Exercise Problems 1–4: Information Theory

Exercise 1

(a) Prove that the information measure is additive: that the information gained from observing the combination of N independent events, whose probabilities are p_i for $i = 1....N$, is the *sum* of the information gained from observing each one of these events separately and in any order.

(b) Calculate the entropy in bits for each of the following random variables:

(i) Pixel values in an image whose possible grey values are all the integers from 0 to 255 with uniform probability.

(ii) Humans classified according to whether they are, or are not, mammals.

(iii) Gender in a tri-sexual insect population whose three genders occur with probabilities 1/4, 1/4, and 1/2.

(iv) A population of persons classified by whether they are older, or not older, than the population's median age.

(c) Consider two independent integer-valued random variables, X and Y. Variable X takes on only the values of the eight integers {1, 2, ..., 8} and does so with uniform probability. Variable Y may take the value of *any* positive integer k, with probabilities $P\{Y = k\} = 2^{-k},\ k = 1, 2, 3, ...$

(i) Which random variable has greater uncertainty? Calculate both entropies $H(X)$ and $H(Y)$.

(ii) What is the joint entropy $H(X, Y)$ of these random variables, and what is their mutual information $I(X; Y)$?

(d) What is the maximum possible entropy H of an alphabet consisting of N different letters? In such a maximum entropy alphabet, what is the probability of its most likely letter? What is the probability of its least likely letter? Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.

Exercise 2

(a) Suppose that women who live beyond the age of 80 outnumber men in the same age group by three to one. How much information, in bits, is gained by learning that a person who lives beyond 80 is male?

(b) Consider n discrete random variables, named $X_1, X_2, ..., X_n$, of which X_i has entropy $H(X_i)$, the largest being $H(X_L)$. What is the upper bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all these random variables, and under what condition will this upper bound be reached? What is the lower bound on the joint entropy $H(X_1, X_2, ..., X_n)$?

(c) Suppose that X is a random variable whose entropy $H(X)$ is 8 bits. Suppose that $Y(X)$ is a deterministic function that takes on a different value for each value of X.

(i) What then is $H(Y)$, the entropy of Y?

(ii) What is $H(Y|X)$, the conditional entropy of Y given X?

(iii) What is $H(X|Y)$, the conditional entropy of X given Y?

(iv) What is $H(X, Y)$, the joint entropy of X and Y?

(v) Suppose now that the deterministic function $Y(X)$ is not invertible; in other words, different values of X may correspond to the same value of $Y(X)$. In that case, what could you say about $H(Y)$?

(vi) In that case, what could you say about $H(X|Y)$?

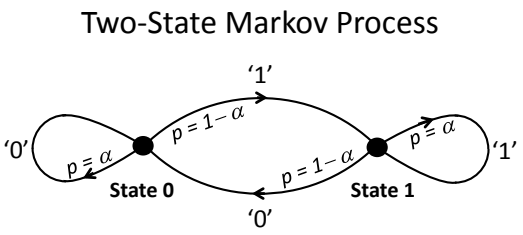
(d) Let the random variable X be five possible symbols $\{\alpha, \beta, \gamma, \delta, \epsilon\}$. Consider two probability distributions $p(x)$ and $q(x)$ over these symbols, and two possible coding schemes $C_1(x)$ and $C_2(x)$ for this random variable:

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
α	1/2	1/2	0	0
β	1/4	1/8	10	100
γ	1/8	1/8	110	101
δ	1/16	1/8	1110	110
ϵ	1/16	1/8	1111	111

1. Calculate $H(p)$, $H(q)$, and relative entropies (Kullback-Leibler distances) $D(p||q)$ and $D(q||p)$.
2. Show that the average codeword length of C_1 under p is equal to $H(p)$, and thus C_1 is optimal for p. Show that C_2 is optimal for q.
3. Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords? By how much does it exceed the entropy $H(p)$? Relate your answer to $D(p||q)$.
4. If we use code C_1 when the distribution is q, by how much does the average codeword length exceed $H(q)$? Relate your answer to $D(q||p)$.

Exercise 3

(a) A two-state Markov process may emit '0' in State 0 or emit '1' in State 1, each with probability α, and return to the same state; or with probability $1-\alpha$ it emits the other symbol and switches to the other state. Thus it tends to be "sticky" or oscillatory, two forms of predictability, depending on α.



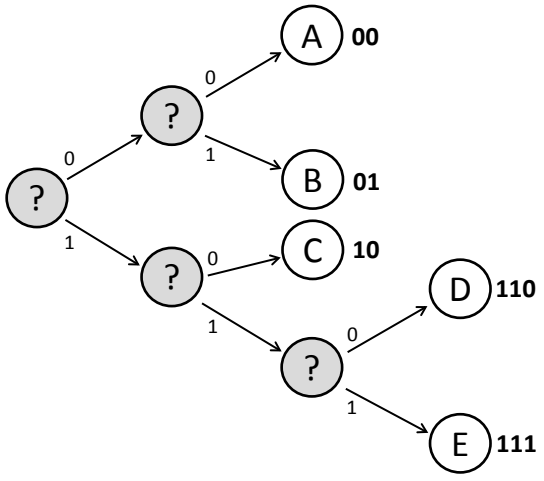
1. What are the state occupancy probabilities for $0 < \alpha < 1$?
2. What are the entropy of State 0, the entropy of State 1, and the overall entropy of this source? Express your answers in terms of α.
3. For what value(s) of α do both forms of predictability disappear? What then is the entropy of this source, in bits per emitted bit?

(b) Consider an alphabet of 8 symbols whose probabilities are as follows:

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

1. If someone has selected one of these symbols and you need to discover which symbol it is by asking 'yes/no' questions that will be truthfully answered, what would be the most efficient sequence of such questions that you could ask in order to discover the selected symbol?
2. By what principle can you claim that each of your proposed questions is maximally informative?
3. On average, how many such questions will need to be asked before the selected symbol is discovered?
4. What is the entropy of the above symbol set?
5. Construct a uniquely decodable prefix code for the symbol set, and explain why it is uniquely decodable and why it has the prefix property.
6. Relate the bits in your prefix code to the 'yes/no' questions that you proposed in 1.

(c) Huffman trees enable construction of uniquely decodable prefix codes with optimal codeword lengths. The five codewords shown here for the alphabet {A,B,C,D,E} form an instantaneous prefix code.



1. Give a probability distribution for the five letters that would result in such a tree;
2. Calculate the entropy of that distribution;
3. Compute the average codeword length for encoding this alphabet. Relate your results to the Source Coding Theorem.

Exercise 4

(a) A fair coin is secretly flipped until the first head occurs. Let X denote the number of flips required. The flipper will truthfully answer any "yes-no" questions about his experiment, and we wish to discover thereby the value of X as efficiently as possible.

(i) What is the most efficient possible sequence of such questions? Justify your answer.

(ii) On average, how many questions should we need to ask? Justify your answer.

(iii) Relate the sequence of questions to the bits in a uniquely decodable prefix code for X.

(b) Consider a binary symmetric communication channel, whose input source is the alphabet $X = \{0, 1\}$ with probabilities $\{0.5, 0.5\}$; output alphabet $Y = \{0, 1\}$; and with channel matrix:

$$\begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

where ϵ is the probability of transmission error.

(i) What is the entropy of the source, $H(X)$?

(ii) What is the probability distribution of the outputs, $p(Y)$, and what is the entropy of this output distribution, $H(Y)$?

(iii) What is the joint probability distribution for the source and the output, $p(X, Y)$, and what is the joint entropy, $H(X, Y)$?

(iv) What is the mutual information of this channel, $I(X; Y)$?

(v) How many values are there for ϵ for which the mutual information of this channel is maximal? What are those values, and what then is the capacity of such a channel in bits?

(vi) For what value of ϵ is the capacity of this channel minimal? What is the channel capacity in that case?

(c) Consider an asymmetric communication channel whose input source is the binary alphabet $X = \{0, 1\}$ with probabilities $\{0.5, 0.5\}$ and whose outputs Y are also this binary alphabet $\{0, 1\}$, but with asymmetric error probabilities. Thus an input 0 is flipped with probability α, but an input 1 is flipped with probability β, giving this channel matrix $p(y_k|x_j)$:

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

(i) Give the probabilities of both outputs, $p(Y = 0)$ and $p(Y = 1)$.

(ii) Give all the values of (α, β) that would maximise the capacity of this channel, and state what that capacity then would be.

(iii) Give three pairs of values of (α, β) that would minimise the capacity of this channel, and state what that capacity would then be.