Fisher Information, Compound Poisson Approximation, and the Poisson Channel

Mokshay Madiman
Department of Statistics
Yale University
New Haven CT, USA
Email: mokshay.madiman@yale.edu

Oliver Johnson
Department of Mathematics
University of Bristol
Bristol, BS8 1TW, UK
Email: O.Johnson@bristol.ac.uk

Ioannis Kontoyiannis
Department of Informatics
Athens University of Economics & Business
Athens 10434, Greece
Email: yiannis@aueb.gr

***Abstract*—Fisher information plays a fundamental role in the analysis of Gaussian noise channels and in the study of Gaussian approximations in probability and statistics. For discrete random variables, the *scaled Fisher information* plays an analogous role in the context of Poisson approximation. Our first results show that it also admits a minimum mean squared error characterization with respect to the Poisson channel, and that it satisfies a monotonicity property that parallels the monotonicity recently established for the central limit theorem in terms of Fisher information. We next turn to the more general case of compound Poisson distributions on the nonnegative integers, and we introduce two new "local information quantities" to play the role of Fisher information in this context. We show that they satisfy subadditivity properties similar to those of classical Fisher information, we derive a minimum mean squared error characterization, and we explore their utility for obtaining compound Poisson approximation bounds.**

I. INTRODUCTION

The study of the distribution P_{S_n} of a finite sum $S_n = \sum_{i=1}^n Y_i$ of random variables $\{Y_i\}$ forms a central part of classical probability theory, and naturally arises in many important applications. For example, if the $\{Y_i\}$ are independent and identically distributed (i.i.d.) with zero mean and variance σ^2, then the central limit theorem (CLT) states that $S_n/(\sigma\sqrt{n})$ converges in distribution to $N(0,1)$, the standard normal, as $n \to \infty$. Moreover, finer Gaussian approximation results give conditions under which $P_{S_n} \approx N(0, n\sigma^2)$.

In 1986, Barron [5] strengthened the classical CLT by showing that, under general assumptions, the distribution of $S_n/(\sigma\sqrt{n})$ converges to $N(0,1)$ in relative entropy. The proof is based on estimates of the Fisher information of $S_n/(\sigma\sqrt{n})$, which acts as a "local" version of the relative entropy. Virtually every approach to this "information-theoretic CLT" to date relies on the more tractable notion of Fisher information as an intermediary; see, e.g., [13], [2], [11].

In the case where the summands $\{Y_i\}$ in $S_n = \sum_{i=1}^n Y_i$ are discrete, the CLT approximation is often not appropriate. E.g., if each Y_i takes values in the set $\mathbb{Z}_+ = \{0, 1, \ldots\}$ of nonnegative integers, P_{S_n} can often be well-approximated by a Poisson distribution. In the simplest example, suppose the $\{Y_i\}$ are i.i.d. Bernoulli($\frac{\lambda}{n}$) random variables; then, for large n, the distribution P_{S_n} approaches Po(λ), the Poisson distribution with parameter λ.

An information-theoretic view of Poisson approximation was recently developed in [17]. Again, the gist of the approach was the use of a discrete version of Fisher information, the *scaled Fisher information* defined in the following section. It was shown there that it plays a role in many ways analogous to the classical continuous Fisher information, and it was demonstrated that it can be used very effectively in providing strong, nonasymptotic Poisson approximation bounds.

In this work we consider the more general problem of *compound Poisson approximation* from an information-theoretic point of view. Let S_n as before denote the sum of random variables $\{Y_i\}$ taking values in $\mathbb{Z}_+$. We find it convenient to write each Y_i as the product $B_i U_i$ of two independent random variables, where B_i is Bernoulli(p_i) and U_i takes values in $\mathbb{N} = \{1, 2, \ldots\}$. This can be done uniquely and without loss of generality, by taking $p_i = \Pr(Y_i \neq 0)$ and U_i having distribution $Q_i(k) = \Pr(Y_i = k)/p_i$ for $k \geq 1$, so that Q_i is simply the conditional distribution of Y_i given that $\{Y_i \geq 1\}$.

The simplest example, which is, in a sense, the very definition of the compound Poisson distribution, is when the $\{Y_i\}$ are i.i.d., with each $Y_i = B_i U_i$ being the product of a Bernoulli(λ/n) random variable B_i and U_i with distribution Q on $\mathbb{N}$, for an arbitrary such Q. Then,

$$S_n = \sum_{i=1}^n B_i U_i \overset{(d)}{=} \sum_{i=1}^{S'_n} U_i, \tag{1}$$

where $S'_n = \sum_{i=1}^n B_i$ has a Binomial($n, \frac{\lambda}{n}$) distribution, and $\overset{(d)}{=}$ denotes equality in distribution. [Throughout, we take the empty sum $\sum_{i=1}^0 [\ldots]$ to be equal to zero.] Since the distribution of S'_n converges to Po(λ) as $n \to \infty$, it is easily seen that P_{S_n} will converge to the distribution of,

$$\sum_{i=1}^Z U_i, \tag{2}$$

where $Z \sim$ Po(λ) is independent of the $\{U_i\}$. This expression is precisely the definition of the *compound Poisson distribution with parameters λ and Q*, denoted by $CP(\lambda, Q)$.

Even if the summands $\{Y_i\}$ are not i.i.d., it is often the case that the distribution P_{S_n} of S_n can be accurately approximated by a compound Poisson distribution. Intuitively, the minimal requirements for such an approximation to hold are that: (i) None of the $\{Y_i\}$ dominate the sum, i.e., the parameters $p_i = \Pr\{Y_i \neq 0\}$ are all appropriately small; and (ii) The $\{Y_i\}$ are only weakly dependent. See [1], [4] and the references therein for general discussions of compound Poisson approximation and its many applications.

In this paper, we focus on the case where the summands are independent, but do not restrict their distributions. An example of the type of result that we prove is the following bound. A proof outline is given in Section III.

Theorem I: [COMPOUND POISSON APPROXIMATION] Consider $S_n = \sum_{i=1}^n B_i U_i$, where the U_i are i.i.d. $\sim Q$ and the B_i are independent Bernoulli(p_i). Then, writing $\lambda = \sum_{i=1}^n p_i$, the relative entropy between the distribution P_{S_n} of S_n and the $CP(\lambda, Q)$ distribution satisfies,

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

Recall that the standardized Fisher information of a random variable X with differentiable density f is,

$$J_N(X) = E\left[\left(\frac{f'(X)}{f(X)} - \frac{g'(X)}{g(X)}\right)^2\right], \tag{3}$$

where g is the density of a normal with the same variance as X. The quantity $J_N(X)$ satisfies the following properties:

(A) $J_N(X)$ is the variance of a zero-mean quantity, namely the (standardized) score function of X.
(B) $J_N(X) = 0$ if and only if $D(f\|g) = 0$, i.e., if and only if X is normal.
(C) $J_N(S_n)$ satisfies a subadditivity property.
(D) If $J_N(X)$ is small, then $D(f\|g)$ is also appropriately small.

In the information-theoretic approach, Gaussian approximations are established by first using property (C) to show that $J_N(S_n/\sqrt{n}) \approx 0$ for large n, and then using (D) to obtain bounds in relative entropy. Note that (D) is a quantitative refinement of (B). For Poisson approximation, the "scaled Fisher information" of [17] plays roughly the same role; in particular, it satisfies properties (A-D).

Similarly, in the more general problem of compound Poisson approximation considered presently, we introduce two "local information" quantities that play corresponding roles in this context. The main difference in their utility is that the analog of property (D) we obtain is in terms of total variation rather than relative entropy. Note that we do not refer to these new local information quantities as "Fisher informations," because, unlike Fisher's information [7], we are not aware of a natural way in which they connect to the efficiency of optimal estimators in parametric inference.

In Section II we briefly review the information-theoretic approach to Poisson approximation, and we give a new interpretation of the scaled Fisher information of [17] involving minimum mean squared error (MMSE) estimation for the Poisson channel. We also prove a monotonicity property for the convergence of the distribution of i.i.d. summands to the Poisson, which is analogous to the recently proved monotonicity of Fisher information in the CLT [3], [20], [25]. Section III contains some of our main approximation bounds, and also generalizations of the MMSE interpretation and the monotonicity property of our local information quantities.

II. POISSON APPROXIMATION

The classical Binomial-to-Poisson convergence result has an information-theoretic interpretation. First, like the normal, the Poisson distribution has a maximum entropy property; for example, in [12] it is shown that it has the highest entropy among all ultra log-concave distributions on $\mathbb{Z}_+$ with mean λ; see also [10], [24]. Second, an information-theoretic approach to Poisson approximation bounds was developed in [17]. This was partly based on the introduction of the following local information quantity:

Definition: Given a $\mathbb{Z}_+$-valued random variable Y with distribution P_Y and mean λ, the score function ρ_Y of Y is,

$$\rho_Y(y) = \frac{(y+1)P_Y(y+1)}{\lambda P_Y(y)} - 1, \tag{4}$$

and *the scaled Fisher information of* Y is defined by,

$$J_\pi(Y) = \lambda E[\rho(Y)]^2, \tag{5}$$

where the random variable $\rho(Y) := \rho_Y(Y)$ is the *score* of Y.

For sums of independent $\mathbb{Z}$-valued random variables, this local information quantity was used in [17] to establish near-optimal Poisson approximation bounds in relative entropy and total variation distance. Previous analogues of Fisher information for discrete random variables [15], [22], [16] suffered from the drawback that they are infinite for random variables with finite support, a problem that is overcome by this $J_\pi(Y)$. Furthermore, $J_\pi(Y)$ satisfies properties (A-D) stated above, as discussed in detail in [17].

We now give an alternative characterization of the scaled Fisher information, related to MMSE estimation for the Poisson channel. This extends to the case of the Poisson channel a similar characterization for the Fisher information J_N developed in the recent work of Guo, Shamai and Verdú [9] for signals in Gaussian noise, and is related to their work on the Poisson channel [8] (which, however, has a somewhat different focus than ours). [See also the earlier work of L.D. Brown in the context of statistical decision theory, discussed in [19], as well as the relevant remarks in [21].]

Theorem II: [MMSE AND SCALED FISHER INFORMATION] Let $X \geq 0$ be a continuous random variable whose value is to be estimated based on the observation Y, where the conditional distribution of Y given X is $\text{Po}(X)$. Then the scaled Fisher information of Y can be expressed as the variance-to-mean ratio of the MMSE estimate of X given Y:

$$J_\pi(Y) = \frac{\text{Var}\{E[X|Y]\}}{E(X)}. \tag{6}$$

Proof: If X has density f supported on $[0,\infty)$, then the distribution P of Y is given by

$$P(y)=\int_0^\infty P(y|x)f(x)dx=\int_0^\infty \frac{e^{-x}x^y f(x)}{y!}dx, \qquad (7)$$

where $P(y|x)\sim$Po(x). This implies that

$$(y+1)P(y+1)=\frac{1}{y!}\int_0^\infty e^{-x}x^{y+1}f(x)dx, \qquad (8)$$

and thus

$$\begin{aligned}\frac{(y+1)P(y+1)}{P(y)} &= \frac{\int_0^\infty e^{-x}x^{y+1}f(x)dx}{\int_0^\infty e^{-x}x^{y}f(x)dx}\\ &= \int_0^\infty x g_y(x)\\ &= E[X|Y=y],\end{aligned} \qquad (9)$$

where $g_y(x)$ is the conditional density of X given Y. Thus,

$$\rho_Y(y)=\frac{E[X|Y=y]}{E(Y)}-1,$$

and substituting this into the definition of J_π proves the desired result, upon noting that $E(X)=E(Y)$. ■

The following convolution identity for the score function of a sum $S_n=X_1+\ldots+X_n$ of independent $\mathbb{Z}_+$-valued random variables was established in [17],

$$\rho_{S_n}(z)=E\Big(\sum_{i=1}^n \frac{\lambda_i}{\lambda}\rho(X_i)\Big|\, S_n=z\Big), \qquad (10)$$

where $E(X_i)=\lambda_i$ and $E(S_n)=\sum_{i=1}^n\lambda_i=\lambda$. As a result, $J_\pi(S_n)$ has a subadditivity property, implying in particular that, when the summands are i.i.d., $J_\pi(S_{2n})\le J_\pi(S_n)$. Theorem III below shows that the sequence $\{J_\pi(S_n)\}$ is in fact monotonic in n. This is analogous to the monotonic decrease of the Fisher information in the CLT [3], [20], [25].

Theorem III: [Monotonicity of scaled Fisher information] Let S_n be the sum of n independent random variables $X_1,X_2,\ldots,X_n$. Write $S_n^{(i)}=\sum_{j\neq i}X_j$ for the leave-one-out sums, and let $\lambda^{(i)}$ denote the mean of $S_n^{(i)}$, for each $i=1,2,\ldots,n$. Then,

$$J_\pi(S_n)\le\frac{1}{n-1}\sum_{i=1}^n\frac{\lambda^{(i)}}{\lambda}J_\pi(S_n^{(i)}), \qquad (11)$$

where λ is the mean of S_n. In particular, when the summands are i.i.d., we have $J_\pi(S_n)\le J_\pi(S_{n-1})$.

Proof: The proof we give here adapts the corresponding technique used in [20]; an alternative proof can be given by combining the characterization of Theorem II with the technique of [25]. In either case, the key idea is Hoeffding's variance drop inequality (see [20] for historical remarks),

$$E\left(\sum_{S\in\mathcal{S}}\psi^{(S)}(X_S)\right)^2\le(n-1)\sum_S E\psi^{(S)}(X_S)^2, \qquad (12)$$

where $\mathcal{S}$ is the collection of subsets of $\{1,\ldots,n\}$ of size $n-1$, $\{\psi^{(S)}\,;\,S\in\mathcal{S}\}$ is an arbitrary collection of square-integrable functions, and $X_S=\sum_{i\in S}X_i$ for any $S\in\mathcal{S}$.

In the present setting, for each $i=1,2,\ldots,n$, write P_i and R_i for the distributions of X_i and $S_n^{(i)}$, respectively, and let F denote the distribution of S_n. Then F can be decomposed as $F(z)=\sum_x P_i(x)R_i(z-x)$, for each $i=1,2,\ldots,n$. Multiplying this with the expression,

$$(n-1)z=\sum_{i=1}^n E(z-Y_i\,|\,Y_1+\ldots+Y_n=z),$$

gives,

$$(n-1)zF(z)=\sum_{i=1}^n\sum_{y_i}P_i(y_i)R_i(z-y_i)(z-y_i). \qquad (13)$$

We can substitute this in (4) to obtain,

$$\begin{aligned}\rho_{S_n}(z) &= \frac{(z+1)F(z+1)}{\lambda F(z)}-1\\ &= \sum_{i=1}^n\sum_{y_i}\frac{P_i(y_i)R_i(z+1-y_i)(z+1-y_i)}{\lambda(n-1)F(z)}-1\\ &= \frac{1}{n-1}\sum_{i=1}^n\sum_{y_i}\frac{P_i(y_i)R_i(z-y_i)}{F(z)}\frac{\lambda^{(i)}}{\lambda}\times\\ &\qquad\times\Big(\frac{(z+1-y_i)R_i(z+1-y_i)}{\lambda^{(i)}R_i(z-y_i)}-1\Big)\\ &= E\Big(\sum_{i=1}^n\frac{\lambda^{(i)}}{\lambda(n-1)}\rho(S_n^{(i)})\Big|\,S_n=z\Big).\end{aligned}$$

Using the conditional Jensen inequality, this implies that $J_\pi(S_n)$ can be bounded as,

$$\begin{aligned}\lambda E\rho(S_n)^2 &\le \lambda E\left(\sum_{i=1}^n\frac{\lambda^{(i)}}{\lambda(n-1)}\rho(S_n^{(i)})\right)^2\\ &\le \lambda(n-1)\sum_{i=1}^n\left(\frac{\lambda^{(i)}}{\lambda(n-1)}\right)^2E\rho(S_n^{(i)})^2\\ &= \frac{1}{n-1}\sum_{i=1}^n\frac{\lambda^{(i)}}{\lambda}J_\pi(S_n^{(i)}),\end{aligned}$$

as claimed. ■

Another way in which scaled Fisher information naturally arises is in connection with a modified logarithmic Sobolev inequality for the Poisson distribution [6]; for an arbitrary distribution P on $\mathbb{Z}_+$ with mean λ and $X\sim P$,

$$D(P\|\text{Po}(\lambda))\le J_\pi(X). \qquad (14)$$

This was combined in [17] with the subadditivity of scaled Fisher information (mentioned above) to obtain the following Poisson approximation bound: If S_n is the sum of n independent Bernoulli(p_i) random variables $\{B_i\}$, then,

$$D(P_{S_n}\|\text{Po}(\lambda))\le\frac{1}{\lambda}\sum_{i=1}^n\frac{p_i^3}{1-p_i}, \qquad (15)$$

where $\lambda = \sum_{i=1}^n p_i$. Theorem I stated in the Introduction generalizes (15) to the compound Poisson case. Combining (15) with Pinsker's inequality, gives a total variation approximation bound, which is near optimal in the regime where $\lambda = O(1)$ and n is large; see [23].

III. Compound Poisson Approximation and Local Informations

In this section we develop an information-theoretic setting within which compound Poisson approximation results can be obtained, generalizing the Poisson approximation results described in the previous section. All of the results below are stated without proof; details will be given in an extended version of the present paper.

Although maximum entropy properties are not the main focus of this work, we note that the compound Poisson can also be seen as a maximum entropy distribution, under certain conditions. [Details will be given in forthcoming work.] Another important characterization of the compound Poisson law is via size-biasing: For any distribution P on $\mathbb{Z}_+$ with mean λ, *the size-biased distribution* $P^\#$ is defined by,

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

[Some authors define $P^\#$ as the above distribution shifted by 1.] If X has distribution P, then we write $X^\#$ for a random variable with distribution $P^\#$. Notice that the score function introduced previously is simply $P^\#(y)/P(y) - 1$.

We also need to define the following *compounding operation*: If X is a $\mathbb{Z}_+$-valued random variable with distribution P, and Q is an arbitrary distribution on $\mathbb{N}$, then *the Q-compound random variable $C_Q X$ with distribution $C_Q P$* is,

$$C_Q X \stackrel{(d)}{=} \sum_{i=1}^{X} U_i,$$

where $\stackrel{(d)}{=}$ denotes equality in distribution as before, and the random variables U_i, $i = 1, 2, \ldots$ are i.i.d. with common distribution Q. Note that $C_Q X \sim CP(\lambda, Q)$ if and only if $X \sim \text{Po}(\lambda)$; therefore, $C_Q X \sim CP(\lambda, Q)$ if and only if $P = P^\#$.

These ideas lead to the following definition of a new local information quantity. Note that it is only defined for Q-compound random variables.

Definition: For a $\mathbb{Z}_+$-valued random variable X with distribution $C_Q P$ and mean λ_X, the *local information $J_{Q,1}(X)$ of X relative to the compound Poisson distribution $CP(\lambda, Q)$* is,

$$J_{Q,1}(X) = \lambda_X E[r_1^2(X)], \tag{16}$$

where the score function r_1 of X is defined by,

$$r_1(x) = \frac{C_Q(P^\#)(x)}{C_Q P(x)} - 1. \tag{17}$$

This definition is motivated by the fact that $P = P^\#$ if and only if P is Poisson, so that $J_{Q,1}(X)$ is identically zero if and only if $X \sim CP(\lambda, Q)$. Note that if $Q = \delta_1$, the compounding operation does nothing, and $J_{Q,1}$ reduces to J_π.

The following property is easily proved using characteristic functions:

Lemma I: $Z \sim CP(\lambda, Q)$ if and only if $Z^\# \stackrel{(d)}{=} Z + U^\#$, where $U \sim Q$ is independent of Z. That is, $C_Q P = CP(\lambda, Q)$ if and only if $(C_Q P)^\# = (C_Q P) \star Q^\#$, where $\star$ is the convolution operation.

We now define another local information quantity in the compound Poisson context.

Definition: For a $\mathbb{Z}_+$-valued random variable X with distribution R and mean λ_X, the *local information $J_{Q,2}(X)$ of X relative to the compound Poisson distribution $CP(\lambda, Q)$* is,

$$J_{Q,2}(X) = \lambda_X E[r_2^2(X)], \tag{18}$$

where the score function r_2 of X is defined by,

$$r_2(x) = \frac{xR(x)}{\lambda_X \sum_u uQ(u)R(x-u)} - 1. \tag{19}$$

Note that again $J_{Q,2}$ reduces to J_π when $Q = \delta_1$. In the simple Poisson case, as we saw, the quantity J_π has a MMSE interpretation, and it satisfies certain subadditivity and monotonicity properties. In the compound case, each of these properties is satisfied by one of $J_{Q,1}$ or $J_{Q,2}$.

The following result shows that the local information $J_{Q,2}$ can be interpreted in terms of MMSE estimation for an appropriate channel.

Theorem IV: [MMSE and $J_{Q,2}$] Let $X \geq 0$ be a continuous random variable whose value is to be estimated based on the observation $Y + V$, suppose that the conditional distribution of Y given X is CP(X, Q), and that $V \sim Q^\#$ is independent of Y. Then,

$$J_{Q,2}(Y) = \frac{\text{Var}\{E[X|Y+V]\}}{E(X)}.$$

The local information quantity $J_{Q,1}$ satisfies a subadditivity relation:

Theorem V: [Subadditivity of $J_{Q,1}$] Suppose the independent random variables $Y_1, Y_2, \ldots, Y_n$ are Q-compound, with each Y_i having mean λ_i, $i = 1, 2, \ldots, n$. Then,

$$J_{Q,1}(Y_1 + Y_2 + \ldots + Y_n) \leq \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} J_{Q,1}(Y_i), \tag{20}$$

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

A corresponding result can be proved for $J_{Q,2}$, but the right-hand side includes additional cross-terms.

In the case of i.i.d. summands, we deduce from Theorem V that $J_{Q,1}(S_n)$ is monotone on doubling of sample size n. As in the normal and Poisson cases, it turns out that $J_{Q,1}(S_n)$ is decreasing in n at every step. The statement and proof of Theorem III easily carry over to this case:

Theorem VI: [Monotonicity of $J_{Q,1}$] Let S_n denote the sum of n independent, Q-compound, random variables

$X_1, X_2, \ldots, X_n$. Write $S_n^{(i)} = \sum_{j \neq i} X_j$ the leave-one-out sums, and let $\lambda^{(i)}$ denote the mean of $S_n^{(i)}$, for each $i = 1, 2, \ldots, n$. Then,

$$J_{Q,1}(S_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} \frac{\lambda^{(i)}}{\lambda} J_{Q,1}(S_n^{(i)}), \tag{21}$$

where λ is the mean of S_n. In particular, when the summands are i.i.d., we have $J_{Q,1}(S_n) \leq J_{Q,1}(S_{n-1})$.

In the special case of Poisson approximation, the logarithmic Sobolev inequality (14) proved in [6] directly relates the relative entropy to the local information quantity J_π. Consequently, the Poisson approximation bounds developed in [17] are proved by combining this result with the subadditivity property of J_π. However, the known logarithmic Sobolev inequalities for compound Poisson distributions [26], [18], only relate the relative entropy to quantities different from $J_{Q,1}$ and $J_{Q,2}$. Instead of developing subadditivity results for those quantities, we build on ideas from Stein's method for compound Poisson approximation and prove the following relationship between the total variation distance and the local informations $J_{Q,1}$ and $J_{Q,2}$.

Theorem VII: [STEIN'S METHOD-LIKE BOUNDS] Let X be a $\mathbb{Z}_+$-valued random variable with distribution P, and let Q be an arbitrary distribution on $\mathbb{N}$ with finite mean q. Then,

$$\|P - CP(\lambda, Q)\|_{\text{TV}} \leq qH(\lambda, Q)\sqrt{\lambda J_{Q,i}(X)}, \tag{22}$$

for each $i = 1, 2$, where $\lambda = E(X)/q$, and $H(\lambda, Q)$ is an explicit constant depending only on λ and Q.

The quantities $H(\lambda, Q)$ arise from the so-called 'magic factors' which appear in Stein's method, and they can be bounded in an easily applicable way. Combining Theorems V and VII leads to very effective approximation bounds in total variation distance; these will be presented in detail in [14].

Finally, we give a short proof outline for the compound Poisson approximation result stated in the Introduction.

Proof of Theorem I: Let $Z' \sim \text{Po}(\lambda)$, where λ is the sum of the p_i, and $S_n' = \sum_{i=1}^n B_i$. Then S_n can also be expressed $S_n = \sum_{i=1}^{S_n'} U_i$, while we can construct a $CP(\lambda, Q)$ random variable Z as $\sum_{i=1}^{Z'} U_i$. Thus $S_n = f(U_1, \ldots, U_n, S_n')$ and $Z = f(U_1, \ldots, U_n, Z')$, where the function f is the same in both places. By the data processing inequality and chain rule,

$$D(P_{S_n} \| CP(\lambda, Q)) \leq D(P_{S_n'} \| \text{Po}(\lambda)),$$

and the result follows from the Poisson approximation bound (15) of [17]. ∎

This data processing argument does not directly extend to the case where the Q_i associated with different summands Y_i are not identical. However, versions of Theorems I, V and VII can be obtained in this case, although there statements are somewhat more complex. These extensions, together with their consequences for compound Poisson approximation bounds, will be given in the forthcoming, longer version [14] of the present work.

REFERENCES

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

[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*, vol. 129, no. 3, pp. 381–390, 2004.

[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.*, vol. 17, no. 4, pp. 975–982 (electronic), 2004.

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

[5] A. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.

[6] S. Bobkov and M. Ledoux, "On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures," *J. Funct. Anal.*, vol. 156, no. 2, pp. 347–365, 1998.

[7] R. A. Fisher, "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, vol. 22, pp. 700–725, 1925.

[8] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *Proc. IEEE Inf. Th. Workshop, San Antonio*, 2004.

[9] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261–1282, April 2005.

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

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

[12] O. Johnson, "Log-concavity and the maximum entropy property of the Poisson distribution," *To appear in Stochastic Processes and their Applications*, 2007. DOI: 10.1016/j.spa.2006.10.006

[13] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 391–409, 2004.

[14] O. Johnson, I. Kontoyiannis, and M. Madiman, "Compound Poisson approximation via local information quantities," *In preparation*, 2007.

[15] I. Johnstone and B. MacGibbon, "Une mesure d'information caractérisant la loi de Poisson," in *Séminaire de Probabilités, XXI*. Berlin: Springer, 1987, pp. 563–573.

[16] 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. 99, no. 1, p. 1, 2001.

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

[18] I. Kontoyiannis and M. Madiman, "Measure concentration for Compound Poisson distributions," *Elect. Comm. Probab.*, vol. 11, paper 5, pp. 45–57, May 2006.

[19] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 1998.

[20] M. Madiman and A. Barron, "The monotonicity of information in the central limit theorem and entropy power inequalities," *Proc. IEEE Intl. Symp. Inform. Th., Seattle*, July 2006.

[21] M. Madiman and A.R. Barron. Generalized entropy power inequalities and monotonicity properties of information. To appear in *IEEE Trans. Inform. Theory*, 2007.

[22] 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.

[23] B. Roos, "Asymptotic and sharp bounds in the Poisson approximation to the Poisson-binomial distribution," *Bernoulli*, vol. 5, no. 6, pp. 1021–1034, 1999.

[24] F. Topsøe, "Maximum entropy versus minimum risk and applications to some classical discrete distributions," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2368–2376, 2002.

[25] 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*, vol. 52, no. 9, pp. 4295–7, September 2006.

[26] L. Wu, "A new modified logarithmic Sobolev inequality for Poisson point processes and several applications," *Probab. Theory Relat. Fields*, vol. 118, pp. 427–438, 2000.