SUT Journal of Mathematics
Vol. 32, No. 2 (1996), 149–156

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, empirical 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 always mean $\log_2$). It is known that $D[P||Q] \geq \frac{1}{2\ln 2}(\sum_{x\in\mathcal{X}} |P(x) - Q(x)|)^2$ holds (see for example [1]). The relative entropy is closely related to mathematical statistics. For example, the log-likelihood ratio can be written as the difference between two relative entropies, and the so-called Fisher information can be expressed in terms of the relative entropy. In this paper, we compute the expected relative entropy between a finite probability distribution 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

$E[D[P_{X^n}||Q]] \le E[D[P_{X^{n-1}}||Q]]$ (see [1]). Actually, however, the following estimate will be found in §3 using a lemma in §2:

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

§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,

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

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

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

$$[g_2(p) - g_1(p)]n^m \to 0 \ (n \to \infty)$$

and

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

where

$$\begin{aligned} p_k &= \binom{n}{k}p^k(1-p)^{(n-k)} \qquad (k=0,1,\ldots,n) \\ g_1(p) &= \sum_{k=1}^{n} p_k f\left(\frac{k}{n}\right) \\ \mu_i &= \sum_{k=0}^{n} p_k(k-np)^i \\ g_2(p) &= \sum_{i=0}^{2m}\frac{f^{(i)}(p)}{i!}\frac{\mu_i}{n^i}. \end{aligned}$$

Note: From the lemma, we have $g_1(p) \approx g_2(p)$, and it is easy to show

$$\begin{aligned} E\left[f\left(\frac{X}{n}\right)\right] &= \sum_{k=0}^{n} p_k f\left(\frac{k}{n}\right) \\ &\approx \sum_{i=0}^{2m} \frac{f^{(i)}(p)}{i!}\frac{\mu_i}{n^i} \\ &\approx f(p) + \frac{f''(p)}{2}\frac{p(1-p)}{n} + \dots . \end{aligned}$$

Proof. Since

$$\begin{aligned} g_1(p) &= \sum_{k=1}^{n} p_k f(\frac{k}{n}) \\ &= \sum_{k=1}^{n} p_k[f(p) + \frac{f'(p)}{1!}(\frac{k}{n} - p) + \dots \\ &\quad + \frac{f^{(2m-1)}(p)}{(2m-1)!}(\frac{k}{n} - p)^{2m-1} + \frac{f^{(2m)}(\theta_{\frac{k}{n}})}{(2m)!}(\frac{k}{n} - p)^{2m}] \end{aligned}$$

with $\theta_{\frac{k}{n}}$ lying between $\frac{k}{n}$ and p, we get with some manipulations

$$\begin{aligned} &[g_2(p) - g_1(p)]n^m \\ &= p_0\left(f(p) + \frac{f'(p)}{1!}(-p) + \dots + \frac{f^{(2m)}(p)}{(2m)!}(-p)^{2m}\right)n^m \\ &\quad + \frac{1}{(2m)!}\frac{1}{n^m}\sum_{k=1}^{n} p_k(f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}}))(k - np)^{2m}. \end{aligned}$$

Since $p_0 n^m = \binom{n}{0} p^0 (1-p)^n n^m = (1-p)^n n^m$, the first part of the right hand side goes to 0 as $n \to \infty$.

The continuity of $f^{(2m)}(x)$ implies

$$\forall \epsilon > 0, \exists \delta > 0; \quad |p - p'| < \delta \Rightarrow |f^{(2m)}(p) - f^{(2m)}(p')| < \epsilon.$$

Hence in the second part:

$$\begin{aligned} &\frac{1}{(2m)!}\frac{1}{n^m}\sum_{k=1}^{n}\left[p_k\left(f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}})\right)(k - np)^{2m}\right] \\ &= \frac{1}{(2m)!}\frac{1}{n^m}\sum_{|\frac{k}{n} - p| < \delta}[\quad] + \frac{1}{(2m)!}\frac{1}{n^m}\sum_{|\frac{k}{n} - p| \geq \delta}[\quad] \\ &= A + B, \end{aligned}$$

we first have

$$\begin{aligned}
|A| &\leq \frac{1}{(2m)!}\frac{1}{n^m}\sum_{|\frac{k}{n}-p|<\delta} p_k \left| f^{(2m)}(p) - f^{(2m)}(\theta_{\frac{k}{n}})\right| (k-np)^{2m} \\
&< \frac{1}{(2m)!}\frac{1}{n^m}\sum_{|\frac{k}{n}-p|<\delta} p_k \epsilon (k-np)^{2m} \\
&\leq \frac{\epsilon}{(2m)!}\frac{1}{n^m}\sum_{k=0}^{n} p_k (k-np)^{2m} \\
&= \frac{\epsilon}{(2m)!}\frac{\mu_{2m}}{n^m}.
\end{aligned}$$

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

$$\begin{aligned}
D_k := \;& D\left[\left(\frac{k}{n}, 1-\frac{k}{n}\right) ||(p, 1-p)\right] \\
\geq \;& \frac{\log e}{2}\left(2|\frac{k}{n} - p|\right)^2 \quad \text{(see §1)},
\end{aligned}$$

hence $\sqrt{\frac{D_k}{2\log e}} \geq |\frac{k}{n} - p| \geq \delta$. Now for large n

$$\begin{aligned}
|B| &\leq \frac{1}{(2m)!}\frac{1}{n^m}\sum_{k:D_k\geq 2\delta^2\log e} p_k |f^{(2m)}(p) - f^{(2m)}\left(\theta_{\frac{k}{n}}\right)|(k-np)^{2m} \\
&\leq \frac{1}{(2m)!}\frac{1}{n^m}\sum_{k:D_k\geq 2\delta^2\log e} p_k \left(|f^{(2m)}(p)| + |f^{(2m)}\left(\theta_{\frac{k}{n}}\right)|\right)(k-np)^{2m} \\
&\leq \frac{1}{(2m)!}\frac{2cn^s}{n^m}\sum_{k:D_k\geq 2\delta^2\log e} p_k n^{2m} \\
&\leq \frac{2c}{(2m)!} n^{s+m}(n+1)^2 2^{-2n\delta^2\log e}.
\end{aligned}$$

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 $\mu_j = O(n^{\lfloor \frac{j}{2} \rfloor})$ ([4]), we can write $g_1(p) = \sum_{i=0}^{2m-2} \frac{f^{(i)}}{i!}\frac{\mu_i}{n^i} + O(n^{-m})$, completing the proof. □

Example 1. Let $f(x) = x \ln x$ and $m = 3$. We can use the lemma since $\max_{x\in[\frac{1}{n},1]} |f^{(6)}(x)| = 4!n^5$. Thus

$$
\begin{aligned}
g_1(p) &= f(p) + \frac{f''(p)}{2!}\frac{p(1-p)}{n} + \frac{f^{(3)}(p)}{3!}\frac{\mu_3}{n^3} + \frac{f^{(4)}(p)}{4!}\frac{\mu_4}{n^4} + O(n^{-3}) \\
&= p\ln p + \frac{1}{2p}\frac{p(1-p)}{n} + \frac{-1}{6p^2}\frac{p(1-p)(1-2p)}{n^2} \\
&\quad + \frac{2}{24p^3}\frac{3p^2(1-p)^2}{n^2} + O(n^{-3}) \\
&= p\ln p + \frac{1-p}{2n} + \frac{(1-p)(1+p)}{12pn^2} + O(n^{-3})
\end{aligned}
$$

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.

$$
\begin{aligned}
&H(X) \\
&= -\sum_{k=0}^{n} p_k \log p_k \\
&= -\sum_{k=0}^{n} p_k \left(\log \binom{n}{k} + k\, \log p + (n-k)\log(1-p)\right) \\
&= -\sum_{k=0}^{n} p_k \left(\log n! - \log k! - \log(n-k)! + k\log p + (n-k)\log(1-p)\right) \\
&= -\log n! - np\log p - n(1-p)\log(1-p) \\
&\quad + \sum_{k=0}^{n} p_k \left(\log k! + \log(n-k)!\right) \\
&= -\log n! - np\log p - n(1-p)\log(1-p) \\
&\quad + \sum_{k=1}^{n} p_k \log k! + \sum_{k=0}^{n-1} p_k \log(n-k)!.
\end{aligned}
$$

In a similar way as in Feller[2, II.9], we may show that there exists $0 \le b_k \le \frac{5}{21}$ such that

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

Then letting $f(x) = \ln x,\ \frac{1}{x},\ \frac{1}{x^3}$ in the lemma and using Example 1, we find with some computations that

$$H(X) = \frac{1}{2}\log\left[2\pi enp(1-p)\right] - (\log e)\left(\frac{(1-2p)^2}{12np(1-p)} + \frac{p^4+(1-p)^4}{24n^2p^2(1-p)^2}\right) + O(\frac{1}{n^3}).$$

§3. Expected Relative Entropy

We prove our main theorem below, using Example 1 (hence our lemma). This theorem states that, for large n, $E\left[D\left[P_{X^n}||Q\right]\right]$ is essentially $\frac{(\ell-1)\log e}{2n}$, 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 $\mathcal{X} = \{1, 2, \ldots, \ell\}$ and let $P_{X^n}(x)$ be the empirical (frequency) distribution corresponding to X^n, then*

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

Proof. The expectation to be computed is given by

$$\begin{aligned} E[D[P_{X^n}||Q]] &= \sum_{(x_1,x_2,\ldots,x_n)\in\mathcal{X}^n} Q^n(x_1, x_2, \ldots, x_n)\ D[P_{x^n}||Q] \\ &= \sum_{P\in\mathcal{P}_n} Q^n(T(P))\ D[P||Q], \end{aligned}$$

where $Q^n(x_1, x_2, \ldots, x_n) = Pr(X_1{=}x_1, X_2{=}x_2, \ldots, X_n{=}x_n)$, $\mathcal{P}_n$ 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 $Q^n(T(P)) = \binom{n}{k_1,k_2,\ldots,k_\ell}Q(1)^{k_1}Q(2)^{k_2}\cdots Q(\ell)^{k_\ell}$, we have

$$\begin{aligned} & E[D[P_{X^n}||Q]] \\ &= \sum_{P\in\mathcal{P}_n} Q^n(T(P))\left(\sum_{i\in\mathcal{X}} P(i)\ \log P(i) - \sum_{i\in\mathcal{X}} P(i)\ \log Q(i)\right) \\ &= -E[H(\frac{K_1}{n}, \frac{K_2}{n}, \ldots, \frac{K_\ell}{n})] - \sum_{i\in\mathcal{X}}\left(\sum_{P\in\mathcal{P}_n} Q^n(T(P))\ P(i)\right)\log Q(i) \\ &= -E[H(\frac{K_1}{n}, \frac{K_2}{n}, \ldots, \frac{K_\ell}{n})] \\ &\quad - \sum_{i\in\mathcal{X}}\left(\sum_{\substack{k_1,k_2,\ldots,k_\ell: \\ k_1+k_2+\ldots+k_\ell=n}}\binom{n}{k_1, k_2, \ldots, k_\ell}Q(1)^{k_1}Q(2)^{k_2}\cdots Q(\ell)^{k_\ell}\frac{k_i}{n}\right)\log Q(i) \\ &= -E[H(\frac{K_1}{n}, \frac{K_2}{n}, \ldots, \frac{K_\ell}{n})] + H(Q). \end{aligned}$$

Note that $P(i) = \frac{K_i}{n}$, $i = 1, \ldots, \ell$, are random variables and $H(\Pi)$ denotes the entropy of the distribution Π. Since $K_i \sim B(n, Q(i))$, we see using Example 1 that

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

Thus

$$\begin{aligned} & -E\left[H(\frac{K_1}{n}, \frac{K_2}{n}, \ldots, \frac{K_\ell}{n})\right] \\ = & E\left[\sum_{i=1}^{\ell}\frac{K_i}{n}\log\frac{K_i}{n}\right] \\ = & \sum_{i=1}^{\ell} E\left[\frac{K_i}{n}\log\frac{K_i}{n}\right] \\ = & \sum_{i=1}^{\ell}\left(Q(i)\log Q(i) + \frac{1-Q(i)}{2n}\log e + \frac{1}{12n^2}(\frac{1}{Q(i)} - Q(i))\log e\right) + O(\frac{1}{n^3}). \end{aligned}$$

Therefore,

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

finishing the proof. □

Acknowledgment

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

References

[1] T. M. Cover and J. A. Thomas, *Elements of Infomation Theory,* New York: John Wiley & Sons, Inc.,1991.

[2] W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol.I 2nd Ed. New York: John Wiley & Sons, Inc.,1968.

[3] O. Frank and J. Öhrvik, Entropy of sums of random digits, *Computational Statistics & Data Analysis* 17 (1994) 177-184.

[4] J. Riordan, Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions, *Ann. Math. Statist.* 8 (1937) 103-111.

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