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 $\lambda$, 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 $p_i = \mathbb{P}(X_i = 1), \lambda = \sum p_i, p_{\text{max}} = \max p_i$, and $S_n = \sum X_i$. We call $S_n$ an $n$-generalized binomial distribution and denote by $B_n(\lambda)$ the set of $n$-generalized binomial distributions with mean $\lambda$. Define the set of generalized binomial distributions $B_{n}(\lambda)$ 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

$$\|P - Q\| = \sum |p_i - q_i|$$

and the information divergence is defined as

$$D(P \| Q) = \sum p_i \log \frac{p_i}{q_i}.$$ 

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 the point probabilities of $b(n, \frac{\lambda}{n})$ 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

$$H(P) = -\sum p_i \log p_i.$$ 

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

II. POISSON’S LAW

Assume $X_1$ and $X_2$ are independent Poisson distributed random variables with intensities $\lambda$ and $\mu$. 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

$$D(X \| \Pi(\lambda)) = \sum_{j=0}^{\infty} p_j \log \left( \frac{p_j}{\lambda^j e^{-\lambda}} \right)$$

$$= \lambda + \sum_{j=0}^{\infty} p_j \left( \frac{1}{\lambda^j} - \frac{1}{j!} \right) - H(X)$$

$$= \lambda - E(X) \log \lambda + E(\log(\lambda)) - H(X)$$

and the derivative with respect to $\lambda$ is $1 - E(\lambda \lambda^{-1}).$ 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

$$D(X_1 + X_2) \leq D(X_1) + D(X_2).$$

Proof: First we observe that

$$D(X_1) + D(X_2) = D(X_1 \| \Pi(\lambda_1)) + D(X_2 \| \Pi(\lambda_2))$$

$$= D((X_1, X_2) \| (\Pi(\lambda_1), \Pi(\lambda_2)))$$

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 $(X_1, X_2) \to X_1 + X_2$. \hfill \Box

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

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

Proof: We have

$$D(S_n) = (1 - p_i) \ln \left( \frac{1 - p_i}{\exp(-p_i)} \right) + p_i \ln \left( \frac{p_i}{\exp(-p_i)} \right)$$

$$= (1 - p_i) \ln(1 - p_i) + p_i$$

$$\leq (1 - p_i)(-p_i) + p_i$$

$$= p_i^2.$$
and, therefore,
\[ D(S_n) \leq \sum_{i=1}^{n} D(X_i) \]
\[ \leq \sum_{i=1}^{n} p_i^2. \]
\[ \square \]

We see that if \( \lambda \) 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 \( D(S_n) \leq \lambda p_{\max} = \frac{\lambda}{2}. \)

Remark 3: The bound can easily be improved by use of the inequality
\[ D(X_i) = (1 - p_i) \ln(1 - p_i) + p_i \]
\[ \leq (1 - p_i) \left( -p_i - \frac{p_i^2}{2} - \frac{p_i^3}{3} \right) + p_i \]
\[ = \frac{1}{2} p_i^2 + \frac{1}{6} p_i^3 + \frac{1}{3} p_i^4 \]
which gives
\[ D(S_n) \leq \sum_{i=1}^{n} D(X_i) \]
\[ = \sum_{i=1}^{n} \left( \frac{p_i^2}{2} + \frac{p_i^3}{6} + \frac{p_i^4}{3} \right) \]
\[ \leq \lambda \left( \frac{p_{\max}^2}{2} + \frac{p_{\max}^3}{6} + \frac{p_{\max}^4}{3} \right). \]

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
\[ s_k^n(x_1, x_2, \ldots, x_n) = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k} \]
defined for \( x_1 > 0, x_2 > 0, \ldots, x_n > 0 \). These functions satisfy the following inequalities:
\[ s_k^n \cdot s_{k+1}^n \leq (s_{k+1}^n)^2. \]
\[ (2) \]
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 \( p_i, p_j i \neq j \) when all other probabilities \( p_k, k \neq i, j \) 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 \( t = (1 - p_1) \) and \( t = \frac{k}{2} \). 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
\[ (p_1 p_2, p_1 (1 - p_2) + (1 - p_1) p_2, (1 - p_1)(1 - p_2)) \]
\[ = \left( \frac{k^2}{4} - t^2, k - \frac{k^2}{2} + 2t^2, \left( 1 - \frac{k^2}{2} \right) - t^2 \right). \]

Therefore, the distribution of \( S_n \) is an affine function of \( t^2 \). Put \( u = t^2 \).

Then we have
\[ \frac{d^2}{dt^2} H(S_n) = \frac{d}{dt} \left( \frac{du}{dt} \frac{d}{du} H(S_n) \right) \]
\[ = 2 \cdot \frac{d}{du} H(S_n) + \left( \frac{du}{dt} \right)^2 \cdot \frac{d^2}{du^2} H(S_n). \]
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 \( b_i = P(X_1 + \cdots + X_n = i) \). Then we have
\[ P(S_n = l) = \left( \frac{k^2}{4} - u \right)b_{l-2} + \left( k - \frac{k^2}{2} + 2u \right)b_{l-1} \]
\[ + \left( 1 - \frac{k^2}{2} - u \right)b_l, \]
and get
\[ \frac{d}{du} H(S_n) = \frac{d}{du} \left( \sum_{i} P(S_n = l) \log P(S_n = l) \right) \]
\[ = \sum_{i} \left( \frac{dP(S_n = l)}{du} \right) (\log P(S_n = l) + 1) \]
\[ = \sum_{i} \left( -b_{l-2} + 2b_{l-1} - b_l \right) \log P(S_n = l) \]
\[ = \sum_{i} \log \left( \frac{P(S_n = l)}{P(S_n = l + 1)} \right) \cdot b_i. \]

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
\[ \frac{P(S_n = l)}{P(S_n = l + 1)} = \frac{s_{l+1}^n}{s_{l}^n} \leq 1 \]
which shows that
\[ \frac{d}{du} H(S_n) \leq 0. \]
\[ \square \]

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 \( m = \lceil \frac{1}{p_{\text{max}}} \rceil \), then
\[ H(S_n) \geq H \left( b \left( m, \frac{\lambda}{m} \right) \right). \]

**Proof:** Let \( K \) be the set of \( n \)-generalized binomial distributions with mean \( \lambda \), with success probabilities \( p_i \), and with \( p_{\text{max}} \leq \frac{\lambda}{m} \). 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 \( i \neq j \), 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.
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 Topsoe and others [16]–[18] to conclude that $b(n, \frac{\lambda}{n})$ converges to a distribution in $dI(B_{\infty}(\lambda))$ in information divergence, without use of the results in Section II.

ACKNOWLEDGMENT

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

REFERENCES