A Note on the Proof of Shannon Inequality

Honghui Wan

National Center for Biotechnology Information
Building 38A, 8th Floor
National Library of Medicine, National Institutes of Health
Bethesda, Maryland 20894, USA
and
Department of Mathematics and Computer Science
Huazhong (Central China) University of Science and Technology
Wuhan, Hubei 430072, China

Abstract. This paper revises Park's proof of Shannon inequality and also gives a new simple proof. **Keywords.** entropy, Shannon inequality Let

$$\begin{aligned} A(n,l) &= \sum_{i=0}^{l} \binom{n}{i}, \\ B(n,l) &= 2^{nH(\frac{l}{n})}, \end{aligned}$$

where $H(x) = -x\log_2(1-x) - (1-x)\log_2 x$. The following is called *Shannon inequality.*

$$A(n,l) \leq B(n,l), \quad (l \leq \frac{n}{2}). \tag{1}$$

The function $H(x)$ is called the *binary entropy function* and is a measure of uncertainty of a random variable. More importantly, $H(x)$ is the correct asymptotic exponent in the number of binary sequences of length n that have no more than l ones. The inequality (1) gives a combinatorial bound on the sum of binomial coefficients and plays a key role in information and communication theory. The original proof of this inequality was very long. Park [2] gave an induction proof of the inequality. However, there is a problem with his proof. In this note, I show why and give a refined proof.Let

$$a_n(l) = \left(\frac{n}{n+1}\right)^n \left(\frac{l+1}{n+1}\right) \left(\frac{l+1}{l}\right)^l.$$

Park [2] based his proof of (1) by assuming the following inequality was true:

$$a_n(l) + a_n(n-l-1) \leq \left(e^{-(n-l-1)} + e^{-l}\right)^{\frac{1}{n+1}}. \tag{2}$$

The inequality (2) is generally not true. In fact, for $n = 7$ and $l = 2$, we have

$$a_7(2) + a_7(4) = 0.931 > \left(e^{-4} + e^{-2}\right)^8 = 0.791.$$

Now we correct Park's proof by induction. Obviously, $A(n,0) = B(n,0) = 1$. When $n = 2l, A(2l,l) < A(2l,2l) = 2^{2l} = B(2l,l)$. The inequality (1) is satisfied in this case. It remains that (1) holds for $(n+1, l+1)$ if it is satisfied for (n,l) and $(n,l+1), l+1 \leq \frac{n}{2}$. In other words, we must prove

$$A(n,l) \leq B(n,l), A(n,l+1) \leq B(n,l+1).$$

Noting that

$$A(n+1,l+1) = A(n,l+1) + A(n,l),$$

we have

$$A(n+1,l+1) \leq B(n,l) + B(n,l+1) = B(n+1,l+1)(a_n(l) + a_n(n-l-1)).$$

Based on the following inequality:

$$\frac{2(n+1)}{e(2n+1)} < \left(\frac{n}{n+1}\right)^n < \frac{2n+1}{2en}, \tag{3}$$

we have $a_n(l) + a_n(n-l-1) < 1$. This completes the proof.(3) is a typical inequality and can be found in [3]. Also, we can directly show (3). The function $f(x) = x + x\ln(1+\frac{x}{2}) - (1+x)\ln(1+x)$ is increasing for $0 < x < 1$ and we have

$$f(\frac{1}{n}) = \frac{1}{n} + \frac{1}{n}\ln(1+\frac{\frac{1}{n}}{2}) - (1+\frac{1}{n})\ln(1+\frac{1}{n}) > f(0) = 0,$$

which yields

$$(\frac{n+1}{n})^n < \frac{e(2n+1)}{2(n+1)}.$$

A similar argument shows

$$(\frac{n}{n+1})^n < \frac{2n+1}{2en}.$$

Now we give an additional proof of (1). The function $g(x) = -\frac{x}{n}\log_2\frac{l}{n} - (1 - \frac{x}{n})\log_2(1-\frac{l}{n})$ is increasing in the interval $0 \leq x \leq l$ where $2l \leq n$ and hence

$$2^{-nH(\frac{l}{n})} = 2^{-ng(\frac{l}{n})} \leq 2^{-ng(\frac{i}{n})} = (\frac{l}{n})^i(1-\frac{l}{n})^{n-i}, \quad (i \leq l).$$

Multiplying by $\binom{n}{i}$ and summing gives

$$2^{-nH(\frac{l}{n})}\sum_{i=0}^{l}\binom{n}{i} \leq \sum_{i=0}^{l}\binom{n}{i}(\frac{l}{n})^i(1-\frac{l}{n})^{n-i} \leq \sum_{i=0}^{n}\binom{n}{i}(\frac{l}{n})^i(1-\frac{l}{n})^{n-i} = 1.$$

which shows (1).

References

[1] C. E. Shannon, *A mathematical theory of communication*, Bell Sys. Tech. Journal, 27 (1948), 379-423, 623-656.

[2] J. H. Park, *IEEE Trans. on Information Theory*, vol. IT-15, pp. 618, September,1969.

[3] G. Polya and G. Szego, *Problems and Theorems in Analysis I (Classics in Mathematics)*. Springer Verlag, 1998.

[4] H. Wan, *On the entropy and complexity of biological sequences*, submitted.

[5] H. Wan and J. C. Wootton, *Axiomatic foundations of complexity functions of biological sequences*, Ann. Comb., 3 (1999), 105-127.

[6] H. Wan and J. C. Wootton, *A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins*, Comput. Chem., 24 (2000), 67 - 88.