Thinning and the Law of Small Numbers

Peter Harremoës
Centrum voor Wiskunde en Informatica
P.O. 94079, 1090 GB Amsterdam
The Nederlands
P.Harremoes@cwi.nl

Oliver Johnson
Dept. Math., Univ. Bristol
University Walk, Bristol, BS8 1TW
United Kingdom
O.Johnson@bristol.ac.uk

Ioannis Kontoyiannis
Athens Univ. of Econ. & Business
Patission 76, Athens 10434
Greece
yiannis@aueb.gr

***Abstract*— The "thinning" operation on a discrete random variable is the natural discrete analog of scaling a continuous variable, i.e., multiplying it by a constant. We examine the role and properties of thinning in the context of information-theoretic inequalities for Poisson approximation. 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 also provided for this limit. A Nash equilibrium is established for a channel game, where Poisson noise and a Poisson input are optimal strategies. Our development partly parallels the development of Gaussian inequalities leading to the information-theoretic version of the central limit theorem.**

I. Introduction

Approximating the distribution of the sum of weakly dependent discrete random variables by a Poisson distribution is a well studied subject in probability; see [1] for an extensive account. Strong connections between these results and information-theoretic techniques were established in [2][3]; see also [4]. For the special case of approximating a Binomial distribution by a Poisson, the sharpest results to date are established via these techniques combined with Pinsker's inequality [5][6][7], at least for most of the parameter values.

Given $\alpha \in (0,1)$ and a discrete random variable Y with distribution P on $\mathbb{N}_0 = \{0,1,2,\ldots\}$, *the α-thinning of P* is the distribution $T_\alpha(P)$ of the sum,

$$\sum_{n=1}^{Y} X_n, \quad \text{where } X_1, X_2 \ldots, X_n \sim \text{i.i.d. Bernoulli}(\alpha), \quad (1)$$

and where Y is assumed to be independent of the $\{X_i\}$. In this work we show that the thinning operation can be used to formulate a version of the law of small numbers, in a way that naturally resembles the classical formulation of the central limit theorem. In particular, the "thinning" law of large numbers we develop gives a Poisson limit theorem for sums of i.i.d. random variables, and not for triangular arrays. These results are shown to hold in total variation as well as in information divergence, and explicit rates of convergence are obtained. Thinning is also shown to be useful in the context of a discrete mutual information game, where the optimal strategies for both sender and jammer are given by the Poisson distribution.

The central limit theorem has been established in the strong sense of information divergence in [8]; see also [9] and the references therein. The main results of this paper can be seen as analogous theorems for Poisson convergence.

II. Thinning

The thinning operation was introduced by Rényi in [10] in connection with the characterization theory of the Poisson process. Let $\alpha \in (0,1)$ and P be a distribution on $\mathbb{N}_0 = \{0,1,2,\ldots\}$. The α-thinning of P is the distribution $T_\alpha(P)$ of the sum (1). An explicit representation of $T_\alpha(P)$ can be given as,

$$T_\alpha(P)(k) = \sum_{l=k}^{\infty} P(l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k}, \quad k \geq 0.$$

It is immediate from the definition that the thinning of a sum of independent random variables is the convolution of the corresponding thinnings.

Example 1: Thinning conserves the set of Bernoulli sums. That is, the thinned version of the distribution of a finite sum of Bernoulli random variables (with possibly different parameters) is also such a sum. This follows from the last remark above together with the observation that the α-thinning of a Bernoulli(p) random variable is the Bernoulli(αp) distribution.

Example 2: Thinning conserves the Poisson law, in that $T_\alpha(Po(\lambda)) = Po(\alpha\lambda)$:

$$\begin{aligned}
T_\alpha(Po(\lambda))(k) &= \sum_{l=k}^{\infty} Po(\lambda, l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k} \\
&= \sum_{l=k}^{\infty} \frac{\lambda^l}{l!} e^{-\lambda} \binom{l}{k} \alpha^k (1-\alpha)^{l-k} \\
&= \frac{e^{-\lambda}}{k!} \alpha^k \lambda^k \sum_{l=k}^{\infty} \frac{\lambda^{l-k}}{(l-k)!} (1-\alpha)^{l-k} \\
&= \frac{e^{-\lambda}}{k!} (\alpha\lambda)^k \sum_{l=0}^{\infty} \frac{(\lambda(1-\alpha))^l}{l!} \\
&= \frac{e^{-\lambda}}{k!} (\alpha\lambda)^k e^{\lambda(1-\alpha)} \\
&= Po(\alpha\lambda, k).
\end{aligned}$$

Similarly, the α-thinning of a geometric distribution with mean λ is a geometric with mean $\alpha\lambda$. And since the sum of n i.i.d. geometric distributions has negative Binomial distribution, the thinning of a negative Binomial is also negative Binomial.

Recall that the mth factorial moment of a random variable X is $fm_m(X) = E\left[X_{[m]}\right] = E\left[X(X-1)\cdots(X-m+1)\right]$. The factorial moments of an α-thinning are easy to calculate:

$$E\left[\left(\sum_{n=1}^{Y} X_n\right)_{[k]}\right] = E\left[E\left[\left(\sum_{n=1}^{Y} X_n\right)_{[k]} \mid Y\right]\right]$$
$$= E\left[\alpha^k (Y)_{[k]}\right] = \alpha^k E\left[Y_{[k]}\right].$$

Thus, thinning scales the factorial moments in the same way that ordinary multiplication scales the ordinary moments.

Next we show that another class of distributions on $\mathbb{N}_0$ that are conserved by thinning is the class of ultra log-concave distributions. Recall that P *is ultra log-concave* if the ratio between P and a Poisson distribution is a (discrete) log-concave function; see [11][12]. In particular, the ultra log-concave class contains all distributions that arise from sums of independent (possibly non-identical) Bernoulli random variables.

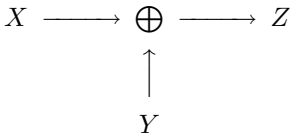
Proposition 3: For any $\alpha \in (0,1)$, the map $P \mapsto T_\alpha(P)$ is injective for P ultra log-concave; that is, if P and Q are ultra log-concave with $T_\alpha P = T_\alpha Q$ then $P = Q$.

Proof: An ultra log-concave distribution is uniquely determined by its (factorial) moments because ultra log-concave distributions satisfy a Cramér-type tail condition. The thinning operation simply scales the factorial moments, so if we know the factorial moments of the thinned distribution we also know the factorial moments of the original distribution. ∎

Note that the α-thinning $T_\alpha(P)$ of a distribution P on $\mathbb{N}_0$ is also a distribution on $\mathbb{N}_0$. We can extend the thinning operation for distributions P of random variables Y on $\mathbb{N}_0/n = \{0, \frac{1}{n}, \frac{2}{n} \ldots\}$, by letting $T_\alpha(P)$ be the distribution of $\frac{1}{n}\sum_{j=1}^{nY} X_j$, where the $\{X_j\}$ are as before. More generally, starting with a random variable X with distribution P on $[0,\infty)$, let P_n denote the uniformly quantized version of P supported on $\mathbb{N}_0$. It is easy to see that, as $n \to \infty$, $T_\alpha(P_n)$ converges to the distribution of αX. In this sense, thinning can be interpreted as a discrete analog of the scaling operation for continuous random variables.

III. A Mutual Information Game

Suppose a transmitter sends a signal X through an additive noise channel $Z = Y + X$, while a jammer adds independent noise Y. The sender wishes to maximize the transmission rate $I(X;Z)$ by choosing an appropriate distribution for X, while the objective of the jammer is to choose Y so that $I(X;Z)$ is minimized:



For continuous random variables X and Y with power constraints of the form $E\left[X^2\right] \le P$ and $E\left[Y^2\right] \le N$, this is a classical problem; see, e.g., [13, p.263][14] and the references therein. In that case, the Gaussian distributions with mean 0 and variances P and N, respectively, form a Nash equilibrium pair, in the sense that neither of the players would benefit by changing her strategy if the other player does not. The entropy power inequality plays an essential role in the proof of the Nash equilibrium condition.

Here we assume that X and Y take values in $\mathbb{N}_0$, and that the strategies of both players are subject to the constraints

$$E[X] \le \lambda_{\text{in}}, \quad E[Y] \le \lambda_{\text{noise}},$$

where λ_{in} and λ_{noise} are positive constants. Moreover, we assume that the distributions of X and Y are both ultra log-concave: $X \in ULC(\lambda), \lambda \le \lambda_{\text{in}}$ and $Y \in ULC(\mu), \mu \le \lambda_{\text{noise}}$, where $ULC(\lambda)$ denotes the class of ultra log-concave distributions on $\mathbb{N}_0$ with mean λ. A similar but more restricted version of this game was considered in [15]. The sets of strategies are not convex, so Von Neumann's classical result on the existence of a game-theoretic equilibrium cannot be used. Nevertheless, our next result states that Poisson distributions form a Nash equilibrium pair for this game. Thus, the Poisson additive noise channel is "worst possible" in this particular class of transmission problems.

Theorem 4: In the above discrete transmission game, the Poisson distribution is the optimal input distribution for Poisson distributed noise: If $Z \sim Po(\lambda_{\text{noise}})$, then:

$$Po(\lambda_{\text{in}}) = \underset{X \in ULC(\lambda),\, \lambda \le \lambda_{\text{in}}}{\arg\max} I(X;Z).$$

Also, if $X \sim Po(\lambda_{\text{in}})$, then the Poisson distribution is the optimal distribution for the jammer, i.e.,

$$Po(\lambda_{\text{noise}}) = \underset{Y \in ULC(\lambda),\, \lambda \le \lambda_{\text{noise}}}{\arg\min} I(X;Z) \ .$$

Thus, the distributions $Po(\lambda_{\text{in}})$, $Po(\lambda_{\text{noise}})$ form a unique Nash equilibrium pair in this discrete transmission game.

Proof: Details will not be given here, but the basic idea in proving the first half of the theorem is to replace X by the sum of two random variables, one with distribution $T_\alpha(X)$ plus a $Po(\lambda_{\text{in}}(1-\alpha))$, so that the sum still has mean less than or equal to λ_{in} and is ultra log-concave. One then shows that the transmission rate increases when α decreases, so that the maximum is attained when X is replaced by a Poisson distribution corresponding to $\alpha = 0$. The second part is proved in a similar manner. The rest of the arguments follow from results in [15][12]. ∎

IV. The Law of Thin Numbers

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. In particular, if $P = \text{Bernoulli}(p)$, then $P^{*n} = \text{Binomial}(n,p)$ and $T_{1/n}(P^{*n}) = \text{Binomial}(n, p/n)$. Therefore, the classical Binomial-to-Poisson convergence can be stated as: If $P = \text{Bernoulli}(p)$, then $T_{1/n}(P^{*n}) \to Po(p)$ as $n \to \infty$. In fact, this result holds in great generality:

Theorem 5 (weak version): Let P be a distribution on $\mathbb{N}_0$ with mean λ. Then $T_{1/n}(P^{*n})$ converges pointwise to $Po(\lambda)$ as $n \to \infty$.

Proof: Note that $T_{1/n}(P^{*n}) = \left(T_{1/n}(P)\right)^{*n}$, and that we have the following elementary inequalities for all α:

$$T_\alpha(P)(0) = \sum_{l=0}^{\infty} P(l)(1-\alpha)^l \geq (1-\alpha)^\lambda$$
$$T_\alpha(P)(1) = \sum_{l=1}^{\infty} P(l)\, l\alpha(1-\alpha)^{l-1}$$
$$T_\alpha(P)(j) \geq 0, \quad j \geq 2.$$

Thus taking $\alpha = 1/n$:

$$\begin{aligned}&\left(T_{1/n}(P)\right)^{*n}(j)\\ &\geq \binom{n}{j}\left(\sum_{l=1}^{\infty} P(l)\, l\alpha(1-\alpha)^{l-1}\right)^j \left((1-\alpha)^\lambda\right)^{n-j}\\ &= \frac{n_{[j]}}{n^j \cdot j!}\left(\sum_{l=1}^{\infty} P(l)\, l\left(1-\frac{1}{n}\right)^{l-1}\right)^j \left(1-\frac{1}{n}\right)^{(n-j)\lambda}.\end{aligned}$$

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

$$\frac{n_{[j]}}{n^j \cdot j!} \to \frac{1}{j!},$$

and

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

and by the monotone convergence theorem,

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

Therefore,

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

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

According to Scheffé's Lemma, pointwise convergence of discrete distributions implies convergence in total variation. Therefore, an immediate corollary is that,

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

Theorem 6 (Thermodynamic version): Let P be a ultra log-concave distribution on $\mathbb{N}_0$ with mean λ. Then,

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

Proof: The distribution $T_{1/n}(P^{*n})$ is ultra log-concave and has mean λ so according to [12, Proof of Theorem 2.5] $H\left(T_{1/n}(P^{*n})\right) \leq H(Po(\lambda))$. The entropy function is lower semi continuous and $T_{1/n}(P^{*n})$ converges to $Po(\lambda)$ so $\liminf H\left(T_{1/n}(P^{*n})\right) \geq H(Po(\lambda))$ which proves the theorem. ∎

By $D(P\|Q)$ we shall denote the usual *information divergence from P to Q*,

$$D(P\|Q) = \sum_j P(j)\log\frac{P(j)}{Q(j)}.$$

For ultra-log concave distributions the thermodynamic version implies convergence in information divergence, which is a much stronger sense of convergence than convergence in total variation. This actually holds in much greater generality:

Theorem 7 (strong version): Let P be a distribution on $\mathbb{N}_0$ with mean λ and $D(P\|Po(\lambda)) < \infty$. Then,

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

Proof: The condition $D(P\|Po(\lambda)) < \infty$ implies that all series in the proof are convergent. According to the data processing inequality,

$$\begin{aligned}&D(P_1 * P_2 * ... * P_n\|Po(\lambda/n) * ... * Po(\lambda/n))\\ &\qquad \leq \sum_{i=1}^{n} D(P_i\|Po(\lambda/n)).\end{aligned}$$

Therefore, it is sufficient to show that, as $n \to \infty$, we have, $n \cdot D\left(T_{1/n}(P)\,\|Po(\lambda/n)\right) \to 0$. Replacing $1/n$ by α, it suffices to show that,

$$\frac{\partial}{\partial\alpha} D(T_\alpha(P)\,\|Po(\alpha\lambda)) \to 0 \quad \text{as } \alpha \downarrow 0.$$

Now, [12, Proposition 3.6] shows that $\frac{\partial}{\partial\alpha}T_\alpha(P)(z) = (zT_\alpha(P)(z) - (z+1)T_\alpha(P)(z+1))/\alpha$ and $\frac{\partial}{\partial\alpha}(Po(\alpha\lambda))(z) = (Po(\alpha\lambda))(z)(z-\alpha\lambda)/(\alpha)$. We deduce that

$$\frac{\partial}{\partial\alpha} D(T_\alpha(P)\,\|Po(\alpha\lambda)) = \lambda D(T_\alpha(P)^{\sim}\,\|T_\alpha(P)),$$

where $T_\alpha(P)^{\sim}(z) = (z+1)T_\alpha(P)(z+1)/(\alpha\lambda)$ is also a distribution.

Since $\lim_{\alpha\to 0} T_\alpha(P)(0) = \lim_{\alpha\to 0} T_\alpha(P)^{\sim}(0) = 1$ and $\lim_{\alpha\to 0} T_\alpha(P)(z) = \lim_{\alpha\to 0} T_\alpha(P)^{\sim}(z) = 0$ for $z \geq 1$, the result follows. ∎

V. Rate of convergence

The weak law of thin numbers only required that the first moment of P be finite, and the strong version also required that the divergence between P and the Poisson be finite. Under the additional condition that P has a finite second moment we also obtain a rate of convergence result.

Proposition 8: Let P be a distribution on $\mathbb{N}_0$ with mean $n\lambda$ and finite second moment. Then,

$$D\left(T_{1/n}(P)\,\|Po(\lambda)\right) \leq \frac{\lambda}{n} + \frac{1}{\lambda n^2}\cdot Var(P).$$

Proof: We have,

$$\begin{aligned}D\left(T_{1/n}(P)\,\|Po(\lambda)\right) &= D\left(\sum_{k=0}^{\infty} P(k)\, Bin(k, 1/n)\,\middle\|\, Po(\lambda)\right)\\ &\leq \sum_{k=0}^{\infty} P(k)\, D(Bin(k,1/n)\,\|Po(\lambda)).\end{aligned}$$

Now, using the fact that the Poisson distributions belong to an exponential family, together with the elementary bound

$D\left(Bin\left(l,p\right)\|Po\left(lp\right)\right)\leq lp^{2}$, we get,

$$\begin{aligned}
&D\left(Bin\left(k,1/n\right)\|Po\left(\lambda\right)\right)\\
&=D\left(Bin\left(k,1/n\right)\|Po\left(k/n\right)\right)+D\left(Po\left(k/n\right)\|Po\left(\lambda\right)\right)\\
&\leq\frac{k}{n^{2}}+\sum_{j=0}^{\infty}Po\left(\frac{k}{n},j\right)\log\frac{\frac{\left(\frac{k}{n}\right)^{j}}{j!}\exp\left(-\frac{k}{n}\right)}{\frac{\lambda^{j}}{j!}\exp\left(-\lambda\right)}\\
&\leq\frac{k}{n^{2}}+\lambda\left(\frac{k}{n\lambda}-1\right)^{2},
\end{aligned}$$

where we have used the elementary inequality $x\log x+1-x\leq x\left(x-1\right)+1-x=\left(x-1\right)^{2}$. Hence,

$$\begin{aligned}
D\left(T_{1/n}\left(P\right)\|Po\left(\lambda\right)\right)&\leq\sum_{k=0}^{\infty}P\left(k\right)\cdot\left(\frac{k}{n^{2}}+\lambda\left(\frac{k}{n\lambda}-1\right)^{2}\right)\\
&=\frac{n\lambda}{n^{2}}+\frac{1}{\lambda n^{2}}\sum_{k=0}^{\infty}P\left(k\right)\cdot\left(k-n\lambda\right)^{2}\\
&=\frac{\lambda}{n}+\frac{Var\left(P\right)}{\lambda n^{2}},
\end{aligned}$$

as claimed. ∎

This gives the following immediate corollary, upon replacing P by P^{*n}:

Corollary 9: Let P be a distribution on $\mathbb{N}_0$ with mean λ and finite second moment. Then,

$$D\left(T_{1/n}\left(P^{*n}\right)\|Po\left(\lambda\right)\right)\leq\frac{1}{n}\left(\lambda+\frac{Var\left(P\right)}{\lambda}\right).$$

Next we turn our attention to asymptotic lower bounds. Let X be a random variable with distribution P and factorial moments $fm_{m}\left(X\right)=E\left(X_{[m]}\right)$. If P is a Poisson distribution with mean λ, then $fm_{m}=\lambda^{m}$. In general, we will have $fm_{m}=\lambda^{m}$ only for a few values of m. Let m_0 denote the first value of m such that $fm_{m}\neq\lambda^{m}$ and put $\gamma=fm_{m_{0}}$. Lower bounds on the rate of convergence are essentially given in terms of m_0 and γ. Using techniques that were developed for the central limit theorem [16], we can obtain that,

$$\liminf_{n\to\infty}n^{2m_{0}-2}D\left(T_{1/n}\left(P^{*n}\right)\middle\|\,Po\left(\lambda\right)\right)\geq m_{0}!\frac{\left(\gamma-\lambda^{m_{0}}\right)^{2}}{2\lambda^{m_{0}}}.$$

We conjecture that this lower bound is asymptotically tight.

VI. Characterizations of the Poisson distribution

The main result of the recent work [12] is that the Poisson distribution is the maximum entropy distribution in the class of ultra log-concave distributions. Above we also saw that the Poisson is the worst noise in a discrete transmission game. Here we shall give some further characterizations of the Poisson law, inspired by analogous results for the Gaussian.

Proposition 10: Let $X\sim P$ be an arbitrary $\mathbb{N}_0$-valued random variable, and write X_α for a random variable with distribution $T_\alpha(P)$. If there exists $\alpha\in(0,1)$ and an independent Poisson random variable Z, such that,

$$X_{\alpha}+Z\sim P,$$

then X has a Poisson distribution.

Proof: Suppose $Z\sim Po(\lambda)$ and note that $\alpha E(X)+\lambda=E(X)$, so that

$$\lambda=(1-\alpha)E\left[X\right]>0.$$

Writing W for a $Po(E(X))$ random variable and W_β for an independent random variable with distribution $T_\beta(Po(E(X))$,

$$X_{\alpha}+W_{1-\alpha}\sim X.$$

Thinning by α and iterating this expression yields,

$$X_{\alpha^{n}}+W_{1-\alpha^{n}}\sim X,$$

for all $n\geq1$, and taking $n\to\infty$ yields the stated result. ∎

Proposition 11: If P is an ultra log-concave distribution such that for all $\alpha\in(0,1)$ there exists an ultra log-concave distribution Q_α with $P=T_\alpha(Q_\alpha)$, then P is a Poisson distribution.

Proof: Let λ and V denote the first two factorial moments of P. Then Proposition 7 gives, for all $n\geq1$,

$$D\left(T_{1/n}\left(Q_{1/n}\right)\|Po\left(\lambda\right)\right)\leq\frac{\lambda+\frac{V-\lambda^{2}+\lambda}{\lambda}}{n}\leq\frac{\lambda+1}{n},$$

and since $T_{1/n}\left(Q_{1/n}\right)=P$ for all n, letting $n\to0$ implies $D\left(P\|Po\left(\lambda\right)\right)=0$. ∎

VII. Compound Thinning

There seems to be a natural generalization of the thinning idea, which parallels the generalization of the Poisson distribution to the compound Poisson. Suppose we start with a random variable $Y\sim P$ with values in $\mathbb{N}_0$. The α-thinned version of Y corresponds to 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.

If, instead, we start with a random variable Y to be "compound-thinned," and we choose and fix a distribution Q on $\mathbb{N}=\{1,2,\ldots\}$ and an $\alpha\in(0,1)$, 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 writing $Y=1+1+\cdots+1$ (Y times), then keeping each one of those 1s with probability α, and replacing each of the 1s that are kept by an independent random sample from Q. This has the corresponding representation,

$$\sum_{n=1}^{Y}X_{n}\xi_{n},\quad X_{i}\sim\text{i.i.d. Bernoulli}(\alpha),\ \xi_{i}\sim\text{i.i.d. }Q,\qquad(2)$$

where the $\{\xi_i\}$ are independent of the $\{X_i\}$, and Y is independent of all the other variables. For fixed α and Q, we write $T_{\alpha,Q}(P)$ for the distribution 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 $CBin(n,\alpha,Q)$, is the distribution of the sum of n i.i.d. random variables, each of which is the product of a Bernoulli(α) 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 (2) with $Y = n$ w.p.1. Then we can express the probabilities of the (α, Q)-thinned version of P as,

$$T_{\alpha,Q}(P)(k) = \sum_{\ell \geq k} P(\ell) \cdot CBin(\ell, \alpha, Q)(k),$$

where $CBin(\ell, \alpha, Q)(k)$ is the probability that a random variable with $CBin(\ell, \alpha, Q)$ distribution equals k.

The following two observations are immediate from the definitions.

1) *Compound Thinning Takes a Bernoulli Sum to a Compound Bernoulli Sum.* If P is the distribution of the Bernoulli sum $\sum_{i=1}^{n} X_i$ where the $\{X_i\}$ are independent Bernoulli(p_i), then $T_{\alpha,Q}(P)$ is the distribution of the "compound Bernoulli sum" $\sum_{i=1}^{n} X_i' \xi_i$ where the $\{X_i'\}$ are independent Bernoulli(αp_i), and the $\{\xi_i\}$ are i.i.d. with distribution Q, independent of the $\{X_i\}$.
2) *Compound Thinning Takes the Poisson to the Compound Poisson.* If $P = Po(\lambda)$, then $T_{\alpha,Q}(P)$ is $CPo(\alpha\lambda, Q)$, i.e., the compound Poisson distribution with rate $\alpha\lambda$ and base distribution Q. Recall that $CPo(\lambda, Q)$ has the representation,

$$CPo(\lambda, Q) \sim \sum_{i=1}^{\Pi_\lambda} \xi_i,$$

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

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

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

where P denotes the Bernoulli(λ) distribution. This convergence remains valid in general, for arbitrary P. The next result generalizes the strong law of thin numbers, and its proof is analogous to that.

Theorem 12: Let P be a distribution on $\mathbb{N}_0$ with mean $\lambda > 0$ and finite variance. Then, for any probability measure Q on $\mathbb{N}$,

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

In fact, the same argument as the proof of the last theorem 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 $Y \sim P$ with respect to Q means that for each of the terms in the expansion $Y = \sum_{i=1}^{Y} 1$, we either accept (with probability α) or reject it (with probability $1 - \alpha$), and we replace terms the accepted terms by a vector randomly sampled from Q. This makes $T_{\alpha,Q}(P)$ itself a probability measure on $\mathbb{R}^d$.

It is somewhat remarkable that the statement *and* proof of Corollary 9 remain entirely unchanged in this case:

Theorem 13: The bound in Corollary 9 remains valid, if we replace the thinning operation $T_{1/n}$ by the compound thinning $T_{1/n,Q}$ with respect to *any* probability measure Q on $\mathbb{R}^d$.

VIII. Discussion

In this paper we have obtained a thinning version of the law of small numbers. This may be termed the "law of thin numbers." The proof of the law of thin numbers relies on the classical law of large numbers. Similarly, the derivation of the convergence rate in the law of thin numbers relies on the central limit theorem. Roughly speaking, this indicates that the level of complexity of the proofs is determined by the number of moments taken into consideration. In this sense, the law of large numbers and the law of thin numbers are "first-order" results, whereas the central limit theorem and the convergence rate to the law of thin numbers are "second-order" results.

IX. Acknowledgement

The authors wish to thank E. Telatar and C. Vignat for hosting a workshop in 2006, where these ideas developed.

References

[1] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*. Oxford Studies in Probability 2, Oxford: Clarendon Press, 1992.
[2] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2039–2041, July 2001.
[3] I. Kontoyiannis, P. Harremoës, and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. Inform. Theory*, vol. 51, pp. 466–472, Feb. 2005.
[4] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *Proc. IEEE Inf. Th. Workshop, San Antonio*, 2004.
[5] I. Csiszár, "Information-type measures of difference of probability distributions andindirectobservations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[6] A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1491–1498, June 2003.
[7] P. Harremoës and P. Ruzankin, "Rate of convergence to Poisson law in terms of information divergence," *IEEE Trans. Inform Theory*, vol. 50, pp. 2145–2149, Sept. 2003.
[8] A. R. Barron, "Entropy and the Central Limit Theorem," *Annals Probab. Theory*, vol. 14, no. 1, pp. 336 – 342, 1986.
[9] O. Johnson, *Information Theory and Central Limit Theorem*. London: Imperial Collage Press, 2004.
[10] A. Rényi, "A characterization of Poisson processes," *Magyar Tud. Akad. Mat. Kutaló Int. Közl.*, vol. 1, pp. 519–527, 1956.
[11] R. Pemantle, "Towards a theory of negative dependence," *J. Math. Phys.*, vol. 41, no. 3, pp. 1371–1390, 2000. Probabilistic techniques in equilibrium and nonequilibrium statistical physics.
[12] O. Johnson, "Log-concavity and the maximum entropy property of the Poisson distribution," *Stochastic Processes and their Applications*, 2007. In press.
[13] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.
[14] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inform. Theory*, vol. IT-47, pp. 3072–3081, Nov. 2001.
[15] P. Harremoës and C. Vignat, "A Nash equilibrium related to the Poisson channel," *Communications in Information and Systems*, vol. 3, pp. 183–190, March 2004.
[16] P. Harremoës, "Lower bounds on divergence in central limit theorem," *Electronic Notes i Discrete Mathematics*, vol. 21, pp. 319–313, Aug. 2005.