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. We show that it also admits a minimum mean squared error characterization, and we prove a monotonicity result analogous to the monotonicity recently established for the central limit theorem in terms of Fisher information. More generally, replacing the Poisson distribution by the richer class of compound Poisson distributions on the non-negative integers, we define two new "local information quantities," which, in many ways, play a role analogous to that of the Fisher information for a continuous random variable. We show that they satisfy subadditivity properties which parallel 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

Consider a sum $S_n = \sum_{i=1}^n Y_i$ of random variables Y_i taking values in the set $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ of non-negative integers. The study of the distribution P_{S_n} of S_n is an important part of classical probability theory, and it arises naturally in applications involving counting. In particular, it is often the case that P_{S_n} can be well-approximated by a compound Poisson distribution. In the simplest case, when the Y_i are i.i.d. Bernoulli($\frac{\lambda}{n}$) random variables, we know that for large n the distribution P_{S_n} approaches the Poisson(λ) distribution.

In the general case, we find it convenient to write each Y_i as the product $B_i U_i$ of two independent random variables, where B_i takes values in $\{0, 1\}$ and U_i takes values in $\mathbb{N} = \{1, 2, \ldots\}$. This can be done uniquely and without loss of generality, by taking B_i to be Bernoulli(p_i) with $p_i = \Pr(Y_i \neq 0)$, and U_i having distribution Q_i on $\mathbb{N}$, where $Q_i(k) = \Pr(Y_i = k)/p_i$ for $k \geq 1$. The distribution of U_i is clearly the conditional distribution of Y_i given that $\{Y_i \geq 1\}$.

If the $\{Y_i\}$ are i.i.d. then so are the $\{U_i\}$, and we write Q for their common distribution. Similarly, the $\{B_i\}$ are i.i.d. Bernoulli, and we denote their common parameter by $\frac{\lambda}{n}$ for some $\lambda > 0$. Then we can write,

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

where $S'_n = \sum_{i=1}^n B_i$ has a Binomial($n, \frac{\lambda}{n}$) distribution, and $\stackrel{(d)}{=}$ denotes equality in distribution. [Throughout, we take the empty sum $\sum_{i=1}^0 [\ldots]$ to be equal to zero.] The sum S_n in (1) is said to have a compound binomial distribution. Since the distribution of S'_n converges to Po(λ), the Poisson distribution with parameter λ, as $n \to \infty$, the distribution of S_n itself will converge to that of the compound sum,

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

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

More generally, 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. Compound Poisson approximation is very widely applicable, see, e.g., [1].

In this paper, we focus on the case where the summands are independent. Although most of the results do not require further restriction of the distributions, for clarity of exposition we only present the details in the case where all the Q_i are identical. An example of the type of result that we prove using information-theoretic ideas is the following bound. The ideas behind its proof are outlined 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,

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

In 1986, Barron [2] proved a relative entropy version of the central limit theorem (CLT). The proof was based on estimates of (the standardized) Fisher information, which acts

as a "local" version of the relative entropy. In fact, virtually every approach to the information-theoretic CLT relies in some way on the (more tractable) notion of Fisher information as an intermediary; see, e.g., [3], [4], [5], and the references therein. Recently, [6] used similar ideas to prove Poisson approximation bounds. Our work builds on the work of these and other authors, and is motivated by information-theoretic tools.

The key methodological idea in this work is to break up the problem into two smaller problems, using a "local information quantity." This is partly motivated by the normal approximation results mentioned above; there, the standardized Fisher information of a random variable X with differentiable density f is,

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

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

(A) J_N is the variance of a zero-mean quantity, namely the (standardized) score function.
(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 satisfies a subadditivity property for sums.
(D) If $J_N(X)$ is small, then $D(f\|g)$ is also appropriately small.

Roughly speaking, the information-theoretic version of the CLT can be proved by first using property (C) to show that $J_N(S_n/\sqrt{n}) \to 0$ as $n \to \infty$, and then using (D) to obtain convergence in relative entropy.

The "scaled Fisher information" J_P of [6] plays the same role for Poisson approximation that J_N plays for normal approximation. In particular, it satisfies properties (A-D). In this paper, we identify two quantities which can play similar roles for compound Poisson approximation; however we relax property (D) in that we only require that these new local information quantities should control total variation distance (and not necessarily relative entropy), and we relax property (C) to allow for modified forms of "subadditivity." We then use these properties to prove compound Poisson approximation bounds.

Note that we do not refer to these new local information quantities as "Fisher informations." This is because they do not naturally arise in the context of parametric inference like Fisher's information does [7], and we are not aware of any analogous interpretations in the present context.

In addition to obtaining approximation bounds, we also demonstrate some results of independent interest. In Section II, where we also review the information-theoretic approach to Poisson approximation, we give a new interpretation of the scaled Fisher information of [6] involving minimum mean square estimation for the Poisson channel. We also prove a monotonicity property for the convergence of the Binomial to the Poisson, which is analogous to the recently proved monotonicity of Fisher information in the CLT [8], [9], [10]. Section III contains our main approximation bounds, and also some results indicating that connections to minimum mean square estimation and monotonicity properties extend in an appropriate fashion to the compound Poisson case.

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 [11] it is shown that it has the highest entropy among all ultra log-concave distributions on $\mathbb{Z}_+$ with mean λ; see also [12], [13]. Second, an information-theoretic approach to Poisson approximation bounds was developed in [6]. 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 is defined by,

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

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

$$J_P(Y) = \lambda E[\rho_Y(Y)]^2. \tag{5}$$

For sums of independent $\mathbb{Z}$-valued random variables, this local information quantity was used in [6] to establish near-optimal Poisson approximation bounds in relative entropy and total variation distance. Previous analogues of Fisher information for discrete random variables [14], [15], [16] suffered from the drawback that they are infinite for random variables with finite support, a problem that is overcome by this $J_P(Y)$. Furthermore, $J_P(Y)$ satisfies properties (A-D) stated above, as discussed in detail in [6].

We now give an alternative characterization of the scaled Fisher information, related to minimum mean square 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 [19], [17] for signals in Gaussian noise. [See also the earlier work of L.D. Brown in the context of statistical decision theory, discussed in [18], as well as the relevant remarks in [9].]

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, and suppose that the conditional distribution of Y given X is Poisson(X). Then the scaled Fisher information $J_P(Y)$ of Y can be expressed as the variance-to-mean ratio of the minimum mean square estimate of X based on Y:

$$J_P(Y) = \frac{\mathrm{Var}\{E[X|Y]\}}{EX}. \tag{6}$$

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

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

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

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

and thus

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

where $g_y(x)$ is the density on $[0,\infty)$ corresponding to the conditional distribution of X given Y. Thus

$$\rho_Y(y) = \frac{E[X|Y=y]}{EY} - 1, \tag{10}$$

and substituting this into the definition of J_P proves the desired result, since $EX = EY$. ∎

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 [6],

$$\rho_{S_n}(z) = E\left[\sum_{i=1}^n \frac{\lambda_i}{\lambda}\rho_{X_i}(X_i)\middle| S_n = z\right], \tag{11}$$

where $E(X_i) = \lambda_i$ and $E(S_n) = \sum_{i=1}^n \lambda_i = \lambda$. As a result, $J_P(S_n)$ has a subadditivity property, implying in particular that, when the summands are i.i.d., then $J_P(S_{2n}) \leq J_P(S_n)$. Theorem III below shows that the sequence $\{J_P(S_n)\}$ is in fact monotonic in n. This fact is analogous to the monotonic decrease of the Fisher information for the normalized sums in the CLT; c.f. [8][9], [10].

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

$$J_P(S_n) \leq \frac{1}{n-1}\sum_{i=1}^n \frac{\lambda^{(i)}}{\lambda} J_P(U^{(i)}), \tag{12}$$

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

Proof: The proof we give here adapts the corresponding technique used in [9]; an alternative proof can be given by combining the characterization of Theorem II with the technique of [10]. In either case, the key idea is Hoeffding's variance drop inequality (see [9] for historical remarks),

$$E\left(\sum_{S\in\mathcal{S}}\psi^{(S)}(X_S)\right)^2 \leq (n-1)\sum_S E\psi^{(S)}(X_S)^2, \tag{13}$$

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 distribution of X_i and $U^{(i)}$, 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,

$$(n-1)zF(z) = \sum_{i=1}^n \sum_{y_i} P_i(y_i)R_i(z-y_i)(z-y_i). \tag{14}$$

We can substitute this in Equation (4) to obtain,

$$\begin{aligned}\rho_{S_n}(z) &= \frac{(z+1)F(z+1)}{\lambda F(z)} - 1 \\ &= \sum_{i=1}^n\sum_{y_i}\frac{P_i(y_i)R_i(z+1-y_i)(z+1-y_i)}{\lambda(n-1)F(z)} - 1 \\ &= \frac{1}{n-1}\sum_{i=1}^n\sum_{y_i}\frac{P_i(y_i)R_i(z-y_i)}{F(z)}\frac{\lambda^{(i)}}{\lambda}\times \\ &\quad \left(\frac{(z+1-y_i)R_i(z+1-y_i)}{\lambda^{(i)}R_i(z-y_i)} - 1\right) \\ &= E\left[\sum_{i=1}^n \frac{\lambda^{(i)}}{\lambda(n-1)}\rho_{U^{(i)}}(U^{(i)})\middle| S_n = z\right].\end{aligned}$$

Using the conditional Jensen inequality, this implies that $J_P(S_n)$ equals,

$$\begin{aligned}&\lambda E\rho_{S_n}(S_n)^2 \\ \leq\ &\lambda E\left(\sum_{i=1}^n \frac{\lambda^{(i)}}{\lambda(n-1)}\rho_{U^{(i)}}(U^{(i)})\right)^2 \\ \leq\ &\lambda(n-1)\sum_{i=1}^n\left(\frac{\lambda^{(i)}}{\lambda(n-1)}\right)^2 E\rho_{U^{(i)}}(U^{(i)})^2 \\ =\ &\frac{1}{n-1}\sum_{i=1}^n\frac{\lambda^{(i)}}{\lambda}J_P(U^{(i)}),\end{aligned}$$

as claimed. ∎

Another way in which scaled Fisher information naturally arises, is in connection with a modified logarithmic Sobolev-type inequality for the Poisson distribution proved in [20]. This states that, for an arbitrary distribution P on $\mathbb{Z}_+$ with mean λ,

$$D(P\|\text{Po}(\lambda)) \leq J_P(X). \tag{15}$$

This was combined in [6] 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 random variables $\{B_i\}$ with corresponding parameters $\{p_i\}$, then,

$$D(P_{S_n}\|\text{Po}(\lambda)) \leq \frac{1}{\lambda}\sum_{i=1}^n \frac{p_i^3}{1-p_i}, \tag{16}$$

where $\lambda = \sum_{i=1}^n p_i$. Our Theorem I stated in the Introduction generalizes this result to the compound Poisson case. Note

that by Pinsker's inequality, (16) gives a total variation approximation bound, which is near optimal in the regime where $\lambda = O(1)$ and n is large; see [21].

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 should mention that the compound Poisson can also be seen as a maximum entropy distribution, at least under certain conditions. [Details will be given in forthcoming work.] Another important characterization of the compound Poisson distribution is via size-biasing. Recall that, 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 size-biasing as the above $P^\#$ 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 a distribution on $\mathbb{N}$, then the random variable $C_Q X$ with distribution $C_Q P$ is defined by,

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

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. We refer to such a random variable as being Q-*compound*. 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 first definition of a new local information quantity. Note that it is only defined for Q-compound random variables.

Definition: Let X be a $\mathbb{Z}_+$-valued random variable with distribution $C_Q P$. Then, *the local information* $J_{Q,1}(X)$ *of* X *relative to the compound Poisson distribution* $CP(\lambda, Q)$, is defined by,

$$J_{Q,1}(X) = \lambda_X E[r_1^2(X)], \tag{17}$$

where λ_X is the mean of X, and the score function r_1 is,

$$r_1(x) = \frac{C_Q(P^\#)(x)}{C_Q P(x)} - 1. \tag{18}$$

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_P.

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: Let X be an $\mathbb{Z}_+$-valued random variable with distribution R. Then, *the local information* $J_{Q,2}(X)$ *of* X *relative to the compound Poisson distribution* $CP(\lambda, Q)$, is defined by,

$$J_{Q,2}(X) = \lambda_X E[r_2^2(X)], \tag{19}$$

where λ_X is the mean of X, and the score function r_2 is,

$$r_2(x) = \frac{xR(x)}{\lambda_X \sum_u uQ(u)R(x-u)} - 1. \tag{20}$$

Note that again $J_{Q,2}$ reduces to J_P when $Q = \delta_1$. In the simple Poisson case, as we saw, the quantity J_P has a minimum mean square estimation 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 minimum mean square 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,

$$J_{Q,2}(Y) = \frac{\text{Var}\{E[X|Y+V]\}}{EX}. \tag{21}$$

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,

$$J_{Q,1}(Y_1 + Y_2 + \ldots + Y_n) \leq \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} J_{Q,1}(Y_i), \tag{22}$$

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 be the sum of n independent, Q-compound, random variables $X_1, X_2, \ldots, X_n$. Write $U^{(i)} = \sum_{j \neq i} X_i$ the leave-one-out sums, and let $\lambda^{(i)}$ denote the mean of $U^{(i)}$, for each $i = 1, 2, \ldots, n$. Then,

$$J_{Q,1}(S_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} \frac{\lambda^{(i)}}{\lambda} J_{Q,1}(U^{(i)}), \tag{23}$$

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 (15) proved in [20] directly relates the relative entropy to the local information quantity J_P. In fact, the Poisson approximation bounds developed in [6] in relative entropy, are proved by combining this result with the subadditivity property of J_P. However, the known logarithmic Sobolev inequalities for compound Poisson distributions [22], [23], 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 for $i = 1, 2$,

$$\|P - CP(\lambda, Q)\|_{\mathrm{TV}} \leq q H(\lambda, Q) \sqrt{\lambda J_{Q,i}(X)}, \tag{24}$$

where $\lambda = E(X)/q$, and $H(\lambda, Q)$ is a constant depending only on λ and Q.

The quantity $H(\lambda, Q)$ arises from the so-called 'magic factors' which appear in Stein's method, and it can easily be bounded in an explicit and easily applicable way. Combining Theorems V and VII leads to very effective approximation bounds in total variation distance.

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

Proof of Theorem I: Let $Z' \sim \mathrm{Po}(\lambda)$, 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,

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

and the result follows from the Poisson approximation bound (16) of [6].

This data processing argument does not directly extend to the case where the Q_i associated with the summands are different. However, versions of Theorems V and VII can be generalized to this case, although they become much more complicated to state. Such extensions, and their consequences for compound Poisson approximation bounds in total variation, may be found in a forthcoming longer version [24] of the present work.

REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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

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

[13] 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.

[14] 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.

[15] 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.

[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] 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.

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

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

[20] 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.

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

[22] 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.

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

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