EXPECTED RELATIVE ENTROPY BETWEEN
A FINITE DISTRIBUTION AND ITS
EMPIRICAL DISTRIBUTION

Syuuji Abe

(Received September 13, 1996)

Abstract. The expected relative entropy (or the expected divergence) between
finite probability distribution $Q$ on $\{1, 2, \ldots, \ell\}$ and its empirical one obtained
from the sample of size $n$ drawn from $Q$ is computed and is found to be given
asymptotically by $(\ell - 1)(\log e)/2n$ which is independent of $Q$. A method to
compute the entropy of the binomial distribution more accurately than before
is also given.

AMS 1991 Mathematics Subject Classification. 62B10, 94A17, 94A15.

Key words and phrases. expected relative entropy, expected divergence, em-
pirical distribution, entropy of the binomial distribution.

§1. Introduction

In information theory, the relative entropy (or divergence) $D[P||Q] := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ plays an important role as a kind of measure of distance
between two probability distributions $P, Q$ on a discrete set $\mathcal{X}$ (log will al-
ways mean $\log_2$). It is known that $D[P||Q] \geq \frac{1}{2\min_{x \in \mathcal{X}} (P(x) - Q(x))^2}$ holds (see for example [1]). The relative entropy is closely related to math-
ematical statistics. For example, the log-likelihood ratio can be written as the
difference between two relative entropies, and the so-called Fisher informa-
tion can be expressed in terms of the relative entropy. In this paper, we
compute the expected relative entropy between a finite probability distri-
bution and its empirical one. Let $X^n = (X_1, X_2, \ldots, X_n)$ be the sample of size
$n$ drawn from the distribution $Q(x)$ on $\mathcal{X} = \{1, 2, \ldots, \ell\}$ and let $P_{X^n}(x)$ be
the empirical (frequency) distribution corresponding to $X^n$. It is known that
\[ E[D[P_X||Q]] \leq E[D[P_{X^{n-1}}||Q]] \text{ (see [1]). Actually, however, the following estimate will be found in §3 using a lemma in §2:} \]

\[
E[D[P_X||Q]] = \left(\frac{(t-1)\log e}{2n} + \frac{\log e}{12} \left(\sum_{x \in \mathcal{X}} \frac{1}{Q(x)} - 1\right) \right) \frac{1}{n^2} + O\left(\frac{1}{n^3}\right).
\]

§2. A Lemma

We prove a lemma which is essential for the proof of the theorem given in §3. The lemma states that for the random variable \( X \) obeying \( B(n,p) \) (the binomial distribution with parameters \( n, p \)) and for sufficiently large \( n \),

\[
E\left[f\left(\frac{X}{n}\right)\right] \approx \sum_{i=0}^{2m-2} \frac{f^{(i)}(p)}{i! n^i} E\left[(X - np)^i\right],
\]

where \( f(x) \) is an arbitrary function such that \( \max_{x \in [\frac{1}{n}, 1]} |f^{(2m)}(x)| \leq cn^s \) for some \( m \geq 1 \), for example \( f(x) = -x \ln x \) that appears in the entropy \( -\sum_{x \in \mathcal{X}} p(x) \ln p(x) \).

**Lemma.** Let \( f(x) \in C^{(2m)}(0,1) \) for some \( m \geq 1 \) and suppose there exist constants \( c \) and \( s \) such that \( \max_{x \in [\frac{1}{n}, 1]} |f^{(2m)}(x)| \leq cn^s \) for any positive integer \( n \). Then for \( 0 < p < 1 \), we have

\[
\left|g_2(p) - g_1(p)\right|n^m \to 0 \quad (n \to \infty)
\]

and

\[
g_1(p) = \sum_{i=0}^{2m-2} \frac{f^{(i)}(p)}{i! n^i} \mu_i + O(n^{-m}),
\]

where

\[
p_k = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \ldots, n)
\]

\[
g_1(p) = \sum_{k=0}^{n} p_k f\left(\frac{k}{n}\right)
\]

\[
\mu_i = \sum_{k=0}^{n} p_k (k - np)^i
\]

\[
g_2(p) = \sum_{i=0}^{2m} \frac{f^{(i)}(p)}{i!} \frac{\mu_i}{n^i}.
\]
Note: From the lemma, we have \( g_1(p) \approx g_2(p) \), and it is easy to show
\[
E \left[ f \left( \frac{X}{n} \right) \right] = \sum_{k=0}^{n} p_k f \left( \frac{k}{n} \right)
\]
\[
\approx \sum_{i=0}^{2m} \frac{f^{(i)}(p) \mu_i}{n^i}
\]
\[
\approx f(p) + \frac{f''(p) p(1-p)}{2} + \ldots.
\]

Proof. Since
\[
g_1(p) = \sum_{k=1}^{n} p_k f \left( \frac{k}{n} \right)
\]
\[
= \sum_{k=1}^{n} p_k \left[ f(p) + \frac{f'(p)}{1!} \left( \frac{k}{n} - p \right) + \ldots \right.
\]
\[
+ \frac{f^{(2m-1)}(p)}{(2m-1)!} \left( \frac{k}{n} - p \right)^{2m-1} + \left. \frac{f^{(2m)}(\theta_{\frac{k}{n}})}{(2m)!} \left( \frac{k}{n} - p \right)^{2m} \right]
\]
with \( \theta_{\frac{k}{n}} \) lying between \( \frac{k}{n} \) and \( p \), we get with some manipulations
\[
[g_2(p) - g_1(p)] n^m
\]
\[
= p_0 \left( f(p) + \frac{f'(p)}{1!} (-p) + \ldots + \frac{f^{(2m)}(p)}{(2m)!} (-p)^{2m} \right) n^m
\]
\[
+ \frac{1}{(2m)!} \frac{1}{n^m} \sum_{k=1}^{n} p_k \left( f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}}) \right) \left( \frac{k}{n} - np \right)^{2m}.
\]
Since \( p_0 n^m = \binom{n}{m} p^m (1-p)^n n^m = (1-p)^n n^m \), the first part of the right hand side goes to 0 as \( n \to \infty \).

The continuity of \( f^{(2m)}(x) \) implies
\[
\forall \epsilon > 0, \exists \delta > 0; \ |p - p'| < \delta \Rightarrow |f^{(2m)}(p) - f^{(2m)}(p')| < \epsilon.
\]

Hence in the second part:
\[
\frac{1}{(2m)!} \frac{1}{n^m} \sum_{k=1}^{n} \left[ p_k \left( f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}}) \right) \left( \frac{k}{n} - np \right)^{2m} \right]
\]
\[
= \frac{1}{(2m)!} \frac{1}{n^m} \sum_{\left| \frac{k}{n} - p \right| < \delta} \left[ \right] + \frac{1}{(2m)!} \frac{1}{n^m} \sum_{\left| \frac{k}{n} - p \right| \geq \delta} \left[ \right]
\]
\[
= A + B,
\]
we first have

\[
|A| \leq \frac{1}{(2m)!} \frac{1}{n^m} \sum_{|\frac{k}{n} - p| < \delta} p_k |f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}})| (k - np)^{2m}
\]

\[
< \frac{1}{(2m)!} \frac{1}{n^m} \sum_{|\frac{k}{n} - p| < \delta} p_k (k - np)^{2m}
\]

\[
\leq \frac{\epsilon}{(2m)!} \frac{1}{n^m} \sum_{k=0}^{n} p_k (k - np)^{2m}
\]

\[
= \frac{\epsilon}{(2m)!} \mu_{2m}.
\]

We know from Riordan [4] that

\[
\mu_{2m} = (2m - 1)(2m - 3) \cdots 3 \cdot 1(p(1-p)n)^m + O(n^{m-1})
\]

and so we obtain \(|A| < \epsilon + O(\frac{1}{n})\). Thus \(A \to 0\) as \(n \to \infty\).

To estimate \(|B|\), we note that, in the case \(|\frac{k}{n} - p| \geq \delta\), we have

\[
D_k := D \left[ \left( \frac{k}{n}, 1 - \frac{k}{n} \right) \left( p, 1 - p \right) \right]
\]

\[
\geq \log e \left( \frac{2}{\left| k - np \right|} \right)^2
\]

(see §1), hence

\[
\sqrt{\frac{D_k}{2 \log e}} \geq |\frac{k}{n} - p| \geq \delta.
\]

Now for large \(n\)

\[
|B| \leq \frac{1}{(2m)!} \frac{1}{n^m} \sum_{k:D_k \geq 2\delta^2 \log e} p_k |f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}})| (k - np)^{2m}
\]

\[
\leq \frac{1}{(2m)!} \frac{1}{n^m} \sum_{k:D_k \geq 2\delta^2 \log e} p_k \left( |f^{(2m)}(p)| + |f^{(2m)}(\theta_{\frac{k}{n}})| \right) (k - np)^{2m}
\]

\[
\leq \frac{1}{(2m)!} \frac{2c{n^s}}{n^m} \sum_{k:D_k \geq 2\delta^2 \log e} p_k n^{2m}
\]

\[
\leq \frac{2c}{(2m)!} n^{s+m}(n + 1)^2 2^{-2sn^2 \log e}.
\]

Here in the last inequality we used

\[
\sum_{k:D_k \geq a} p_k \leq (n + 1)^2 2^{-an}
\]

(see Theorem 12.2.1 in [1]). Thus \(B \to 0\) as \(n \to \infty\). And \([g_2(p) - g_1(p)]n^m \to 0\) \((n \to \infty)\), hence \(g_1(p) = g_2(p) + o(n^{-m})\). Recalling \(\mu_j = O(n^{\frac{j}{2}}) ([4])\), we can write \(g_1(p) = \sum_{i=0}^{2m-2} \frac{f^{(i)}(p)}{i!} \mu_i \frac{n^i}{m^i} + O(n^{-m})\), completing the proof. \(\square\)
Example 1. Let $f(x) = x \ln x$ and $m = 3$. We can use the lemma since $\max_{x \in [\frac{1}{2}, 1]} |f^{(6)}(x)| = 4! n^5$. Thus

\[
g_1(p) = f(p) + \frac{f''(p)}{2!} \frac{p(1-p)}{n} + \frac{f^{(3)}(p)}{3!} \frac{\mu_3}{n^3} + \frac{f^{(4)}(p)}{4!} \frac{\mu_4}{n^4} + O(n^{-3})
\]
\[
= p \ln p + \frac{1}{2p} \frac{p(1-p)}{n} + \frac{-1}{6p^2} \frac{p(1-p)(1-2p)}{n^2}
\]
\[
+ \frac{2}{24p^3} \frac{3p^2(1-p)^2}{n^2} + O(n^{-3})
\]
\[
= p \ln p + \frac{1-p}{2n} + \frac{(1-p)(1+p)}{12pn^2} + O(n^{-3})
\]

Example 2 [entropy of the binomial distribution].

Frank and Öhrvik[3] computed the entropy of the binomial distribution. Here we observe it in more detail using the lemma.

\[
H(X) = - \sum_{k=0}^{n} p_k \log p_k
\]
\[
= - \sum_{k=0}^{n} p_k \left( \log \left( \frac{n}{k} \right) + k \log p + (n-k) \log (1-p) \right)
\]
\[
= - \sum_{k=0}^{n} p_k (\log n! - k \log k! - (n-k)! \log (n-k)! + k \log p + (n-k) \log (1-p))
\]
\[
= - \log n! - np \log p - n(1-p) \log (1-p)
\]
\[
+ \sum_{k=0}^{n} p_k (\log k! + (n-k)!)
\]
\[
= - \log n! - np \log p - n(1-p) \log (1-p)
\]
\[
+ \sum_{k=1}^{n} p_k \log k! + \sum_{k=0}^{n-1} p_k \log (n-k)!
\]

In a similar way as in Feller[2, II.9], we may show that there exists $0 \leq b_k \leq \frac{5}{27}$ such that

\[
\ln k! = \frac{1}{2} \ln 2\pi + (k + \frac{1}{2}) \ln k - k + \left( \frac{1}{12k} - b_k \frac{1-b_k}{360k^3} \right) \quad (k \geq 1).
\]

Then letting $f(x) = \ln x$, $\frac{1}{x}$, $\frac{1}{x^2}$ in the lemma and using Example 1, we find with some computations that

\[
H(X) \approx \frac{1}{2} \log [2\pi n p (1-p)] - \log e \left( \frac{(1-2p)^2}{12np(1-p)} + \frac{p^4 + (1-p)^4}{24n^2p^2(1-p)^2} \right) + O\left( \frac{1}{n^3} \right).
\]
§3. Expected Relative Entropy

We prove our main theorem below, using Example 1 (hence our lemma). This theorem states that, for large $n$, $E[D[P_{X^n}||Q]]$ is essentially $(\ell - 1) \log e - \frac{\log e}{2n}$, in inverse proportion to the sample size $n$ and not dependent on the true distribution.

**Theorem.** Let $X^n = (X_1, X_2, \ldots, X_n)$ be the sample of size $n$ drawn from the distribution $Q(x)$ on $X = \{1, 2, \ldots, \ell\}$ and let $P_{X^n}(x)$ be the empirical (frequency) distribution corresponding to $X^n$, then

$$E[D[P_{X^n}||Q]] = (\ell - 1) \log e - \frac{\log e}{2n} + O\left(\frac{1}{n^2}\right).$$

**Proof.** The expectation to be computed is given by

$$E[D[P_{X^n}||Q]] = \sum_{(x_1, x_2, \ldots, x_n) \in X^n} Q^n(x_1, x_2, \ldots, x_n) D[P_{X^n}||Q]$$

$$= \sum_{P \in P_n} Q^n(T(P)) D[P||Q],$$

where $Q^n(x_1, x_2, \ldots, x_n) = Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, $P_n$ is the set of all possible empirical distributions, $Q^n(T(P))$ denotes the probability that the empirical distribution becomes exactly $P$. Since the empirical distribution $P$ is written as $(\frac{k_1}{n}, \frac{k_2}{n}, \ldots, \frac{k_\ell}{n})$ and $Q^n(T(P)) = \frac{n}{k_1} Q(1)^{k_1} Q(2)^{k_2} \cdots Q(\ell)^{k_\ell}$, we have

$$E[D[P_{X^n}||Q]]$$

$$= \sum_{P \in P_n} Q^n(T(P)) \left(\sum_{i \in X} P(i) \log P(i) - \sum_{i \in X} P(i) \log Q(i)\right)$$

$$= -E[H(K_1, K_2, \ldots, K_\ell)] - \sum_{i \in X} \left(\sum_{P \in P_n} Q^n(T(P)) P(i)\right) \log Q(i)$$

$$= -E[H(K_1, K_2, \ldots, K_\ell)]$$

$$- \sum_{i \in X} \left(\sum_{k_1, k_2, \ldots, k_\ell} \frac{n}{k_1+k_2+\ldots+k_\ell} Q(1)^{k_1} Q(2)^{k_2} \cdots Q(\ell)^{k_\ell} \frac{k_i}{n}\right) \log Q(i)$$

$$= -E[H(K_1, K_2, \ldots, K_\ell)] + H(Q).$$
Note that $P(i) = \frac{K_i}{n}, i = 1, \ldots, \ell$, are random variables and $H(II)$ denotes the entropy of the distribution II. Since $K_i \sim B(n, Q(i))$, we see using Example 1 that

$$
E \left[ \frac{K_i}{n} \log \frac{K_i}{n} \right] = \sum_{k=0}^{n} p_k \frac{k}{n} \log \frac{k}{n} = \frac{Q(i) \log Q(i) + 1}{2n} \log e + \frac{1}{12n^2} \left( \frac{1}{Q(i)} - Q(i) \right) \log e + O\left( \frac{1}{n^3} \right).
$$

Thus

$$
-E \left[ \frac{K_1}{n} \log \frac{K_1}{n} \right] = \sum_{i=1}^{\ell} \frac{K_i}{n} \log \frac{K_i}{n} = \sum_{i=1}^{\ell} \left( Q(i) \log Q(i) + \frac{1 - Q(i)}{2n} \log e + \frac{1}{12n^2} \left( \frac{1}{Q(i)} - Q(i) \right) \log e \right) + O\left( \frac{1}{n^3} \right).
$$

Therefore,

$$
E \left[ D[P_{X^n}||Q] \right] = \frac{1}{2n} \sum_{x \in \mathcal{X}} \frac{1}{Q(x)} \log e + \frac{1}{12} \left( \sum_{x \in \mathcal{X}} \frac{1}{Q(x)} - 1 \right) \log e + O\left( \frac{1}{n^3} \right),
$$

finishing the proof.

**Acknowledgment**

The author is grateful to Professor Yasuichi Horibe for his valuable remarks and helpful advices.

**References**


[4] J. Riordan, Moment recurrence relations for binomial, Poisson and hypergeometric

Syuuji Abe
Department of Applied Mathematics,
Science University of Tokyo
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162, Japan