Binomial and Poisson Distributions as Maximum Entropy Distributions

Peter Harremoës

***Abstract*—The binomial and the Poisson distributions are shown to be maximum entropy distributions of suitably defined sets. Poisson's law is considered as a case of entropy maximization, and also convergence in information divergence is established.**

***Index Terms*—Binomial distribution, entropy, generalized binomial distribution, information divergence, Poisson distribution, Poisson's law.**

I. Introduction

We shall use $\Pi(\lambda)$ to denote the Poisson distribution with mean λ, and $b(n, p)$ to denote the binomial distribution with parameters (n, p). We will not distinguish between a random variable and its distribution in the notation. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent Bernoulli random variables, i.e., random variables with range $\{0, 1\}$. Define the *success probabilities* by $p_i = P(X_i = 1)$, $\lambda = \sum_1^n p_i$, $p_{\max} = \max_i p_i$, and $S_n = \sum_1^n X_i$. We call S_n an *n-generalized binomial distribution* and denote by $B_n(\lambda)$ the set of n-generalized binomial distributions with mean λ. Define *the set of generalized binomial distributions* $B_\infty(\lambda)$ as the union $\bigcup B_n(\lambda)$ of all n-generalized binomial distributions.

Let P and Q be probability measures on $\{0, 1, 2, \ldots\}$ with point probabilities p_i and q_i, $i = 0, 1, 2, \ldots$. Then the total variation between the distributions is defined as

$$\|P - Q\| = \sum_{i=0}^{\infty} |p_i - q_i|$$

and the information divergence is defined as

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

The basic properties of the information divergence are described, for instance, in [1].

The convergence of the point probabilities of $b(n, \frac{\lambda}{n})$ to the point probabilities of $\Pi(\lambda)$ was established by Poisson. Convergence in total variation was studied by Prohorov [2] for the binomial distribution. Convergence of more general distributions are studied in [3]–[6]. See Steele [7] for a survey on the subject and further references. Information divergence does not define a metric but is related to total variation via *Pinsker's inequality* $\frac{1}{2}\|P - Q\|^2 \leq D(P\|Q)$ proved by Csiszár [8] and others. If $(Q_n)_{n\in\mathbb{N}}$ is a sequence of probability distributions, we say that $(Q_n)_{n\in\mathbb{N}}$ converges to Q in *information divergence* if $D(Q_n\|Q) \to 0$ for $n \to \infty$. In Section II, it is shown that $b(n, \frac{\lambda}{n})$ converges to $\Pi(\lambda)$ in information divergence, and the proof is at least as simple as the proof of convergence in total variation. Pinsker's inequality shows that convergence in information divergence is a stronger condition than convergence in total variation. The use of information divergence also fits better together with the idea of maximum-likelihood estimation known from statistics.

The entropy of P is defined by

$$H(P) = -\sum_{i=0}^{\infty} p_i \log p_i.$$

If Ω is a set of distributions we define $H(\Omega) = \sup_{P\in\Omega}(H(P))$.

Manuscript received October 11, 2000.

The author is with Aurehøj Amtsgymnasium, Københavns Amt, DK-2860 Soborg, Denmark (e-mail: moes@post7.tele.dk).

Communicated by I. Csiszár, Associate Editor for Shannon Theory.

Publisher Item Identifier S 0018-9448(01)04417-0.

In Section III, it is shown that both the binomial distributions and the Poisson distributions are maximum-entropy distributions on sets of generalized binomial distributions. Also Poisson's law is shown to be closely related to the maximum-entropy principle in the sense that Poisson's law can be formulated as "the entropy increases to its maximum." In this sense, these results are closely related to results about the central limit theorem obtained by Barron in [9] and Takano in [10].

II. Poisson's Law

Assume X_1 and X_2 are independent Poisson distributed random variables with intensities λ and μ. Then $X_1 + X_2$ is a Poisson distributed random variable with intensity $\lambda+\mu$, which shows that Poisson distributions are infinitely divisible. Let X be a random variable with values in $\{0, 1, 2, \ldots\}$ and with point probabilities p_i. Then

$$\begin{aligned} D(X\|\Pi(\lambda)) &= \sum_{j=0}^{\infty} p_j \log\left(\frac{p_j}{\frac{\lambda^j}{j!}e^{-\lambda}}\right) \\ &= \lambda + \sum_{j=0}^{\infty} p_j \log\left(\frac{j!}{\lambda^j}\right) - H(X) \\ &= \lambda - E(X)\log\lambda + E(\log(X!)) - H(X) \end{aligned}$$

and the derivative with respect to λ is $1-\frac{E(X)}{\lambda}$. Therefore, $D(X\|\Pi(\lambda))$ is minimal for $\lambda = E(X)$. Equivalently, $\lambda = E(X)$ is the maximum-likelihood estimate given an empirical distribution according to X. Now it is convenient to define

$$D(X) = \min_{\lambda} D(X\|\Pi(\lambda)).$$

If total variation is used to measure the difference between the distributions, the maximum-likelihood estimate is not the nearest distribution. In [11]–[13], bounds on the total variation between the distribution of X and the nearest Poisson distribution are given.

Lemma 1: For independent random variables X_1 and X_2 we have

$$D(X_1 + X_2) \leq D(X_1) + D(X_2). \tag{1}$$

Proof: First we observe that

$$\begin{aligned} D(X_1) + D(X_2) &= D(X_1\|\Pi(\lambda_1)) + D(X_2\|\Pi(\lambda_2)) \\ &= D((X_1, X_2)\|(\Pi(\lambda_1), \Pi(\lambda_2))) \end{aligned}$$

where $\Pi(\lambda_1)$ and $\Pi(\lambda_2)$ are considered as independent Poisson distributions. The inequality (1) is obtained by data reduction of the map $(X_1, X_2) \to X_1 + X_2$. □

Theorem 2: Let $X_1, X_2 \cdots X_n$ be a sequence of independent Bernoulli random variables. Define $p_i = P(X_i = 1)$, $\lambda = \sum^n p_i$, and $S_n = \sum^n X_i$. Then

$$D(S_n) \leq \sum_{i=1}^{n} p_i^2 \leq \lambda \cdot p_{\max}.$$

Proof: We have

$$\begin{aligned} D(X_i) &= (1-p_i)\ln\left(\frac{1-p_i}{\exp(-p_i)}\right) + p_i \ln\left(\frac{p_i}{p_i\exp(-p_i)}\right) \\ &= (1-p_i)\ln(1-p_i) + p_i \\ &\leq (1-p_i)(-p_i) + p_i \\ &= p_i^2. \end{aligned}$$

and, therefore,

$$D(S_n) \le \sum_{i=1}^{n} D(X_i) \le \sum_{i=1}^{n} p_i^2. \qquad \square$$

We see that if λ is fixed and $p_{\max}$ converges to 0 then $D(S_n)$ converges to 0, which is Poisson's law. If the Bernoulli random variables are identically distributed we get $D(S_n) \le \lambda p_{\max} = \frac{\lambda^2}{n}$.

Remark 3: The bound can easily be improved by use of the inequality

$$\begin{aligned} D(X_i) &= (1-p_i)\ln(1-p_i) + p_i \\ &\le (1-p_i)\left(-p_i - \frac{p_i^2}{2} - \frac{p_i^3}{3}\right) + p_i \\ &= \frac{1}{2}p_i^2 + \frac{1}{6}p_i^3 + \frac{1}{3}p_i^4 \end{aligned}$$

which gives

$$\begin{aligned} D(S_n) &\le \sum_{i=1}^{n} D(X_i) \\ &= \sum_{i=1}^{n}\left(\frac{p_i^2}{2} + \frac{p_i^3}{6} + \frac{p_i^4}{3}\right) \\ &\le \lambda \cdot \left(\frac{p_{\max}}{2} + \frac{p_{\max}^2}{6} + \frac{p_{\max}^3}{3}\right). \end{aligned}$$

III. Maximum-Entropy Distributions

In order to study the entropy of generalized binomial distributions, we need the following lemma which is a strengthening of a result obtained by Shepp and Olkin [14, Lemma 1] . Basically, we use the same proof technique as these authors. We shall need the elementary symmetric functions

$$s_k^n(x_1, x_2, \ldots, x_n) = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} x_{i_1} \cdot x_{i_2} \cdot \cdots \cdot x_{i_k}$$

defined for $x_1 > 0,\ x_2 > 0, \ldots, x_n > 0$. These functions satisfy the following inequalities:

$$s_l^n \cdot s_{l+2}^n \le (s_{l+1}^n)^2. \tag{2}$$

A proof of (2) can be found in [15, Sec. 2.22].

Lemma 4: The entropy $H(S_n)$ is a strictly concave function of p_i, p_j $i \ne j$ when all other probabilities p_k, $k \ne i, j$ are kept fixed and $E(S_n)$ is fixed.

Proof: Without loss of generality we can assume that $i = 1$ and $j = 2$. When the other probabilities are kept fixed we have $p_1 + p_2 = k$ for some constant k. Define $t = p_1 - \frac{k}{2}$. We have to show that $\frac{d^2}{dt^2} H(S_n) < 0$. The distribution of $X_1 + X_2$ is given by the point probabilities

$$\begin{aligned} &(p_1p_2,\ p_1(1-p_2) + (1-p_1)p_2,\ (1-p_1)(1-p_2)) \\ &\qquad = \left(\frac{k^2}{4} - t^2,\ k - \frac{k^2}{2} + 2t^2,\ \left(1-\frac{k}{2}\right)^2 - t^2\right). \end{aligned}$$

Therefore, the distribution of S_n is an affine function of t^2. Put $u = t^2$. Then we have

$$\begin{aligned} \frac{d^2}{dt^2} H(S_n) &= \frac{d}{dt}\left(\frac{du}{dt} \cdot \frac{d}{du} H(S_n)\right) \\ &= 2 \cdot \frac{d}{du} H(S_n) + \left(\frac{du}{dt}\right)^2 \cdot \frac{d^2}{du^2} H(S_n). \end{aligned}$$

The last term is negative by concavity of the entropy function. We shall show that also the first term $\frac{d}{du} H(S_n)$ is less than or equal to 0.

Define $b_l = P(X_3 + \cdots + X_n = l)$. Then we have

$$\begin{aligned} P(S_n = l) = \left(\frac{k^2}{4} - u\right) b_{l-2} + \left(k - \frac{k^2}{2} + 2u\right) b_{l-1} \\ + \left(\left(1 - \frac{k}{2}\right)^2 - u\right) b_l \end{aligned}$$

and get

$$\begin{aligned} \frac{d}{du} H(S_n) &= -\frac{d}{du}\left(\sum_l P(S_n = l)\log P(S_n = l)\right) \\ &= -\sum_l \frac{dP(S_n = l)}{du}(\log P(S_n = l) + 1) \\ &= -\sum_l (-b_{l-2} + 2b_{l-1} - b_l)\log P(S_n = l) \\ &= \sum_l \log\left(\frac{P(S_n = l) \cdot P(S_n = l+2)}{P(S_n = l+1)^2}\right) \cdot b_l. \end{aligned}$$

Now

$$P(S_n = l) = s_l^n\left(\frac{p_1}{1-p_1}, \frac{p_2}{1-p_2}, \ldots, \frac{p_n}{1-p_n}\right) \cdot \prod_k (1 - p_k)$$

and using (2) gives

$$\frac{P(S_n = l) \cdot P(S_n = l+2)}{P(S_n = l+1)^2} = \frac{s_l^n \cdot s_{l+2}^n}{(s_{l+1}^n)^2} \le 1$$

which shows that

$$\frac{d}{du} H(S_n) \le 0. \qquad \square$$

The lemma gives more evidence to the following conjecture stated by Shepp and Olkin [14, p. 4].

Conjecture 5: The entropy $H(S_n)$ is a concave function of the vector $(p_1, p_2, \ldots, p_n)$.

Theorem 6: If $m = \lfloor \frac{\lambda}{p_{\max}} \rfloor$, then

$$H(S_n) \ge H\left(b\left(m, \frac{\lambda}{m}\right)\right).$$

Proof: Let K be the set of n-generalized binomial distributions with mean λ, with success probabilities q_i, and with $q_{\max} \le \frac{\lambda}{m}$. Then there exists a generalized binomial distribution $R \in K$ with success probabilities r_i where $H(R) = \min_{P \in K} H(P)$. If there were two success probabilities r_i and r_j in $]0; \frac{\lambda}{m}[$ with $i \ne j$, then the generalized binomial distribution with the same success probabilities except r_i replaced by $r_i \pm \varepsilon$ and r_j replaced by $r_j \mp \varepsilon$ would have lower entropy than R for some small number ε. Therefore, there is at most one success probability in $]0\,;\frac{\lambda}{m}[$. Assume $r_j\in]0;\,\frac{\lambda}{m}[$. Let l be the number of success probabilities r_j with $r_j=\frac{\lambda}{m}$. Then we have $l\frac{\lambda}{m}+r_j=\lambda$ which is not possible. Hence, all $r_i\in\{0,\,\frac{\lambda}{m}\}$ and R is a binomial distribution with parameters $(m,\,\frac{\lambda}{m})$, and the result follows. □

Theorem 7: The binomial distribution $b(n,\,\frac{\lambda}{n})$ is the maximum-entropy distribution in $B_n(\lambda)$, and for λ fixed $H(b(n,\,\frac{\lambda}{n}))$ is increasing and

$$H\left(b\left(n,\,\frac{\lambda}{n}\right)\right)\nearrow H\left(B_\infty(\lambda)\right),\qquad \text{for } n\to\infty.$$

Proof: There exists an n-generalized binomial distribution R with success probabilities r_i where $H(R)=H(B_n(\lambda))$. By Lemma 4 and symmetry we have that $r_i=r_j$ for all $i,\,j$. Therefore, R is a binomial distribution with parameters $(n,\,\frac{\lambda}{n})$. To see that the sequence $(H(b(n,\,\frac{\lambda}{n})))_{n\in\mathbb{N}}$ is increasing we note that the sequence of sets $(B_n(\lambda))_{n\in\mathbb{N}}$ is increasing. □

Theorem 8: The Poisson distribution $\Pi(\lambda)$ satisfies

$$H\left(\Pi(\lambda)\right)=H\left(B_\infty(\lambda)\right).$$

Proof: The geometric distribution is the maximum-entropy distribution among distributions with mean λ (see [16]), so the Poisson distribution has entropy less than the entropy of the geometric distribution, which is finite.

We have to show that

$$H\left(b\left(n,\,\frac{\lambda}{n}\right)\right)\to H\left(\Pi(\lambda)\right),\qquad \text{for } n\to\infty.$$

We know that

$$D\left(b\left(n,\,\frac{\lambda}{n}\right)\middle\|\Pi(\lambda)\right)\to 0,\qquad \text{for } n\to\infty$$

which implies $b(n,\,\frac{\lambda}{n},\,j)\to\Pi(\lambda,\,j)$ for $n\to\infty$ for all j. Further, we have

$$H\left(b\left(n,\,\frac{\lambda}{n}\right)\right)+D\left(b\left(n,\,\frac{\lambda}{n}\right)\right) = -\sum_{j=0}^{\infty} b\left(n,\,\frac{\lambda}{n},\,j\right)\log\left(\Pi(\lambda,j)\right)$$

so it is sufficient to show that

$$-\sum_{j=0}^{\infty} b\left(n,\,\frac{\lambda}{n},\,j\right)\log\left(\Pi(\lambda,\,j)\right)\to H\left(\Pi(\lambda)\right),\qquad \text{for } n\to\infty.$$

Now

$$\begin{aligned} b\left(n,\,\frac{\lambda}{n},\,j\right)&=\frac{n!}{j!(n-j)!}\left(\frac{\lambda}{n}\right)^j\left(1-\frac{\lambda}{n}\right)^{n-j}\\ &\le\frac{n!}{n^j(n-j)!}\cdot\frac{\lambda^j}{j!}\exp(-\lambda)\cdot\exp(\lambda)\\ &\le\Pi(\lambda,\,j)\cdot\exp(\lambda)\end{aligned}$$

and, therefore,

$$-b\left(n,\,\frac{\lambda}{n},\,j\right)\log\left(\Pi(\lambda,\,j)\right)\le-\exp(\lambda)\cdot\Pi(\lambda,\,j)\log\left(\Pi(\lambda,\,j)\right)$$

which is an integrable upper bounding function with respect to the counting measure. □

Remark 9: None of the sets $B_n(\lambda)$, $n=3,4,5,\ldots,\infty$ are convex. If the sets $B_n(\lambda)$ had been convex, we could have used Theorem 7 together with general results on entropy maximization obtained by Topsøe and others [16]–[18] to conclude that $b(n,\,\frac{\lambda}{n})$ converges to a distribution in $cl(B_\infty(\lambda))$ in information divergence, without use of the results in Section II.

ACKNOWLEDGMENT

The author wants to thank Flemming Topsøe, Department of Mathematics, University of Copenhagen, for useful discussions.

REFERENCES

[1] W. Ochs, "Basic properties of the generalized Bolzmann–Gibbs–Shannon entropy," *Rep. Math. Phys.*, vol. 9, pp. 135–155, 1976.
[2] Y. V. Prohorov, "Asymptotic behavior of the binomial distribution" (in Russian), *Usp. Mat. Nauk.*, vol. 8, no. 3(55), pp. 135–142, 1953.
[3] L. Le Cam, "An approximation theorem for the Poisson binomial distribution," *Pacif. J. Math.*, vol. 19, no. 3, pp. 1181–1197, 1960.
[4] ——, *Asymptotic Methods in Statistical Theory*. New York: Springer-Verlag, 1986.
[5] S. Y. Shorgin, "Approximation of a generalized binomial distribution," *Theory Probab. Appl.*, vol. 22, pp. 846–850, 1977.
[6] E. L. Presman, "Approximation in variation of the distribution of a sum of independent Bernoulli variables with a Poisson law," *Theory Probab. Appl.*, vol. 20, no. 2, pp. 417–422, 1985.
[7] J. M. Steele, "Le Cam's inequality and Poisson approximatons," *Amer. Math. Monthly*, vol. 101, pp. 48–54, Jan. 1994.
[8] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[9] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336–342, 1986.
[10] S. Takano, "Convergence of entropy in the central limit theorem," *Yokohama Math. J.*, vol. 35, pp. 143–148, 1987.
[11] R. J. Serfling, "A general Poisson approximation theorem," *Ann. Probab.*, vol. 3, pp. 726–731, 1975.
[12] ——, "Some elementary results on Poisson approximation to the distribution of a sum of dependent random variables," *SIAM Rev.*, vol. 2, pp. 567–579, 1978.
[13] P. Deheuvels and D. Pfeifer, "A semigroup approach to Poisson approximation," *Ann. Probab.*, vol. 14, pp. 663–676, 1986.
[14] L. A. Shepp and J. Olkin, "Entropy of the sum of independent Bernoulli random variables and of the multidimensional distribution," Stanford Univ., Stanford, CA, Tech. Rep. 131, July 1978.
[15] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. Cambridge, U.K.: Cambridge Univ. Press, 1964.
[16] F. Topsøe, "Information theoretic optimization technics," *Kybernetika*, vol. 15, no. 1, 1979.
[17] ——, "Game theoretical equilibrium, maximum entropy and minimum information discrimination," in *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari and G. Demoments, Eds. Dordrecht/Boston/London: Kluwer, 1993, pp. 15–23.
[18] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, 1975.