Applied Mathematics E-Notes, 1(2001), 73-76 ©
Available free at mirror sites of http://math2.math.nthu.edu.tw/∼amen/

Optimal Discrete Entropy *

Aditya Shastri and Rekha Govil[†‡]

Received 24 January 2001

Abstract

In this paper we obtain a discrete version of Shannon's classical theorem that when the probabilities are frequencies the entropy function attains its maximum value when probabilities are as equal as possible.

In this paper we address the following question: *When does the entropy function attain its maximum values if the probabilities arise from frequencies?* Every probability distribution has some 'uncertainty' associated with it. The concept of 'entropy' was introduced by Shannon [3] in 1948 to provide a quantitative measure of this uncertainty. According to the maximum-entropy principle formulated by Jaynes [1], it is of considerable interest to maximize entropy among probability distributions, so that the one having the maximum entropy is the 'most unbiased', 'least prejudiced' or 'most uniform'. For specific situations where we need to maximize entropies, the readers may consult [2].

Consider a discrete probability distribution where the sample space S consists of a finite number of elements, say $S = \{a_1, a_2, ..., a_k\}$, and with each elementary event $\{a_i\}$ we have a probability p_i associated with it satisfying the following conditions : (i) $p_1, ..., p_k \geq 0$, and (ii) $p_1 + p_2 + ... + p_k = 1$. Shannon in 1948, defined the entropy for this probability distribution as

$$H(p_1, p_2, ..., p_k) = -\sum_{i=1}^{k} p_i \log p_i.$$

A special case of the above distribution is of particular interest where all outcomes are equally likely. In that case, it follows that $p_i = 1/k$. It is well known, through a classical theorem of Shannon, that $H(p_1, p_2, ..., p_k)$ attains its maximum value $\log k$ when all probabilities are equal to $1/k$.

The above result is based on the assumption that variables can assume arbitrary values in the interval $(0, 1)$. There are occasions, however, when this need not be the case. In the most classical example of probability, we have a set A of cardinality N

*Mathematics Subject Classifications: 94A17

[†]Department of Computer Science, Banasthali University, P. O. Banasthali Vidyapith - 304 022, India

[‡]This research was partly carried out while the first author was visiting School of Mathematics, Tata Institute of Fundamental Research, Mumbai, India. Research supported in part by the Mahindra Foundation.

which is partitioned into $A_1 \cup A_2 \cup ... \cup A_k$ where $n_i = |A_i|$ for $i = 1, 2, ..., k$ and $N = n_1 + n_2 + ... + n_k$. Here $A_1, ..., A_k$ are interpreted as the partitions of the sample space A and $n_1, ..., n_k$ are *frequencies* of occurrences of A_i's. Therefore, the probability p_i of an event A_i is given by n_i/N where N is the total number of outcomes and n_i is the number of outcomes favorable to A_i. Here, the entropy needs not assume its maximum value $\log k$ since the value $p_i = 1/k$ may be forbidden for some i.

In this note, we will obtain the extreme values of the entropy function in the discrete case where the probabilities may assume only rational values n_i/N corresponding to different partitions of the set A into nonempty subsets A_i's. All the notations and terminologies are standard and we presume elementary basic knowledge of probability. One of the fundamental requirements of any entropy function is that it should be a symmetric function of its arguments. That certainly is true for the discrete case as the value of entropy depends only on the sizes of A_i's. Moreover, the following two results are also true.

THEOREM 1. The entropy function $H(A_1, A_2, ..., A_k)$ assumes its maximum value for the partitions for which the sizes of the subsets $A_1, ..., A_k$ differ by at most 1.

THEOREM 2. The entropy function $H(A_1, A_2, ..., A_k)$ assumes its minimum value for the partitions for which all subsets but exactly one have cardinality 1.

As we have already observed that the entropy for the discrete case depends only on the sizes of A_i's, i.e. it is a symmetric function of its arguments:

$$H(A_1, A_2, ..., A_k) = H(A_{i_1}, A_{i_2}, ..., A_{i_k})$$

if $(i_1, i_2, ..., i_k)$ is a permutation of $(1, 2, ..., k)$. Thus, we are led to say that two partitions $\pi_1 = (A_1, A_2, ..., A_k)$ and $\pi_2 = (B_1, B_2, ..., B_k)$ of a set A into k nonempty subsets are equivalent if the multisets $\{|A_i||1 \leq i \leq k\}$ and $\{|B_j||1 \leq j \leq k\}$ are identical. The above relation is easily seen to be an equivalence relation, which divides all the partitions into equivalence classes. The entropy function remains constant on each equivalence class.

In what follows we record a simple yet important observation that there is a *unique* equivalence class with part sizes differing by at most 1 and probabilities p_i's are close to $1/k$.

LEMMA 3.1. There exists a *unique* equivalence class such that every partition $\pi = (A_1, A_2, ..., A_k)$ of the class satisfies $||A_i| - |A_j|| \leq 1$ for $1 \leq i, j \leq k$ and $||A_i|/N - 1/k| < 1/N$ for $1 \leq i \leq k$.

The proof is elementary and is thus omitted.

The proof of our theorem 1 is reminiscent of the proof of Sperner's classical theorem that the number of subsets of an n-set such that no subset is contained in another can be at most $C^n_{\lceil n/2 \rceil}$. For if there are sets of small (large) size they can be replaced by sets of size one larger (smaller) without violating the condition and the theorem follows. We shall also show that if the partition has two sets having sizes differing by at least 2, one can then remove one element from the larger set and include it in the smaller set resulting in a net increase in the entropy. It will then follow that the optimum is attained over the equivalence class defined in Lemma 3.1 which remains fixed with respect to the operation just described.

We now turn to the proof of Theorem 1. The cases $k = 1$ and $k = N$ are trivial. When $k = 1$, $H_{max}(A) = H_{min}(A) = 0$. When $k = N$, all probabilities are equal to $1/N$ leading to an optimal value $H_{max}(A_1, A_2, ...A_N) = H_{min}(A_1, A_2, ..., A_N) = \log_2 N$, where all A_i's are singletons, as in the classical case. Furthermore, the case $k = N - 1$ can also be disposed of at once by observing that there is only one part of size 2 and all the remaining parts are singletons implying

$$H_{max}(A_1, A_2, ..., A_{N-1}) = H_{min}(A_1, A_2, ..., A_{N-1}) = \log_2 N - \frac{2}{N}.$$

Therefore, in what follows we shall assume $2 \leq k \leq N - 2$. Let $\pi = (A_1, A_2, ..., A_k)$ be a partition, $2 \leq k \leq N - 2$, having two sets say A_i and A_j, $i \neq j$, such that $|A_i| - |A_j| \geq 2$. If $a \in A_i$ then define a new partition $\pi' = (A'_1, A'_2, ..., A'_k)$, where $A'_m = A_m$ for $m \neq i, j$, $A'_i = A_i - \{a\}$ and $A'_j = A_j \cup \{a\}$. It follows that

$$\begin{aligned} H(\pi') - H(\pi) &= -\left(p_i - \frac{1}{N}\right)\log_2\left(p_i - \frac{1}{N}\right) \\ &\quad -\left(p_j + \frac{1}{N}\right)\log_2\left(p_j + \frac{1}{N}\right) + p_i \log_2 p_i + p_j \log_2 p_j \\ &= \frac{1}{N}\left(\log_2\left(p_i - \frac{1}{N}\right) - \log_2\left(p_j + \frac{1}{N}\right)\right) + p_i \log_2 p_i \\ &\quad -p_i \log_2\left(p_i - \frac{1}{N}\right) + p_j \log_2 p_j - p_j \log_2\left(p_j + \frac{1}{N}\right). \qquad (1) \end{aligned}$$

Since $|A_i| - |A_j| \geq 2$ and $p_i - p_j \geq 2/N$, hence $p_i - 1/N \geq p_j + 1/N$. Therefore, the first term in (1) is non-negative and we have

$$H(\pi') - H(\pi) \geq p_i \log_2 p_i - p_i \log_2\left(p_i - \frac{1}{N}\right) + p_j \log_2 p_j - p_j \log_2\left(p_j + \frac{1}{N}\right).$$

Simplifying and using $Np_i = |A_i|$ and $Np_j = |A_j|$, we get

$$H(\pi') - H(\pi) \geq p_i \left(\log_2 \frac{|A_i|}{|A_i| - 1}\right) + p_j \left(\log_2 \frac{|A_j|}{|A_j| + 1}\right). \qquad (2)$$

Observe that $\log_2 x \geq 1$ if $x \geq 1$, and $|A_j|/(|A_j|+1) \geq 1/2$ which implies $\log_2 |A_j|/(|A_j|+1) \geq -1$. Putting these values in (2) we obtain

$$H(\pi') - H(\pi) \geq p_i - p_j$$

and the theorem follows.

As for the proof of Theorem 2, one can define a local operation on any given partition by moving an element from the smaller set into a bigger set. This operation does not increase the entropy; and thereby deduce that the entropy is minimized when all but one set have cardinality 1.

As an immediate corollary, the extreme values of the entropy, where a n-set is partitioned into k subsets, is given by

$$H_{max} = -\frac{1}{N}\left\{r(m+1)\log_2\frac{m+1}{N} + (k-r)\log_2\frac{m}{N}\right\},$$

and

$$H_{min} = -\frac{1}{N}\left\{(k-1)\log_2\frac{1}{N} + (N-k+1)\log_2\left(1-\frac{k-1}{N}\right)\right\},$$

where $N = mk + r$ for $0 \leq r < k$.

Theorem 1 and Theorem 2 are discrete analogues of Shannon's classical theorem that the entropy is maximized when all probabilities are equal. Note that Theorem 1 and Theorem 2 do not follow from Shannon's theorem. It would be interesting to look for more cases where the probability variation is somewhat restricted and compute the corresponding extreme values of the entropy functions.

References

[1] E. T. Jaynes, Information Theory and Statistical Mechanics, Physical Reviews 106(1957), 620-630.

[2] J. N. Kapur, Maximum Entropy Models in Science and Engineering, Wiley Eastern Limited, New Delhi, 1987.

[3] C. E. Shannon, A Mathematical Theory of Communication, Bell System Tech. J. 27(1948), 379-423, 623-659.