Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications 117 (2007) 791–802
stochastic processes and their applications
www.elsevier.com/locate/spa

Log-concavity and the maximum entropy property of the Poisson distribution

Oliver Johnson*

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Rd, Cambridge, CB3 0WB, UK

Received 7 June 2006; received in revised form 11 October 2006; accepted 16 October 2006
Available online 7 November 2006

Abstract

We prove that the Poisson distribution maximises entropy in the class of ultra log–concave distributions, extending a result of Harremoës. The proof uses ideas concerning log-concavity, and a semigroup action involving adding Poisson variables and thinning. We go on to show that the entropy is a concave function along this semigroup.

Keywords: Log-concavity; Maximum entropy; Poisson distribution; Thinning; Ultra log-concavity

1. Maximum entropy distributions

It is well-known that the distributions which maximise entropy under certain very natural conditions take a simple form. For example, among random variables with fixed mean and variance the entropy is maximised by the normal distribution. Similarly, for random variables with positive support and fixed mean, the entropy is maximised by the exponential distribution. The standard technique for proving such results uses the Gibbs inequality, and establishes the fact that, given a function $R(\cdot)$ and fixing $\mathbb{E}R(X)$, the maximum entropy density is of the form $\alpha \exp(-\beta R(x))$ for constants α and β.

Example 1.1. Fix mean μ and variance σ^2 and write ϕ_{μ,σ^2} for the density of $Z_{\mu,\sigma^2} \sim N(\mu, \sigma^2)$. For random variable Y with density p_Y write $\Lambda(Y) = -\int p_Y(y) \log \phi_{\mu,\sigma^2}(y) \mathrm{d}y$. Then for any

* Corresponding address: Bristol University, Maths Department, University Walk, BS8 1TW Bristol, UK. Tel.: +44 117 9288 632; fax: +44 117 9287 999.
E-mail address: O.Johnson@bristol.ac.uk.

0304-4149/$ - see front matter
doi:10.1016/j.spa.2006.10.006

random variable X with mean μ, variance σ^2 and density p_X,

$$\Lambda(X) = -\int p_X(x)\log\phi_{\mu,\sigma^2}(x)\mathrm{d}x = \int p_X(x)\left(\frac{\log(2\pi\sigma^2)}{2} + \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x$$
$$= -\int \phi_{\mu,\sigma^2}(x)\log\phi_{\mu,\sigma^2}(x)\mathrm{d}x = \Lambda(Z_{\mu,\sigma^2}). \quad (1)$$

This means that, for any random variable X with mean μ and variance σ^2, the entropy H satisfies $H(X) \le H(Z_{\mu,\sigma^2})$, since Eq. (1) gives that $\Lambda(X) = \Lambda(Z_{\mu,\sigma^2}) = H(Z_{\mu,\sigma^2})$,

$$-H(X) + H(Z_{\mu,\sigma^2}) = \int p_X(x)\log p_X(x)\mathrm{d}x - \int p_X(x)\log\phi_{\mu,\sigma^2}(x)\mathrm{d}x. \quad (2)$$

This expression is the relative entropy $D(X \parallel Z_{\mu,\sigma^2})$, and is positive by the Gibbs inequality (see Eq. (18) below), with equality holding if and only if $p_X \equiv \phi_{\mu,\sigma^2}$.

This maximum entropy result can be regarded as the first stage in understanding the Central Limit theorem as a result concerning maximum entropy. Note that both the class of variables with mean μ and variance σ^2 (over which the entropy is maximised) and the maximum entropy variables Z_{μ,σ^2} are well-behaved on convolution. Further, the normalized sum of IID copies of any random variable X in this class converges in total variation to the maximum entropy distribution Z_{μ,σ^2}. The main theorem of Barron [2] extends this to prove convergence in relative entropy, assuming that $H(X) > -\infty$.

However, for functions R where $\mathbb{E}R(X)$ is not so well-behaved on convolution, the situation is more complicated. Examples of such random variables, for which we would hope to prove limit laws of a similar kind, include the Poisson and Cauchy families. In particular, we would like to understand the "Law of Small Numbers" convergence to the Poisson distribution as a maximum entropy result. Harremoës proved in [7] that the Poisson random variables Z_λ (with mass function $\Pi_\lambda(x) = \mathrm{e}^{-\lambda}\lambda^x/x!$ and mean λ) do satisfy a natural maximum entropy property.

Definition 1.2. For each $\lambda \ge 0$ and $n \ge 1$ define the classes

$$B_n(\lambda) = \left\{ S : \mathbb{E}S = \lambda,\ S = \sum_{i=1}^{n} X_i,\ \text{where } X_i \text{ are independent Bernoulli variables} \right\},$$

and $B_\infty(\lambda) = \bigcup_n B_n(\lambda)$.

Theorem 1.3 (*[7, Theorem 8]*). *For each $\lambda \ge 0$, the entropy of any random variable in class $B_\infty(\lambda)$ is less than or equal to the entropy of a Poisson random variable Z_λ:*

$$\sup_{S \in B_\infty(\lambda)} H(S) = H(Z_\lambda).$$

Note that Shepp and Olkin [18] and Mateev [14] also showed that the maximum entropy distribution in the class $B_n(\lambda)$ is Binomial$(n, \lambda/n)$.

In this paper, we show how this maximum entropy property relates to the property of log-concavity, and give an alternative proof, which shows that Z_λ is the maximum entropy random variable in a larger class $\mathbf{ULC}(\lambda)$.

2. Log-concavity and main theorem

First, recall the following definition:

Definition 2.1. A non-negative sequence $(u(i), i \geq 0)$ is log–concave if, for all $i \geq 1$,

$$u(i)^2 \geq u(i+1)u(i-1). \tag{3}$$

We say that a random variable V taking values in $\mathbb{Z}_+$ is log–concave if its probability mass function $P_V(i) = \mathbb{P}(V = i)$ forms a log–concave sequence. Any random variable $S \in B_\infty$ is log–concave, which is a corollary of the following theorem (see for example Theorem 1.2 on P.394 of [11]).

Theorem 2.2. *The convolution of any two log–concave sequences is log–concave.*

Among random variables, the extreme cases of log-concavity are given by the geometric family — that is, geometric probability mass functions are the only ones which achieve equality in Eq. (3) for all i. The argument of Example 1.1 shows that discrete entropy is maximised under a mean constraint by the geometric distribution. Hence, in the class of log–concave random variables with a given mean, the geometric is both the extreme and the maximum entropy distribution.

Unfortunately, the sum of two geometric distributions is a negative binomial distribution, which has a mass function which is log–concave but no longer achieves equality in (3). This means that under the condition of log-concavity the extreme cases and the maximum entropy family are not well-behaved under convolution. This suggests that log-concavity alone is too weak a condition to motivate an entropy-theoretic understanding of the Law of Small Numbers.

A more restrictive condition than log-concavity is ultra log-concavity, defined as follows:

Definition 2.3. A non-negative sequence $(u(i), i \geq 0)$ is ultra log–concave if the sequence $(u(i)i!, i \geq 0)$ is log–concave. That is, for all $i \geq 1$,

$$iu(i)^2 \geq (i+1)u(i+1)u(i-1). \tag{4}$$

Note that in Pemantle [16], Liggett [13], and Wang and Yeh [22], this property is referred to as 'ultra log-concavity of order ∞' — see Eq. (7) below for the definition of ultra log-concavity of order n.

An equivalent characterization of ultra log-concavity is that for any λ, the sequence of ratios $(u(i)/\Pi_\lambda(i))$ is log–concave. This makes it clear that among probability mass functions the extreme cases of ultra log-concavity, in the sense of equality holding in Eq. (4) for each i, are exactly the Poisson family, which is preserved on convolution.

Definition 2.4. For any $\lambda \geq 0$, define $\mathbf{ULC}(\lambda)$ to be the class of random variables V with mean $\mathbb{E}V = \lambda$ such that probability mass function P_V is ultra log–concave, that is

$$iP_V(i)^2 \geq (i+1)P_V(i+1)P_V(i-1), \quad \text{for all } i \geq 1. \tag{5}$$

An equivalent characterization of the class $\mathbf{ULC}(\lambda)$ is that the scaled score function introduced in [10] is decreasing, that is

$$\rho_V(i) = \frac{(i+1)P_V(i+1)}{\lambda P_V(i)} - 1 \text{ is a decreasing function in } i. \tag{6}$$

In Section 3 we discuss properties of the class $\mathbf{ULC}(\lambda)$. For example, Lemma 3.1 shows that (as for Harremoës's $B_\infty(\lambda)$) the $\mathbf{ULC}(\lambda)$ are well-behaved on convolution, and that $B_\infty(\lambda) \subset \mathbf{ULC}(\lambda)$, with $Z_\lambda \in \mathbf{ULC}(\lambda)$.

The main theorem of this paper is as follows:

Theorem 2.5. *For any $\lambda \geq 0$, if $X \in \mathbf{ULC}(\lambda)$ then the entropy of X satisfies*

$$H(X) \leq H(Z_\lambda),$$

with equality if and only if $X \sim Z_\lambda$.

We argue that this result gives the discrete analogue of the maximum entropy property of the normal distribution described in Example 1.1, since both the class $\mathbf{ULC}(\lambda)$ and the family Z_λ of maximum entropy random variables are preserved on convolution, and since $\mathbf{ULC}(\lambda)$ has another desirable property, that of "accumulation". That is, suppose we fix λ and take a triangular array of random variables $\{X_i^{(n)}\}$, where for $i = 1, \ldots, n$ the $X_i^{(n)}$ are IID and in $\mathbf{ULC}(\lambda/n)$. The techniques of [10] can be extended to show that as $n \to \infty$ the sum $X_1^{(n)} + \cdots + X_n^{(n)}$ converges to Z_λ in total variation (and indeed in relative entropy).

It is natural to wonder whether Theorem 2.5 is optimal, or whether for each λ there exists a strictly larger class $C(\lambda)$ such that (i) the $C(\lambda)$ are well-behaved on convolution (ii) Z_λ is the maximum entropy random variable in each $C(\lambda)$ (iii) accumulation holds. We do not offer a complete answer to this question though, as discussed above, the class of log–concave variables is too large and fails both conditions (i) and (ii).

For larger classes $C(\lambda)$, again consider a triangular array where $\{X_i^{(n)}\} \in C(\lambda/n)$. Write $p_n = \mathbb{P}(X_i^{(n)} > 0)$ and Q_n for the conditional distribution $Q_n(x) = \mathbb{P}(X_i^{(n)} = x | X_i^{(n)} > 0)$. If the classes $C(\lambda)$ are large enough that we can find a subsequence (n_k) such that $Q_{n_k} \to Q$ and $\mathbb{E}Q_{n_k} \to \mathbb{E}Q$, then the sum $X_1^{(n)} + \cdots + X_n^{(n)}$ converges to a compound Poisson distribution $CP(\lambda/\mathbb{E}Q, Q)$. Thus, if $C(\lambda)$ are large enough that we can find a limit $Q \not\equiv \delta_1$ then the limit is not equal to Z_λ and so the property of accumulation fails. (Note that for $X \in \mathbf{ULC}(\lambda)$ the $\mathbb{P}(X \geq 2 | X > 0) \leq (\exp(\lambda) - \lambda - 1)/\lambda$, so the only limiting conditional distribution is indeed δ_1.)

The proof of Theorem 2.5 is given in Sections 3 and 4, and is based on a family of maps $(\mathbf{U}_\alpha)$ which we introduce in Definition 4.1 below. This map mimics the role played by the Ornstein-Uhlenbeck semigroup in the normal case. In the normal case, differentiating along this semigroup shows that the probability densities satisfy a partial differential equation, the heat equation, and hence that the derivative of relative entropy is the Fisher information (a fact referred to as the de Bruijn identity — see [2]). This property is used by Stam [19] and Blachman [3] to prove the Entropy Power Inequality, which gives a sharp bound on the behaviour of continuous entropy on convolution. It is possible that a version of $\mathbf{U}_\alpha$ may give a similar result for discrete entropy.

As α varies between 1 and 0, the map $\mathbf{U}_\alpha$ interpolates between a given random variable X and a Poisson random variable with the same mean. By establishing monotonicity properties with respect to α, the maximum entropy result, Theorem 2.5, follows. The action of $\mathbf{U}_\alpha$ is to thin X and then to add an independent Poisson random variable to it. In Section 4, we use $\mathbf{U}_\alpha$ to establish the maximum entropy property of the Poisson distribution. The key expression is Eq. (8), which shows that the resulting probabilities satisfy an analogue of the heat equation.

We abuse terminology slightly in referring to $\mathbf{U}_\alpha$ as a semigroup; in fact (see Eq. (12) below) $\mathbf{U}_{\alpha_1} \circ \mathbf{U}_{\alpha_2} = \mathbf{U}_{\alpha_1\alpha_2}$, so we would require a reparametrization $\mathbf{W}_\theta = \mathbf{U}_{\exp(-\theta)}$ reminiscent of Bakry and Émery [1] to obtain the more familiar relation that $\mathbf{W}_{\theta_1} \circ \mathbf{W}_{\theta_2} = \mathbf{W}_{\theta_1+\theta_2}$. However,

in Section 5, we argue that $\mathbf{U}_\alpha$ has the 'right' parametrization, by proving Theorem 5.1 which shows that $H(\mathbf{U}_\alpha X)$ is not only monotonically decreasing in α, but is indeed a concave function of α. We prove this by writing $H(\mathbf{U}_\alpha X) = \Lambda(\mathbf{U}_\alpha X) - D(\mathbf{U}_\alpha X \parallel Z_\lambda)$, and differentiating both terms.

In contrast to conventions in Information theory, throughout the paper entropy is defined using logarithms to base e. However, scaling by a factor of log 2 restores the standard definitions.

3. Properties of ULC(λ) and definitions of maps

In this section, we first note some results concerning properties of the classes $\mathbf{ULC}(\lambda)$, before defining actions of addition and thinning that will be used to prove the main results of the paper.

Lemma 3.1. *For any $\lambda \geq 0$ and $\mu \geq 0$:*

(1) *If $V \in \mathbf{ULC}(\lambda)$ then it is log–concave.*
(2) *The Poisson random variable $Z_\lambda \in \mathbf{ULC}(\lambda)$.*
(3) *The classes are closed on convolution: that is for independent $U \in \mathbf{ULC}(\lambda)$ and $V \in \mathbf{ULC}(\mu)$, the sum $U + V \in \mathbf{ULC}(\lambda + \mu)$.*
(4) $B_\infty(\lambda) \subset \mathbf{ULC}(\lambda)$.

Proof. Parts (1) and (2) follow from the definitions. Theorem 1 of Walkup [21] implies that part (3) holds, though a more direct proof is given by Theorem 2 of Liggett [13]. Part (4) follows from part (3), since any Bernoulli(p) mass function scaled by Π_p is supported only on 2 points, so belongs to $\mathbf{ULC}(p)$. □

We can give an alternative proof of part (3) of Lemma 3.1, using ideas of negative association developed by Efron [6] and by Joag-Dev and Proschan [8]. The key result is that if U and V are log–concave random variables, then for any decreasing function ϕ

$$\mathbb{E}[\phi(U, V)|U + V = w] \text{ is a decreasing function of } w.$$

Now, the Lemma on P. 471 of [10] shows that, writing $\alpha = \mathbb{E}U/(\mathbb{E}U + \mathbb{E}V)$ and using the score function of Eq. (6), for independent U and V:

$$\rho_{U+V}(w) = \mathbb{E}[\alpha\rho_U(U) + (1-\alpha)\rho_V(V)|U + V = w],$$

so that if ρ_U and ρ_V are decreasing, then so is ρ_{U+V}.

Remark 3.2. For each n, the Poisson mass function Π_λ is not supported on $[0, n]$ and hence $Z_\lambda \notin B_n(\lambda)$, so that $Z_\lambda \notin B_\infty(\lambda)$. Indeed, we can see that the class of ultra log–concave random variables is non-trivially larger than the class of Bernoulli sums. For all random variables $V \in B_n(\lambda)$, the Newton inequalities (see for example Theorem 1.1 of Niculescu [15]) imply that the scaled mass function $P_V(i)/\binom{n}{i}$ is log–concave, so that for all $i \geq 1$:

$$\frac{i\,P_V(i)^2}{(i+1)\,P_V(i+1)\,P_V(i+1)} \geq \frac{n-i+1}{n-i}. \tag{7}$$

This is the property referred to by Pemantle [16] and Liggett [13] as "ultra log-concavity of order n", and is strictly more restrictive than simply ultra log-concavity which (see Eq. (5)) only requires a lower bound of 1 on the right-hand side.

Next we introduce the maps $\mathbf{S}_\beta$ and $\mathbf{T}_\alpha$ that will be key to our results.

Definition 3.3. Define the maps $\mathbf{S}_\beta$ and $\mathbf{T}_\alpha$ which act as follows:

(1) For any $\beta \geq 0$, define the map $\mathbf{S}_\beta$ that maps random variable X to random variable

$$\mathbf{S}_\beta X \sim X + Z_\beta,$$

where Z_β is a Poisson (β) random variable independent of X.

(2) For any $0 \leq \alpha \leq 1$, define the map $\mathbf{T}_\alpha$ that maps random variable X to random variable

$$\mathbf{T}_\alpha X \sim \sum_{i=1}^{X} B_i(\alpha),$$

where $B_i(\alpha)$ are Bernoulli (α) random variables, independent of each other and of X. This is the thinning operation introduced by Rényi [17].

We now show how these maps interact:

Lemma 3.4. *For any $0 \leq \alpha, \alpha_1, \alpha_2 \leq 1$ and for any $\beta, \beta_1, \beta_2 \geq 0$, the maps defined in Definition 3.3 satisfy:*

(1) $\mathbf{S}_{\beta_1} \circ \mathbf{S}_{\beta_2} = \mathbf{S}_{\beta_2} \circ \mathbf{S}_{\beta_1} = \mathbf{S}_{\beta_1+\beta_2}$.
(2) $\mathbf{T}_{\alpha_1} \circ \mathbf{T}_{\alpha_2} = \mathbf{T}_{\alpha_2} \circ \mathbf{T}_{\alpha_1} = \mathbf{T}_{\alpha_1\alpha_2}$.
(3) $\mathbf{T}_\alpha \circ \mathbf{S}_\beta = \mathbf{S}_{\alpha\beta} \circ \mathbf{T}_\alpha$.

Proof. Part (1) follows immediately from the definition. To prove part (2), we write $B_i(\alpha_1\alpha_2) = B_i(\alpha_1)B_i(\alpha_2)$ where $B_i(\alpha_1)$ and $B_i(\alpha_2)$ are independent, then for any X

$$\mathbf{T}_{\alpha_1\alpha_2} X \sim \sum_{i=1}^{X} B_i(\alpha_1)B_i(\alpha_2) = \sum_{i:B_i(\alpha_1)=1, i\leq X} B_i(\alpha_2) \sim \sum_{i=1}^{\mathbf{T}_{\alpha_1}X} B_i(\alpha_2).$$

Part (3) uses the fact that the sum of a Poisson number of IID Bernoulli random variables is itself Poisson. This means that for any X

$$\begin{aligned}\left(\mathbf{T}_\alpha \circ \mathbf{S}_\beta\right) X &\sim \sum_{i=1}^{\mathbf{S}_\beta X} B_i(\alpha) \\ &= \left(\sum_{i=1}^{X} B_i(\alpha)\right) + \left(\sum_{i=X+1}^{X+Z_\beta} B_i(\alpha)\right) \sim \mathbf{T}_\alpha X + Z_{\alpha\beta} \sim \left(\mathbf{S}_{\alpha\beta} \circ \mathbf{T}_\alpha\right) X,\end{aligned}$$

as required. □

Definition 3.5. Define the two-parameter family of maps

$$\mathbf{V}_{\alpha,\beta} = \mathbf{S}_\beta \circ \mathbf{T}_\alpha, \quad \text{for } 0 \leq \alpha \leq 1, \beta > 0.$$

As in Stam [19] and Blachman [3], we will differentiate along this family of maps, and see that the resulting probabilities satisfy a partial differential–difference equation.

Proposition 3.6. *Given X with mean λ, writing $P_\alpha(z) = \mathbb{P}(\mathbf{V}_{\alpha,f(\alpha)}X = z)$, then*

$$\frac{\partial}{\partial\alpha} P_\alpha(z) = g(\alpha)(P_\alpha(z) - P_\alpha(z-1)) - \frac{1}{\alpha}((z+1)P_\alpha(z+1) - zP_\alpha(z)), \tag{8}$$

where $g(\alpha) = f(\alpha)/\alpha - f'(\alpha)$. Equivalently, $f(\alpha) = \alpha f(1) + \alpha \int_\alpha^1 g(\beta)/\beta \mathrm{d}\beta$.

Proof. We consider probability generating functions (pgfs). Notice that

$$\mathbb{P}(\mathbf{T}_\alpha X = z) = \sum_{x \geq z} \mathbb{P}(X = x) \binom{x}{z} \alpha^z (1-\alpha)^{x-z},$$

so that if X has pgf $G_X(t) = \sum \mathbb{P}(X = x) t^x$, then $\mathbf{T}_\alpha X$ has pgf $\sum_z t^z \sum_{x \geq z} \mathbb{P}(X = x) \binom{x}{z} \alpha^z (1-\alpha)^{x-z} = \sum_x \mathbb{P}(X = x) \sum_{z=0}^x \binom{x}{z} (t\alpha)^z (1-\alpha)^{x-z} = G_X(t\alpha + 1 - \alpha)$.

If Y has pgf $G_Y(t)$ then $\mathbf{S}_\beta Y$ has pgf $G_Y(t) \exp(\beta(t-1))$. Overall then, $\mathbf{V}_{\alpha, f(\alpha)} X$ has pgf

$$G_\alpha(t) = G_X(t\alpha + (1-\alpha)) \exp(f(\alpha)(t-1)), \tag{9}$$

which satisfies

$$\frac{\partial}{\partial \alpha} G_\alpha(t) = (1-t) \left(\frac{1}{\alpha} \frac{\partial}{\partial t} G_\alpha(t) - G_\alpha(t) g(\alpha) \right),$$

and comparing coefficients the result follows. □

We now prove that both maps $\mathbf{S}_\beta$ and $\mathbf{T}_\alpha$ preserve ultra log-concavity.

Proposition 3.7. *If X is an ultra log–concave random variable then for any $\alpha \in [0,1]$ and $\beta \geq 0$ random variables $\mathbf{S}_\beta X$ and $\mathbf{T}_\alpha X$ are both ultra log–concave, and hence so is $\mathbf{V}_{\alpha,\beta} X$.*

Proof. The first result follows by part (3) of Lemma 3.1. We prove the second result using the case $f(\alpha) \equiv 0$ of Proposition 3.6, which tells us that writing $P_\alpha(x) = \mathbb{P}(\mathbf{T}_\alpha X = x)$, the derivative

$$\frac{\partial}{\partial \alpha} P_\alpha(x) = \frac{1}{\alpha} (x P_\alpha(x) - (x+1) P_\alpha(x+1)). \tag{10}$$

Writing $g_\alpha(z) = z P_\alpha(z)^2 - (z+1) P_\alpha(z+1) P_\alpha(z-1)$, Eq. (10) gives that for each z,

$$\begin{aligned} \frac{\partial}{\partial \alpha} g_\alpha(z) &= 2z \frac{g_\alpha(z)}{\alpha} + \frac{z+1}{\alpha} ((z+2) P_\alpha(z+2) P_\alpha(z-1) - z P_\alpha(z) P_\alpha(z+1)) \\ &= \left(2z - \frac{(z+2) P_\alpha(z+2)}{P_\alpha(z+1)} \right) \frac{g_\alpha(z)}{\alpha} - \frac{z P_\alpha(z)}{\alpha P_\alpha(z+1)} g_\alpha(z+1). \end{aligned} \tag{11}$$

We know that P_α is ultra log–concave for $\alpha = 1$, and will show that this holds for smaller values of α. Suppose that for some α, P_α is ultra log–concave, so for each z, $g_\alpha(z) \geq 0$. If for some z, $g_\alpha(z) = 0$ then since $g_\alpha(z+1) \geq 0$, Eq. (11) simplifies to give $\frac{\partial}{\partial \alpha} g_\alpha(z) \leq 0$. This means (by continuity) that there is no value of z for which $g_\alpha(z)$ can become negative as α gets smaller, so ultra log-concavity is preserved. □

4. Maximum entropy result for the Poisson distribution

We now prove the maximum entropy property of the Poisson distribution within the class $\mathbf{ULC}(\lambda)$. We choose a one-parameter family of maps $(\mathbf{U}_\alpha)$, which have the property that they preserve the mean λ.

Definition 4.1. Given mean $\lambda \geq 0$ and $0 \leq \alpha \leq 1$, define the combined map

$$\mathbf{U}_\alpha = \mathbf{V}_{\alpha, \lambda(1-\alpha)}.$$

Equivalently $\mathbf{U}_\alpha = \mathbf{S}_{\lambda(1-\alpha)} \circ \mathbf{T}_\alpha$ or $\mathbf{U}_\alpha = \mathbf{T}_\alpha \circ \mathbf{S}_{\lambda(1/\alpha - 1)}$.

Note that the maps $\mathbf{U}_\alpha$ have a semigroup-like structure — by Lemma 3.4 we know that $(\mathbf{S}_{\lambda(1-\alpha_1)} \circ \mathbf{T}_{\alpha_1}) \circ (\mathbf{S}_{\lambda(1-\alpha_2)} \circ \mathbf{T}_{\alpha_2}) = (\mathbf{S}_{\lambda(1-\alpha_1)} \circ \mathbf{S}_{\lambda\alpha_1(1-\alpha_2)}) \circ (\mathbf{T}_{\alpha_1} \circ \mathbf{T}_{\alpha_2}) = \mathbf{S}_{\lambda(1-\alpha_1\alpha_2)} \circ \mathbf{T}_{\alpha_1\alpha_2}$.

That is, we know that

$$\mathbf{U}_{\alpha_1} \circ \mathbf{U}_{\alpha_2} = \mathbf{U}_{\alpha_1\alpha_2}. \tag{12}$$

Eq. (8) can be simplified with the introduction of some helpful notation. Define Δ and its adjoint Δ^* by $\Delta p(x) = p(x+1) - p(x)$ and $\Delta^* q(x) = q(x-1) - q(x)$. These maps Δ and Δ^* are indeed adjoint since for any functions p, q:

$$\begin{aligned}\sum_x (\Delta p(x))q(x) &= \sum_x (p(x+1) - p(x))q(x) = \sum_x p(x)(q(x-1) - q(x)) \\ &= \sum_x p(x)(\Delta^* q(x)).\end{aligned} \tag{13}$$

We write $\rho_\alpha(z)$ for $\rho_{\mathbf{U}_\alpha X}(z) = (z+1)P_\alpha(z+1)/\lambda P_\alpha(z) - 1$. Then, noting that $(z+1)P_\alpha(z+1)/\lambda - P_\alpha(z) = P_\alpha(z)\rho_\alpha(z) = \Pi_\lambda(z)(P_\alpha(z+1)/\Pi_\lambda(z+1) - P_\alpha(z)/\Pi_\lambda(z))$, we can give two alternative reformulations of Eq. (8) in the case where $\mathbf{V}_{\alpha, f(\alpha)} = \mathbf{U}_\alpha$.

Corollary 4.2. *Writing $P_\alpha(z) = \mathbb{P}(\mathbf{U}_\alpha X = z)$:*

$$\frac{\partial}{\partial\alpha} P_\alpha(z) = \frac{\lambda}{\alpha}\Delta^*(P_\alpha(z)\rho_\alpha(z)). \tag{14}$$

Secondly, in a form more reminiscent of the heat equation:

$$\frac{\partial}{\partial\alpha} P_\alpha(z) = \frac{\lambda}{\alpha}\Delta^*\left(\Pi_\lambda(z)\Delta\left(\frac{P_\alpha(z)}{\Pi_\lambda(z)}\right)\right).$$

Note that we can also view $\mathbf{U}_\alpha$ as the action of the M/M/∞ queue. In particular Eq. (8), representing the evolution of probabilities under $\mathbf{U}_\alpha$, is the adjoint of

$$Lf(z) = -\lambda\Delta\Delta^* f(z) + (z-\lambda)\Delta^* f(z),$$

representing the evolution of functions. This equation is the polarised form of the infinitesimal generator of the M/M/∞ queue, as described in Section 1.1 of Chafaï [4]. Chafaï uses this equation to prove a number of inequalities concerning generalized entropy functionals.

Proof of Theorem 2.5. Given random variable X with mass function P_X, we define $\Lambda(X) = -\sum_x P_X(x)\log\Pi_\lambda(x)$. Notice that (as remarked by Topsøe [20]), the conditions required in Example 1.1 can be weakened. If $\Lambda(X) \le \Lambda(Z_\lambda) = H(Z_\lambda)$ then adapting Eq. (2) gives that $-H(X) + H(Z_\lambda) \ge -H(X) + \Lambda(X) = D(X \parallel Z_\lambda) \ge 0$, and we can deduce the maximum entropy property.

We will in fact show that if $X \in \mathbf{ULC}(\lambda)$ then $\Lambda(\mathbf{U}_\alpha X)$ is an decreasing function of α. In particular, since $\mathbf{U}_0 X \sim Z_\lambda$, and $\mathbf{U}_1 X \sim X$, we deduce that $\Lambda(X) \le \Lambda(Z_\lambda)$. (A similar technique of controlling the sign of the derivative is used by Blachman [3] and Stam [19] to prove the Entropy Power Inequality.)

We simply differentiate and use Eqs. (13) and (14). Note that

$$\begin{aligned}\frac{\partial}{\partial\alpha}\Lambda(\mathbf{U}_\alpha X) &= -\frac{\lambda}{\alpha}\sum_z \Delta^*(P_\alpha(z)\rho_\alpha(z))\log\Pi_\lambda(z) \\ &= -\frac{\lambda}{\alpha}\sum_z P_\alpha(z)\rho_\alpha(z)\Delta(\log\Pi_\lambda(z)) \\ &= \frac{\lambda}{\alpha}\sum_z P_\alpha(z)\rho_\alpha(z)\log\left(\frac{z+1}{\lambda}\right).\end{aligned} \tag{15}$$

By assumption $X \in \mathbf{ULC}(\lambda)$, so by Proposition 3.7 $\mathbf{U}_\alpha X \in \mathbf{ULC}(\lambda)$, which is equivalent to saying that the score function $\rho_\alpha(z)$ is decreasing in z. Further, note that $\sum_z P_\alpha(z)\rho_\alpha(z) = 0$. Since $\log((z+1)/\lambda)$ is increasing in z (a fact which is equivalent to saying that the Poisson mass function $\Pi_\lambda(z)$ is itself log–concave), $\frac{\partial}{\partial\alpha}\Lambda(\mathbf{U}_\alpha X)$ is negative by Chebyshev's rearrangement lemma, since it is the covariance of a decreasing and increasing function.

In fact, $\Lambda(\mathbf{U}_\alpha X)$ is strictly decreasing in α, unless X is Poisson. This follows since equality holds in Eq. (15) if and only if $\rho_\alpha(z) \equiv 0$, which characterizes the Poisson distribution. □

5. Concavity of entropy along the semigroup

In fact, rather than just showing that the Poisson distribution has a maximum entropy property, in this section we establish a stronger result, as follows.

Theorem 5.1. *If $X \in \mathbf{ULC}(\lambda)$, then the entropy of $\mathbf{U}_\alpha X$ is a decreasing and concave function of α, that is*

$$\frac{\partial}{\partial\alpha} H(\mathbf{U}_\alpha X) \le 0 \quad \text{and} \quad \frac{\partial^2}{\partial\alpha^2} H(\mathbf{U}_\alpha X) \le 0,$$

with equality if and only if $X \sim \Pi_\lambda$.

Proof. The proof is contained in the remainder of this section, and involves writing $H(\mathbf{U}_\alpha X) = \Lambda(\mathbf{U}_\alpha X) - D(\mathbf{U}_\alpha X \parallel Z_\lambda)$, and differentiating both terms.

We have already shown in Eq. (15) that $\Lambda(\mathbf{U}_\alpha X)$ is decreasing in α. We show in Lemma 5.3 that it is concave in α, and in Lemmas 5.2 and 5.5 respectively we show that $D(\mathbf{U}_\alpha X \parallel Z_\lambda)$ is increasing and convex. Some of the proofs of these lemmas are merely sketched, since they involve long algebraic manipulations using Eq. (14). □

In the case of continuous random variables, Costa [5] uses the concavity of the entropy power on addition of an independent normal variable (a stronger result than concavity of entropy itself) to prove a version of the Entropy Power Inequality. We regard Theorem 5.1 as the first stage in a similar proof of a discrete form of the Entropy Power Inequality.

Lemma 5.2. *For X with mean λ, $D(\mathbf{U}_\alpha X \parallel Z_\lambda)$ is an increasing function of α.*

Proof. We use Eq. (14). Note that (omitting arguments for the sake of brevity):

$$\frac{\partial}{\partial\alpha}\sum P_\alpha \log\left(\frac{P_\alpha}{\Pi_\lambda}\right) = \sum \frac{\partial P_\alpha}{\partial\alpha}\log\left(\frac{P_\alpha}{\Pi_\lambda}\right) + \sum\frac{\partial P_\alpha}{\partial\alpha} = \sum \frac{\partial P_\alpha}{\partial\alpha}\log\left(\frac{P_\alpha}{\Pi_\lambda}\right).$$

This means that

$$\begin{aligned}\frac{\partial}{\partial\alpha} D(\mathbf{U}_\alpha X \parallel Z_\lambda) &= \frac{\lambda}{\alpha}\sum_z \Delta^*(P_\alpha(z)\rho_\alpha(z)) \log\left(\frac{P_\alpha(z)}{\Pi_\lambda(z)}\right)\\ &= \frac{\lambda}{\alpha}\sum_z P_\alpha(z)\rho_\alpha(z)\log\left(\frac{P_\alpha(z+1)\Pi_\lambda(z)}{P_\alpha(z)\Pi_\lambda(z+1)}\right)\\ &= \frac{\lambda}{\alpha}\sum_z P_\alpha(z)\rho_\alpha(z)\log(1+\rho_\alpha(z)). \qquad (16)\end{aligned}$$

Now, as in [10], we write $\widetilde{P}_\alpha(z) = (z+1)P_\alpha(z+1)/\lambda$. $\widetilde{P}_\alpha$ is often referred to as the size-biased version of P_α, and is a probability mass function because $\mathbf{U}_\alpha$ fixes the mean. Notice that

$\rho_\alpha(z) = \widetilde{P}_\alpha(z)/P_\alpha(z) - 1$, so that we can rewrite Eq. (16) as

$$\frac{\lambda}{\alpha}\sum_z (\widetilde{P}_\alpha(z) - P_\alpha(z)) \log\left(\frac{\widetilde{P}_\alpha(z)}{P_\alpha(z)}\right) = \frac{\lambda}{\alpha}(D(P_\alpha \parallel \widetilde{P}_\alpha) + D(\widetilde{P}_\alpha \parallel P_\alpha)) \geq 0. \tag{17}$$

This quantity is a symmetrised version of the relative entropy, and was originally introduced by Kullback and Leibler in [12]. □

Lemma 5.3. *Using the definitions above, if $X \in \mathbf{ULC}(\lambda)$ then $\Lambda(\mathbf{U}_\alpha X)$ is a concave function of α. It is strictly concave unless X is Poisson.*

Sketch Proof. Using Eqs. (14) and (15), it can be shown that

$$\frac{\partial^2}{\partial\alpha^2}\Lambda(\mathbf{U}_\alpha X) = \frac{\lambda^2}{\alpha^2}\sum_z P_\alpha(z)\rho_\alpha(z)\left(\frac{z}{\lambda}\log\left(\frac{z+1}{z}\right) - \log\left(\frac{z+2}{z+1}\right)\right).$$

Now, the result follows in the same way as before, since for any λ the function $z/\lambda\log((z+1)/z) - \log((z+2)/(z+1))$ is increasing, so $\frac{\partial^2}{\partial\alpha^2}\Lambda(\mathbf{U}_\alpha X) \leq 0$. □

Taking a further derivative of Eq. (17), we can show that (the proof is omitted for the sake of brevity):

Lemma 5.4. *The relative entropy $D(\mathbf{U}_\alpha X \parallel Z_\lambda)$ satisfies*

$$\frac{\partial^2}{\partial\alpha^2}\sum_z P_\alpha(z)\log\left(\frac{P_\alpha(z)}{\Pi_\lambda(z)}\right) = \frac{\lambda^2}{\alpha^2}\sum_z (\widetilde{\widetilde{P}}_\alpha(z) - 2\widetilde{P}_\alpha(z) + P_\alpha(z))\log\left(\frac{\widetilde{\widetilde{P}}_\alpha(z)P_\alpha(z)}{\widetilde{P}_\alpha(z)^2}\right) + \sum_z P_\alpha(z)\left(\frac{1}{P_\alpha(z)}\frac{\partial P_\alpha}{\partial\alpha}(z)\right)^2$$

where $\widetilde{P}_\alpha(z) = (z+1)P_\alpha(z+1)/\lambda$ and $\widetilde{\widetilde{P}}_\alpha(z) = (z+2)(z+1)P_\alpha(z+2)/\lambda^2$.

Lemma 5.5. *For X with mean λ and $\mathrm{Var}\,X \leq \lambda$, $D(\mathbf{U}_\alpha X \parallel Z_\lambda)$ is a convex function of α. It is a strictly convex function unless X is Poisson.*

Proof. Notice that the map $\mathbf{T}_\alpha$ scales the rth falling moment of X by α^r. This means that $\mathrm{Var}\,\mathbf{T}_\alpha X = \alpha^2\mathrm{Var}\,X + \alpha(1-\alpha)\lambda$, so that $\mathrm{Var}\,\mathbf{U}_\alpha X = \alpha^2\mathrm{Var}\,X + \lambda(1-\alpha^2)$. Hence, the condition $\mathrm{Var}\,X \leq \lambda$ implies that for all α, $\mathrm{Var}\,\mathbf{U}_\alpha X \leq \lambda$. Equivalently, $S := \sum_z \widetilde{\widetilde{P}}_\alpha(z) = \mathbb{E}(\mathbf{U}_\alpha X)(\mathbf{U}_\alpha X - 1)/\lambda^2 < 1$.

We will use the log–sum inequality, which is equivalent to the Gibbs inequality and states that for positive sequences (a_i) and (b_i) (not necessarily summing to 1),

$$D(a_i \parallel b_i) = \sum_i a_i \log(a_i/b_i) \geq \left(\sum_i a_i\right)\log\left(\frac{\sum_i a_i}{\sum_i b_i}\right). \tag{18}$$

Since $\log u \leq u-1$, this simplifies further to give $D(a_i \parallel b_i) \geq (\sum_i a_i)(\log(\sum_i a_i) + 1 - \sum_i b_i)$.

We express the first term of Lemma 5.4 as a sum of relative entropies, and recall that $\sum_z P_\alpha(z) = 1$ and $\sum_z \widetilde{P}_\alpha(z) = 1$, simplifying the bounds on the second and third terms:

$$\frac{\lambda^2}{\alpha^2}\left(D\left(\widetilde{\widetilde{P}}_\alpha \middle\| \frac{\widetilde{P}_\alpha^2}{P_\alpha}\right) + 2D\left(\widetilde{P}_\alpha \middle\| \frac{\widetilde{\widetilde{P}}_\alpha P_\alpha}{\widetilde{P}_\alpha}\right) + D\left(P_\alpha \middle\| \frac{\widetilde{P}^2}{\widetilde{\widetilde{P}}}\right)\right)$$

$$\geq \frac{\lambda^2}{\alpha^2}\left(S\log S + S - S\sum_z \frac{(z+1)^2 P_\alpha(z+1)^2}{\lambda^2 P_\alpha(z)}\right.$$

$$\left. + \ 2 - 2\sum_z \frac{(z+1)P_\alpha(z+1)P_\alpha(z-1)}{\lambda P_\alpha(z)} + 1 - \sum_z \frac{(z-1)P_\alpha(z-1)^2}{zP_\alpha(z)}\right). \tag{19}$$

Using Eq. (8) we can expand the second (Fisher) term of Lemma 5.4 as

$$= \frac{\lambda^2}{\alpha^2}\left(-3 - \frac{\mathbb{E}(\mathbf{U}_\alpha X)^2}{\lambda^2} + \sum_z \frac{(z+1)^2 P_\alpha(z+1)^2}{\lambda^2 P_\alpha(z)}\right.$$

$$\left. + 2\sum_z \frac{(z+1)P_\alpha(z+1)P_\alpha(z-1)}{\lambda P_\alpha(z)} + \sum_z \frac{P_\alpha(z-1)^2}{P_\alpha(z)}\right). \tag{20}$$

Adding Eq. (19) and (20), and since $S = \mathbb{E}(\mathbf{U}_\alpha X)^2/\lambda^2 - 1/\lambda$, we deduce that

$$\frac{\partial^2}{\partial\alpha^2} D(\mathbf{U}_\alpha X \parallel \Pi_\lambda)$$

$$\geq \frac{\lambda^2}{\alpha^2}\left(S\log S + (1-S)\sum_z \frac{(z+1)^2 P_\alpha(z+1)^2}{\lambda^2 P_\alpha(z)} + \sum_z \frac{P_\alpha(z-1)^2}{zP_\alpha(z)} - \frac{1}{\lambda}\right). \tag{21}$$

Finally we exploit Cramér–Rao type relations which bound the two remaining quadratic terms from below. Firstly, as in [10]:

$$0 \leq \sum_z P_\alpha(z)\left(\frac{(z+1)P_\alpha(z+1)}{\lambda P_\alpha(z)} - 1\right)^2 = \sum_z \frac{(z+1)^2 P_\alpha(z+1)^2}{\lambda^2 P_\alpha(z)} - 1. \tag{22}$$

Similarly, a weighted version of the Fisher information term of Johnstone and MacGibbon [9] gives that:

$$0 \leq \sum_z P_\alpha(z) z\left(\frac{P_\alpha(z-1)}{zP_\alpha(z)} - \frac{1}{\lambda}\right)^2 = \sum_z \frac{P_\alpha(z-1)^2}{zP_\alpha(z)} - \frac{1}{\lambda}. \tag{23}$$

(Note that in Eqs. (22) and (23), equality holds if and only if $P_\alpha \equiv \Pi_\lambda$). Substituting Eqs. (22) and (23) in Eq. (21), we deduce that

$$\frac{\partial^2}{\partial\alpha^2} D(\mathbf{U}_\alpha X \parallel Z_\lambda) \geq \frac{\lambda^2}{\alpha^2}(S\log S + 1 - S) \geq 0,$$

with equality if and only if $S = 1$. □

Combining these lemmas, the proof of Theorem 5.1 is complete, since ultra log-concavity of X implies that $\mathrm{Var}\, X \leq \mathbb{E}X$, as $\sum_x P_X(x)x((x+1)P_X(x+1)/P_X(x) - \lambda) \leq 0$ since it is again the covariance of an increasing and decreasing function.

Acknowledgments

The author would like to thank Christophe Vignat, Ioannis Kontoyiannis, Peter Harremoës, Andrew Barron and Mokshay Madiman for useful discussions concerning this paper, and would like to thank EPFL Lausanne and Yale University for financial support on visits to these colleagues. The author would also like to thank Djalil Chafaï for explaining the connection with the $M/M/\infty$ queue, and two anonymous referees for their very helpful comments, including a simplified proof of Proposition 3.7.

References

[1] D. Bakry, M. Émery, Diffusions hypercontractives, in: Séminaire de probabilités, XIX, 1983/84, in: Lecture Notes in Math., vol. 1123, Springer, Berlin, 1985, pp. 177–206.
[2] A.R. Barron, Entropy and the central limit theorem, Ann. Probab. 14 (1) (1986) 336–342.
[3] N.M. Blachman, The convolution inequality for entropy powers, IEEE Trans. Inform. Theory 11 (1965) 267–271.
[4] D. Chafaï, Binomial-Poisson entropic inequalities and the $M/M/\infty$ queue, ESAIM Probab. Stat. 10 (2006) 317–339.
[5] M.H.M. Costa, A new entropy power inequality, IEEE Trans. Inform. Theory 31 (6) (1985) 751–760.
[6] B. Efron, Increasing properties of Pólya frequency functions, Ann. Math. Statist. 33 (1965) 272–279.
[7] P. Harremoës, Binomial and Poisson distributions as maximum entropy distributions, IEEE Trans. Inform. Theory 47 (5) (2001) 2039–2041.
[8] K. Joag-Dev, F. Proschan, Negative association of random variables with applications, Ann. Statist. 11 (1983) 286–295.
[9] I. Johnstone, B. MacGibbon, Une mesure d'information caractérisant la loi de Poisson, in: Séminaire de Probabilités, XXI, Springer, Berlin, 1987, pp. 563–573.
[10] I. Kontoyiannis, P. Harremoës, O.T. Johnson, Entropy and the law of small numbers, IEEE Trans. Inform. Theory 51 (2) (2005) 466–472.
[11] S. Karlin, Total Positivity, Stanford University Press, Stanford CA, 1968.
[12] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
[13] T.M. Liggett, Ultra logconcave sequences and negative dependence, J. Combin. Theory Ser. A 79 (2) (1997) 315–325.
[14] P. Mateev, The entropy of the multinomial distribution, Teor. Verojyatnost. i Primenen. 23 (1) (1978) 196–198.
[15] C.P. Niculescu, A new look at Newton's inequalities, JIPAM. J. Inequal. Pure Appl. Math. 1 (2) (2000) Article 17. See also http://jipam.vu.edu.au/.
[16] R. Pemantle, Towards a theory of negative dependence, J. Math. Phys. 41 (3) (2000) 1371–1390. Probabilistic Techniques in Equilibrium and Nonequilibrium Statistical Physics.
[17] A. Rényi, A characterization of Poisson processes, Magyar Tud. Akad. Mat. Kutató Int. Közl. 1 (1956) 519–527.
[18] L.A. Shepp, I. Olkin, Entropy of the sum of independent Bernoulli random variables and of the multinomial distribution, in: Contributions to Probability, Academic Press, New York, 1981, pp. 201–206.
[19] A.J. Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon, Inform. Control 2 (1959) 101–112.
[20] F. Topsøe, Maximum entropy versus minimum risk and applications to some classical discrete distributions, IEEE Trans. Inform. Theory 48 (8) (2002) 2368–2376.
[21] D.W. Walkup, Pólya sequences, binomial convolution and the union of random sets, J. Appl. Probab. 13 (1) (1976) 76–85.
[22] Y. Wang, Y.-N. Yeh, Log-concavity and LC-positivity, J. Combin. Theory Ser. A, 2006 (in press). Available at: arXiv:math.CO/0504164.