1. Information and Entropy

We recall

Theorem 1 (Shannon's First Theorem)**.** *Let X be a set with n elements and $n \geq 1$, $q \geq 2$, $P = (p_1, \cdots, p_n)$ a probability distribution on X. Then*

$$H_q(P) \leq \bar{L}(P) < H_q(P) + 1$$

where $\bar{L}(P)$ denotes the minimal average length of codewords used to encode the elements of X with an alphabet of q symbols, and

$$H_q(P) = -\sum_{i=1}^{n} p_i \log_q(p_i).$$

As noted before, Shannon's idea is to postulate $-\log_2(p(x_i))$ as the **information content** of the value x_i: we receive more information from an unexpected result than from an expected one. So information content is also identified with **unexpectedness**.
The entropy function is then the expected value of the information of x, considered as a random variable over X, and it may be also interpreted as a measure of uncertainty: if an experiment has n possible results with probability distribution $p = (p_1, \cdots, p_n)$, $H(p)$ is the expected value of the uncertainty in the outcome of the experiment.

In order to enhance the significance of Shannon's theorem, we sketch a more probabilistic deduction:
Let x be a random variable taking values $\{x_1, \cdots, x_n\}$ with probability distribution $p(x = x_i) = p_i$, or, more precisely, an infinite sequence $x(t)$ of independent and equally distributed random variables. The probability of a sequence sequence of values of $x(t)$, for $1 \leq t \leq M$, containing m_i occurences of x_i (ie, we make M independent "observations" of x, obtaining a sample of m_1 x_1, etc.) is $\prod_{i=1}^{n} p_i^{m_i}$.
The Law of large numbers asserts that, given arbitrarily small constants $\varepsilon > 0$ and $\delta > 0$, for M sufficiently large we have, with probability larger than $1 - \delta$

$$\left| \frac{m_i}{M} - p_i \right| < \varepsilon, \ \forall 1 \leq i \leq n;$$

therefore, outside of a set of small measure, each sequence of M observations has probability approximately $K = \prod_{i=1}^{n} p_i^{Mp_i}$, and the information content of one such sequence is, again approximately,

$$\log_2 \left(\frac{1}{K} \right) = MH(P),$$

ie, the average number of symbols needed to code each observation is $H(P)$.

Exercise 2. *Let $\{p_1, \cdots, p_m\}$ be a probability distribution and $p^* = \max_i \{p_i : 1 \leq i \leq m\}$. Prove that*

a) $H_2(p_1, \cdots, p_m) \geq H_2(p^*)$;
b) $H_2(p_1, \cdots, p_m) \geq -\log_2(p^*)$;
c) $H_2(p_1, \cdots, p_m) \geq 2(1 - p^*)$.

1.1. **Conditional Entropy.** Let X and Y denote random variables taking values, respectively, $x_k : 1 \leq k \leq n$ and $y_j : 1 \leq j \leq m$, with probability distributions $p(X = x_k)$ and $p(Y = y_j)$; when there is no serious risk of confusion we'll use the simpler notation $p(x_k)$, etc.
Let $p(x_k|y_j)$ (or, in the more complete notation, $p(X = x_k|Y = y_j)$) denote the **conditional probability** that X takes the value x_k given that Y takes the value y_j. We have

$$p(x_k|y_j) = \frac{p(x_k, y_j)}{p(y_j},$$

where $p(x_k, y_j)$ is the **joint probability** that X takes the value x_k and Y takes the value y_j.

If the value y_j is given, the **conditional entropy** of X is

$$H(X|y_j) = -\sum_k p(x_k|y_j) \log(p(x_k|y_j))$$

and the conditional entropy $H(X|Y)$ is its expected value

$$H(X|Y) = \sum_j H(X, y_j)p(y_j.$$

Following the interpretation of Entropy, $H(X|Y)$ is the uncertainty about X that remains after the knowledge of Y and

$$I(X;Y) = H(X) - H(X|Y)$$

called the **Mutual Information** of the two random variables, is the information about one of the variables given by the other.

Reversing the roles of the random variables, we have definitions of $p(y_j|x_k)$, $H(Y|X)$, etc.

Exercise 3. *Show that $I(X;Y) = I(Y;X)$.*

Example 4. *B_1 and B_2 are identical boxes; the first contains s white balls and t red balls, while the second contains t white balls and s red balls. We take a ball from one of the boxes; the random variable X takes the value i if the box B_i is chosen; the random variable Y takes the value w if a white ball is chosen and the value r otherwise.*
*It is easy to see that the probability distribution of both variables is uniform and so $H(X) = H(Y) = 1$. On the other hand we have conditional probabilities (**HW**)*

$$p((1|w) = p(2|r) = \frac{s}{2(s+t)}, \qquad p((2|r) = p(1|w) = \frac{t}{2(s+t)};$$

Exercise 5. *Compute the general expression of $H(X|Y)$ and $I(X;Y)$ and their values for some concrete values of s and t.*

Although the notions of conditional entropy and information apply in general to the relation between random variables, we will focus on their significance in the context of transmission of information through a communication channel.
The precise definition of a mathematical model of communication channel will be discussed later, but for now it's enough to take X and Y to be, respectively, the input and output of a transmission and that $p(y_j|x_k)$ is a given probability that the input value x_k is received as y_j.
The simplest example is the **Binary Symmetric Channel**, where X and Y both take values 0 and 1 and the $p(y_j|x_k)$ are

$$p(0|1) = p(1|0) = \rho, \qquad p(0|0) = p(1|1) = 1 - \rho,$$

for some $0 < \rho < 0.5$.

1.1.1. *Further properties of Conditional Entropy.* The entropy of the joint distribution is $H(X,Y) = -\sum_{k,j} p(x_k, y_j) \log(p(x_k, y_j))$. An application of the defining formulas gives

Proposition 6. $H(X|Y) = H(X,Y) - H(Y)$.

Proof. (**HW**). □

We have also, as a simple consequence of Gibbs Lemma,

Lemma 7. *For any random variables X and Y,*

$$H(X,Y) \leq H(X) + H(Y),$$

with equality if and only if X and Y are independent.

Proof. (**HW**). □

Corollary 8. *$H(X|Y) \leq H(X)$, with equality if and only if X and Y are independent.*

Proof. (**HW**) □

Given the interpretation of conditional entropy for transmission of information, it is natural to relate it to the probability of error in the transmission, ie the probability that $X \neq Y$. The next inequality does just this.

Proposition 9 (Fano's inequality)**.** *If X and Y are random variables taking values in the same set $\{x_1, \cdots, x_m\}$ and $p_e = p(X \neq Y)$, then*

$$H(X|Y) \leq H(p_e) + p_e \log(m-1).$$

This will be an easy corollary of the next theorem. We start with the statement of a particular case of Jensen's inequality, which will be used frequently:

Lemma 10 (Jensen's inequality (particular case))**.** *Let X and Y be random variables taking values in finite sets $\{x_i :\leq i \leq m\}$ and $\{y_j : 1 \leq j \leq v\}$, respectively, $f(x,y)$ a joint probability distribution and $g(x,y)$ a positive function. We have*

$$\sum_{i,j} f(x_i, y_j) \log(g(x_i, y_j)) \leq \log\left(\sum_{i,j} f(x_i, y_j) g(x_i, y_j)\right),$$

with equality if and only if there exist x_0, y_0 such that $f(x_0, y_0) = 1$.

Proof. This is a consequence of the convexity of $-\log$ (**HW**). □

Theorem 11. *Let X, Y, Z denote discrete random variables, taking values in finite sets, and $a(z) = \sum_{i,j} p(y_j)p(z_k|x_i, y_j)$. Then*

$$H(X|Y) \leq H(Z) + \sum_k p(z_k) \log(a(z_k)).$$

Proof. We have

$$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log\left(\frac{1}{p(x_i|y_j)}\right) = \sum_{i,j,k} p(x_i, y_j, z_k) \log\left(\frac{1}{p(x_i|y_j)}\right) =$$

$$= \sum_k p(z_k) \sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(z_k)} \log\left(\frac{1}{p(x_i|y_j)}\right).$$

For each z_k, $\frac{p(x_i,y_j,z_k)}{p(z_k)}$ determines a probability distribution on (X, Y); applying Jensen's inequality, we get

$$H(X|Y) \leq \sum_k p(z_k) \log\left(\frac{1}{p(z_k)} \sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(x_i|y_j)}\right) =$$

$$= H(Z) + \sum_k p(z_k) \log\left(\sum_{i,j} \frac{p(x_i, y_j, z_k)}{p(x_i|y_j)}\right).$$

We notice now that

$$\frac{p(x_i, y_j, z_k)}{p(x_i|y_j)} = p(y_j)p(z_k|x_i, y_j).$$

□

Exercise 12. *Check the details of the proof.*

The proof of Fano's inequality follows taking $z = 0$ if $X = Y$ and $z = 1$ if $X \neq Y$ (**HW**).

We end with some properties of mutual information that will be relevant to its interpretation and application in the context of communication channels. Their proofs are almost direct applications of Jensen's inequality. Notice that the same base for logarithms is being used everywhere.

Theorem 13. *If X, Y, Z are discrete random variables with values in finite sets,*

$$I((X, Y); Z) \geq I(Y; Z),$$

with equality if and only if $p(z_k|x_i, y_j) = p(z_k|y_j)$ for all values with $p(x_i, y_j, z_k) > 0$.

Proof. (**HW**): we have

$$I((X, Y); Z) = \sum_{x_i,y_j,z_k} p(x_i, y_j, z_k) \log\left(\frac{p(z_k|(x_i, y_j))}{p(z_k)}\right),$$

and

$$I(Y; Z) = \sum_{y_j,z_k} p(y_j, z_k) \log\left(\frac{p(z_k|y_j)}{p(z_k)}\right) = \sum_{x_i,y_j,z_k} p(x_i, y_j, z_k) \log\left(\frac{p(z_k|y_j)}{p(z_k)}\right).$$

Apply Jensen's inequality. □

Although the inequality is not surprising, given the interpretation of mutual information, the condition for equality is more important: it means that the sequence (X, Y, Z) is a **Markov chain**; informally, Z depends on X only through the dependence of Y on X. This implies (see the problem at the end)

that the reversed sequence (Z, Y, X) is also a Markov chain and, consequently,

Corollary 14. *If (X, Y, Z) is a Markov chain, then*

$$I(X;Z) \leq I(X;Y), \qquad I(X;Z) \leq I(Y;Z).$$

Proof. (**HW**).

□

We will come to this point later when we consider the coding and communication process. In this context, (X, Y, Z) is a Markov chain because Y is the output of the transmission, through a communication channel C_1, of an input X, and Z is the output of the transmission of Y through another communication channel C_2.

1.2. **Supplementary Results and Problems.**

1.2.1. *Characterization and properties of Entropy.*

Remark 15. *Recall that in all the formulas we use the convention $0 \log(0) = 0$.*

Proposition 16 (Characterization of Entropy)**.** *Let*

$$\Delta = \{(p_i)_{i\geq 1} : 0 \leq p_i \leq 1; \exists N : p_i = 0 \, \forall i > N; \sum_{i\geq 1} p_i = 1\}.$$

We have

$$\Delta = \bigcup_{N\in\mathbb{N}} \{x = (x_1, \cdots, x_N) \in \mathbb{R}^N : 0 \leq x_i \leq 1; \sum_{i\geq 1} x_i = 1\}\}$$

*and Δ is a metric space under the metric that extends, for each N, the usual metric of $\mathbb{R}^N$ (**HW**).*
Let $\Phi : \Delta \to \mathbb{R}$ be a function; for each sequence $(p_i) \in \Delta$, we will write $\Phi(p_1, \cdots, p_N)$ to denote $\Phi((p_i)_{i\geq 1})$.
Then Φ satisfies properties

1. *Continuity;*
2. $\Phi(\frac{1}{n}, \cdots, \frac{1}{n}) < \Phi(\frac{1}{n+1}, \cdots, \frac{1}{n+1})$ $\forall n \in \mathbb{N}$*;*
3. *For any $n, k \in \mathbb{N}$ and $b_1, \cdots, b_k \in \mathbb{N}$ such that $\sum_{j=1}^k b_j = n$,*

$$\Phi(\frac{1}{n}, \cdots, \frac{1}{n}) = \Phi(\frac{b_1}{n}, \cdots, \frac{b_k}{n}) + \sum_{j=1}^{k} \frac{b_j}{n} \Phi(\frac{1}{b_j}, \cdots, \frac{1}{b_j});$$

if and only if there exists $q > 1$ such that

$$\Phi(p_1, \cdots, p_N) = H_q(p_1, \cdots, p_N) = -\sum_{i=1}^{N} p_i \log_q(p_i).$$

Proof. The fact that the functions H_q satisfy properties 1., 2. and 3. is an easy exercise (**HW**).
In the proof of the converse, we will use the simplified notation $g(n) = \Phi(\frac{1}{n}, \cdots, \frac{1}{n})$. The proof proceeds through a sequence of steps, whose details are left as exercises (**HW**):

i) For any positive s and m, apply 3. with $n = m^s$, $b_j = m$ and $k = m^{s-1}$ to conclude that $g(m^s) = sg(m)$ and deduce from 2. that $g(n)$ is a strictly positive and strictly growing function;

ii) For fixed m and arbitrary positive integers r and t, choose s such that $m^s \leq r^t < m^{s+1}$; this implies

$$\frac{s}{t} \leq \frac{\log_2(r)}{\log_2(m)} < \frac{s+1}{t},$$

and on the other hand, using i), that

$$\frac{s}{t} \leq \frac{g(r)}{g(m)} < \frac{s+1}{t};$$

we conclude that

$$-\frac{1}{t} \leq \frac{g(r)}{g(m)} - \frac{\log_2(r)}{\log_2(m)} < \frac{1}{t},$$

and so that

$$\frac{g(r)}{g(m)} = \frac{\log_2(r)}{\log_2(m)}$$

which implies (why?) that there exists a positive C such that $g(r) = C\log_2(r)$ for all $r \in \mathbb{N}$ or, equivalently, that there exists a $q > 1$ such that $g(r) = \log_q(r)$ for all $r \in \mathbb{N}$;

iii) we may thus rewrite property 3. as

$$\Phi(\frac{b_1}{n}, \cdots, \frac{b_k}{n}) = -\sum_{j=1}^{k} \frac{b_j}{n} \log_q(\frac{b_j}{n})$$

and, as any $(p_i)_{i\geq 21} \in \Delta$ with rational entries may written in the form $(\frac{b_1}{n}, \cdots, \frac{b_k}{n})$, we conclude that $\Phi(p_1, \cdots, p_n) = H_q(p_1, \cdots, p_n)$ if the p_i are rational;

iv) by continuity the equality holds for all $(p_i)_{i\geq 21} \in \Delta$.

□

Problem 17. *Formulate a convincing (for yourself...) interpretation of property* 3. *in terms of uncertainty/information of an experiment.*

Problem 18. *Let* $H_q(p) = -p\log_q(p) - (1-p)\log_q(1-p)$.

a) *Study the function with respect to convexity.*

b) *Let* $m = \lfloor pn \rfloor$*; prove the inequality* $\sum_{j=0}^{\lfloor pn \rfloor} \binom{n}{j} \leq q^{nH_q(m/n)} \leq q^{nH_q(p)}$*, for any* $n \in \mathbb{N}$.

Hint: *Apply Newton's Binomial formula to* $1 = \left(\frac{m}{n} + (1 - \frac{m}{n})\right)^n$.

Problem 19. *Show that if*

$$0 \leq p_i \leq 1, \sum_i p_i = 1 \text{ and } 0 \leq t_i \leq 1, \sum_i t_i \leq 1$$

with $1 \leq i \leq n$, *then*

$$-\sum_i p_i \log_q(p_i) \leq -\sum_i p_i \log_q(t_i).$$

Problem 20. *Show that* $H(P)$ *is concave in the set of all probability distributions: if* $0 \leq t \leq 1$ *and*

$$P = (p_1, \cdots, p_n), \qquad Q = (q_1, \cdots, q_n),$$

then

$$H(tP + (1-t)Q) \geq tH(P) + (1-t)H(Q).$$

Problem 21. *Fix* n *and* q.

a) *Suppose that* $(p_1, \cdots, p_n)$ *satisfies, for a certain* $\varepsilon > 0$, $p_1 > p_2 + 2\varepsilon$ *and let* $(r_1, \cdots, r_n)$ *be defined as*

$$r_1 = p_1 - \varepsilon,\ r_2 = p_2 + \varepsilon,\ r_j = p_j\ \forall 3 \leq j \leq n.$$

Show that $H_q(p_1, \cdots, p_n) \leq H_q(r_1, \cdots, r_n)$, *ie, a variation on the probability distribution that tends to approach any equalization of probabilities implies an increase of entropy.*

b) *Let* $A = [a_{ij}]$ *be a* $n \times n$ *doubly stochastic matrix:* $a_{ij} \geq 0$ *for all* $1 \leq i, j \leq n$ *and*

$$\sum_{i=1}^{n} a_{ij} = 1\ \forall 1 \leq j \leq n, \qquad \sum_{j=1}^{n} a_{ij} = 1\ \forall 1 \leq i \leq n.$$

Define $(r_1, \cdots, r_n) = (p_1, \cdots, p_n)A$. *Show that* $H_q(p_1, \cdots, p_n) \leq H_q(r_1, \cdots, r_n)$. *Verify that a) is a particular case.*

Hint *for b):* $H_q(r_1, \cdots, r_n) = -\sum_i \sum_j p_i a_{ij} \log_q(r_j)$; *use the convexity of* $-log()$ *and apply Gibbs's inequality.*

1.2.2. *Convexity properties of Mutual Information.* The next two results state that the mutual information between two random variables has convexity properties:

Proposition 22. *Let the forward channel probabilities* $p(y|x)$ *be fixed and* $p_1(x)$ *and* $p_2(x)$ *be probability distributions on input random variables* X_1 *and* X_2, *respectively; let* Y_1 *and* Y_2 *be the corresponding output random variables. For* $0 \leq t \leq 1$, *let* X *be the input random variable with probability distribution* $p(x) = tp_1(x) + (1-t)p_2(x)$ *and* Y *the corresponding output. Then*

$$tI(X_1; Y_1) + (1-t)I(X_2; Y_2) \leq I(X; Y).$$

Proof. (**HW**). □

Proposition 23. *Let $p(x)$ be a fixed probability distribution on the input variable X, and $p_1(y|x)$ and $p_2(y|x)$ be two sets of forward channel probabilities; denote as Y_1 and Y_2 the corresponding output variables. Then, for $0 \leq t \leq 1$ the forward channel probabilities*

$$p(y|x) = tp_1(y|x) + (1-t)p_2(y|x)$$

with output Y satifies

$$I(X;Y) \leq tI(X;Y_1) + (1-t)I(X;Y_2).$$

Proof. (**HW**). □

1.2.3. *Markov Chains.*

Problem 24. *Prove that (X,Y,Z) is a Markov chain if and only if (Z,Y,X) is.*

Problem 25. *Let, for $i \geq 0$, X_i, be a sequence of random variables, all with values $\{0,1\}$, such that X_0 has probability distribution $P(X_0 = 0) = a$ and $P(X_0 = 1) = 1-a$, and*

$$p(X_{i+1} = 0|X_i = 0) = p(X_{i+1} = 1|X_i = 1) = 1-\rho, \qquad p(X_{i+1} = 1|X_i = 0) = p(X_{i+1} = 0|X_i = 1) = \rho.$$

Determine the probability distribution of X_n and $\lim_{n\to+\infty} I(X_n;X_0)$.

Problem 26. *Let, for $i \geq 0$, X_i, be a sequence of random variables, all with values $\{1,\cdots,m\}$, and that, for any i,*

$$p(X_{i+1} = k|X_i = j) = a_{jk};$$

suppose that X_0 has probability distribution $p = \{p_1,\cdots,p_m\}$ satisfying

$$p_k = \sum_{j=1}^{m} p_j a_{jk},$$

*(p is **stable** with respect to $A = [a_{jk}]$).*

a) *Show that the X_i have all probability distribution p.*

b) *Define the entropy of the Markov chain X_i as*

$$H = \lim_{n\to+\infty} \frac{1}{n} H((X_0, X_1, \cdots, X_{n-1})).$$

Show that

$$H = -\sum_{j,k} p_k a_{jk} \log(a_{jk}).$$