Basics of Shannon Entropy

M. G. Nadkarni

Introduction: Already in his article 'The Entropy Concept in Probability Theory', published in 1953 and reproduced in his book ' mathematical foundations of information theory' A. I Khinchine states

"As of the present, a unified exposition of the theory of entropy can be found only in specialized articles and monographs dealing with the transmission of information. Although the study of entropy has actually evolved into an important and interesting chapter of general theory of of probability, a presentation of it in this general theoretical setting has so far been lacking.
This article is a first attempt at such a presentation. ..... There is no doubt that in the years to come the study of entropy will become a permanent part of probability theory; the work I have done seems to me to be necessary stage in the developement of this study."

The situation remains the same today, nearly sixty years after Khinchine's article, and many very successful and highly recommended books on basic probability even of recent origin fail to give any systematic account Shannon entropy, although Khinchine has more than half done the job.[1] A discussion of this notion continues to remains a part of books on special topics like ergodic theory or information theory. On the other hand some of these books on probability have room for a discussion of financial mathematics, although the idea of entropy seems far more basic than the notion of ' no arbitrage opportunity'. It is equally true that most scientists learn the notion of entropy appropriate to their field either through self study or from their colleagues, but very rarely they learn Shannon's formulation of entropy in their under-graduate course in probability or statistics. Their appreciation of the concept would increase and come at at an earlier period of learning if the Shannon entropy was part of the under-graduate syllabus. Also those under-graduate and post-graduates students of

[1]See for example 1 G. R. Grimmett and D. R. Stirzaker: Probability and Random Processes, 3rd edition, Y. S. Chow and H. Teicher: Probability Theory, K. R. Parthasarathy: Introduction to Probability and Measure, K. B. Athreya and S. N. Lahiri, Measure Theory, Probability Theory

science who choose to go to other areas will stand to benefit by learning the basics of entropy which are easy to grasp and apply.

The purpose of this article is therefore to reiterate Khinchine, and to set forth the basics of Shannon entropy and some of its applications in an elementary (finite) probability model, suitable for college and university students and teachers in India.

Section 1:Uncertainity, Information

Consider a game of tennis between two players A and B. Let us assume that the probability that A wins is 0.99 and the probability that B wins is 0.01. If A wins then an event of very high probability has occurred and one is not 'surprised' by this event, while if one is informed that B has won, then there is much 'surprise' since an event of very small probability has occurred. In other words, the occurrence of the event 'A wins' has much smaller information content than the occurrence of the event 'B wins'. The function $-\log_e x = \log x, 0 < x \leq 1$ seems tailor made to measure this information content, since $-\log x$ is large when x is positive and close to 0 and small when x is close to 1. Let us therefore say that information gained if A wins is $-\log(0.99)$ while the information gained is $-\log(0.01)$ if A looses.

We write W and L respectively for the events that A wins or A looses, so that $\{L, W\}$ is the sample space of the experiment of a game of tennis between A and B. The function

$$X(W) = -\log(0.99), X(L) = -\log(0.01)$$

is a random variable on this sample space whose expected value is

$$-(0.99)\log(0.99) - (0.01)\log(0.01)$$

which is rather small. This quantity thus measures the overall degree of uncertainity (entropy) in the scheme of a game of tennis between A and B.

Assume now that the probability that A wins is half , so that the probability that B wins is also half. In analogy with the above discussion we may say that the information gained if A wins is $-\log(\frac{1}{2})$, which is same as the information gained if B wins. The outcome of game in this situation is 'most uncertain', and the uncertainity of the scheme is measured by the

$$-\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) = \log 2$$

which is also called the entropy of the scheme, a concept which we develop systematically in the rest of this article.

Section 2:The Function H

The function

$$f(x) = x \log x, 0 < x \leq 1, f(0) = 0$$

is continuous and convex. This fact will be used in what follows very crucially.

A vector $(p_1, p_2, \cdots, p_n) \in \mathbb{R}^n$ is called a probability vector if $p_i \geq 0$ for each i and $\sum_{i=1}^{n} p_i = 1$.The quantity

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

is called the entropy of the probability vector $(p_1, p_2, \cdots, p_n)$

Suppose $(\Omega, \mathcal{B}, P)$ is a probability space and $\{P_1, P_2, \cdots, P_n\}$ are pairwise disjoint events in this probability space such that their union is all of Ω. In other words $\mathbb{P} = \{P_1, P_2, \cdots, P_n\}$ is a partition of Ω with each $P_i \in \mathcal{B}$. If we write $p_i = P(P_i)$, then $(p_1, p_2, \cdots, p_n)$ is a probability vector with entropy $H(p_1, p_2, \cdots, p_n)$. We denote this entropy also as $H(\mathbb{P})$, so that

$$H(\mathbb{P}) = -\sum_{i=1}^{n} P(P_i) \log P(P_i) = H(p_1, p_2, \cdots, p_n)$$

Section 3:Properties of the function H

1) $H(p_1, p_2, \cdots, p_n) \geq 0$, with $H(p_1, p_2, \cdots, p_n) = 0$ if and only if some $p_i = 1$ (in which case the remaining p's are all zero).

2) H is a concave function, equivalently $-H$ is a convex function. This mean that if $\vec{p} = (p_1, p_2, \cdots, p_n), \vec{q} = (q_1, q_2, \cdots, q_n)$ are two probability vectors and if a, b are non-negative real numbers with $a + b = 1$, then

$$aH(\vec{p}) + bH(\vec{q}) \leq H(a \cdot \vec{p} + b \cdot \vec{q})$$

Since the function $f = x \log x, 0 < x \leq 1, f(0) = 0$ is convex we note that

$$af(p_i) + bf(q_i) \geq f(ap_i + bq_i)$$

Summing over i,

$$a\sum_{i=1}^{n} f(p_i) + b\sum_{i=1}^{n} f(q_i) \geq \sum_{i=1}^{n} f(ap_i + bq_i)$$

$$-a\sum_{i=1}^{n} f(p_i) - b\sum_{i=1}^{n} f(q_i) \leq -\sum_{i=1}^{n} f(ap_i + bq_i)$$

whence,

$$aH(\vec{p}) + bH(\vec{q}) \leq H(a\vec{p} + b\vec{q})$$

as was to be proved.

3) H attains its maximum value at the vector $(\frac{1}{n}, \frac{1}{n}, \cdot, \frac{1}{n})$, equivalently, for all probability vectors $(p_1, p_2, \cdots, p_n)$,

$$H(p_1, p_2, \cdots, p_n) \leq H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$$

This is also a consequence of the convexity of the function f. Indeed, we have

$$f(\sum_{i=1}^{n} \frac{1}{n} p_i) \leq \sum_{i=1}^{n} \frac{1}{n} f(p_i)$$

$$\frac{1}{n} \log \frac{1}{n} \leq \sum_{i=1}^{n} \frac{1}{n} f(p_i)$$

$$-\log n \leq -H(p_1, p_2, \cdots, p_n)$$

$$H(p_1, p_2, \cdots, p_n) \leq H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$$

as was to be proved.

4) We now define conditional entropy. Let $\mathbb{P} = (P_1, P_2, \cdots, P_m)$, $\mathbb{Q} = (Q_1, Q_2, \cdots, Q_n)$ be two partitions of Ω. Write $p_i = P(A_i), q_j = P(Q_j), p_{i,j} = P(P_i \cap Q_j)$. The conditional probability of P_i given Q_j is given by

$$\frac{P(P_i \cap Q_j)}{P(Q_j)} = \frac{p_{i,j}}{q_j}$$

where the ratio is interpreted to be 0 if $q_j = 0$. We write $P(P_i \mid Q_j)$ as $q_{i,j}$ Note that $\sum_{i=1}^{m} q_{i,j} = 1$ and that $p_{i,j} = q_{i,j}q_j$.

The conditional entropy of $\mathbb{P}$ given Q_j, $H(\mathbb{P} \mid Q_j)$ is defined to be

$$H(\mathbb{P} \mid Q_j) = H(q_{1,j}, q_{2,j}, \cdots, q_{m,j}).$$

The conditional entropy of the partition $\mathbb{P}$ given $\mathbb{Q}$ is defined as follows:

$$H(\mathbb{P} \mid \mathbb{Q}) = \sum_{j=1}^{n} P(Q_j)H(\mathbb{P} \mid Q_j) = \sum_{j=1}^{n} q_j H(q_{1,j}, q_{2,j}, \cdots, q_{n,j})$$

The concavity of the function H immediately implies the important inequality:

$$H(\mathbb{P} \mid \mathbb{Q}) \leq H(\mathbb{P})$$

.

We will write $\mathbb{P} \vee \mathbb{Q}$ to denote the superposition of the partitions $\mathbb{P}$ and $\mathbb{Q}$

$$\mathbb{P} \vee \mathbb{Q} = \{P_i \cap Q_j : 1 \leq i \leq m, 1 \leq j \leq n\}$$

If $\mathbb{P}$ and $\mathbb{Q}$ are independent partitions, i.e., if $P(P_i \cap Q_j) = P(P_i)P(Q_j)$ for all $1 \leq i \leq m, 1 \leq j \leq n$ then

$$H(\mathbb{P} \vee \mathbb{Q}) = H(\mathbb{P}) + H(\mathbb{Q})$$

Indeed we then have,

$$-H(\mathbb{P} \vee \mathbb{Q}) = \sum_{i=1}^{m}\sum_{j=1}^{n} p_i q_j \log p_i q_j$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} p_i q_j (\log p_i + \log q_j)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} p_i q_j \log p_i + \sum_{i=1}^{m}\sum_{j=1}^{n} p_i q_j \log q_j$$

$$= \sum_{i=1}^{m} p_i \log p_i + \sum_{j=1}^{n} q_j \log q_j$$

$$-H(\mathbb{P}) - H(\mathbb{Q})$$

as was the claim.

In general, whether $\mathbb{P}$ and $\mathbb{Q}$ are independent or not, we have:

Theorem 1: $H(\mathbb{P} \vee \mathbb{Q}) = H(\mathbb{Q}) + H(\mathbb{P} \mid \mathbb{Q})$

Proof Recall that $p_{i,j} = q_{i,j}q_j, \sum_{i=1}^{m} q_{i,j} = 1$.

$$\begin{aligned} H(\mathbb{P} \vee \mathbb{Q}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} p_{i,j} \log p_{i,j} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}q_j(\log q_{i,j} + \log q_j) \\ &= \sum_{j=1}^{n} q_j \sum_{i=1}^{m} q_{i,j} \log q_j + \sum_{j=1}^{n} q_j \sum_{i=1}^{m} q_{i,j} \log q_{i,j} \\ &= H(\mathbb{Q}) + H(\mathbb{P} \mid \mathbb{Q}) \end{aligned}$$

as was to be proved.

Section 4:Maximum Entropy and Boltzmann Distribution

We have seen that $H(p_1, p_2, \cdots, p_n)$ is maximum when each p_i is $\frac{1}{n}$, i.e., when the distribution $(p_1, p_2, \cdots, p_n)$ is uniform. We will generalise this to the case when the probabilities p_i's are constrained.

Let $x_1, x_2, \cdots, x_n$ be real numbers with x_1 and x_n being the minimum and maximum among these. Also, let α be a real number. Then it is clear that there is a probability vector $(p_1, p_2, \cdots, p_n)$ such that $\sum_{i=1}^{n} p_i x_i = \alpha$ if and only if $x_1 \leq \alpha \leq x_n$. However, such a probability vector need not be unique (unless $n = 1$ or $n = 2$ and x_1, and x_2 are distinct or α is one of $\{x_1, x_n\}$). The notion of entropy allows us to make a unique choice among such vectors.

Theorem *Let $x_1x_2, x_3, \cdots, x_n$ be real numbers, with x_1 and x_n the minimum and maximum among these. Let $x_1 \leq \alpha \leq x_n$. Then the function*

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

defined on the space of probability vectors in $\mathbb{R}^n$ *admits a unique maximum under the constraint*

$$\sum_{i=1}^{n} p_i x_i = \alpha.$$

Proof We use the method of Lagrange's multipliers. Consider the function

$$g(p_1, p_2, \cdots, p_n) = \sum_{i=1}^{n} p_i \log p_i + \lambda(1 - \sum_{i=1}^{n} p_i) + \mu(\alpha - \sum_{i=1}^{n} p_i x_i),$$

where λ and μ are real parameters. We maximise g. We have

$$\frac{\partial g}{\partial p_i} = 1 + \log p_i - \lambda - \mu x_i$$

If we set this partial derivative equal to zero and solve for p_i, we get

$$p_i = \exp\{\mu x_i + \lambda - 1\} = \exp\{\lambda - 1\} \exp\{\mu x_i\}.$$

Since $\sum_{i=1}^{n} p_i = 1$, we get

$$\exp\{\lambda - 1\} = \frac{1}{\sum_{i=1}^{n} \exp\{\mu x_i\}},$$

whence

$$p_i = \frac{\exp\{\mu x_i\}}{\sum_{i=1}^{n} \exp\{\mu x_i\}}.$$

We will now show that there is a unique μ such that the corresponding p_i's will satisfy $\sum_{i=1}^{n} p_i x_i = \alpha$. If $\alpha = x_1$ or x_n, then the unique distribution is clearly $p_1 = 1$ or $p_n = 1$ respectively. We therefore assume that $x_1 < \alpha < x_n$.

Consider the function

$$M(\mu) = \sum_{i=1}^{n} \frac{\exp\{\mu x_i\}}{\sum_{i=1}^{n} \exp\{\mu x_i\}} x_i,$$

which is the mean of the quantities $x_1, x_2, \cdots, x_n$ with weights $p_i = \frac{\exp\{\mu x_i\}}{\sum_{i=1}^{n} \exp\{\mu x_i\}}, i = 1, 2, \cdots, n$

A calculation shows that

$$\frac{dM}{d\mu} = \sum_{i=1}^{n} x_i^2 p_i - (\sum_{i=1}^{n} x_i p_i)^2,$$

where $p_i = \frac{\exp\{\mu x_i\}}{\sum_{i=1}^{n} \exp\{\mu x_i\}}$. The derivative $\frac{dM}{d\mu}$ is the variance of the distribution $(p_1, p_2, \cdots, p_n)$ on $x_1, x_2, \cdots, x_n$ hence positive. Thus the function M is continuous and strictly increasing. Further, a careful look at M shows that $\lim_{\mu\to-\infty} M(\mu) = x_1$ and $\lim_{\mu\to\infty} M(\mu) = x_n$. Hence by intermediate value theorem there is a unique μ , say μ_0 , such that such that $M(\mu_0) = \alpha$. The theorem follows.

The unique distribution $p_i = \frac{\exp\{\mu_0 x_i\}}{\sum_{i=1}^{n} \exp\{\mu_0 x_i\}}$, $i = 1, 2, \cdots, n$, which maximises the entropy $H(p_1, p_2, \cdots, p_n)$ under the constraint $\sum p_i x_i = \alpha$ is called Boltzmann distribution. It depends on $x_1, x_2, \cdots, x_n$ and α. The μ_0 is uniquely determined by these.

Section 5:Applications

Application to Wealth Distribution. Consider a population of N individuals with total wealth W. If this wealth is distributed uniformly among the N individuals, each individual will get W/N share of wealth. However this is not the real life situation. Let $w_1 \leq w_2 \leq \cdots \leq w_N$ be the wealths of the N individuals in the increasing order. The graph of the map $\frac{l}{N} \to \frac{\sum_{i=1}^{l} w_i}{W}, 1 \leq l \leq N$ is called Lorentz curve in economics. If the wealth is equally distributed then for each i $w_i = \frac{W}{N}$. The Lorenz curve in that case is the line $y = x$ confined to unit square. However in reality the Lorenz curve is below the line $y = x$, increasing from $(0, 0)$ to $(1, 1)$. Let A be the area of between the line
$y = x$ and the Lorenz curve. Let $B = \frac{1}{2}$ be the area below the line $y = x$ in the unit square. The ratio $\frac{A}{B} = 2A$ is called the Gini index in economics. It is one of the earliest measure of inequality. Gini index is 0 in case the wealth is equally distributed among the population. It is one if all the wealth is owned by one individual. In general the Gini index lies between zero and one. Closer it is to 0, more equitable is the wealth distribution.

Let us suppose that there are n possible levels of wealth of an individual, say $L_1 < L_2 < \cdots < L_n$, and that for each i, there are k_i individuals each with wealth L_i. We have

$$\sum_{i=1}^{n} k_i L_i = W, \quad \sum_{i=1}^{n} k_i = N.$$

$$\sum_{i=1}^{n} \frac{n_i}{N} L_i = \frac{W}{N}, \quad \sum_{i=1}^{n} \frac{k_i}{N} = 1.$$

Given the requirement that the wealth of an individual can be only one of the possibilities $L_1, L_2, \cdots, L_n$, and given that W and N are fixed, is there a choice of

$k_1, k_2, \cdots, k_n$ which is most well distributed in some sense? One possible answer is to choose $k_1, k_2, \cdots, k_n$ such that $\frac{k_1}{N}, \frac{k_2}{N}, \cdots, \frac{k_n}{N}$ is close to the Boltzmann distribution for $x = L_1, x_2 = L_2, \cdots, x_n = L_n$ and $\alpha = \frac{W}{N}$.

Application to Measurable Dynamics.

Let $(\Omega, \mathcal{B}, P)$ be a probability space an let T be a measurable transformation from Ω into Ω. T is said to be measure preserving if for all $A \in \mathcal{B}$, $P(T^{-1}(A)) = P(A)$. If S and T are two measure preserving transformations on Ω, we say that S and T are isomorphic or equivalent if there is an invertible measure preserving transformation σ on Ω such that the equality $\sigma S \sigma^{-1} = T$ holds a.e.

When are two m.p.t. isomorphic ? As shown by Kolmogorov the notion of entropy allows us to define a very strong numerical isomorphism invariant for the class of measure preserving transformations on a probability space. In the following we borrow from the book on Ergodic Theory by Cornfeld, Fomin and Sinai.

Let k and l be integers, $k \leq l$ and let $\mathbb{P}$ be a finite partition of Ω. Write

$$\mathbb{P}_k^l = T^{-k}\mathbb{P} \vee T^{-k-1}\mathbb{P} \vee \cdots \vee T^{-l}\mathbb{P}$$

Write

$$H_n = H(\mathbb{P}_0^{n-1})$$

Note that since T is measure preserving

$$H_n = H(T^{-m}\mathbb{P}_0^{n-1})$$

for all $n \geq 0$. We have,

$$H_{m+n} \leq H_m + H_n.$$

Indeed,

$$H_{m+n} = H(\mathbb{P}_0^{m+n-1}) \leq H(\mathbb{P}_0^{m-1}) + H(\mathbb{P}_m^{m+n-1}) = H_m + H_n$$

Thus, since the sequence $H_n, n = 1, 2, \cdots$ non-negative and subadditive, the sequence $\frac{1}{n}H_n, n = 1, 2, \cdots$ has a limit as $n \to \infty$. The limit is bounded by $H(\mathbb{P})$ as is easy to see. Write

$$h(T, \mathbb{P}) = \lim_{n \to \infty} \frac{1}{n} H_n$$

$$h(T) = \sup_{\mathbb{P}} h(T, \mathbb{P}),$$

where the supremum is taken over all finite partitions $\mathbb{P}$ of Ω.

If σ is an invertible measure preserving transformation on Ω, then it is clear that

$$H(T, \mathbb{P}) = H(\sigma T \sigma^{-1}, \sigma\mathbb{P}).$$

If $\mathbb{P}$ runs over all finite partitions of Ω, $\sigma\mathbb{P}$ also runs over all finite partitions of Ω, since σ is invertible. It follows that

$$\sup_{\mathbb{P}} h(T, \mathbb{P}) = \sup_{\mathbb{P}} h(\sigma T \sigma^{-1}, \mathbb{P}),$$

equivalently, $h(T) = h(\sigma T \sigma^{-1})$, i.e, $h(T)$ is an isomorphism invariant in the class of all measure preserving transformations on $(\Omega, \mathcal{B}, P)$.

A finite partition is said to be generating partition for T if $\cup_{n=1}^{\infty} T^{-n}\mathbb{P}$ generates the the σ-algebra $\mathcal{B}$. A basic theorem of Kolmogorov and Sinai states that if $\mathbb{P}$ is generating for T then $h(T) = h(T, P)$. We will use this theorem below to calculate entropies of irrational rotation and Bernoulli shifts.

Examples:

1)(Periodic maps). Say that T is periodic with period k if for all $x \in \Omega$, $T^k x = x$, and k is the smallest positive such k. Then for any finite partition $\mathbb{P}$ of Ω, $T^k\mathbb{P} = \mathbb{P}$, hence the sequence $H_n, n = 1, 2, \cdots$ is periodic with period k. It follows that $h(T, \mathbb{P}) = \lim_{n\to\infty} \frac{H_n}{n} = 0$. Consequently $h(T) = sup_{\mathbb{P}} h(T, \mathbb{P}) = 0$.

2)(Irrational rotation) Let $\Omega = \{z :| \ z \ |= 1\}$ the unit circle, equipped with its Borel σ-algebra and the normalised Haar measure. Let $\alpha = \exp\{i2\pi\beta\}$, where β is irrational. Define $Tz = \alpha z, z \in \Omega$. Then T is measure preserving and invertible transformation on Ω. We show that T has entropy zero. Let $\mathbb{P} = \{A, \Omega - A\}$ where A is the part of Ω in the upper half plane. Since β is irrational, by choosing n sufficiently large we can make the arc $A \cap T^n B$ as small as we please. The images of such arcs under powers of T generate the Borel σ-algebra of Ω. Hence by Kolmogorov-Sinai Theorem mentioned above we see that $h(T) = h(T, \mathbb{P})$. We now show that $h(T, \mathbb{P}) = 0$. We note that the partition $\vee_{i=0}^{n-1} T^i\mathbb{P}$ is made up of arcs whose end points form the set $T^i\{-1\}, T^i\{1\}, 0 \leq i \leq n-1$ so that there are $2n$ arcs in the partition $\vee_{i=0}^{n-1} T^i\mathbb{P}$. Hence we have $H(\vee_{i=0}^{n-1} T^i\mathbb{P}) \leq \log 2n$, whence

$$h(T) = \lim_{n\to\infty} \frac{1}{n} H(\vee_{i=0}^{n-1} T^i\mathbb{P}) \leq \lim_{n\to\infty} \frac{\log 2n}{n} = 0.$$

3) (Bernoulli Shift). Let $Y = \{y_1, y_2, \cdots, y_k\}$ be a finite set with a probability distribution on it given by $p(y_i) = p_i, \sum_{i=1}^{k} p_i = 1$. Let $\Omega = \Pi_{j=0}^{\infty} Y_j, \;\; Y_j = Y$ for all j. Equip Ω with its Borel σ-algebra $\mathcal{B}$ and product probability P. The map T defined on Ω

$$T(\omega_1, \omega_2, \cdots) = (\omega_2, \omega_3, \cdots), (\omega_1, \omega_2, \cdots) \in \Omega,$$

is called Bernoulli shift (non-invertible). It preserves P. We show that its entropy, i.e., $h(T)$, is equal to $-\sum_{j=1}^{k} p_j \log p_j$.

Let $\mathbb{P} = \{P_1, P_2, \cdots, P_k\}$, where $P_i = \{y_i\} \times \Pi_{j=2}^{\infty} Y_j$. Then $\cup_{j=0}^{\infty} T^{-j}(\mathbb{P})$ generates the σ-algebra $\mathcal{B}$. Hence by Kolmogorov-Sinai theorem mentioned above it is enough to calculate $h(T, \mathbb{P})$. An general element of $\vee_{j=0}^{n-1} T^{-j}\mathbb{P}$ is of the form

$$(\omega_0, \omega_1, \omega_2, \cdots, \omega_{n-1}) \times \Pi_{i=n}^{\infty} Y_i, Y_i = Y, i \geq n-1, \omega_j \in Y, j = 0, 1, 2, \cdots, n-1,$$

with probability

$$\Pi_{j=0}^{n-1} p(\omega_j)$$

We have

$$H(\vee_{j=0}^{n-1} T^{-j}\mathbb{P}) = -\sum_{(\omega_0, \omega_1, \omega_2, \cdots, \omega_{n-1})} \Pi_{j=0}^{n-1} p(\omega_j) \log(\Pi_{j=0}^{n-1} p(\omega_j))$$

$$= -\sum_{(\omega_0, \omega_1, \omega_2, \cdots, \omega_{n-1})} \Pi_{j=0}^{n-1} p(\omega_j) (\sum_{j=0}^{n-1} \log p(\omega_j))$$

$$= n \sum_{i=1}^{k} p_i \log p_i = nH(p_1, p_2, \cdots, p_k) = H_n$$

Clearly, then

$$h(T) = \lim_{n \to} \frac{1}{n} H_n = H(p_1, p_2, \cdots, p_k)$$

Entropy and Large Deviation

We follow S. R. S. Varadhan [2] and illustrate the connection between entropy and large deviation by considering the simplest example of binomial distribution.

We have

$$P(n,k) = \binom{n}{k}2^{-n} = \frac{n!}{k!(n-k)!}2^{-n}$$

Starling's approximation formula says that

$$n! \asymp \sqrt{2\pi}\exp\{-n\}n^{n+\frac{1}{2}}.$$

We have

$$P(n,k) \asymp \frac{\sqrt{2\pi}\exp(-n)n^{n+\frac{1}{2}}2^{-n}}{\sqrt{2\pi}\exp(-k)k^{k+\frac{1}{2}}\sqrt{2\pi}\exp -(n-k)(n-k)^{n-k+\frac{1}{2}}}$$

$$\log P(n,k) \asymp -\frac{1}{2}\log(2\pi)+(n+\frac{1}{2})\log n - n\log 2 - (k+\frac{1}{2})\log k - (n-k+\frac{1}{2})\log(n-k)$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log n - (n-k+\frac{1}{2})\log(1-\frac{k}{n})$$
$$-(k+\frac{1}{2})\log\frac{k}{n} - n\log 2$$

Assume now that k and n tend to infinity in such a way that $\frac{k}{n} \to x$. Then

$$\log P(n,k) \ -n[\log 2 + xlogx + (1-x)\log(1-x)] + o(n)$$
$$\asymp -nH(x) + o(n)$$

where

$$H(x) = \log 2 + x\log x + (1-x)\log(1-x).$$

This function H is one of the simplest example of what in theory of large deviation is called rate function, and as one can see, it is intimately related to the the notion of entropy introduced above.

Section 6:A Characterization of the Function H

Let Ω_n denote the space of probability vectors in $\mathbb{R}^n$ and let Ω denote the union of Ω_n over all n. Let H be a function on Ω with values in non-negative real numbers. Suppose that H satisfies the following properties:

$(1) H(p_1, p_2, \cdots, p_n) = H(p_1, p_2, \cdots, p_n, 0)$

(2) H is symmetric, i.e., for any permutation σ of $\{1, 2, \cdots, n\}$

$$H(p_{\sigma(1)}, p_{\sigma(2)}, \cdots, p_{\sigma(n)}) = H(p_1, p_2, \cdots, p_n),$$

(3) for all n,

$$H(p_1, p_2, \cdots, p_n) \leq H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$$

(4) Let

$$(p_{1,1}, p_{1,2}, \ldots, p_{1,n}, p_{2,1}, p_{2,2}, \cdots, p_{2,n}, \cdots, p_{m,1}, p_{m,2}, \cdots, p_{m,n})$$

be a probability vector with $m \cdot n$ entries, equivalently, let $[p_{i,j}], 1 \leq i \leq m, 1 \leq j \leq n$ be an $m \times n$ matrix which is a joint distribution of two random variables X and Y, taking values in the segments $\{1, 2, \cdots, m\}, \{1, 2, \cdots, n\}$ respectively. Let $\sum_{i=1}^{m} p_{i,j} = q_j$, so that $(q_1, q_2, \cdots, q_n)$ is the distribution of Y. Write $q_{i,j} = \frac{p_{i,j}}{q_j}$, the conditional probability that $X = i$ given that $Y = j$, so that $q_{i,j}, i = 1, 2, \cdots, m$ is a probability distribution on $1, 2, \cdots, m$. Let us now bring the function H into consideration. It is required that H satisfies the equality:

$$H(p_{1,1}, p_{2,1}, \cdots, p_{m,1}, \cdots, p_{m,1}, p_{m,2}, \cdots, p_{m,n})$$

$$= H(q_1, q_2, \cdots, q_n) + \sum_{j=1}^{n} q_j H(q_{1,j}, q_{2,j}, \cdots, q_{m,j})$$

It is a consequence of this requirement that if X and Y are independent random variables, equivalently if for each pair (i, j), $p_{i,j} = p_i \cdot q_j$ then

$$H(p_{1,1}, \cdots, p_{m,n}) = H(p_1, p_2, \cdots, p_m) + H(q_1, q_2, \cdots, q_n)$$

(5) H is continuous.

Note that the function $H_1(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i, (p_1, p_2, \cdots, p_n) \in \Omega$, satisfies the above five conditions. We have

Theorem: *Any function H on Ω into the set of non-negative real numbers satisfying (1) -(5) above is a constant multiple of H_1.*

Proof If H identically zero function, then $H = 0 \times H_1$ and there is nothing to prove. Hence we assume that H does not vanish identically. Write $L(n) = H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$. Then

$$L(n) = H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}, 0) \leq H(\frac{1}{n+1}, \frac{1}{n+1}, \cdots, \frac{1}{n+1}) = L(n+1).$$

From the comment following condition (4) it is easy to see that for all positive integers m, n

$$L(m \cdot n) = L(m) + L(n)$$

Fix positive integers m and n, and let t be the largest positive integer such that

$$m^t \leq n \leq m^{t+1}.$$

Then

$$tL(m) \leq L(n) \leq (t+1)L(m)$$

The function log also satisfies $\log m \cdot n = \log m + \log n$, we have similarly

$$t \log m \leq \log n \leq (t+1) \log m.$$

Thus $\frac{\log n}{logm}$ and $\frac{L(n)}{L(m)}$ lie in the same interval $[t, t+1]$, whence

$$| \frac{L(n)}{L(m)} - \frac{\log n}{\log m} | \leq 1.$$

Dividing by $\frac{\log n}{L(m)}$, we get

$$| \frac{L(n)}{\log n} - \frac{L(m)}{\log m} | \leq \frac{L(m)}{\log n}$$

For fixed m, letting n tend to infinity we see that $\frac{L(n)}{\log n}$ tends to $\frac{L(m)}{\log m}$, whence $\frac{L(m)}{\log m}$ is independent of m, hence a constant, which proves the theorem for positive integers.

Assume now that $(p_1, p_2, \cdots, p_n)$ is a probability vector with rational entries. Assume without loss of generality that $p_i = \frac{g_i}{g}$, $g_1 + g_2 + \cdots + g_n = g$. Consider two random variables X and Y with values respectively in the initial segments $[1, g]$ and $[1, n]$ of positive integers, such that

$$P(X = u, Y = 1) = \frac{1}{g} \text{ if } 1 \leq u \leq g_1,$$

$$P(X = u, Y = 1) = 0, \quad \text{otherwise}$$

$$P(X = u, Y = 2) = \frac{1}{g} \text{ if } g_1 < u \leq g_1 + g_2,$$
$$P(X = u, Y = 2) = 0, \quad \text{otherwise},$$

$\vdots$

$$P(X = u, Y = n) = \frac{1}{g} \text{ if } \sum_{j=1}^{n-1} g_j < u \leq \sum_{j=1}^{n} g_j$$
$$P(X = u, Y = g_n) = 0, \quad \text{otherwise.}$$

If $p_{u,j} = P(X = u, Y = j)$, then it is easy to see that

$$P(Y = j) = \sum_{u=1}^{g} p_{u,j} = \frac{g_j}{g}, 1 \leq j \leq n$$

$$\frac{p_{u,j}}{P(Y = j)} = \frac{1}{g_j}, \text{ if } \sum_{v=1}^{j-1} g_v < u \leq \sum_{v=1}^{j} g_j, = 0 \text{ otherwise}$$

We have

$$H(p_{1,1}, p_{1,2}, \cdots, p_{g,n}) = H(\frac{1}{g}, \frac{1}{g}, \cdots, \frac{1}{g}),$$

and by condition (4) this is equal to

$$H(\frac{g_1}{g}, \frac{g_2}{g} \cdots \frac{g_n}{g}) + \sum_{i=1}^{n} \frac{g_i}{g} H(\frac{1}{g_i}, \frac{1}{g_i}, \cdots, \frac{1}{g_i})$$

Since $H(\frac{1}{n}, \frac{1}{n}, \cdots \frac{1}{n}) = K \log n$, we at once get

$$H(\frac{g_1}{g}, \frac{g_2}{g} \cdots \frac{g_n}{g}) = -K \sum_{i=1}^{n} \frac{g_i}{g} \log \frac{g_i}{g} = K H_1(\frac{g_1}{g}, \frac{g_2}{g}, \cdots, \frac{g_n}{g})$$

Since the function H is continuous (condition 5), we see that $H = H_1$ on all of Ω and the proof is complete.