MR1930296

[35] A. M. Tulino and S. Verdú. Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof. *IEEE Trans. Inform. Theory*, 52(9):4295–4297, September 2006. MR2298559

[36] W. Vervaat. Upper bounds for the distance in total variation between the binomial or negative binomial and the Poisson distribution. *Statistica Neerlandica*, 23:79–86, 1969. MR0242235

[37] L. Wu. A new modified logarithmic Sobolev inequality for Poisson point processes and several applications. *Probab. Theory Related Fields*, 118(3):427-438, 2000. MR1800540

[38] Y. Yu. On the entropy of compound distributions on nonnegative integers. *IEEE Trans. Inform. Theory*, 55:3645–3650, August 2009. MR2598065