Thinning, Entropy, and the Law of Thin Numbers

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

***Abstract*—Rényi's *thinning* operation on a discrete random variable is a natural discrete analog of the scaling operation for continuous random variables. The properties of thinning are investigated in an information-theoretic context, especially in connection with information-theoretic inequalities related to Poisson approximation results. The classical Binomial-to-Poisson convergence (sometimes referred to as the "law of small numbers") is seen to be a special case of a thinning limit theorem for convolutions of discrete distributions. A rate of convergence is provided for this limit, and nonasymptotic bounds are also established. This development parallels, in part, the development of Gaussian inequalities leading to the information-theoretic version of the central limit theorem. In particular, a "thinning Markov chain" is introduced, and it is shown to play a role analogous to that of the Ornstein-Uhlenbeck process in connection to the entropy power inequality.**

***Index Terms*—Binomial distribution, compound Poisson distribution, entropy, information divergence, law of small numbers, law of thin numbers, Poisson distribution, Poisson-Charlier polynomials, thinning.**

I. Introduction

APPROXIMATING the distribution of a sum of weakly dependent discrete random variables by a Poisson distribution is an important and well-studied problem in probability; see [2] and the references therein for an extensive account. Strong connections between these results and information-theoretic techniques were established [18], [28]. In particular, for the special case of approximating a binomial distribution by a Poisson, some of the sharpest results to date are established using a combination of the techniques [18], [28], and Pinsker's inequality [10], [13], [22]. Earlier work on information-theoretic bounds for Poisson approximation is reported in [42], [25], [34].

The thinning operation, which we define next, was introduced by Rényi in [35], who used it to provide an alternative characterization of Poisson measures.

Definition 1.1: Given $\alpha \in [0,1]$ and a discrete random variable X with distribution P on $\mathbb{N}_0 = \{0,1,\ldots\}$, the *$\alpha$-thinning of P* is the distribution $T_\alpha(P)$ of the sum

$$\sum_{x=1}^{X} B_x, \text{ where } B_1, B_2 \ldots \sim \text{i.i.d. } \mathrm{Bern}(\alpha) \tag{1}$$

Manuscript received June 03, 2009; revised April 22, 2010. Date of current version August 18, 2010. The work of P. Harremoës was supported by a grant from the Danish Natural Science Research Council and by the European Pascal Network of Excellence. The work of I. Kontoyiannis was supported in part by a Marie Curie International Outgoing Fellowship, PIOF-GA-2009-235837. The material in this paper was presented in part at the 2007 IEEE International Symposium on Information Theory, Nice, France, June 2007.

P. Harremoës is with Copenhagen Business College, 1175 Copenhagen, Denmark (e-mail: harremoes@ieee.org).

O. Johnson is with the Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom (e-mail: O.Johnson@bristol.ac.uk).

I. Kontoyiannis is with the Department of Informatics, Athens University of Economics and Business, Athens 10434, Greece (e-mail: yiannis@aueb.gr).

Communicated by E. Ordentlich, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2010.2053893

where the random variables $\{B_x\}$ are independent and identically distributed (i.i.d.) each with a Bernoulli distribution with parameter α, denoted $\mathrm{Bern}(\alpha)$, and also independent of X. [As usual, we take the empty sum $\sum_{x=1}^{0}(\cdot)$ to be equal to zero.] An explicit representation of $T_\alpha(P)$ can be given as

$$T_\alpha(P)(z) = \sum_{x=z}^{\infty} P(x) \binom{x}{z} \alpha^z (1-\alpha)^{x-z}, \quad z \geq 0. \tag{2}$$

When it causes no ambiguity, the thinned distribution $T_\alpha(P)$ is written simply $T_\alpha P$.

For any random variable X with distribution P on $\mathbb{N}_0$, we write P^{*n} for the n-fold convolution of P with itself, i.e., the distribution of the sum of n i.i.d. copies of X. For example, if $P \sim \mathrm{Bern}(p)$, then $P^{*n} \sim \mathrm{Bin}(n,p)$, the binomial distribution with parameters n and p. It is easy to see that its $(1/n)$-thinning, $T_{1/n}(P^{*n})$, is simply $\mathrm{Bin}(n, p/n)$; see Example 2.2. Therefore, the classical Binomial-to-Poisson convergence result—sometimes referred to as the "law of small numbers"—can be phrased as saying that, if $P \sim \mathrm{Bern}(p)$, then

$$T_{1/n}(P^{*n}) \to \mathrm{Po}(p), \quad \text{as } n \to \infty \tag{3}$$

where $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with parameter $\lambda > 0$.

One of the main points of this paper is to show that this result holds for very wide class of distributions P, and to provide conditions under which several stronger and more general versions of (3) can be obtained. We refer to results of the form (3) as laws of thin numbers.

Section II contains numerous examples that illustrate how particular families of random variables behave on thinning, and it also introduces some of the particular classes of random variables that will be considered in the rest of the paper. In Sections III and IV several versions of the law of thin numbers are formulated; first for i.i.d. random variables in Section III, and then for general classes of (not necessarily independent or identically distributed) random variables in Section IV. For example, in the simplest case where $Y_1, Y_2, \ldots$ are i.i.d. with distribution P on $\mathbb{N}_0$ and with mean λ, so that the distribution of their sum, $S_n = Y_1 + Y_2 + \cdots + Y_n$, is P^{*n}, Theorem 3.3 shows that

$$D\left(T_{1/n}(P^{*n}) || \mathrm{Po}(\lambda)\right) \to 0, \quad \text{as } n \to \infty \tag{4}$$

as long as $D(P||\mathrm{Po}(\lambda)) < \infty$, where, as usual, $D(P||Q)$ denotes the *information divergence*, or *relative entropy*, from P to Q,[1]

$$D(P||Q) = \sum_{z=0}^{\infty} P(z) \log \frac{P(z)}{Q(z)}.$$

[1]Throughout the paper, $\log$ denotes the natural logarithm to base e, and we adopt the usual convention that $0 \log 0 = 0$.

Note that, unlike most classical Poisson convergence results, the law of thin numbers in (4) proves a Poisson limit theorem for the sum of a single sequence of random variables, rather than for a triangular array.

It may be illuminating to compare the result (4) with the information-theoretic version of the central limit theorem (CLT); see, e.g., [3] and [23]. Suppose $Y_1, Y_2, \ldots$ are i.i.d. continuous random variables with density f on $\mathbb{R}$, and with zero mean and unit variance. Then the density of their sum $S_n = Y_1 + Y_2 + \cdots + Y_n$, is the n-fold convolution f^{*n} of f with itself. Write Σ_α for the standard scaling operation in the CLT regime: If a continuous random variable X has density f, then $\Sigma_\alpha(f)$ is the density of the scaled random variable $\sqrt{\alpha}X$, and, in particular, the density of the standardized sum $\frac{1}{\sqrt{n}}S_n$ is $\Sigma_{1/n}(f^{*n})$. The information-theoretic CLT states that, if $D(f\|\phi) < \infty$, we have

$$D\left(\Sigma_{1/n}(f^{*n})\|\phi\right) \to 0, \quad \text{as } n \to \infty \tag{5}$$

where ϕ is the standard Normal density. Note the close analogy between the statements of the law of thin numbers in (4) and the CLT in (5).

Before describing the rest of our results, we mention that there is a significant thread in the literature on thinning limit theorems and associated results for point processes. Convergence theorems of the "law of thin numbers" type, as in (3) and (4), were first examined in the context of queueing theory by Palm [33] and Khinchin [26], while more general results were established by Grigelionis [17]. See the discussion in the text, [12, pp. 146–166], for details and historical remarks; also see the comments following Section IV, Theorem 4.1. More specifically, this line of work considered asymptotic results, primarily in the sense of weak convergence, for the distribution of a superposition of the sample paths of independent (or appropriately weakly dependent) point processes. Here we take a different direction and, instead of considering the full infinite-dimensional distribution of a point process, we focus on finer results—e.g., convergence in information divergence and nonasymptotic bounds—for the one-dimensional distribution of the thinned sum of integer-valued random variables.

With these goals in mind, before examining the finite-n behavior of $T_{1/n}(P^{*n})$, in Section V we study a simpler but related problem, on the convergence of a continuous-time "thinning" Markov chain on $\mathbb{N}_0$, which acts as the $M/M/\infty$ queue. It is well known that this Markov chain has the Poisson law as its unique invariant measure (see for example, [31] or [32]). In Theorem 5.1 we characterize precisely the rate at which it converges to the Poisson law in terms of the χ^2 distance, which leads to an upper bound on its convergence in information divergence. A new characterization of the Poisson distribution in terms of thinning is also obtained. The main technical tool used here is based on an examination of the L^2 properties of the Poisson-Charlier polynomials in the thinning context. In the present context, as described by Chafaï [7], this Markov chain plays a role parallel to that of the Ornstein-Uhlenbeck process in the context of Gaussian convergence and the entropy power inequality [37], [38], [29].

In Section VI we give both asymptotic and finite-n bounds on the rate of convergence for the law of thin numbers. Specifically, we employ the scaled Fisher information functional introduced in [28] to give precise, explicit bounds on the divergence $D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right)$. An example of the type of result we prove is the following: Suppose X is an ultra bounded (see Section II, Definition 2.1) random variable, with distribution P, mean λ, and finite variance $\sigma^2 \neq \lambda$. Then

$$\limsup_{n\to\infty} n^2 D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right) \leq 2c^2$$

for a nonzero constant c we explicitly identify; cf. Corollary 6.1.

Similarly, in Section VIII we give both finite-n and asymptotic bounds on the law of small numbers in terms of the total variation distance, $\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\|$, between $T_{1/n}(P^{*n})$ and the $\mathrm{Po}(\lambda)$ distribution. In particular, Theorem 8.1 states that if $X \sim P$ has mean λ and finite variance σ^2, then, for all n

$$\left\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\right\| \leq \frac{1}{n2^{1/2}} + \frac{\sigma}{n^{1/2}} \min\left\{1, \frac{1}{2\lambda^{1/2}}\right\}.$$

[Corresponding lower bounds will be presented in the companion paper [21].]

A closer examination of the monotonicity properties of the scaled Fisher information in relation to the thinning operation is described in Section VII. Finally, Section IX shows how the idea of thinning can be extended to compound Poisson distributions. TheAppendix contains the proofs of some of the more technical results.

Finally we mention that, after the announcement of (weaker and somewhat more specialized versions of) the present results in [20], Yu [41] also obtained some interesting, related results. In particular, he showed that the conditions of the strong and thermodynamic versions of the law of thin numbers (see Theorems 3.3 and 3.2) can be weakened, and he also provided conditions under which the convergence in these limit theorems is monotonic in n.

II. Examples of Thinning and Distribution Classes

This section contains several examples of the thinning operation, statements of its more basic properties, and the definitions of some important classes of distributions that will play a central role in the rest of this paper. The proofs of all the lemmas and propositions of this section are given in the Appendix.

Note, first, two important properties of thinning that are immediate from its definition:

1. The thinning of a sum of independent random variables is the convolution of the corresponding thinnings.
2. For all $\alpha, \beta \in [0,1]$ and any distribution P on $\mathbb{N}_0$, we have the following.

$$T_\alpha(T_\beta(P)) = T_{\alpha\beta}(P). \tag{6}$$

Example 2.1: Thinning preserves the Poisson family of laws, in that $T_\alpha(\mathrm{Po}(\lambda)) = \mathrm{Po}(\alpha\lambda)$. This follows from (2), since

$$\begin{aligned} T_\alpha(\mathrm{Po}(\lambda))(z) &= \sum_{x=z}^{\infty} \mathrm{Po}(\lambda, x)\binom{x}{z}\alpha^z(1-\alpha)^{x-z} \\ &= \sum_{x=z}^{\infty} \frac{\lambda^x}{x!}e^{-\lambda}\binom{x}{z}\alpha^z(1-\alpha)^{x-z} \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{-\lambda}}{z!}(\alpha\lambda)^z \sum_{x=z}^{\infty} \frac{(\lambda(1-\alpha))^{x-z}}{(x-z)!} \\
&= \frac{e^{-\lambda}}{z!}(\alpha\lambda)^z e^{\lambda(1-\alpha)} \\
&= \mathrm{Po}(\alpha\lambda, z)
\end{aligned}
$$

where $\mathrm{Po}(\lambda, x) = e^{-\lambda}\lambda^x/x!$, $x \geq 0$, denotes the Poisson mass function.

As it turns out, the factorial moments of a thinned distribution are easier to work with than ordinary moments. Recall that the kth factorial moment of X is $E[X^{\underline{k}}]$, where $x^{\underline{k}}$ denotes the falling factorial

$$x^{\underline{k}} = x(x-1)\cdots(x-k+1) = \frac{x!}{(x-k)!}.$$

The factorial moments of an α-thinning are easy as given by the following result.

Lemma 2.1: For any random variable Y with distribution P on $\mathbb{N}_0$ and for $\alpha \in (0,1)$, writing Y_α for a random variable with distribution $T_\alpha P$:

$$E[Y_\alpha^{\underline{k}}] = \alpha^k E[Y^{\underline{k}}]. \quad \text{for all } k. \tag{7}$$

That is, thinning scales factorial moments in the same way as ordinary multiplication scales ordinary moments.

We will use the following result, which is a multinomial version of Vandermonde's identity and is easily proved by induction. The details are omitted.

Lemma 2.2: The falling factorial satisfies the multinomial expansion; that is, for any positive integer y, all integers $x_1, x_2, \ldots, x_y$, and any $k \geq 1$, the factorial

$$\left(\sum_{i=1}^{y} x_i\right)^{\underline{k}}$$

equals

$$\sum_{\substack{k_1,k_2\ldots,k_y \\ k_1+k_2+\cdots+k_y=k}} \binom{k}{k_1 \quad k_2 \quad \cdots \quad k_y} \prod_{i=1}^{y} x_i^{\underline{k_i}}.$$

The following is a basic regularity property of the thinning operation.

Proposition 2.1: For any $\alpha \in (0,1)$, the map $P \mapsto T_\alpha(P)$ is injective.

Example 2.2: Thinning preserves the class of Bernoulli sums. That is, the thinned version of the distribution of a finite sum of independent Bernoulli random variables (with possibly different parameters) is also such a sum. This follows from property 1 stated in the beginning of this section, combined with the observation that the α-thinning of the $\mathrm{Bern}(p)$ distribution is the $\mathrm{Bern}(\alpha p)$ distribution. In particular, thinning preserves the binomial family: $T_\alpha(\mathrm{Bin}(n,p)) = \mathrm{Bin}(n, \alpha p)$.

Example 2.3: Thinning by α transforms a geometric distribution with mean λ into a geometric distribution with mean $\alpha\lambda$. Recalling that the geometric distribution with mean λ has point probabilities

$$\mathrm{Geo}(\lambda, x) = \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^x, \quad x = 0, 1, \ldots$$

using (2)

$$
\begin{aligned}
T_\alpha \mathrm{Geo}(\lambda)(z) &= \sum_{x=z}^{\infty} \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^x \binom{x}{z} \alpha^z (1-\alpha)^{x-z} \\
&= \frac{1}{(1+\lambda)z!}\left(\frac{\alpha\lambda}{1+\lambda}\right)^z \sum_{x=z}^{\infty} \left(\frac{\lambda(1-\alpha)}{1+\lambda}\right)^{x-z} x^{\underline{z}} \\
&= \frac{1}{(1+\lambda)}\left(\frac{\alpha\lambda}{1+\lambda}\right)^z \left(1 - \frac{\lambda(1-\alpha)}{1+\lambda}\right)^{-z-1} \\
&= \mathrm{Geo}(\alpha\lambda, z).
\end{aligned}
$$

The sum of n i.i.d. geometrics has a negative binomial distribution. Thus, in view of this example and property 1 stated in the beginning of this section, the thinning of a negative binomial distribution is also negative binomial.

Partly motivated by these examples, we describe certain classes of random variables (some of which are new). These appear as natural technical assumptions in the subsequent development of our results. The reader may prefer to skip the remainder of this section and only refer back to the definitions when necessary.

Definition 2.1:

1) A *Bernoulli sum* is a distribution that can be obtained from the sum of finitely many independent Bernoulli random variables with possibly different parameters. The class of Bernoulli sums with mean λ is denoted by $Ber(\lambda)$ and the the union $\cup_{\mu \leq \lambda} Ber(\mu)$ is denoted by $Ber^{\leq}(\lambda)$.
2) A distribution P satisfying

$$\log \frac{P(j)}{\mathrm{Po}(\lambda, j)} \geq \frac{1}{2}\log \frac{P(j-1)}{\mathrm{Po}(\lambda, j-1)} + \frac{1}{2}\log \frac{P(j+1)}{\mathrm{Po}(\lambda, j+1)} \tag{8}$$

is said to be *ultra log-concave* (ULC); cf. [24]. The set of ultra log-concave distributions with mean λ shall be denoted $ULC(\lambda)$, and we also write $ULC^{\leq}(\lambda)$ for the union $\cup_{\mu \leq \lambda} ULC(\mu)$. Note that (8) is satisfied for a single value of $\lambda > 0$ if and only if it is satisfied for all $\lambda > 0$.
3) The distribution of a random variable X that satisfies $E[X^{\underline{k+1}}] \leq \lambda E[X^{\underline{k}}]$ for all $k \geq 0$ will be said to be *ultra bounded* (UB) with ratio λ. The set of ultra bounded distributions with this ratio is denoted $UB(\lambda)$.
4) The distribution of a random variable X satisfying $E[X^{\underline{k}}] \leq \lambda^k$ for all $k \geq 0$ will be said to be *Poisson bounded* (PB) with ratio λ. The set of Poisson bounded distributions with this ratio is denoted $PB(\lambda)$.
5) A random variable will be said to be ULC, UB, or PB, if its distribution is ULC, UB or PB, respectively.

First we mention some simple relationships between these classes. Walkup [39] showed that if $X \sim P \in ULC(\lambda)$ and $Y \sim Q \in ULC(\mu)$ then $X + Y \sim P * Q \in ULC(\lambda + \mu)$. Hence, $Ber(\lambda) \subseteq ULC(\lambda)$. In [24] it was shown that, if $P \in ULC(\lambda)$, then $T_\alpha P \in ULC(\alpha\lambda)$. Clearly, $UB(\lambda) \subseteq PB(\lambda)$. Further, P is Poisson bounded if and only if the α-thinning $T_\alpha P$ is Poisson bounded, for some $\alpha > 0$. The same holds for ultra boundedness.

Proposition 2.2: In the notation of Definition 2.1, the class $ULC(\lambda) \subseteq UB(\lambda)$. That is, if the distribution of X is in $ULC(\lambda)$ then $E[X^{\underline{k+1}}] \leq \lambda E[X^{\underline{k}}]$.

The next result states that the PB and UB properties are preserved on summing and thinning.

Proposition 2.3:

a) If $X \sim P \in PB(\lambda)$ and $Y \sim Q \in PB(\mu)$ are independent, then $X+Y \sim P*Q \in PB(\lambda+\mu)$ and $T_\alpha P \in PB(\alpha\lambda)$.
b) If $X \sim P \in UB(\lambda)$ and $Y \sim Q \in UB(\mu)$ are independent, then $X+Y \sim P*Q \in UB(\lambda+\mu)$ and $T_\alpha P \in UB(\alpha\lambda)$.

Formally, the above discussion can be summarized as

$$Ber^{\leq}(\lambda) \subseteq ULC^{\leq}(\lambda) \subseteq UB(\lambda) \subseteq PB(\lambda).$$

Finally, we note that each of these classes of distributions is "thinning-convex," i.e., if P and Q are element of a set then $T_\alpha(P) * T_{1-\alpha}(Q)$ is also an element of the same set. In particular, thinning maps each of these sets into itself, since $T_\alpha(P) = T_\alpha(P) * T_{1-\alpha}(\delta_0)$ where δ_0, the point mass at zero, has $\delta_0 \in Ber^{\leq}(\lambda)$.

III. Laws of Thin Numbers: The I.I.D. Case

In this section we state and prove three versions of the law of thin numbers, under appropriate conditions; recall the relevant discussion in the Introduction. Theorem 3.1 proves convergence in total variation, Theorem 3.2 in entropy, and Theorem 3.3 in information divergence.

Recall that the total variation distance $||P-Q||$ between two probability distributions P, Q on $\mathbb{N}_0$ is

$$\begin{aligned}||P-Q|| &:= \sup_{B\subset\mathbb{N}_0} |P(B)-Q(B)| \\ &= \frac{1}{2}\sum_{k\geq 0}|P(k)-Q(k)|.\end{aligned}$$

Theorem 3.1 (Weak Version): For any distribution P on $\mathbb{N}_0$ with mean λ

$$||T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)|| \to 0, \quad n\to\infty.$$

Proof: In view of Scheffé's lemma, pointwise convergence of discrete distributions is equivalent to convergence in total variation, so it suffices to show that, $T_{1/n}(P^{*n})(z)$ converges to $e^{-\lambda}\lambda^z/z!$, for all $z \geq 0$.

Note that $T_{1/n}(P^{*n}) = (T_{1/n}(P))^{*n}$, and that (2) implies the following elementary bounds for all α, using Jensen's inequality:

$$\begin{aligned}T_\alpha(P)(0) &= \sum_{x=0}^{\infty} P(x)(1-\alpha)^x \geq (1-\alpha)^\lambda \\ T_\alpha(P)(1) &= \sum_{x=1}^{\infty} P(x)x\alpha(1-\alpha)^{x-1}. \end{aligned} \tag{9}$$

Since for i.i.d. variables Y_i, the probability $\Pr\{Y_1+\cdots+Y_n = z\} \geq \binom{n}{z}\Pr\{Y_1=1\}^z P\{Y_1=0\}^{n-z}$, taking $\alpha = 1/n$ we obtain that the convolution

$$(T_{1/n}(P))^{*n}(z)$$

is bounded below by

$$\begin{aligned}&\binom{n}{z}\left(\sum_{x=1}^{\infty}P(x)\frac{x}{n}\left(1-\frac{1}{n}\right)^{x-1}\right)^z\left(\left(1-\frac{1}{n}\right)^\lambda\right)^{n-z} \\ &\quad = \frac{n^{\underline{z}}}{n^z\, z!}\left(\sum_{x=1}^{\infty}P(x)x\left(1-\frac{1}{n}\right)^{x-1}\right)^z\left(1-\frac{1}{n}\right)^{(n-z)\lambda}.\end{aligned}$$

Now, for any fixed value of z and n tending to infinity

$$\frac{n^{\underline{z}}}{n^z\, z!} \to \frac{1}{z!}$$

and

$$\left(1-\frac{1}{n}\right)^{(n-z)\lambda} \to e^{-\lambda}$$

and by monotone convergence

$$\sum_{x=1}^{\infty}P(x)x\left(1-\frac{1}{n}\right)^{x-1} \to \lambda.$$

Therefore

$$\liminf_{n\to\infty}(T_{1/n}(P))^{*n}(z) \geq \mathrm{Po}(\lambda,z).$$

Since all $(T_{1/n}(P))^{*n}$ are probability mass functions and so is $\mathrm{Po}(\lambda)$, the above $\liminf$ is necessarily a limit. ∎

As usual, the entropy of a probability distribution P on $\mathbb{N}_0$ is defined by

$$H(P) = -\sum_{k\geq 0}P(k)\log P(k).$$

Recall that the entropy of the Poisson distribution cannot be expressed in closed form, although useful bounds do exist; see, e.g., [1] and the references therein.

Theorem 3.2 (Thermodynamic Version): If P is any distribution on $\mathbb{N}_0$ which is Poisson bounded with mean λ, then

$$H\left(T_{1/n}(P^{*n})\right) \to H(\mathrm{Po}(\lambda)), \quad \text{as } n\to\infty.$$

Proof: The distribution $T_{1/n}(P^{*n})$ converges pointwise to the Poisson distribution so, by dominated convergence, it is sufficient to prove that $-T_{1/n}(P^{*n})(x)\log\left(T_{1/n}(P^{*n})(x)\right)$ is dominated by a summable function. This easily follows from the simple bound in the following lemma. ∎

Lemma 3.1: Suppose P is Poisson bounded with ratio μ. Then, $P(x) \leq \mathrm{Po}(\mu,x)\cdot e^\mu$, for all $x \geq 0$.

Proof: Note that, for all x

$$P(x)x^{\underline{k}} \leq \sum_{x=0}^{\infty}P(x)x^{\underline{k}} \leq \mu^k$$

so that, in particular, $P(x)x^{\underline{x}} \leq \mu^x$, and, $P(x) \leq \frac{\mu^x}{x!} = \mathrm{Po}(\mu,x)\, e^\mu$. ∎

From the proof of [24, Theorem 2.5] we know that, $H\left(T_{1/n}(P^{*n})\right) \leq H(\mathrm{Po}(\lambda))$ if P is ultra log-concave, so for such distributions the theorem states that the entropy converges

to its maximum. For ultra log-concave distributions the thermodynamic version also implies convergence in information divergence. This also holds for Poisson bounded distributions, which is easily proved using dominated convergence. As shown in the next theorem, convergence in information divergence can be established under quite general conditions.

Theorem 3.3 (Strong Version): For any distribution P on $\mathbb{N}_0$ with mean λ and $D(P||\mathrm{Po}(\lambda)) < \infty$

$$D\left(T_{1/n}(P^{*n})||\mathrm{Po}(\lambda)\right) \to 0, \quad \text{as } n \to \infty.$$

The proof of Theorem 3.3 is given in the Appendix; it is based on a straightforward but somewhat technical application of the following general bound.

Proposition 3.1: Let X be a random variable with distribution P on $\mathbb{N}_0$ and with finite mean λ/α, for some $\alpha \in (0,1)$. If $D(P||\mathrm{Po}(\lambda/\alpha)) < \infty$, then

$$D(T_\alpha(P)||\mathrm{Po}(\lambda)) \le \frac{\alpha^2}{2(1-\alpha)} + E\left[\alpha X \log\left(\frac{\alpha X}{\lambda}\right)\right] \quad (10)$$

where the right-hand side is always finite.

Proof: First note that, since P has finite mean, its entropy is bounded by the entropy of a geometric with the same mean, which is finite, so $H(P)$ is finite. Therefore, the divergence $D(P||\mathrm{Po}(\lambda/\alpha))$ can be expanded as

$$\begin{aligned} &E\left[\log\left(\frac{P(X)}{\mathrm{Po}(\lambda/\alpha, X)}\right)\right] \\ &= E[\log(X!)] + \frac{\lambda}{\alpha} - H(P) - \frac{\lambda}{\alpha}\log\left(\frac{\lambda}{\alpha}\right) \\ &\ge \frac{1}{2}E[\log^+(2\pi X)] + E[X\log X] - H(P) - \frac{\lambda}{\alpha}\log\left(\frac{\lambda}{\alpha}\right) \end{aligned} \quad (11)$$

where the last inequality follows from the Stirling bound

$$\log(x!) \ge \frac{1}{2}\log^+(2\pi x) + x\log x - x$$

and $\log^+(x)$ denotes the function $\log\max\{x,1\}$. Since the divergence $D(P||\mathrm{Po}(\lambda))$ is finite, the bound (11) implies that $E[X\log X]$ is finite. [Recall the convention that $0\log 0 = 0$.]

Also note that the representation of $T_\alpha(P)$ in (2) can be written as,

$$T_\alpha P(z) = \sum_{x=0}^{\infty} P(x)\Pr\{\mathrm{Bin}(x,\alpha) = z\}.$$

Using this and the joint convexity of information divergence in its two arguments (see, e.g., [9, Th. 2.7.2]), the divergence of interest can be bounded as

$$\begin{aligned} &D(T_\alpha(P)||\mathrm{Po}(\lambda)) \\ &\quad = D\left(\sum_{x=0}^{\infty} P(x)\mathrm{Bin}(x,\alpha)\middle\| \sum_{x=0}^{\infty} P(x)\mathrm{Po}(\lambda)\right) \\ &\quad \le \sum_{x=0}^{\infty} P(x) D(\mathrm{Bin}(x,\alpha)||\mathrm{Po}(\lambda)) \end{aligned} \quad (12)$$

where the first term (corresponding to $x = 0$) equals λ. Since the Poisson measures form an exponential family, they satisfy a Pythagorean identity [11] which, together with the bound

$$D(\mathrm{Bin}(x,p)||\mathrm{Po}(xp)) \le \frac{p^2}{2(1-p)} \quad (13)$$

see, e.g., [22] or [28], gives, for each $x \ge 1$

$$\begin{aligned} &D(\mathrm{Bin}(x,\alpha)||\mathrm{Po}(\lambda)) \\ &\quad = D(\mathrm{Bin}(x,\alpha)||\mathrm{Po}(\alpha x)) + D(\mathrm{Po}(\alpha x)||\mathrm{Po}(\lambda)) \\ &\quad \le \frac{\alpha^2}{2(1-\alpha)} + \sum_{j=0}^{\infty} \mathrm{Po}(\alpha x, j)\log\left(\frac{(\alpha x)^j \exp(-\alpha x)/j!}{\lambda^j \exp(-\lambda)/j!}\right) \\ &\quad = \frac{\alpha^2}{2(1-\alpha)} + \left(\alpha x \log\left(\frac{\alpha x}{\lambda}\right) - \alpha x + \lambda\right). \end{aligned}$$

Since the final bound clearly remains valid for $x = 0$, substituting it into (12) gives (10). ∎

IV. Laws of Thin Numbers: The Non-I.I.D. Case

In this section we state and prove more general versions of the law of thin numbers, for sequences of random variables that are not necessarily independent or identically distributed. Although some of the results in this section are strict generalizations of Theorems 3.1 and 3.3, their proofs are different.

We begin by showing that, using a general proof technique introduced in [28], the weak law of thin numbers can be established under weaker conditions than those in Theorem 3.1. The main idea is to use the data-processing inequality on the total variation distance between an appropriate pair of distributions.

Theorem 4.1 (Weak Version, Non-I.I.D.): Let $P_1, P_2, \ldots$ be an arbitrary sequence of distributions on $\mathbb{N}_0$, and write $P^{(n)} = P_1 * P_2 * \cdots * P_n$ for the convolution of the first n of them. Then,

$$\left\| T_{1/n}(P^{(n)}) - \mathrm{Po}(\lambda) \right\| \to 0, \quad n \to \infty,$$

as long as the following three conditions are satisfied as $n \to \infty$:

a) $a_n = \max_{1\le i\le n}[1 - T_{1/n}P_i(0)] \to 0$;
b) $b_n = \sum_{i=1}^n \left[1 - T_{1/n}P_i(0)\right] \to \lambda$;
c) $c_n = \sum_{i=1}^n \left[1 - T_{1/n}P_i(0) - T_{1/n}P_i(1)\right] \to 0$.

Note that Theorem 4.1 can be viewed as a one-dimensional version of Grigelionis' Theorem 1 in [17]; recall the relevant comments in the Introduction. Recently, Schuhmacher [36] established nonasymptotic, quantitative versions of this result, in terms of the Barbour-Brown distance, which metrizes weak convergence in the space of probability measures of point processes. As the information divergence is a finer functional than the Barbour-Brown distance, Schuhmacher's results are not directly comparable with the finite-n bounds we obtain in Propositions 3.1, 4.1, and Corollary 6.1.

Before giving the proof of the theorem, we state a simple lemma on a well-known bound for $\|\mathrm{Po}(\lambda) - \mathrm{Po}(\mu)\|$. Its short proof is included for completeness.

Lemma 4.1: For any $\lambda, \mu > 0$,

$$\|\mathrm{Po}(\lambda) - \mathrm{Po}(\mu)\| \le 1 - e^{-|\lambda-\mu|} \le |\lambda - \mu|.$$

Proof: Suppose, without loss of generality, that $\lambda > \mu$, and define two independent random variables $X \sim \mathrm{Po}(\mu)$ and $Z \sim \mathrm{Po}(\lambda - \mu)$, so that, $Y = X + Z \sim \mathrm{Po}(\lambda)$. Then, by the coupling inequality [30]

$$\begin{aligned}\|\mathrm{Po}(\lambda) - \mathrm{Po}(\mu)\| &\le \Pr\{X \neq Y\} \\ &= \Pr\{Z \neq 0\} \\ &= 1 - e^{-(\lambda-\mu)}.\end{aligned}$$

The second inequality in the lemma is trivial. ∎

Proof of Theorem 4.1: First we introduce some convenient notation. Let $X_1, X_2, \ldots$ be independent random variables with $X_i \sim P_i$ for all i; for each $n \ge 1$, let $Y_1^{(n)}, Y_2^{(n)}, \ldots$ be independent random variables with $Y_i^{(n)} \sim T_{1/n}P_i$ for all i; and similarly let $Z_1^{(n)}, Z_2^{(n)}, \ldots$ be independent $\mathrm{Po}(\lambda_i^{(n)})$ random variables, where $\lambda_i^{(n)} = T_{1/n}P_i(1)$, for $i, n \ge 1$. Also we define the sums, $S_n = \sum_{i=1}^n Y_i^{(n)}$ and $T_n = \sum_{i=1}^n Z_i^{(n)}$, and note that, $S_n \sim P^{(n)}$, and $T_n \sim \mathrm{Po}(\lambda^{(n)})$, where $\lambda^{(n)} = \sum_{i=1}^n \lambda_i^{(n)}$, for all $n \ge 1$.

Note that $\lambda^{(n)} \to \lambda$ as $n \to \infty$, since

$$\lambda^{(n)} = \sum_{i=1}^n \lambda_i^{(n)} = b_n - c_n$$

and, by assumption, $b_n \to \lambda$ and $c_n \to 0$, as $n \to \infty$.

With these definitions in place, we approximate

$$\begin{aligned}&\left\|T_{1/n}(P^{(n)}) - \mathrm{Po}(\lambda)\right\| \\ &\quad\le \|T_{1/n}(P^{(n)}) - \mathrm{Po}(\lambda^{(n)})\| + \|\mathrm{Po}(\lambda^{(n)}) - \mathrm{Po}(\lambda)\|, \quad (14)\end{aligned}$$

where, by Lemma 4.1, the second term is bounded by $|\lambda^{(n)} - \lambda|$ which vanishes as $n \to \infty$. Therefore, it suffices to show that the first term in (14) goes to zero. For that term

$$\begin{aligned}&\left\|T_{1/n}(P^{(n)}) - \mathrm{Po}(\lambda^{(n)})\right\| \\ &\quad= \|P_{S_n} - P_{T_n}\| \\ &\quad\le \|P_{\{Y_i^{(n)}\}} - P_{\{Z_i^{(n)}\}}\| \\ &\quad\le \sum_{i=1}^n \left\|T_{1/n}P_i - \mathrm{Po}(\lambda_i^{(n)})\right\| \\ &\quad\le \sum_{i=1}^n \Big[\left\|T_{1/n}P_i - \mathrm{Bern}\left(\lambda_i^{(n)}\right)\right\| \\ &\qquad\qquad + \left\|\mathrm{Bern}(\lambda_i^{(n)}) - \mathrm{Po}(\lambda_i^{(n)})\right\|\Big]\end{aligned}$$

where the first inequality above follows from the fact that, being an f-divergence, the total variation distance satisfies the data-processing inequality [11]; the second inequality comes from the well-known bound on the total variation distance between two product measures as the sum of the distances between their respective marginals; and the third bound is simply the triangle inequality.

Finally, noting that, for any random variable $X \sim P$, $\|P - \mathrm{Bern}(P(1))\| = \Pr\{X \ge 2\}$, and also recalling the simple estimate

$$\|\mathrm{Bern}(p) - \mathrm{Po}(p)\| = p(1 - e^{-p}) \le p^2$$

yields

$$\begin{aligned}&\left\|T_{1/n}(P^{(n)}) - \mathrm{Po}(\lambda^{(n)})\right\| \\ &\quad\le c_n + \sum_{i=1}^n \left(\lambda_i^{(n)}\right)^2 \\ &\quad\le c_n + \lambda^{(n)} \max_{1\le i\le n} \lambda_i^{(n)} \\ &\quad\le c_n + \lambda^{(n)} a_n\end{aligned}$$

and, by assumption, this converges to zero as $n \to \infty$, completing the proof. ∎

Recall that, in the i.i.d. case, the weak law of thin numbers only required the first moment of P to be finite, while the strong version also required that the divergence from P to the Poisson distribution be finite. For a sum of independent, non-identically distributed random variables with finite *second* moments, Proposition 3.1 can be used as in the proof of Theorem 3.3 to prove the following result. Note that the precise conditions required are somewhat analogous to those in Theorem 4.1.

Theorem 4.2 (Strong Version, Non-i.i.d.): Let $P_1, P_2, \ldots$ be an arbitrary sequence of distributions on $\mathbb{N}_0$, where each P_i has finite mean λ_i and finite variance. Writing $P^{(n)}$ for the convolution $P_1 * P_2 * \cdots * P_n$, we have

$$D\left(T_{1/n}(P^{(n)})\middle\|\mathrm{Po}(\lambda)\right) \to 0, \quad n \to \infty$$

as long as the following two conditions are satisfied:

a) $\lambda^{(n)} = \frac{1}{n}\sum_{i=1}^n \lambda_i \to \lambda$, as $n \to \infty$;
b) $\sum_{i=1}^\infty \frac{1}{i^2} E(X_i^2) < \infty$.

The proof of Theorem 4.2 is given in the Appendix, and it is based on Proposition 3.1. It turns out that under the additional condition of finite second moments, the proof of Proposition 3.1 can be refined to produce a stronger upper bound on the divergence.

Proposition 4.1: If P is a distribution on $\mathbb{N}_0$ with mean λ/α and variance $\sigma^2 < \infty$, for some $\alpha \in (0, 1)$, then

$$D(T_\alpha(P)\|\mathrm{Po}(\lambda)) \le \alpha^2\left(\frac{1}{2(1-\alpha)} + \frac{\sigma^2}{\lambda}\right). \quad (15)$$

Proof: Recall that in the proof of Proposition 3.1 it was shown that

$$D(T_\alpha(P)\|\mathrm{Po}(\lambda)) \le \sum_{x=0}^\infty P(x) D(\mathrm{Bin}(x,\alpha)\|\mathrm{Po}(\lambda)) \quad (16)$$

where

$$\begin{aligned}&D(\mathrm{Bin}(x,\alpha)\|\mathrm{Po}(\lambda)) \\ &\quad\le \frac{\alpha^2}{2(1-\alpha)} + \lambda\left(\frac{\alpha x}{\lambda}\log\left(\frac{\alpha x}{\lambda}\right) - \frac{\alpha x}{\lambda} + 1\right) \\ &\quad\le \frac{\alpha^2}{2(1-\alpha)} + \lambda\left(\frac{\alpha x}{\lambda} - 1\right)^2 \quad (17)\end{aligned}$$

and where in the last step above we used the simple bound $y\log y - y + 1 \le y(y-1) - y + 1 = (y-1)^2$, for $y > 0$. Substituting (17) into (16) yields

$$\begin{aligned} &D(T_\alpha(P)\|\mathrm{Po}(\lambda)) \\ &\quad\le \sum_{x=0}^{\infty} P(x)\left(\frac{\alpha^2}{2(1-\alpha)} + \lambda\left(\frac{\alpha x}{\lambda} - 1\right)^2\right) \\ &\quad= \frac{\alpha^2}{2(1-\alpha)} + \frac{\alpha^2}{\lambda}\sum_{x=0}^{\infty} P(x)\left(x - \frac{\lambda}{\alpha}\right)^2 \\ &\quad= \frac{\alpha^2}{2(1-\alpha)} + \frac{\alpha^2\sigma^2}{\lambda} \end{aligned}$$

as claimed. ∎

Using the bound (15) instead of Proposition 3.1, the following more general version of the law of thin numbers can be established.

Theorem 4.3 (Strong Version, Non-I.I.D.): Let $\{X_i\}$ be a sequence of (not necessarily independent or identically distributed) random variables on $\mathbb{N}_0$, and write $P^{(n)}$ for the distribution of the partial sum $S_n = X_1 + X_2 + \cdots + X_n$, $n \ge 1$. Assume that the $\{X_i\}$ have finite means and variances, and that:

a) They are "uniformly ultra bounded," in that, $\mathrm{Var}(X_i) \le CE(X_i)$ for all i, with a common $C < \infty$;
b) Their means satisfy $E(S_n) \to \infty$ as $n \to \infty$;
c) Their covariances satisfy

$$\lim_{n\to\infty} \frac{\sum_{1\le i<j\le n} \mathrm{Cov}(X_i, X_j)}{(E(S_n))^2} = 0.$$

If in fact $E(X_i) = \lambda > 0$ for all i, then,

$$\lim_{n\to\infty} D(T_{1/n}(P^{(n)})\|\mathrm{Po}(\lambda)) = 0.$$

More generally

$$\lim_{n\to\infty} D\left(T_{\alpha_n}(P^{(n)})\right\|\mathrm{Po}(\lambda)) = 0, \text{ where } \alpha_n = \lambda/E(S_n).$$

Proof: Obviously it suffices to prove the general statement. Proposition 4.1 applied to $P^{(n)}$ gives

$$\begin{aligned} &D\left(T_{\alpha_n}(P^{(n)})\right\|\mathrm{Po}(\lambda)) \\ &\quad\le \alpha_n^2\left(\frac{1}{2(1-\alpha_n)} + \frac{\mathrm{Var}(S_n)}{\lambda}\right) \\ &\quad= \frac{\alpha_n^2}{2(1-\alpha_n)} + \frac{\lambda\mathrm{Var}(S_n)}{(E(S_n))^2} \\ &\quad= \frac{\alpha_n^2}{2(1-\alpha_n)} + \frac{\lambda}{(E(S_n))^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &\qquad+ \frac{2\lambda}{(E(S_n))^2}\sum_{1\le i<j\le n}\mathrm{Cov}(X_i, X_j). \end{aligned}$$

The first and third terms tend to zero by assumptions (b) and (c), respectively. And using assumption (a), the second term is bounded above by

$$\frac{\lambda}{(E(S_n))^2}\,CE(S_n)$$

which also tends to zero by assumption (b). ∎

V. The Thinning Markov Chain

Before examining the rate of convergence in the law of thin numbers, we consider a related and somewhat simpler problem for a Markov chain. Several of the results in this section may be of independent interest. The Markov chain we will discuss represents the evolution of the $M/M/\infty$ queue. Its properties were used by Chafaï [7] to develop a family of inequalities which extends the logarithmic Sobolev inequalities of Bobkov and Ledoux [5], and other authors. Chafaï argues in [7, Sec. 1.2, 1.3] that this process is a natural discrete analog of the Ornstein-Uhlenbeck process associated with the Gaussian distribution.

Definition 5.1: Let P be a distribution on $\mathbb{N}_0$. For any $\alpha \in [0,1]$ and $\lambda > 0$, we write $U_\alpha^\lambda(P)$ for the distribution

$$U_\alpha^\lambda(P) = T_\alpha(P) * \mathrm{Po}((1-\alpha)\lambda).$$

For simplicity, $U_\alpha^\lambda(P)$ is often written simply as $U_\alpha^\lambda P$.

We note that $U_\alpha^\lambda U_\beta^\lambda = U_{\alpha\beta}^\lambda$, and that, obviously, U_α^λ maps probability distributions to probability distributions. Therefore, if for a *fixed* λ we define $Q^t = U_{e^{-t}}^\lambda$ for all $t \ge 0$, the collection $\{Q^t; t \ge 0\}$ of linear operators on the space of probability measures on $\mathbb{N}_0$ defines a Markov transition semigroup. Specifically, for $i, j \in \mathbb{N}_0$, the transition probabilities

$$\begin{aligned} Q_{ij}^t &= (Q^t(\delta_i))(j) \\ &= \left(U_{e^{-t}}^\lambda(\delta_i)\right)(j) \\ &= \left(T_{e^{-t}}(\delta_i) * \mathrm{Po}((1-e^{-t})\lambda)\right)(j) \\ &= \Pr\{\mathrm{Bin}(i, e^{-t}) + \mathrm{Po}((1-e^{-t})\lambda) = j\} \end{aligned}$$

define a continuous-time Markov chain $\{Z_t; t \ge 0\}$ on $\mathbb{N}_0$. It is intuitively clear that, as $\alpha \downarrow 0$ (or, equivalently, $t \to \infty$), the distribution $U_\alpha^\lambda P$ should converge to the $\mathrm{Po}(\lambda)$ distribution. Indeed, the following two well-known results (see for example, [31] or [32]) state that $\{Z_t\}$ is ergodic, with unique invariant measure $\mathrm{Po}(\lambda)$. Theorem 5.1 gives the rate at which it converges to $\mathrm{Po}(\lambda)$ in terms of the moments of the underlying distribution. This complements results such as [7, Th. 3.1], which gives a bound in terms of the tightest value of the log-Sobolev constant.

Proposition 5.1: For any distribution P on $\mathbb{N}_0$, $U_\alpha^\lambda(P)$ converges in total variation to $\mathrm{Po}(\lambda)$, as $\alpha \downarrow 0$.

Proof: From the definition of $U_\alpha^\lambda(P)$

$$\begin{aligned} &\left\|U_\alpha^\lambda(P) - \mathrm{Po}(\lambda)\right\| \\ &\quad= \|T_\alpha(P) * \mathrm{Po}((1-\alpha)\lambda) - \mathrm{Po}(\lambda)\| \\ &\quad= \|(T_\alpha(P) - \mathrm{Po}(\alpha\lambda)) * \mathrm{Po}((1-\alpha)\lambda)\| \\ &\quad\le \|T_\alpha(P) - \mathrm{Po}(\alpha\lambda)\| \end{aligned} \tag{18}$$

$$= \frac{1}{2}|(1 - T_\alpha(P)(0)) - (1 - \mathrm{Po}(\alpha\lambda, 0))| + \frac{1}{2}\sum_{x=1}^{\infty} |T_\alpha(P)(x) - \mathrm{Po}(\alpha\lambda, x)|$$

$$\leq \frac{1}{2}[(1 - T_\alpha(P)(0)) + (1 - \mathrm{Po}(\alpha\lambda, 0))] + \frac{1}{2}\sum_{x=1}^{\infty}(T_\alpha(P)(x) + \mathrm{Po}(\alpha\lambda, x)) \quad (19)$$

$$= 2 - T_\alpha(P)(0) - \mathrm{Po}(\alpha\lambda, 0) \quad (20)$$

where (18) follows from the fact that convolution with any distribution is a contraction with respect to the L^1 norm, (19) follows from the triangle inequality, and (20) converges to zero because of the bound (9). ∎

Using this, we can give a characterization of the Poisson distribution.

Corollary 5.1: Let P denote a discrete distribution with mean λ. If $P = U_\alpha^\lambda(P)$ for some $\alpha \in (0,1)$, then $P = \mathrm{Po}(\lambda)$. That is, $\mathrm{Po}(\lambda)$ is the unique invariant measure of the Markov chain $\{Z_t\}$, and, moreover,

$$D(U_\alpha^\lambda(P)||\mathrm{Po}(\lambda)) \to 0, \quad \text{as } \alpha \downarrow 0,$$

if and only if $D(U_\alpha^\lambda(P)||\mathrm{Po}(\lambda)) < \infty$ for some $\alpha > 0$.

Proof: Assume that $P = U_\alpha^\lambda(P)$. Then for any n, $P = U_{\alpha^n}^\lambda(P)$, so for any $\epsilon > 0$, by Proposition 5.1, $\|P - \mathrm{Po}(\lambda)\| = \left\|U_{\alpha^n}^\lambda(P) - \mathrm{Po}(\lambda)\right\| \leq \epsilon$ for n sufficiently large. The strengthened convergence of $D\left(U_\alpha^\lambda(P)||\mathrm{Po}(\lambda)\right)$ to zero if $D\left(U_\alpha^\lambda(P)||\mathrm{Po}(\lambda)\right) < \infty$ can be proved using standard arguments along the lines of the corresponding discrete-time results in [15], [4], and [19]. ∎

Next we shall study the rate of convergence of $U_\alpha^\lambda(P)$ to the Poisson distribution. It is easy to check that the Markov chain $\{Z_t\}$ is in fact *reversible* with respect to its invariant measure $\mathrm{Po}(\lambda)$. Therefore, the natural setting for the study of its convergence is the L^2 space of functions $f : \mathbb{N}_0 \to \mathbb{R}$ such that, $E[f(Z)^2] < \infty$ for $Z \sim \mathrm{Po}(\lambda)$. This space is also endowed with the usual inner product

$$\langle f, g\rangle = E[f(Z)g(Z)], \quad \text{for } Z \sim \mathrm{Po}(\lambda), f, g \in L^2$$

and the linear operators U_α^λ act on functions $f \in L^2$ by mapping each f into

$$(U_\alpha^\lambda f)(x) = E[f(Z_{\alpha,\lambda,x})] \quad \text{for } Z_{\alpha,\lambda,x} \sim U_\alpha^\lambda(\delta_x).$$

In other words

$$\left(U_\alpha^\lambda f\right)(x) = E[Z_{\log(1/\alpha)}|Z_0 = x], \quad x \in \mathbb{N}_0.$$

The reversibility of $\{Z_t\}$ with respect to $\mathrm{Po}(\lambda)$ implies that U_α^λ is a self-adjoint linear operator on L^2, therefore, its eigenvectors are orthogonal functions. In this context, we introduce the Poisson-Charlier family of orthogonal polynomials C_k^λ; see [8] for a broad introduction.

Definition 5.2: For given λ, the Poisson-Charlier polynomial of order k is given by

$$C_k^\lambda(x) = \frac{1}{(\lambda^k k!)^{1/2}} \sum_{\ell=0}^{k} (-\lambda)^{k-\ell} \binom{k}{\ell} x^{\underline{\ell}}.$$

Some well-known properties of the Poisson-Charlier polynomials are listed in the following lemma without proof. Note that their exact form depends on the chosen normalization; other authors present similar results, but with different normalizations.

Lemma 5.1: For any λ, μ, k and ℓ:

1)
$$\langle C_k^\lambda, C_\ell^\lambda\rangle = \delta_{k\ell} \quad (21)$$

2)
$$C_{k+1}^\lambda(x) = \frac{xC_k^\lambda(x-1) - \lambda C_k^\lambda(x)}{(\lambda(k+1))^{1/2}} \quad (22)$$

3)
$$C_k^\lambda(x+1) - C_k^\lambda(x) = \left(\frac{k}{\lambda}\right)^{1/2} C_{k-1}^\lambda(x) \quad (23)$$

4)
$$C_k^{\lambda+\mu}(x+y) = \sum_{\ell=0}^{k}\left[\left(\binom{k}{\ell}\alpha^\ell(1-\alpha)^{k-\ell}\right)^{1/2} C_\ell^\lambda(x) C_{k-\ell}^\mu(y)\right] \quad (24)$$

where $\alpha = \lambda/(\lambda+\mu)$.

Observe that, since the Poisson-Charlier polynomials form an orthonormal set, any function $f \in L^2$ can be expanded as

$$f(x) = \sum_{k=0}^{\infty} \left\langle f, C_k^\lambda\right\rangle C_k^\lambda(x). \quad (25)$$

It will be convenient to be able to translate between factorial moments and the "Poisson-Charlier moments," $E\left[C_k^\lambda(X)\right]$. For example, if $X \sim \mathrm{Po}(\lambda)$, then taking $\ell = 0$ in (21) shows that $E[C_k^\lambda(X)] = 0$ for all $k \geq 1$. More generally, the following proposition shows that the role of the Poisson-Charlier moments with respect to the Markov chain $\{Z_t\}$ is analogous to the role played by the factorial moments with respect to the pure thinning operation; cf. Lemma 2.1. Its proof, given in the Appendix, is similar to that of Lemma 2.1.

Proposition 5.2: Let $X \sim P$ be a random variable with mean λ and write $X_{\alpha,\lambda}$ for a random variable with distribution $U_\alpha^\lambda(P)$. Then

$$E[C_k^\lambda(X_{\alpha,\lambda})] = \alpha^k E[C_k^\lambda(X)].$$

If we replace α by $\exp(-t)$ and assume that the thinning Markov chain $\{Z_t\}$ has initial distribution $Z_0 \sim P$ with mean λ, then, Proposition 5.2 states that

$$E[C_k^\lambda(Z_t)] = e^{-kt} E[C_k^\lambda(Z_0)]$$

that is, the Poisson-Charlier moments of Z_t tend to zero like $\exp(-kt)E[C_k^\lambda(Z_0)]$. Similarly, expanding any function $f \in L^2$ in terms of Poisson-Charlier polynomials, $f(x) = \sum_{k=0}^\infty \langle f, C_k^\lambda\rangle C_k^\lambda(x)$, and using Proposition 5.2

$$E[f(Z_t)] = E\left[\sum_{k=0}^\infty \langle f, C_k^\lambda\rangle C_k^\lambda(Z_t)\right] = \sum_{k=0}^\infty \exp(-kt)\langle f, C_k^\lambda\rangle E\left[C_k^\lambda(X)\right].$$

Thus, the rate of convergence of $\{Z_t\}$ will be dominated by the term corresponding to $E\left[C_\kappa^\lambda(X)\right]$, where κ is the first $k \geq 1$ such that $E\left[C_k^\lambda(X)\right] \neq 0$.

The following proposition (proved in the Appendix) will be used in the proof of Theorem 5.1, which shows that this is indeed the right rate in terms of the χ^2 distance. Note that there is no restriction on the mean of $X \sim P$ in the proposition.

Proposition 5.3: If $X \sim P$ is Poisson bounded, then the likelihood ratio $P/\mathrm{Po}(\lambda)$ can be expanded as

$$\frac{P(x)}{\mathrm{Po}(\lambda,x)} = \sum_{k=0}^\infty E\left[C_k^\lambda(X)\right] C_k^\lambda(x), \quad x \geq 0.$$

Assuming $X \sim P \in PB(\lambda)$, combining Propositions 5.2 and 5.3, we obtain that

$$\begin{aligned}\frac{U_\alpha^\lambda P(x)}{\mathrm{Po}(\lambda,x)} &= \sum_{k=0}^\infty E\left[C_k^\lambda(X_{\alpha,\lambda})\right] C_k^\lambda(x)\\ &= 1 + \sum_{k=\kappa}^\infty \alpha^k E\left[C_k^\lambda(X)\right] C_k^\lambda(x)\\ &= 1 + \alpha^\kappa \sum_{k=\kappa}^\infty \alpha^{k-\kappa} E\left[C_k^\lambda(X)\right] C_k^\lambda(x) \quad (26)\end{aligned}$$

where, as before, κ denotes the first integer $k \geq 1$ such that $E\left[C_k^\lambda(X)\right] \neq 0$. This sum can be viewed as a discrete analog of the well-known Edgeworth expansion for the distribution of a continuous random variable. A technical disadvantage of both this and the standard Edgeworth expansion is that, although the sum converges in L^2, truncating it to a finite number of terms in general produces an expression which may take negative values. By a more detailed analysis we shall see in the following two sections how to get around this problem.

For now, we determine the rate of convergence of $U_\alpha^\lambda P$ to $\mathrm{Po}(\lambda)$ in terms of the χ^2 distance between $U_\alpha^\lambda P$ and $\mathrm{Po}(\lambda)$; recall the definition of the χ^2 distance between two probability distributions P and Q on $\mathbb{N}_0$

$$\chi^2(P,Q) = \sum_{x=0}^\infty Q(x)\left(\frac{P(x)}{Q(x)} - 1\right)^2.$$

Theorem 5.1: If $X \sim P$ is Poisson bounded, then the distance $\chi^2\left(U_\alpha^\lambda P, \mathrm{Po}(\lambda)\right)$ is finite for all $\alpha \in [0,1]$ and

$$\frac{\chi^2\left(U_\alpha^\lambda P, \mathrm{Po}(\lambda)\right)}{\alpha^{2\kappa}} \to E\left[C_\kappa^\lambda(X)\right]^2, \quad \text{as } \alpha \downarrow 0$$

where κ denotes the smallest $k > 0$ such that $E\left[C_k^\lambda(X)\right] \neq 0$.

Proof: The proof is based on a Hilbert space argument using the fact that the Poisson-Charlier polynomials are orthogonal. Suppose $X \sim P \in PB(\mu)$. Using Proposition 5.3

$$\begin{aligned}&\chi^2\left(U_\alpha^\lambda P, \mathrm{Po}(\lambda)\right)\\ &= \sum_{x=0}^\infty \mathrm{Po}(\lambda,x)\left(\frac{U_\alpha^\lambda P(x)}{\mathrm{Po}(\lambda,x)} - 1\right)^2\\ &= \sum_{x=0}^\infty \mathrm{Po}(\lambda,x)\left(\sum_{k=\kappa}^\infty \alpha^k E\left[C_k^\lambda(X)\right] C_k^\lambda(x)\right)^2\\ &= \sum_{k=\kappa}^\infty \alpha^{2k} E\left[C_k^\lambda(X)\right]^2\end{aligned}$$

where the last step follows from the orthogonality relation (22). For $\alpha = 1$ we have

$$\begin{aligned}\chi^2(P, \mathrm{Po}(\lambda)) &= \sum_{x=0}^\infty \mathrm{Po}(\lambda,x)\left(\frac{P(x)}{\mathrm{Po}(\lambda,x)} - 1\right)^2\\ &= \sum_{x=0}^\infty \mathrm{Po}(\lambda,x)\left(\frac{P(x)}{\mathrm{Po}(\lambda,x)}\right)^2 - 1\end{aligned}$$

which is finite. From the previous expansion we see that $\chi^2\left(U_\alpha^\lambda P, \mathrm{Po}(\lambda)\right)$ is increasing in α, which implies the finiteness claim. Moreover, that expansion has $\alpha^{2\kappa} E\left[C_\kappa^\lambda(X)\right]^2$ as its dominant term, implying the stated limit. ∎

Theorem 5.1 readily leads to upper bounds on the rate of convergence in terms of information divergence via the standard bound

$$D(P\|Q) \leq \log(1 + \chi^2(P,Q)) \leq \chi^2(P,Q)$$

which follows from direct applications of Jensen's inequality. Furthermore, replacing this bound by the well-known approximation [11]

$$D(P\|Q) \approx \frac{1}{2}\chi^2(P,Q)$$

gives the estimate

$$D\left(U_\alpha^\lambda P \| \mathrm{Po}(\lambda)\right) \approx \alpha^{2\kappa}\frac{E\left[C_\kappa^\lambda(X)\right]^2}{2} = \frac{E\left[C_\kappa^\lambda\left(U_\alpha^\lambda X\right)\right]^2}{2}.$$

We shall later prove that, in certain cases, this approximation can indeed be rigorously justified.

VI. The Rate of Convergence in the Strong Law of Thin Numbers

Let $X \sim P$ be a random variable on $\mathbb{N}_0$ with mean λ. In Theorem 3.3 we showed that, if $D(P\|\mathrm{Po}(\lambda))$ is finite, then

$$D(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)) \to 0, \quad \text{as } n \to \infty. \quad (27)$$

If P also has finite variance σ^2, then Proposition 4.1 implies that, for all $n \geq 2$

$$D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right) \leq \frac{\sigma^2}{n\lambda} + \frac{1}{n^2} \quad (28)$$

suggesting a convergence rate of order $1/n$. In this section, we prove more precise upper bounds on the rate of convergence in the strong law of thin numbers (27). For example, if X is an ultra bounded random variable with $\sigma^2 \neq \lambda$, then we show that in fact

$$\limsup_{n\to\infty} n^2 D\left(T_{1/n}(P^{*n})||\mathrm{Po}(\lambda)\right) \leq 2c^2$$

where $c = E\left[C_2^\lambda(X)\right] = (\sigma^2 - \lambda)/(\lambda\sqrt{2}) \neq 0$. This follows from the more general result of Corollary 6.1; its proof is based on a detailed analysis of the scaled Fisher information introduced in [28]. We begin by briefly reviewing some properties of the scaled Fisher information.

Definition 6.1: The scaled Fisher information of a random variable $X \sim P$ with mean λ, is defined by

$$K(X) = K(P) = \lambda E\left[\rho_X(X)^2\right]$$

where ρ_X denotes the scaled score function

$$\rho_X(x) = \frac{(x+1)P(x+1)}{\lambda P(x)} - 1.$$

In [28, Prop. 2] it was shown, using a logarithmic Sobolev inequality of Bobkov and Ledoux [5], that for any $X \sim P$

$$D(P||\mathrm{Po}(\lambda)) \leq K(X) \tag{29}$$

under mild conditions on the support of P. Also, [28, Prop. 3] states that $K(X)$ satisfies a subadditivity property: For independent random variables $X_1, X_2, \ldots, X_n$

$$K\left(\sum_{i=1}^n X_i\right) \leq \sum_{i=1}^n \frac{E[X_i]}{\lambda} K(X_i) \tag{30}$$

where $\lambda = \sum_i E(X_i)$. In particular, recalling that the thinning of a convolution is the convolution of the corresponding thinnings, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean λ then the bounds in (29) and (30) imply

$$D\left(T_{1/n}(P^{*n})||\mathrm{Po}(\lambda)\right) \leq K(T_{1/n}(P)). \tag{31}$$

Therefore, our next goal is to determine the rate at which $K(T_\alpha(X))$ tends to 0 for α tending to 0. We begin with the following proposition; its proof is given in Appendix.

Proposition 6.1: If $X \sim P$ is Poisson bounded, then P admits the representation

$$P(x) = \frac{1}{x!}\sum_{\ell=0}^{\infty}(-1)^\ell \frac{E[X^{\underline{x+\ell}}]}{\ell!}.$$

Moreover, the truncated sum from $\ell = 0$ to m is an upper bound for $P(x)$ if m is even, and a lower bound if m is odd.

An important consequence of this proposition is that the probability $T_\alpha P(x)$ tends to zero like α^x, as $\alpha \downarrow 0$. Moreover, it leads to the following asymptotic result for the scaled Fisher information, also proved in theAppendix.

Theorem 6.1: Suppose $X \sim P$ has mean λ and it is ultra bounded with ratio λ. Let κ denote the smallest integer $k \geq 1$ such that $E\left[C_k^\lambda(X)\right] \neq 0$. Then

$$\lim_{\alpha\to 0} \frac{K(T_\alpha P)}{\alpha^\kappa} = \kappa c^2$$

where $c = E\left[C_\kappa^\lambda(X)\right]$.

Combining Theorem 6.1 with (31) immediately yields the following.

Corollary 6.1: Suppose $X \sim P$ has mean λ and it is ultra bounded with ratio λ. Let κ denote the smallest integer $k \geq 1$ such that $E\left[C_k^\lambda(X)\right] \neq 0$. Then

$$\limsup_{n\to\infty} n^\kappa D(T_{1/n}(P^{*n})||\mathrm{Po}(\lambda)) \leq \kappa c^2$$

where $c = E\left[C_\kappa^\lambda(X)\right]$.

VII. Monotonicity Results for the Scaled Fisher Information

In this section we establish a finer result for the behavior of the scaled Fisher information upon thinning, and use that to deduce a stronger finite-n upper bound for the strong law of thin numbers. Specifically, if $X \sim P$ is ULC with mean λ, and X_α denotes a random variable with distribution $T_\alpha P$, we will show that $K(X_\alpha) \leq \alpha^2 K(X)$. This implies that, for all ULC random variables X, we have the following finite-n version of the strong law of thin numbers

$$D\left(T_{1/n}(P^{*n})||\mathrm{Po}(\lambda)\right) \leq \frac{K(X)}{n^2}.$$

Note that, unlike the more general result in (28) which gives a bound of order $1/n$, the above bound is of order $1/n^2$, as long as X is ULC.

The key observation for these results is in the following lemma.

Lemma 7.1: Suppose X is a ULC random variable with distribution P and mean λ. For any $\alpha \in (0,1)$, write X_α for a random variable with distribution $T_\alpha P$. Then the derivative of $K(X_\alpha)/\alpha$ with respect to α satisfies

$$\frac{\partial}{\partial\alpha}\left(\frac{K(X_\alpha)}{\alpha}\right) = \frac{1}{\alpha^2} S(X_\alpha),$$

where, for a random variable Y with mass function Q and mean μ, we define

$$S(Y) = \sum_{y=0}^{\infty}\left[\frac{Q(y+1)(y+1)}{\mu Q(y)}\left(\frac{Q(y+1)(y+1)}{Q(y)} - \frac{Q(y+2)(y+2)}{Q(y+1)}\right)^2\right].$$

Proof: This result follows on using the expression for the derivative of $T_\alpha P$ arising as the case $f(\alpha) = g(\alpha) = 0$ in [24, Prop. 3.6], that is

$$\frac{\partial}{\partial\alpha}(T_\alpha P)(x) = \frac{1}{\alpha}[x(T_\alpha P)(x) - (x+1)(T_\alpha P)(x+1)].$$

Using this, for each x we deduce that the derivative

$$\frac{\partial}{\partial\alpha}\left(\frac{((T_\alpha P)(x+1))^2(x+1)^2}{\alpha^2(T_\alpha P)(x)\lambda}\right)$$

can be expressed as the sum of

$$\frac{(T_\alpha P)(x+1)(x+1)}{\alpha^3\lambda}\left(\frac{(T_\alpha P)(x+1)(x+1)}{(T_\alpha P)(x)} - \frac{(T_\alpha P)(x+2)(x+2)}{(T_\alpha P)(x+1)}\right)^2$$

plus

$$\frac{1}{\alpha^3\lambda}\left(\frac{((T_\alpha P)(x+1))^2(x+1)^2 x}{(T_\alpha P)(x)} - \frac{((T_\alpha P)(x+2))^2(x+2)^2(x+1)}{(T_\alpha P)(x+1)}\right).$$

The result follows (with the term-by-term differentiation of the infinite sum justified) if the sum of these terms in x is absolutely convergent. The first terms are positive, and their sum is absolutely convergent to S by assumption. The second terms form a collapsing sum, which is absolutely convergent assuming that

$$\sum_{x=0}^{\infty}\frac{((T_\alpha P)(x+1))^2(x+1)^2 x}{(T_\alpha P)(x)} < \infty.$$

Note that, for any ULC distribution Q, by definition we have for all x, $(x+1)Q(x+1)/Q(x) \le xQ(x)/Q(x-1)$, so that the above sum is bounded above by,

$$\frac{(T_\alpha P)(1)}{(T_\alpha P)(0)}\sum_x (T_\alpha P)(x+1)(x+1)x,$$

which is finite by Proposition 2.2. ∎

We now deduce the following theorem, which parallels, respectively, [41, Th. 8], where a corresponding result is proved for the information divergence.

Theorem 7.1: Let $X \sim P$ be a ULC random variable with mean λ. Write X_α for a random variable with distribution $T_\alpha P$.

(i)

$$K(X_\alpha) \le \alpha^2 K(X), \quad \alpha \in (0,1); \tag{32}$$

(ii)

$$D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right) \le \frac{K(X)}{n^2}, \quad n \ge 2. \tag{33}$$

Proof: The first part follows from the observation that $K(T_\alpha X)/\alpha^2$ is increasing in α, since, by Lemma 7.1, its derivative is $(S(T_\alpha X) - K(T_\alpha X))/\alpha^3$. Taking $g(y) = P(y+1)(y+1)/P(y)$ in the more technical Lemma 7.2 below, we deduce that $S(Y) \ge K(Y)$ for any random variable Y, and this proves (i). Then (ii) immediately follows from (i) combined with the earlier bound (31), upon recalling that thinning preserves the ULC property [24]. ∎

Consider the finite difference operator Δ defined by

$$(\Delta g)(x) = g(x+1) - g(x)$$

for functions $g : \mathbb{N}_0 \to \mathbb{R}$. We require a result suggested by relevant results in [6], [27]. Its proof is given in the Appendix.

Lemma 7.2: Let Y be ULC random variable with distribution P on $\mathbb{N}_0$. Then for any function g, defining $\mu = \sum_y P(y)g(y)$,

$$\sum_{y=0}^{\infty} P(y)(g(y)-\mu)^2 \le \sum_{y=0}^{\infty} P(y+1)(y+1)\Delta g(y)^2.$$

VIII. Bounds in Total Variation

In this section, we show that a modified version of the argument used in the proof of Proposition 4.1 gives an upper bound to the rate of convergence in the weak law of small numbers. If $X \sim P$ has mean λ and variance σ^2, then combining the bound (15) of Proposition 4.1 with Pinsker's inequality we obtain

$$\left\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\right\| \le \left(\frac{1}{2n^2(1-n^{-1})} + \frac{\sigma^2}{n\lambda}\right)^{1/2} \tag{34}$$

which gives an upper bound of order $n^{-1/2}$. From the asymptotic upper bound on information divergence, Corollary 6.1, we know that one should be able to obtain upper bounds of order n^{-1}. Here we derive an upper bound on total variation using the same technique used in the proof of Proposition 4.1.

Theorem 8.1: Let P be a distribution on $\mathbb{N}_0$ with finite mean λ and variance σ^2. Then

$$\left\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\right\| \le \frac{1}{n2^{1/2}} + \frac{\sigma}{n^{1/2}}\min\left\{1, \frac{1}{2\lambda^{1/2}}\right\}$$

for all $n \ge 2$.

The proof uses the following simple bound, which follows easily from a result of Yannaros, [40, Theorem 2.3]; the details are omitted.

Lemma 8.1: For any $\lambda > 0$, $m \ge 1$ and $t \in (0, 1/2]$, we have,

$$\|\mathrm{Bin}(m,t) - \mathrm{Po}(\lambda)\| \le t2^{-1/2} + |mt-\lambda|\min\left\{1, \frac{1}{2\lambda^{1/2}}\right\}.$$

Proof: The first inequality in the proof of Proposition 4.1 remains valid due to the convexity of the total variation norm (since it is an f-divergence). The next equality becomes an inequality triangle, and it is justified by the triangle, and we have

$$\begin{aligned}
&\left\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\right\| \\
&= \frac{1}{2}\sum_{x\ge 0}\left|\sum_{y\ge 0} P^{*n}(y)\left[\Pr\{\mathrm{Bin}(y,1/n) = x\} - \mathrm{Po}(\lambda,x)\right]\right| \\
&\le \sum_{y\ge 0} P^{*n}(y)\frac{1}{2}\sum_x \left|\Pr\{\mathrm{Bin}(y,1/n) = x\} - \mathrm{Po}(\lambda,x)\right| \\
&= \sum_{y\ge 0} P^{*n}(y)\|\mathrm{Bin}(y,1/n) - \mathrm{Po}(\lambda)\|.
\end{aligned}$$

And using Lemma 8.1 leads to

$$\begin{aligned}&\left\|T_{1/n}(P^{*n}) - \mathrm{Po}(\lambda)\right\| \\ &\le \sum_{y\ge 0} P^{*n}(y)\|\mathrm{Bin}(y,1/n) - \mathrm{Po}(\lambda)\|. \\ &= \sum_{y\ge 0} P^{*n}(y)\left(\frac{1}{n2^{1/2}} + \left|\frac{y}{n} - \lambda\right| \min\left\{1, \frac{1}{2\lambda^{1/2}}\right\}\right)\end{aligned}$$

and the result follows by an application of Hölder's inequality. ■

IX. Compound Thinning

There is a natural generalization of the thinning operation, via a process which closely parallels the generalization of the Poisson distribution to the compound Poisson. Starting with a random variable $Y \sim P$ with values in $\mathbb{N}_0$, the α-thinned version of Y is obtained by writing $Y = 1+1+\cdots+1$ (Y times), and then keeping each of these 1s with probability α, independently of all the others; cf. (1) above.

More generally, we choose and fix a "compounding" distribution Q on $\mathbb{N} = \{1, 2, \ldots\}$. Given $Y \sim P$ on $\mathbb{N}_0$ and $\alpha \in [0,1]$, then the compound α-thinned version of Y with respect to Q, or, for short, the (α, Q)-thinned version of Y, is the random variable which results from first thinning Y as above and then replacing of the 1s that are kept by an independent random sample from Q

$$\sum_{n=1}^{Y} B_n\xi_n, \quad B_i \sim \text{i.i.d. } \mathrm{Bern}(\alpha), \quad \xi_i \sim \text{i.i.d. } Q \tag{35}$$

where all the random variables involved are independent. For fixed α and Q, we write $T_{\alpha,Q}(P)$ for the distribution of the (α, Q)-thinned version of $Y \sim P$. Then $T_{\alpha,Q}(P)$ can be expressed as a mixture of "compound binomials" in the same way as $T_\alpha(P)$ is a mixture of binomials. The compound binomial distribution with parameters n, α, Q, denoted $\mathrm{CBin}(n, \alpha, Q)$, is the distribution of the sum of n i.i.d. random variables, each of which is the product of a $\mathrm{Bern}(\alpha)$ random variable and an independent $\xi \sim Q$ random variable. In other words, it is the (α, Q)-thinned version of the point mass at n, i.e., the distribution of (35) with $Y = n$ w.p.1. Then we can express the probabilities of the (α, Q)-thinned version of P as, $T_{\alpha,Q}(P)(k) = \sum_{\ell>k} P(\ell)\Pr\{\mathrm{CBin}(\ell, \alpha, Q) = k\}$.

The following two observations are immediate from the definitions.

1) Compound thinning maps a Bernoulli sum into a compound Bernoulli sum: If P is the distribution of the Bernoulli sum $\sum_{i=1}^n B_i$ where the B_i are independent $\mathrm{Bern}(p_i)$, then $T_{\alpha,Q}(P)$ is the distribution of the "compound Bernoulli sum," $\sum_{i=1}^n B_i'\xi_i$ where the B_i' are independent $\mathrm{Bern}(\alpha p_i)$, and the ξ_i are i.i.d. with distribution Q, independent of the B_i.
2) Compound thinning maps the Poisson to the compound Poisson distribution, that is, $T_{\alpha,Q}(\mathrm{Po}(\lambda)) = \mathrm{CPo}(\alpha\lambda, Q)$, the compound Poisson distribution with rate $\alpha\lambda$ and compounding distribution Q. Recall that $\mathrm{CPo}(\lambda, Q)$ is defined as the distribution of

$$\sum_{i=1}^{\Pi_\lambda} \xi_i$$

where the ξ_i are as before, and Π_λ is a $\mathrm{Po}(\lambda)$ random variable that is independent of the ξ_i.

Perhaps the most natural way in which the compound Poisson distribution arises is as the limit of compound binomials. That is, $\mathrm{CBin}(n, \lambda/n, Q) \to \mathrm{CPo}(\lambda, Q)$, as $n \to \infty$, or, equivalently

$$T_{1/n,Q}(\mathrm{Bin}(n,\lambda)) = T_{1/n,Q}(P^{*n}) \to \mathrm{CPo}(\lambda, Q)$$

where P denotes the $\mathrm{Bern}(\lambda)$ distribution.

As with the strong law of thin numbers, this result remains true for general distributions P, and the convergence can be established in the sense of information divergence.

Theorem 9.1: Let P be a distribution on $\mathbb{N}_0$ with mean $\lambda > 0$ and finite variance σ^2. Then, for any probability measure Q on $\mathbb{N}$

$$D\left(T_{1/n,Q}(P^{*n})\|\mathrm{CPo}(\lambda, Q)\right) \to 0, \quad \text{as } n \to \infty$$

as long as $D(P\|\mathrm{Po}(\lambda)) < \infty$.

The proof is very similar to that of Theorem 3.3 and thus omitted. In fact, the same argument as that proof works for non-integer-valued compounding. That is, if Q is an *arbitrary* probability measure on $\mathbb{R}^d$, then compound thinning a $\mathbb{N}_0$-valued random variable $Y \sim P$ as in (35) gives a probability measure $T_{\alpha,Q}(P)$ on $\mathbb{R}^d$.

It is somewhat remarkable that the statement *and* proof of most of our results concerning the information divergence remain essentially unchanged in this case. For example, we easily obtain the following analog of Proposition 4.1.

Proposition 9.1: If P is a distribution on $\mathbb{N}_0$ with mean λ/α and variance $\sigma^2 < \infty$, for some $\alpha \in (0,1)$, then, for any probability measure Q on $\mathbb{R}^d$

$$D(T_{\alpha,Q}(P)\|\mathrm{CPo}(\lambda, Q)) \le \alpha^2\left(\frac{1}{2(1-\alpha)} + \frac{\sigma^2}{\lambda}\right).$$

The details of the argument of the proof of the proposition are straightforward extensions of the corresponding proof of Proposition 4.1.

Appendix

Proof of Lemma 2.1: Simply apply Lemma 2.2 to Definition 1.1 with $Y \sim P$, to obtain

$$\begin{aligned}E\left[Y_\alpha^{\underline{k}}\right] &= E\left[\left(\sum_{x=1}^{Y} B_x\right)^{\underline{k}}\right] \\ &= E\left\{E\left[\left(\sum_{x=1}^{Y} B_x\right)^{\underline{k}} \middle| Y\right]\right\} \\ &= E\left\{E\left[\sum_{k_x\in\{0,1\}, \sum k_x = k} k! \prod_{x=1}^{Y} B^{\underline{k_x}}_x \middle| Y\right]\right\} \\ &= E\left[\binom{Y}{k} k!\alpha^k\right] \\ &= \alpha^k E[Y^{\underline{k}}]\end{aligned}$$

using the fact that the sequence of factorial moments of the $\mathrm{Bern}(\alpha)$ distribution are $\{1, \alpha, 0, 0, \ldots\}$. ∎

Proof of Proposition 2.1: Assume that $T_{\alpha_0}P = T_{\alpha_0}Q$ for a given $\alpha_0 > 0$. Then, recalling the property stated in (6), it follows that, $T_\alpha P = T_\alpha Q$ for all $\alpha \in [0, \alpha_0]$. In particular, $T_\alpha P(0) = T_\alpha Q(0)$ for all $\alpha \in [0, \alpha_0]$, i.e.

$$\sum_{x=0}^{\infty} P(x)(1-\alpha)^x = \sum_{x=0}^{\infty} Q(x)(1-\alpha)^x$$

for all $\alpha \in [0, \alpha_0]$, which is only possible if $P(x) = Q(x)$ for all $x \geq 0$. ∎

Proof of Proposition 2.2: Note that the expectation

$$\sum_{x=0}^{\infty} P(x) x^{\underline{k}} \left(\frac{(x+1)P(x+1)}{\lambda P(x)} - 1 \right) \leq 0$$

by the Chebyshev rearrangement lemma, since it is the covariance between an increasing and a decreasing function. Rearranging this inequality gives

$$\begin{aligned} E[X^{\underline{k+1}}] &= \sum_{x=0}^{\infty} P(x+1)(x+1)^{\underline{k+1}} \leq \lambda \sum_{x=0}^{\infty} P(x) x^{\underline{k}} \\ &= \lambda E[X^{\underline{k}}] \end{aligned}$$

as required. ∎

Proof of Proposition 2.3: For part (a), using Lemma 2.2, we have

$$\begin{aligned} E[(X+Y)^{\underline{k}}] &= E\left[\sum_{\ell=0}^{k} \binom{k}{\ell} X^{\underline{\ell}} Y^{\underline{k-\ell}} \right] \\ &= \sum_{\ell=0}^{k} \binom{k}{\ell} E[X^{\underline{\ell}}] E[Y^{\underline{k-\ell}}] \\ &\leq \sum_{\ell=0}^{k} \binom{k}{\ell} \lambda^\ell \mu^{k-\ell} \\ &= (\lambda+\mu)^k. \end{aligned}$$

It is straightforward to check, using Lemma 2.1, that $T_\alpha P \in PB(\alpha\lambda)$.

To prove part (b), using Lemma 2.2, Pascal's identity and relabelling, yields

$$\begin{aligned} &E[(X+Y)^{\underline{k+1}}] \\ &= E\left[\sum_{\ell=0}^{k+1} \binom{k+1}{\ell} X^{\underline{\ell}} Y^{\underline{k+1-\ell}} \right] \\ &= E\left[\sum_{\ell=0}^{k+1} \left(\binom{k}{\ell-1} + \binom{k}{\ell} \right) X^{\underline{\ell}} Y^{\underline{k+1-\ell}} \right] \\ &= \sum_{\ell=0}^{k+1} \binom{k}{\ell-1} E\left[X^{\underline{\ell}} Y^{\underline{k+1-\ell}} \right] + \sum_{\ell=0}^{k+1} \binom{k}{\ell} E\left[X^{\underline{\ell}} Y^{\underline{k+1-\ell}} \right] \\ &= \sum_{\ell=0}^{k} \binom{k}{\ell} E\left[X^{\underline{\ell+1}} \right] E\left[Y^{\underline{k-\ell}} \right] \\ &\quad + \sum_{\ell=0}^{k} \binom{k}{\ell} E\left[X^{\underline{\ell}} \right] E\left[Y^{\underline{k+1-\ell}} \right] \\ &\leq \sum_{\ell=0}^{k} \binom{k}{\ell} \lambda E\left[X^{\underline{\ell}} \right] E\left[Y^{\underline{k-\ell}} \right] \\ &\quad + \sum_{\ell=0}^{k} \binom{k}{\ell} E\left[X^{\underline{\ell}} \right] \mu E\left[Y^{\underline{k-\ell}} \right] \\ &= (\lambda+\mu) E[(X+Y)^{\underline{k}}]. \end{aligned}$$

The second property is easily checked using Lemma 2.1. ∎

Proof of Theorem 3.3: In order to apply Proposition 3.1 with P^{*n} in place of P and $\alpha = 1/n$, we need to check that $D(P^{*n}\|\mathrm{Po}(n\lambda))$ is finite. Let S_n denote the sum of n i.i.d. random variables $X_i \sim P$, so that P^{*n} is the distribution of S_n. Similarly, $\mathrm{Po}(n\lambda)$ is the sum of n independent $\mathrm{Po}(\lambda)$ variables. Therefore, using the data-processing inequality [11] as in [28] implies that $D(P^{*n}\|\mathrm{Po}(n\lambda)) \leq nD(P\|\mathrm{Po}(\lambda))$, which is finite by assumption.

Proposition 3.1 gives

$$\begin{aligned} D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right) &\leq \frac{1}{2n^2(1-1/n)} \\ &\quad + E[(S_n/n)\log(S_n/n)] - \lambda\log\lambda. \end{aligned}$$

The law of large numbers implies that, $S_n/n \to \lambda$ a.s., so $(S_n/n)\log(S_n/n) \to \lambda\log\lambda$ a.s., as $n \to \infty$. Therefore, to complete the proof it suffices to show that $(S_n/n)\log(S_n/n)$ converges to $\lambda\log\lambda$ also in L^1, or, equivalently, that the sequence $\{T_n = (S_n/n)\log(S_n/n)\}$ is uniformly integrable. We will actually show that the nonnegative random variables T_n are bounded above by a different uniformly integrable sequence. Indeed, by the log-sum inequality

$$\begin{aligned} T_n &= \sum_{i=1}^{n} \frac{X_i}{n} \log\left(\frac{\sum_{i=1}^{n} \frac{X_i}{n}}{\sum_{i=1}^{n} \frac{1}{n}} \right) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} X_i \log X_i. \end{aligned} \tag{36}$$

Arguing as in the beginning of the proof of Proposition 3.1 shows that the mean $\mu = E[X_i \log X_i]$ is finite, so the law of large numbers implies that the averages in (36) converge to μ a.s. and in L^1. Hence, they form a uniformly integrable sequence; this implies that the T_n are also uniformly integrable, completing the proof. ∎

Proof of Theorem 4.2: The proof is similar to that of Theorem 3.3, so some details are omitted. For each $n \geq 1$, let $\lambda^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i$ and write $S_n = \sum_{i=1}^{n} X_i$, where the random variables X_i are independent, with each $X_i \sim P_i$.

First, to see that $D\left(P^{(n)}\|\mathrm{Po}(n\lambda^{(n)})\right)$ is finite, applying the data-processing inequality [11] as in [28] gives $D(P^{(n)}\|\mathrm{Po}(n\lambda^{(n)})) \leq \sum_{i=1}^{n} D(P_i\|\mathrm{Po}(\lambda_i))$, and it is easy to check that each of these terms is finite because all P_i have finite second moments. As before, Proposition 3.1 gives

$$\begin{aligned} D\left(T_{1/n}(P^{(n)})\|\mathrm{Po}(\lambda^{(n)})\right) &\leq \frac{1}{2n^2(1-1/n)} \\ &\quad + E[(S_n/n)\log(S_n/n)] - \lambda^{(n)}\log\lambda^{(n)}. \end{aligned} \tag{37}$$

Letting $Y_i = X_i - \lambda_i$ for each i, the independent random variables Y_i have zero mean and

$$\sum_{i=1}^{\infty} \frac{1}{i^2} E(Y_i^2) = \sum_{i=1}^{\infty} \frac{1}{i^2} \mathrm{Var}(X_i^2) \leq \sum_{i=1}^{\infty} \frac{1}{i^2} E(X_i^2)$$

which is finite by assumption (b). Then, by the general version of the law of large numbers on [14, p. 239] $\frac{1}{n}\sum_{i=1}^{n} Y_i \to 0$, a.s., and hence, by assumption (a), $S_n/n \to \lambda$ a.s., so that also, $(S_n/n)\log(S_n/n) \to \lambda \log \lambda$ a.s., as $n \to \infty$. Moreover, since $(x \log x)^{4/3} \leq x^2$ for every integer $x \geq 1$, we have

$$\begin{aligned} E&\left\{ \left(\frac{S_n}{n} \log \frac{S_n}{n} \right)^{4/3} \right\} \\ &\leq E\left\{ \left(\frac{S_n}{n} \right)^2 \right\} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} E(X_i^2) + \frac{1}{n^2} \sum_{1 \leq i \neq j \leq n} E(X_i) E(X_j) \\ &\leq \sum_{i=1}^{n} \frac{1}{i^2} E(X_i^2) + (\lambda^{(n)})^2 \end{aligned}$$

which is uniformly bounded over n by our assumptions. Therefore, the sequence $\{(S_n/n)\log(S_n/n)\}$ is bounded in L^p with $p = 4/3 > 1$, which implies that it is uniformly integrable, therefore it converges to $\lambda \log \lambda$ also in L^1, so that, $D\left(T_{1/n}(P^{(n)}) \| \mathrm{Po}(\lambda^{(n)})\right) \to 0$ as $n \to \infty$.

Finally, recalling once more that the Poisson measures form an exponential family, they satisfy a Pythagorean identity [11], so that

$$\begin{aligned} D&\left(T_{1/n}(P^{(n)}) \| \mathrm{Po}(\lambda)\right) \\ &= D\left(T_{1/n}(P^{(n)}) \| \mathrm{Po}(\lambda^{(n)})\right) + D\left(\mathrm{Po}(\lambda^{(n)}) \| \mathrm{Po}(\lambda)\right) \end{aligned}$$

where the first term was just shown to go to zero as $n \to \infty$, and the second term is actually equal to

$$\lambda^{(n)} \log \frac{\lambda^{(n)}}{\lambda} + \lambda - \lambda^{(n)},$$

which also vanishes as $n \to \infty$ by assumption (a). ∎

Proof of Proposition 5.2: Let X_α and Z denote independent random variables with distributions $T_\alpha P$ and $\mathrm{Po}((1-\alpha)\lambda)$, respectively. Then from the definitions, and using Lemmas 2.2 and 2.1

$$\begin{aligned} E&\left[C_k^\lambda(X_{\alpha,\lambda})\right] \\ &= \frac{1}{(\lambda^k k!)^{1/2}} \sum_{\ell=0}^{k} (-\lambda)^{k-\ell} \binom{k}{\ell} E\left\{(X_\alpha + Z)^{\underline{\ell}}\right\} \\ &= \frac{1}{(\lambda^k k!)^{1/2}} \sum_{\ell=0}^{k} \Bigg\{ (-\lambda)^{k-\ell} \binom{k}{\ell} \sum_{m=0}^{\ell} \binom{\ell}{m} \\ &\quad \times E(X_\alpha^{\underline{m}}) E(Z^{\underline{\ell-m}}) \Bigg\} \\ &= \frac{1}{(\lambda^k k!)^{1/2}} \sum_{\ell=0}^{k} \Bigg\{ (-\lambda)^{k-\ell} \binom{k}{\ell} \sum_{m=0}^{\ell} \binom{\ell}{m} \alpha^m \\ &\quad \times E(X^{\underline{m}}) ((1-\alpha)\lambda)^{\ell-m} \Bigg\}, \end{aligned}$$

where we have used the fact that the factorial moments of a $\mathrm{Po}(t)$ random variable Z_t satisfy, $E[Z_t^{\underline{n}}] = t^n$. Simplifying and interchanging the two sums,

$$\begin{aligned} E&\left[C_k^\lambda(X_{\alpha,\lambda})\right] \\ &= \frac{1}{(\lambda^k k!)^{1/2}} \sum_{m=0}^{k} \Bigg\{ \binom{k}{m} \alpha^m E(X^{\underline{m}}) \\ &\quad \times \sum_{\ell=m}^{k} \binom{k-m}{\ell-m} (-\lambda)^{k-\ell} ((1-\alpha)\lambda)^{\underline{\ell-m}} \Bigg\} \\ &= \frac{1}{(\lambda^k k!)^{1/2}} \sum_{m=0}^{k} \binom{k}{m} \alpha^m E(X^{\underline{m}}) (-\alpha\lambda)^{k-m} \\ &= \alpha^k E[C_k^\lambda(X)], \end{aligned}$$

as claimed. ∎

Proof of Proposition 5.3: First we have to prove that $P/\mathrm{Po}(\lambda) \in L^2$. Assume P is Poisson bounded with ratio μ, say. Using the bound in Lemma 3.1

$$\begin{aligned} \sum_{x=0}^{\infty} &\mathrm{Po}(\lambda, x) \left(\frac{P(x)}{\mathrm{Po}(\lambda, x)} \right)^2 \\ &\leq \sum_{x=0}^{\infty} \mathrm{Po}(\lambda, x) \left(\frac{\mathrm{Po}(\mu, x)\, e^{\mu}}{\mathrm{Po}(\lambda, x)} \right)^2 \\ &= e^{\lambda} \sum_{x=0}^{\infty} \frac{(\mu^2/\lambda)^x}{x!} \\ &= e^{\lambda + \mu^2/\lambda} \end{aligned}$$

which is finite.

Now, recalling the general expansion (25), it suffices to show that $\left\langle P/\mathrm{Po}(\lambda), C_k^\lambda \right\rangle = E\left[C_k^\lambda(X)\right]$. Indeed, for $Z \sim \mathrm{Po}(\lambda)$

$$\left\langle \frac{P}{\mathrm{Po}(\lambda)}, C_k^\lambda \right\rangle = E\left(\frac{P(Z)}{\mathrm{Po}(\lambda, Z)} C_k^\lambda(Z) \right) = E\left[C_k^\lambda(X)\right]$$

as required. ∎

Proof of Proposition 6.1: We need the following simple lemma; for a proof see, e.g., [16].

Lemma A.1: If

$$F(m, x) = \sum_{\ell=0}^{m} \binom{x}{\ell} (-1)^\ell$$

then

$$\begin{cases} F(m, x) \geq \delta_x & \text{for } m \text{ even}, \\ F(m, x) \leq \delta_x & \text{for } m \text{ odd}. \end{cases}$$

Turning to the proof of Proposition 6.1, assume $X \sim P$ is Poisson bounded with ratio λ. Then the series in the statement converges, since

$$\frac{1}{x!} \sum_{\ell=0}^{\infty} \left| (-1)^\ell \frac{E[X^{\underline{x+\ell}}]}{\ell!} \right| \leq \frac{1}{x!} \sum_{\ell=0}^{\infty} \frac{\lambda^{x+\ell}}{\ell!} < \infty.$$

For m even we have

$$\delta_{z-x} \le \sum_{\ell=0}^{m} \binom{z-x}{\ell}(-1)^{\ell}$$

therefore

$$\binom{z}{x}\delta_{z-x} \le \sum_{\ell=0}^{m}\binom{z}{x}\binom{z-x}{\ell}(-1)^{\ell} = \frac{1}{x!}\sum_{\ell=0}^{m}(-1)^{\ell}\frac{z^{\underline{x+\ell}}}{\ell!}.$$

Multiplying by $P(z)$ and summing over z

$$\begin{aligned} P(x) &= \sum_{z=0}^{\infty} P(z)\binom{z}{x}\delta_{z-x} \\ &\le \sum_{z=0}^{\infty} P(z)\frac{1}{x!}\sum_{\ell=0}^{m}(-1)^{\ell}\frac{z^{\underline{x+\ell}}}{\ell!} \\ &= \frac{1}{x!}\sum_{\ell=0}^{m}(-1)^{\ell}\frac{E[X^{\underline{x+\ell}}]}{\ell!}. \end{aligned}$$

A similar argument holds for m odd. ∎

Proof of Theorem 6.1: Let X_α have distribution $T_\alpha P$. Using Lemma 2.1, Proposition 6.1, and the fact that X is ultra bounded, for small enough α the score function of X_α can be bounded as

$$\begin{aligned} \rho_{X_\alpha}(z) &= \frac{(z+1)T_\alpha P(z+1)}{\alpha\lambda T_\alpha P(z)} - 1 \\ &\le \frac{(z+1)E[X_\alpha^{\underline{z+1}}]/(z+1)!}{\alpha\lambda\left(E[X_\alpha^{\underline{z}}] - E[X_\alpha^{\underline{z+1}}]\right)/z!} - 1 \\ &= \frac{\alpha^{z+1}E[X^{\underline{z+1}}]}{\alpha\lambda(\alpha^{z}E[X^{\underline{z}}] - \alpha^{z+1}E[X^{\underline{z+1}}])} - 1 \\ &= \left[\frac{\lambda E[X^{\underline{z}}]}{E[X^{\underline{z+1}}]} - \lambda\alpha\right]^{-1} - 1 \\ &\le [1-\lambda\alpha]^{-1} - 1 \\ &= \frac{\alpha\lambda}{1-\alpha\lambda}. \end{aligned}$$

Since the lower bound $\rho_{X_\alpha}(z) \ge -1$ is obvious, it follows that,

$$\rho_{X_\alpha}(z)^2 \le 1, \quad \text{for all } \alpha > 0 \text{ small enough.} \tag{38}$$

We express $K(T_\alpha P)$ in three terms

$$\begin{aligned} K(T_\alpha P) = {} & \lambda\alpha\sum_{z=0}^{\kappa-2} T_\alpha P(z)\rho_{X_\alpha}(z)^2 \\ & + \lambda\alpha T_\alpha P(\kappa-1)\rho_{X_\alpha}(\kappa-1)^2 \\ & + \lambda\alpha\sum_{z=\kappa}^{\infty} T_\alpha P(z)\rho_{X_\alpha}(z)^2. \end{aligned} \tag{39}$$

For the third term note that, applying Markov's inequality to the function $f(x) = x(x-1)\cdots(x-\kappa+1)$, which increases on the integers, we obtain,

$$\Pr\{X_\alpha \ge \kappa\} \le \frac{E[X_\alpha^{\underline{\kappa}}]}{\kappa!} = \frac{\alpha^{\kappa}E[X^{\underline{\kappa}}]}{\kappa!} \le \frac{(\alpha\lambda)^{\kappa}}{\kappa!}.$$

Therefore, using this and (38), for small enough $\alpha > 0$ the third term in (39) is bounded above by

$$\alpha\lambda\frac{(\alpha\lambda)^{\kappa}}{\kappa!}$$

which, divided by α^{κ}, tends to zero as $\alpha \to 0$.

For the other two terms we use the full expansion of Proposition 6.1, together with Lemma 2.1, to obtain a more accurate expression for the score function

$$\begin{aligned} &\rho_{X_\alpha}(z) \\ &= \frac{(z+1)T_\alpha P(z+1) - \alpha\lambda T_\alpha P(z)}{\alpha\lambda T_\alpha P(z)} \\ &= \frac{\frac{1}{z!}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{E[X_\alpha^{\underline{z+1+\ell}}]}{\ell!} - \alpha\lambda\frac{1}{z!}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{E[X_\alpha^{\underline{z+\ell}}]}{\ell!}}{\alpha\lambda\frac{1}{z!}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{E[X_\alpha^{\underline{z+\ell}}]}{\ell!}} \\ &= \frac{\sum_{\ell=0}^{\infty}(-1)^{\ell}\alpha^{\ell}(E[X^{\underline{z+1+\ell}}] - \lambda E[X^{\underline{z+\ell}}])/\ell!}{\lambda\sum_{\ell=0}^{\infty}(-1)^{\ell}\alpha^{\ell}E[X^{\underline{z+\ell}}]/\ell!}. \end{aligned}$$

Since, by assumption, $E\left[X^{\underline{z+1+\ell}}\right] - \lambda E\left[X^{\underline{z+\ell}}\right] = 0$ for $z + \ell < \kappa - 1$, the first terms in the series in the numerator above vanish. Therefore, $\rho_{X_\alpha}(z)$ equals

$$\alpha^{\kappa-z-1}\frac{\sum_{\ell=0}^{\infty}(-1)^{\ell+\kappa-z-1}\frac{\alpha^{\ell}(E[X^{\underline{\ell+z}}] - \lambda E[X^{\underline{\ell+z-1}}])}{(\ell+\kappa-z-1)!}}{\lambda\sum_{\ell=0}^{\infty}(-1)^{\ell}\alpha^{\ell}E[X^{\underline{z+\ell}}]/\ell!}.$$

For $z \le \kappa - 2$, the numerator and denominator above are both bounded functions of α, and the denominator is bounded away from zero (because of the term corresponding to $\ell = 0$). Therefore, for each $0 \le z \le \kappa - 2$, the score function $\rho_{X_\alpha}(z)$ is of order $\alpha^{\kappa-z-1}$. For the first term in (39) we thus have

$$\begin{aligned} \lambda\alpha\sum_{z=0}^{\kappa-2} T_\alpha P(z)\rho_{X_\alpha}(z)^2 &= \alpha\sum_{z=0}^{\kappa-2} O(\alpha^{z})O(\alpha^{2\kappa-2z-2}) \\ &= O(\alpha^{\kappa+1}) \end{aligned}$$

which, again, when divided by α^{κ}, tends to zero as $\alpha \to 0$.

Thus only the second term in (39) contributes. For this term, we similarly obtain that the limit

$$\lim_{\alpha\to 0}\rho_{X_\alpha}(\kappa-1)$$

equals

$$\begin{aligned} &\lim_{\alpha\to 0}\frac{\sum_{j=0}^{\infty}(-1)^{j}\alpha^{j}(E[X^{\underline{j+\kappa}}] - \lambda E[X^{\underline{j+\kappa-1}}])/j!}{\lambda\sum_{j=0}^{\infty}(-1)^{j}\alpha^{j}E[X^{\underline{j+\kappa-1}}]/j!} \\ &\quad = \frac{E[X^{\underline{\kappa}}] - \lambda E[X^{\underline{\kappa-1}}]}{\lambda E[X^{\underline{\kappa-1}}]} \\ &\quad = \frac{E[X^{\underline{\kappa}}] - \lambda^{\kappa}}{\lambda^{\kappa}} \end{aligned} \tag{40}$$

and the limit

$$\lim_{\alpha\to 0}\frac{\alpha\lambda T_\alpha P(\kappa-1)}{\alpha^\kappa}$$

equals

$$\begin{aligned}\lim_{\alpha\to 0}&\frac{\alpha\lambda\sum_{j=0}^{\infty}(-1)^j E\left[(T_\alpha X)^{\underline{\kappa-1+j}}\right]/j!}{(\kappa-1)!\,\alpha^\kappa}\\ &=\lim_{\alpha\to 0}\frac{\lambda\sum_{j=0}^{\infty}(-1)^j\alpha^j E[X^{\underline{\kappa-1+j}}]/j!}{(\kappa-1)!}\\ &=\frac{\lambda^\kappa}{(\kappa-1)!}.\end{aligned}\tag{41}$$

Finally, combining the above limits with (39) yields

$$\begin{aligned}\lim_{\alpha\to 0}\frac{K(T_\alpha X)}{\alpha^\kappa}&=\frac{\lambda^\kappa}{(\kappa-1)!}\left(\frac{E[X^{\underline{\kappa}}]-\lambda^\kappa}{\lambda^\kappa}\right)^2\\ &=\kappa\left(\frac{E[X^{\underline{\kappa}}]-\lambda^\kappa}{\lambda^{\kappa/2}(\kappa!)^{1/2}}\right)^2\\ &=\kappa E\left[C_\kappa^\lambda(X)\right]^2,\end{aligned}$$

as claimed. ∎

Proof of Lemma 7.2: The key is to observe that for Y ULC, since $P(y+1)(y+1)/P(y)$ is decreasing in y, and y is increasing in y, there exists an integer y_0 such that $P(y+1)(y+1)\le y_0P(y)$ for $y\ge y_0$ and $P(y+1)(y+1)\ge y_0P(y)$ for $y<y_0$. Hence:

$$\begin{aligned}&\sum_{y=z+1}^{\infty}P(y)(y-y_0)\\ &=P(z+1)(z+1)+\sum_{y=z+1}^{\infty}(P(y+1)(y+1)-y_0P(y))\\ &\le(z+1)P(z+1),\quad\text{for } z\ge y_0,\end{aligned}$$

and,

$$\begin{aligned}&\sum_{y=0}^{z}P(y)(y_0-y)\\ &=P(z+1)(z+1)-\sum_{y=0}^{z}(P(y+1)(y+1)-y_0P(y))\\ &\le(z+1)P(z+1),\text{for } z\le y_0-1.\end{aligned}$$

Further, by Cauchy-Schwarz, for $y\ge y_0$

$$\begin{aligned}(g(y)-g(y_0))^2&=\left(\sum_{z=y_0}^{y-1}\Delta g(z)\right)^2\\ &\le(y-y_0)\left(\sum_{z=y_0}^{y-1}\Delta g(z)^2\right)\end{aligned}\tag{42}$$

while for $y\le y_0-1$

$$\begin{aligned}(g(y)-g(y_0))^2&=\left(\sum_{z=y}^{y_0-1}\Delta g(z)\right)^2\\ &\le(y_0-y)\left(\sum_{z=y_0}^{y-1}\Delta g(z)^2\right).\end{aligned}\tag{43}$$

This means that (with the reversal of order of summation justified by Fubini, since all the terms have the same sign)

$$\begin{aligned}&\sum_{y=0}^{\infty}P(y)(g(y)-\mu)^2\\ &\le\sum_{y=0}^{\infty}P(y)(g(y)-g(y_0))^2\\ &=\sum_{y=0}^{y_0-1}P(y)(g(y)-g(y_0))^2+\sum_{y=y_0}^{\infty}P(y)(g(y)-g(y_0))^2\\ &\le\sum_{y=0}^{y_0-1}P(y)(y_0-y)\left(\sum_{z=y}^{y_0-1}\Delta g(z)^2\right)\\ &\quad+\sum_{y=y_0}^{\infty}P(y)(y-y_0)\left(\sum_{z=y_0}^{y-1}\Delta g(z)^2\right)\qquad(44)\\ &\le\sum_{z=0}^{y_0-1}\Delta g(z)^2\left(\sum_{y=0}^{z}P(y)(y_0-y)\right)\\ &\quad+\sum_{z=y_0}^{\infty}\Delta g(z)^2\left(\sum_{y=z+1}^{\infty}P(y)(y-y_0)\right)\\ &\le\sum_{z=0}^{\infty}(\Delta g)(z)^2P(z+1)(z+1)\qquad(45)\end{aligned}$$

and the result holds. Note that the inequality in (44) follows by (42) and (43), and the inequality in (45) by the discussion above. ∎

ACKNOWLEDGMENT

The authors wish to thank E. Telatar and C. Vignat for hosting a small workshop in January 2006, during which some of these ideas developed. J. Swart also provided us with useful comments.

REFERENCES

[1] J. A. Adell, A. Lekuona, and Y. Yu, "Sharp bounds on the entropy of the Poisson law and related quantities," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2299–2306, 2010.

[2] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*. Oxford, U.K.: Clarendon , 1992.

[3] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab. Theory*, vol. 14, no. 1, pp. 336–342, 1986.

[4] A. R. Barron, "Limits of information, Markov chains, and projections," in *Proc. 2000 Int. Symp. Inf. Theory*, 2000, pp. 25–25.

[5] S. G. 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.

[6] A. A. Borovkov and S. A. Utev, "An inequality and a characterization of the normal distribution connected with it," *Teor. Veroyatnost. i Primenen*, vol. 28, no. 2, pp. 209–218, 1983.

[7] D. Chafaï, "Binomial-Poisson entropic inequalities and the $\mathrm{M}/\mathrm{M}/\infty$ queue," *ESAIM Probab. Statist.*, vol. 10, pp. 317–339, 2006.
[8] T. S. Chihara, *An Introduction to Orthogonal Polynomials*. New York: Gordon and Breach, 1978.
[9] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[10] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[11] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Found. Trends in Commun. Inf. Theory*, vol. 1, pp. 1–111, 2004.
[12] D. J. Daley and D. Vere-Jones, *An Introduction To The Theory of Point Processes*, Second ed. New York: Springer, 2008, vol. II.
[13] A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1491–1498, Jun. 2003.
[14] W. Feller, *An Introduction to Probability Theory and Its Applications*, second ed. New York: Wiley, 1971, vol. II.
[15] J. Fritz, "An information-theoretical proof of limit theorems for reversible Markov processes," in *Proc. 6th Prague Conf. Inf. Theory, Statist. Decision Functions, Random Processes*, Prague, Sep. 1971.
[16] L. Gerber, "An extension of Bernoulli's inequality," *Amer. Math. Monthly*, vol. 75, pp. 875–876, 1968.
[17] B. Grigelionis, "The convergence of stepwise random processes to a Poisson process," *Teor. Verojatnost. i Primenen*, vol. 8, pp. 189–194, 1963.
[18] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. Inf. Theory*, vol. IT-47, pp. 2039–2041, Jul. 2001.
[19] P. Harremoës and K. K. Holst, "Convergence of Markov chains in information divergence," *J. Theoret. Probab.*, vol. 22, no. 1, pp. 186–202, 2009.
[20] O. Johnson and I. Kontoyiannis, "Thinning and the law of small numbers," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 1491–1495.
[21] P. Harremoës, O. Johnson, and I. Kontoyiannis, "Thinning and information projections," *Preparation*, 2010.
[22] P. Harremoës and P. Ruzankin, "Rate of convergence to Poisson law in terms of information divergence," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2145–2149, 2004.
[23] O. Johnson, *Information Theory and Central Limit Theorem*. London, U.K.: Imperial College Press, 2004.
[24] O. Johnson, "Log-concavity and the maximum entropy property of the Poisson distribution," *Stochastic Process. Appl.*, vol. 117, no. 6, pp. 791–82, 2006.
[25] I. M. Johnstone and B. MacGibbon, "Une mesure d'information caractérisant la loi de Poisson," *Séminaire de Probab., XXI*, pp. 563–573, 1987.
[26] A. Y. Khintchine, *Mathematical Methods in the Theory of Queueing*. New York: Hafner, 1960.
[27] C. A. J. Klaassen, "On an inequality of Chernoff," *Ann. Probab.*, vol. 13, no. 3, pp. 966–974, 1985.
[28] I. Kontoyiannis, P. Harremoës, and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. Inf. Theory*, vol. IT-51, pp. 466–472, 2005.
[29] E. H. Lieb, "Proof of an entropy conjecture by Wehrl," *Commun. Math. Phys.*, vol. 62, pp. 35–41, 1978.
[30] T. Lindvall, *Lectures on the Coupling Method*. New York: Wiley , 1992.
[31] D. Mejzler, "A note on Erlang's formulas," *Israel J. Math.*, vol. 3, pp. 157–162, 1965.
[32] G. F. Newell, "The $M/G/\infty$ queue," *SIAM J. Appl. Math.*, vol. 14, pp. 86–88, 1966.
[33] C. Palm, "Intensitätsschwankungen im Fernsprechverkehr," *Ericsson Technics No*, vol. 44, pp. 189–189, 1943.
[34] R.-D. Reiss, *Approximate Distributions of Order Statistics*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1989.
[35] A. Rényi, "A characterization of Poisson processes," *Magyar Tud. Akad. Mat. Kutaló Int. Közl.*, vol. 1, pp. 519–527, 1956.
[36] D. Schuhmacher, "Distance estimates for dependent superpositions of point processes," *Stochastic Process. Appl.*, vol. 115, no. 11, pp. 1819–1837, 2005.
[37] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 379–423, pp. 623–656, 1948.
[38] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, Jun. 1959.
[39] D. W. Walkup, "Pólya sequences, binomial convolution and the union of random sets," *J. Appl. Probabil.*, vol. 13, no. 1, pp. 76–85, 1976.
[40] N. Yannaros, "Poisson approximation for random sums of Bernoulli random variables," *Statist. Probab. Lett.*, vol. 11, no. 2, pp. 161–165, 1991.
[41] Y. Yu, "Monotonic convergence in an information-theoretic law of small numbers," *IEEE Trans. Inf. Theory*, vol. 55, pp. 5412–5422, 2009.
[42] V. M. Zolotarev, "Probability metrics," *Teor. Veroyatnost. i Primenen*, vol. 28, no. 2, pp. 264–287, 1983.

Peter Harremoës (M'00) received the B.Sc. degree in mathematics in 1984, the Exam. Art. degree in archaeology in 1985, and the M.Sc. degree in mathematics in 1988, all from the University of Copenhagen, Denmark. He received the Ph.D. degree in natural sciences in 1993 from Roskilde University, Denmark.

From 1993 to 1998, he worked as a mountaineer. From 1998 to 2000, he held various teaching positions in mathematics. From 2001 to 2006, he was a Postdoctoral Fellow with the University of Copenhagen, with a longer visit to Zentrum für Interdiziplinäre Forschung, Bielefeld, Germany, 2003. From 2006 to 2009, he was affiliated with Centrum Wiskunde and Informatica, Amsterdam, The Netherlands, under the European Pascal Network of Excellence. Since then he has been affiliated with Niels Brock, Copenhagen Business College, Denmark.

Dr. Harremoës has been Editor-in-Chief of the journal *Entropy* since 2007.

Oliver Johnson received the B.A. degree in 1995, Part III Mathematics in 1996 and the Ph.D. degree in 2000, all from the University of Cambridge.

He was a Clayton Research Fellow of Christ's College and Max Newman Research Fellow of Cambridge University until 2006, during which time he published the book *Information Theory and the Central Limit Theorem* in 2004. Since 2006, he has been Lecturer in Statistics at Bristol University, U.K..

Ioannis Kontoyiannis (S'94–M'98–SM'08) was born in Athens, Greece, in 1972. He received the B.Sc. degree in mathematics in 1992 from Imperial College (University of London), U.K., and in 1993 he obtained a distinction in Part III of the Cambridge University Pure Mathematics Tripos. He received the M.S. degree in statistics and the Ph.D. degree in electrical engineering, both from Stanford University, Stanford, CA, in 1997 and 1998, respectively.

Between June and December 1995, he was with IBM Research working on a NASA-IBM satellite image processing and compression project. From June 1998 to August 2001, we was an Assistant Professor with the Department of Statistics, Purdue University, West Lafayette, IN (and also, by courtesy, with the Department of Mathematics, and the School of Electrical and Computer Engineering). From August 2000 until July 2005, he was an Assistant, then Associate Professor, with the Division of Applied Mathematics and with the Department of Computer Science, Brown University, Providence, RI. Since March 2005, he has been an Associate Professor with the Department of Informatics, Athens University of Economics and Business.

Dr. Kontoyiannis was awarded the Manning endowed Assistant Professorship in 2002, and was awarded an honorary Master of Arts degree Ad Eundem, in 2005, both by Brown University. In 2004, he was awarded a Sloan Foundation Research Fellowship. Currently he serves as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. His research interests include data compression, applied probability, information theory, statistics, simulation, and mathematical biology.