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
UK
O.Johnson@bristol.ac.uk

Ioannis Kontoyiannis
Dept. Informatics, Athens Univ. Econ.
Patission 76, Athens 10434
Greece
yiannis@aueb.gr

***Abstract*— The Law of Small Numbers is is normally formulated using triangular arreys. We replace the triangular arrey by a new mathematical construction and prove an information theoretic version of the Law of Small Numbers. By this approach the relation between the Law of Smal Numbers and the information Theoretic version of the Central Limit Theorem is emphasized. These ideas are also used for the compound Poisson distributions.**

I. Introduction

Approximation by Poisson distributions is a well studied subject and the most complete presentation can be found in [1]. Later the connection between information theory has been established in [2] and [3]. For most values of the parameters the best bounds on total variation between a binomial distribution and a Poisson distribution with the same mean have been proved by ideas from information theory via Pinsker's inequality [4], [5] and [6]. Here we shall see that the idea of thinning can be used to formulate a Law of Small Numbers (Poisson's Law) in a way that resembles the i.i.d. formulation of the Central Limit Theorem. Results of this kind will be termed Laws of Thin Numbers. Convergence in the Central Limit Theorem has been established in the strong sense that information divergence converges to zero [7] and [8]. The main result of this paper is a similar theorem with convergence to the Poisson distribution.

II. Thinning

The idea of thinning was introduced by Rényi in [9] in order to characterize Poisson processes. Let P denote a distribution on $\mathbb{N}_0$. The α-thinning of P is the distribution $T(P)$ given by

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

If $X_1, X_2, X_3, \ldots$ are independent identically distributed Bernoulli random variables with success probability and Y has distribution P independent of $X_1, X_2, \ldots$then the distribution of

$$\sum_{n=1}^{Y}X_n$$

has distribution $T_\alpha(P)$ .Obviously the tinning of an independent sum of random variables is the convolution of tinnings.

Example 1: The thinning conserves the set of Bernoulli sums. To see this we just remark that the α-thinning of a Bernoulli random variable with success probability p is a Bernoulli random variable with success probability p.

Example 2: Thinning conserves the Poisson distributions.

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

Similarly, the α-thinning of a geometric distribution with mean λ is a geometric distribution with mean $\alpha\lambda$. A sum of n independent identically distributed geometric distributions is a negative binomial distribution. Thus a thinning of a negative binomial distribution is negative binomial.

The factorial moments of an α-thinning are easy to calculate

$$\begin{aligned}E\left(\left(\sum_{n=1}^{Y}X_n\right)_{[k]}\right)&=E\left(E\left(\left(\sum_{n=1}^{Y}X_n\right)_{[k]}\mid Y\right)\right)\\&=E\left(\alpha^k(Y)_{[k]}\right)\\&=\alpha^kE\left((Y)_{[k]}\right).\end{aligned}$$

Thus thinning scales the factorial moments in the same way as ordinary multiplication scales the ordinary moments. A distribution on $\mathbb{N}_0$ is said to be ultra log-concave if its density with respect to a Poisson distribution is discrete log-concave[10].

Proposition 3: In the class of ultra-log concave distributions the thinning operation $T_\alpha, \alpha\in]0;1[$ is injective.

Proof: An ultra log concave distribution is uniquely determined by its (factorial) moments because ultra log concave distributions satisfies a Cramér condition. The thinning operation just scales the moments so if we know the factorial moments of the thinned distribution we also know the factorial moments of the original distribution. $\blacksquare$

For a distribution P on $\mathbb{N}_0$ the α-thinning operation gives a distribution on $\mathbb{N}_0$. We note that $\mathbb{N}_0 \subseteq \mathbb{N}_0/n \subseteq \mathbb{R}$ and that the thinning also can be made on the larger set of grid points $\mathbb{N}_0/n$. If X has distribution P then the distribution of $T\alpha(P)$ converges weakly to the distribution of X according to the law of large numbers. Similarly, if a positive random variable X with continuous distribution is quantized according to a uniform quantization and then thinned then the thinned distribution approximately equals the distribution of X. In this sense thinning is a discrete analog of scaling of a continuous random variable.

III. The sender and the jammer

We consider the following setup: an emitter transmits a signal X through a channel; a jammer sends noise Y on the same additive channel, and the received signal is $Z = X + Y$. The signals X and Y are assumed to be independent. The emitter wants to maximize the transmission rate i.e. the mutual information $I(X;Z)$ by choosing an appropriate distribution of X. Conversely, the objective of the jammer is to choose the distribution of Y such that the transmission rate is minimized.

$$\begin{tikzcd} X \arrow[r] & \bigoplus \arrow[r] & Z \\ & Y \arrow[u] & \end{tikzcd}$$

For continuous random variables X and Y with power constraints of the form $E(X^2) \leq P$ and $E(Y^2) \leq N$, this problem has been studied by T. Cover and S. Diggavi in [11, Exercise 1, p. 263] and in more detail in [12]. In this case the Gaussian distributions with mean 0 and variances P and N respectively form a Nash equilibrium pair, in the sense that none of the players has any benefit of changing his strategy if the other player does not change his strategy either. The Entropy Power Inequality plays an essential role in the proof of the Nash equilibrium condition.

Here we shall assume that the strategies of both players are subject to constraints

$$E(X) \leq \lambda_{\text{in}}, \quad E(Y) \leq \lambda_{\text{noise}}$$

where λ_{in} and λ_{noise} are positive constants. We shall consider the case when the messages X and Y are supposed to be ultra log-concave: $X \in ULC(\lambda), \lambda \leq \lambda_{\text{in}}$ and $Y \in ULC(\mu), \mu \leq \lambda_{\text{noise}}$. This game will be called *a discrete transmission game*. A similar but more restricted game was considered in [13]. The sets of strategies are not convex so Von Neumann's classical result on existence of a game theoretic equilibrium cannot be used. The main result in this paper is that the Poisson distributions form a Nash equilibrium pair of the discrete transmission game. Thus the *Poisson channel*, i.e. an information channel with Poisson distributed noise, is in a natural way related to the discrete transmission game.

Theorem 4: In the discrete transmission game the Poisson distribution is the optimal input distribution for Poisson distributed noise

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

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

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

Thus the pair of Poisson distributions is a unique Nash equilibrium pair in the 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 $T_\alpha(X)$ plus the Poisson distribution $Po(\lambda_{\text{in}}(1-\alpha))$ so that the sum still has mean less than or equal to λ_{in} and still 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 the same manner. The details can be filled in using results from [13] and [10]. $\blacksquare$

IV. The Law of Thin Numbers

By use of the thinning operation it is possible to formulate an i.i.d. version of the Law of Small Numbers.

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 $Po(\lambda)$ for $n \to \infty$.

Proof: First we note that $T_{1/n}(P^{*n}) = \left(T_{1/n}(P)\right)^{*n}$. For $\alpha = 1/\alpha$ we have the inequalities

$$T_{1/n}(P)(0) = \sum_{l=0}^{\infty} P(l)(1-\alpha)^l \geq (1-\alpha)^\lambda$$

$\blacksquare$

and

$$T_{1/n}(P)(1) = \sum_{l=1}^{\infty} P(l)\, l\alpha (1-\alpha)^{l-1}$$

and $T_{1/n}(P)(j) \geq 0$ for $j \geq 2$. Thus

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

For a fixed value of j and n tending to infinity we have

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

and

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

and by Lebesgue's Theorem on monotone convergence

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

Thus

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

The sequence $\left(T_{1/n}(P)\right)^{*n}$ is a sequence of discrete probability measures and $Po(\lambda)$ is a discrete probability measure so for any value of j $\left(T_{1/n}(P)\right)^{*n}(j)$ converges converges to $Po(\lambda, j)$ for n tending to infinity.

According to Sheffe's Lemma pointwise convergence implies convergence in total variation. An even stronger kind of convergence can be obtained. By $D(P\|Q)$ we shall denote the *information divergence from P to Q* defined by

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

where P and Q are discrete probability measures.

Theorem 6 (strong version): Let P be a distribution on $\mathbb{N}_0$ with mean λ and $D(P\|Po(\lambda)) < \infty$. Then

$$T_{1/n}\left(P^{*n}\right) \xrightarrow{I} Po(\lambda) \text{ for } n \to \infty.$$

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

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

■

Hence it is sufficient to show that

$$n \cdot D\left(T_{1/n}(P) \| Po(\lambda/n)\right)$$

converges to zero for n tending to infinity. We put $\alpha = 1/n$ and have to prove that

$$\frac{\partial}{\partial\alpha} D(T_\alpha(P) \| Po(\alpha\lambda))$$

converges to zero for α tending to zero. Now

$$\begin{aligned}
&D(T_\alpha(P) \| Po(\alpha\lambda)) \\
&= \sum_{k=0}^{\infty} \left(\begin{array}{c} \left(\sum_{l=k}^{\infty} P(l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k}\right) \cdot \\ \log \frac{\sum_{l=k}^{\infty} P(l)\binom{l}{k}\alpha^k(1-\alpha)^{l-k}}{\frac{(\alpha\lambda)^k}{k!}\exp(-\alpha\lambda)} \end{array} \right) \\
&= \alpha\lambda + \sum_{k=0}^{\infty} \left(\begin{array}{c} \left(\sum_{l=k}^{\infty} P(l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k}\right) \cdot \\ \log\left(\sum_{l=k}^{\infty} P(l) \frac{l_{[k]}}{\lambda^k} (1-\alpha)^{l-k}\right) \end{array} \right).
\end{aligned}$$

The derivative is

$$\begin{aligned}
&\frac{\partial}{\partial\alpha} D(T_\alpha(P) \| Po(\alpha\lambda)) \\
&= \lambda + \sum_{k=1}^{\infty} \left(\begin{array}{c} \left(\sum_{l=k}^{\infty} P(l) \binom{l}{k} k\alpha^{k-1} (1-\alpha)^{l-k}\right) \\ \cdot \log\left(\sum_{l=k}^{\infty} P(l) \frac{l_{[k]}}{\lambda^k} (1-\alpha)^{l-k}\right) \end{array} \right) \\
&- \sum_{k=0}^{\infty} \left(\begin{array}{c} \left(\sum_{l=k+1}^{\infty} P(l) \binom{l}{k} \alpha^k (l-k)(1-\alpha)^{l-k-1}\right) \cdot \\ \log\left(\sum_{l=k}^{\infty} P(l) \frac{l_{[k]}}{\lambda^k} (1-\alpha)^{l-k}\right) \end{array} \right) \\
&- \sum_{k=0}^{\infty} \left(\begin{array}{c} \left(\sum_{l=k}^{\infty} P(l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k}\right) \cdot \\ \frac{\sum_{l=k}^{\infty} P(l) \frac{l_{[k+1]}}{\lambda^k} (1-\alpha)^{l-k-1}}{\sum_{l=k}^{\infty} P(l) \frac{l_{[k]}}{\lambda^k} (1-\alpha)^{l-k}} \end{array} \right)
\end{aligned}$$

For α tending to zero this expression tends to

$$\begin{aligned}
&\lambda + \left(\sum_{l=1}^{\infty} P(l)\, l\right) \log\left(\sum_{l=1}^{\infty} P(l) \frac{l_{[1]}}{\lambda}\right) \\
&- \left(\sum_{l=1}^{\infty} P(l)\, l\right) \log\left(\sum_{l=0}^{\infty} P(l)\right) \\
&- \sum_{l=0}^{\infty} P(l) \frac{\sum_{l=0}^{\infty} P(l)\, l}{\sum_{l=0}^{\infty} P(l)} = 0.
\end{aligned}$$

V. Rate of convergence

In the Law of Small Numbers the weak version only required a condition on the first moment and the strong version also required that the divergence is finite. If we have a condition on the second moment we can get bounds on the rate of convergence.

Proposition 7: 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

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

We use that the Poisson distributions belong to an exponential family and the elementary bound $D(Bi(l,p) \| Po(lp)) \leq lp^2$ and get

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

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

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

■

Theorem 8: Let P be a distribution on $\mathbb{N}_0$ with mean λ and finite second moment. Then the rate of convergence of $T_{1/n}(P^{*n})$ to $Po(\lambda)$ is upper bounded by $o(1/n)$.

Proof: We have

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

■

Next we turn our attention to asymptotic lower bounds. Let X be a random variable with distribution P and factorial moments $fm_m(X)=E\left(X_{[m]}\right)$ were $X_{[m]}$ denotes the falling factorial $X(X-1)\ldots(X-m+1)$. If P is a Poisson distribution with mean λ then $fm_m=\lambda^m$. In general $fm_m=\lambda^m$ will only hold 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 [14] we get the asymptotic lower bound

$$\lim_{n\to\infty}\inf n^{2m_0-2}D\left(T_{1/n}\left(\sum_{j=1}^{n}X_j\right)\|Po(\lambda)\right)\geq m_0!\frac{(\gamma-\lambda^{m_0})^2}{2\lambda^{m_0}}.$$

We conjecture that this asymptotic lower bound is also an asymptotic upper bound.

VI. Characterizations of the Poisson distribution

The main result in [10] was that the Poisson distribution is the maximum entropy distribution in the class of ultra log-concave distributions. We have already seen that the Poisson distribution plays the role as the capacity achieving distribution and most noisy noise when a tranmitted signal gets disturbed by a jammer. Here we shall give some further characterizations inspired by similar characterizations of the Gaussian distribution.

Proposition 9: Let X denote a discrete random variable and let Y denote an independent Poisson random variable. Assume that

$$T_\alpha(X)+T_\beta(Y)\sim X$$

where $\alpha,\beta\in]0;1[$. Then X has a Poisson distribution.

Proof: First we observe that $\alpha E(X)+\beta E(Y)=E(X)$ so that $E(Y)>0$ and

$$\beta=(1-\alpha)\frac{E(X)}{E(Y)}.$$

Without loss of generality we may assume that $E(X)=E(Y)$ and $\beta=1-\alpha$. Put $P_\alpha(X)=T_\alpha(X)+T_{1-\alpha}(Y)$. Then $P_\alpha(X)\sim X$ and $P_\alpha^n(X)\sim X$. Now,

$$\begin{aligned}X&\sim P_\alpha^n(X)\\&=P_{\alpha^n}(X)\\&=T_{\alpha^n}(x)+T_{1-\alpha^n}(Y).\end{aligned}$$

The right hand side converges to a Poisson distribution so that X must be Poisson distributed. ■

Proposition 10: If P is a ultra-log concave distribution such that for all $\alpha\in]0;1[$ there exists a ultra-log concave distribution Q_α such that $P=T_\alpha(Q_\alpha)$. Then P is a Poisson distribution.

Proof: Assume that P is a ultra-log concave distribution such that for all such that for all $\alpha\in]0;1[$ there exists a ultra-log concave distribution Q_α such that $P=T_\alpha(Q_\alpha)$.

Let λ and V denote the first two factorial moments of P. Then

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

which is proved as in the first proof of Theorem 8. Now, $T_{1/n}\left(Q_{1/n}\right)=P$ for all n so we get $D(P\|Po(\lambda))=0$ and $P=Po(\lambda)$. ■

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. In words it can be explained as follows. 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 one of those 1s with probability α, independently of all the others. This leads to the representation

$$\sum_{i=1}^{Y}X_i,\quad X_i, i.i.d.\sim Bern(\alpha).$$

Now suppose 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 $\alpha\in[0,1]$. 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 then substituting a random sample from Q for each of the 1s that are kept. This has the corresponding representation $\sum_{i=1}^{Y}X_i\xi_i$, where the X_i are i.i.d. Bern(α), and the ξ_i are i.i.d. $\sim Q$, independent of the X_i.

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 $Ber(\alpha)$ random variable and an independent $\xi \sim Q$ random variable. In other words, it's the (α, Q)-thinned version of the point mass at n, i.e., the distribution of (VII) with $Y = n$ w.p.1. Then we can express the probabilities of the (α, Q)-thinned version of P as, $T_{\alpha,Q}(P)(k) = \sum_{\ell \geq k} P(\ell) CBin(\ell, \alpha, Q)(k)$, where $CBin(\ell, \alpha, Q)(k)$ is the probability that a random variable with $CBin(\ell, \alpha, Q)$ distribution takes the value k.

The following two observations are immediate from the definitions.

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 Bern(p_i), then $T_{\alpha,Q}(P)$ is the distribution of the "compound Bernoulli sum" sum $\sum_{i=1}^{n} X_i' \xi_i$ where the X_i' are independent Bern(αp_i), and the ξ_i are i.i.d. with distribution Q, independent of the X_i.

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 ξ_i.

The compound-Poisson-equivalent to the "Law of Small Numbers" is the fact that $CBin(n, \lambda/n, Q) \approx CPo(\lambda, Q)$ for large n. Equivalently, this says that

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

where P denotes the Bern(λ) distribution.

Theorem 11: 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 $\mathbb{N}$,

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

One final note. In the above arguments there is nothing that restricts us to integer-valued compounding. That is, suppose Q is now 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 those terms that are accepted we replace by a vector randomly sampled from Q. This makes $T_{\alpha,Q}(P)$ itself a probability measure on $\mathbb{R}^d$.

It is quite amazing that the statement *and* proof of Theorem 5 remain entirely unchanged in this case:

Theorem 12: The bounds in Theorem 8 remain valid as stated if we allow Q to be an arbitrary probability measure 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. In order to prove the Law of Thin Numbers we have used the Law of Large Numbers. In order to calculate the rate of convergence we had to use the Central Limit Theorem. Hvis clearly points to the fact that the complexity of the proofs are mainly determined by the number of moments taken into consideration. One may say that Law of Large Numbers and Law of Thin Numbers are convergence theorems of first order whereas the Central Limit Theorem is a theorem of second order. To calculate rate of convergence normally increase the order by one.

IX. Acknowledgement

The authors want to thank Emre Telatar and Christophe Vignat for hosting a small workshop January 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. IT-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. IT-51, no. 2, pp. 466–472, 2005.

[4] I. Csiszár, "Information-type measures of difference of probability distributions andindirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[5] A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's Inequality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1491–1498, June 2003.

[6] P. Harremoës and P. Ruzankin, "Rate of convergence to Poisson law in terms of information divergence." Accepted for publication, 2003.

[7] A. R. Barron, "Entropy and the Central Limit Theorem," *Annals Probab. Theory*, vol. 14, no. 1, pp. 336 – 342, 1986.

[8] O. Johnson, *Information Theory and Central Limit Theorem*. London: Imperial Collage Press, 2004. To appear.

[9] A. Rényi, "A characterization of Poisson processes," *Magyar Tud. Akad. Mat. Kutaló Int. Közl.*, vol. 1, pp. 519–527, 1956.

[10] O. Johnson, "Log-concavity and the maximum entropy property of the Poisson distribution," *Stochastic Processes and their Applications*, 2006. To appear.

[11] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[12] 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.

[13] 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.

[14] P. Harremoës, "Lower bounds on divergence in central limit theorem." To appear, 2004.