Solutions to Information Theory Exercise Problems 1–4

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.

Solution:

(a) The information measure assigns $\log_2(p)$ bits to the observation of an event whose probability is p. The joint probability of a combination of N independent events whose probabilities are $p_1....p_N$ is $\prod_{i=1}^{N} p_i$. Thus the information content of such a combination is:

$$\log_2(\prod_{i=1}^{N} p_i) = \log_2(p_1) + \log_2(p_2) + \cdots + \log_2(p_N)$$

which is the sum of the information content of all of the separate events.

(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-sexed 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.

Solution:

(b) By definition, $H = -\sum_i p_i \log_2 p_i$ is the entropy in bits for a discrete random variable distributed over states whose probabilities are p_i. So:

(i) In this case each $p_i = 1/256$ and the ensemble entropy summation extends over 256 such equiprobable grey values, so $H = -(256)(1/256)(-8) = 8$ bits.

(ii) Since all humans are in this category (humans $\subset$ mammals), there is no uncertainty about this classification and hence the entropy is 0 bits.

(iii) The entropy of this distribution is $-(1/4)(-2) - (1/4)(-2) - (1/2)(-1) = 1.5$ bits.

(iv) By the definition of median, both classes have probability 0.5, so the entropy is 1 bit.

(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 $I(X; Y)$?

Solution:

(c)

(i) Surprisingly, there is greater uncertainty about random variable X which is just any one of the first 8 integers, than about Y which can be *any* positive integer. The uniform probability distribution over the eight possibilities for X means that this random variable has entropy $H(X) = 3$ bits. But the rapidly decaying probability distribution for random variable Y has entropy

$$H(Y) = -\lim_{N\to\infty} \sum_{k=1}^{N} 2^{-k} \log_2(2^{-k})$$

and this series is known (Slide 53) to converge to just 2 bits.

(ii) Since random variables X and Y are independent, their joint entropy $H(X, Y)$ is $H(X) + H(Y) = 5$ bits, and their mutual information is $I(X; Y) = 0$ bits.

(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.

Solution:

(d) The maximum possible entropy of an alphabet consisting of N different letters is $H = \log_2 N$. This is only achieved if the probability of every letter is $1/N$. Thus $1/N$ is the probability of both the "most likely" and the "least likely" letter.

Fixed length codes are inefficient for alphabets whose letters are not equiprobable because the cost of coding improbable letters is the same as that of coding more probable ones. It is more efficient to allocate fewer bits to coding the more probable letters, and to make up for the reduced address space of such short strings of bits by making longer codes for the less probable letters. In other words, a variable-length code. An example is Morse Code, in which the most probable English letter, e, is coded by a single dot.

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?

Solution:

(a) Rewriting "live beyond the age of 80" simply as "old", we have the conditional probabilities $p(\text{female}|\text{old}) = 3p(\text{male}|\text{old})$ and also of course $p(\text{female}|\text{old})+p(\text{male}|\text{old})= 1$. It follows that $p(\text{male}|\text{old})= 1/4$. The amount of information (in bits) gained from an observation is $-\log_2$ of its probability. Thus the information gained by such an observation is 2 bits worth.

(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 $H(X_1, X_2, ..., X_n)$?

Solution:

(b) The upper bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of the random variables is:

$$H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$$

This upper bound is reached only in the case that all the random variables are independent. The lower bound on joint entropy is $H(X_L)$, the largest of the marginal entropies of any of the random variables (see Slides 26 – 27 and the Venn diagram). In summary,

$$H(X_L) \leq H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$$

(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 $H(X|Y)$?

Solution:

(i) The entropy of Y: $H(Y) = 8$ bits also.

(ii) The conditional entropy of Y given X: $H(Y|X) = 0$ because of determinism.

(iii) The conditional entropy of X given Y: $H(X|Y) = 0$ also.

(iv) The joint entropy $H(X,Y) = H(X) + H(Y|X) = 8$ bits.

(v) Since now different values of X may correspond to the same value of $Y(X)$, the new distribution of Y has lost entropy and so $H(Y) < 8$ bits.

(vi) Now knowledge of Y no longer determines X, and so the conditional entropy $H(X|Y)$ is no longer zero: $H(X|Y) > 0$.

(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 $D(q||p)$.

   Solution:

   $H(p) = -\sum_i p_i \log_2 p_i = 1/2 + 2/4 + 3/8 + 4/16 + 4/16 = 1\frac{7}{8}$ bits

   $H(q) = -\sum_i q_i \log_2 q_i = 1/2 + 4 \times 3/8 = 2$ bits

   $D(p||q) = \sum_i p_i \log_2(p_i/q_i) = 1/2 \times \log_2(1) + 1/4 \times \log_2(2) + 1/8 \times \log_2(1) + 2 \times 1/16 \times \log_2(1/2) = 0 + 1/4 + 0 - 1/8 = \frac{1}{8}$ bit.

   $D(q||p) = \sum_i q_i \log_2(q_i/p_i) = 1/2 \times \log_2(1) + 1/8 \times \log_2(1/2) + 1/8 \times \log_2(1) + 2 \times 1/8 \times \log_2(2) = 0 - 1/8 + 0 + 1/4 = \frac{1}{8}$ bit.

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.

   Solution:

   The average codeword length of C_1 (weighting codeword lengths in bits by their symbol probabilities under p) is: $1/2 + 2/4 + 3/8 + 4/16 + 4/16 = 1\frac{7}{8}$ bits. This equals the entropy $H(p)$, so C_1 is optimal for p.

   The average codeword length of C_2 (weighting codeword lengths in bits by their symbol probabilities under q) is: $1/2 + 3/8 + 3/8 + 3/8 + 3/8 = 2$ bits. This equals the entropy $H(q)$, so 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 $D(p||q)$.

   Solution:

   The average codeword length of C_2 when the distribution is p is: $1/2+3/4+3/8+3/16+3/16 = 2$ bits. This exceeds the entropy $H(p) = 1\frac{7}{8}$ by $\frac{1}{8}$ bit, which is, as expected, the miscoding cost given by $D(p||q)$ for source coding with the wrong probability distribution.

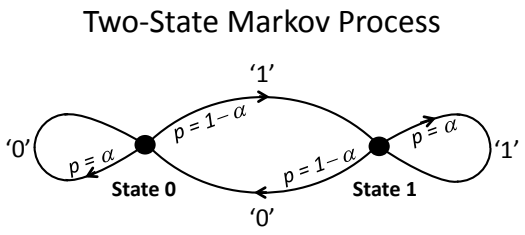
4. If we use code C_1 when the distribution is q, by how much does the average codeword length exceed $H(q)$? Relate your answer to $D(q||p)$.

   Solution:

   The average codeword length of C_1 when the distribution is q is: $1/2+2/8+3/8+4/8+4/8 = 17/8 = 2\frac{1}{8}$ bit. This exceeds entropy $H(q) = 2$ by $\frac{1}{8}$ bit, which is, as expected, the miscoding cost given by $D(q||p)$ for source coding with the wrong probability distribution.

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 $0 < \alpha < 1$?

   Solution:

   State occupancy probabilities: $p(\text{State 0}) = 0.5$ and $p(\text{State 1}) = 0.5$ regardless of the value of α, as both switches are equiprobable and the flow paths between the states are symmetrical.

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.

   Solution:

   From the definition of entropy in terms of event probabilities, the entropy of State 0 is $H(\alpha) = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$. State 1 has the same entropy. The overall entropy of a Markov process combines the entropy of each of its States weighted by their occupancy probabilities. Thus we get again the same expression $H(\alpha) = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$ for the overall entropy of this source, since its two States are equiprobable.

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?

   Solution:

   If $\alpha = 0.5$, the process is neither "sticky" nor oscillatory. It is then just a single-state Bernoulli process, and its entropy is maximum, at $H(\alpha) = 1$ bit per bit emitted.

(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?

   Solution:

   For this symbol distribution, the most efficient sequence of questions to ask (until a 'yes' is obtained) would actually be just: (1) Is it A? (2) Is it B? (3) Is it C? (etc), – quite unlike the original "game of 20 questions".

2. By what principle can you claim that each of your proposed questions is maximally informative?

   Solution:

   Each such 1-bit question is maximally informative because the remaining uncertainty is reduced by half (1 bit).

3. On average, how many such questions will need to be asked before the selected symbol is discovered?

   Solution:

   The probability of terminating successfully after exactly N questions is 2^{-N}. At most seven questions might need to be asked. The weighted average of the interrogation durations is:

$$\frac{1}{2} + (2)(\frac{1}{4}) + (3)(\frac{1}{8}) + (4)(\frac{1}{16}) + (5)(\frac{1}{32}) + (6)(\frac{1}{64}) + (7)(\frac{2}{128}) = 1\frac{126}{128}$$

   In other words, on average just slightly less than two questions need to be asked in order to learn which of the 8 symbols it is.

4. What is the entropy of the above symbol set?

   Solution:

   The entropy of the above symbol set is calculated by the same formula, but over all 8 states (whereas at most 7 questions needed to be asked):

$$H = -\sum_{i=1}^{8} p_i \log_2 p_i = 1\frac{126}{128}$$

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.

   Solution:

   A natural code book to use would be the following:

A	B	C	D	E	F	G	H
1	01	001	0001	00001	000001	0000001	0000000

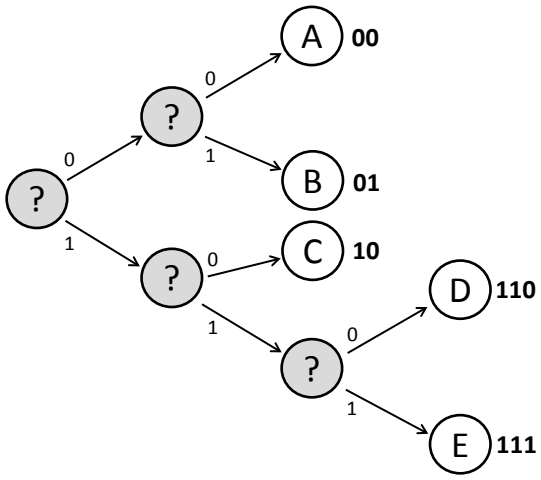
   It is uniquely decodable because each code corresponds to a unique letter rather than any possible combination of letters; and it has the prefix property because the code for no letter could be confused as the prefix for another letter.

6. Relate the bits in your prefix code to the 'yes/no' questions that you proposed in 1.

   Solution:

   The bit strings in the above prefix code for each letter can be interpreted as the history of answers to the 'yes/no' questions.

(*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.

Solutions:

(No codeword is the start of any other codeword, so they are instantaneously decodable: no "punctuation" is required to demarcate codewords within a string of bits.)

1. A probability distribution for the symbols that would generate such a Huffman tree is:

A	B	C	D	E
1/4	1/4	1/4	1/8	1/8

2. The entropy of this probability distribution is:

$$H = -\sum_i p_i \log_2 p_i = 3 \cdot (1/4) \cdot 2 + 2 \cdot (1/8) \cdot 3 = 2.25 \text{ bits.}$$

3. Average codeword length, weighting by each letter's probability, is $3 \cdot (1/4) \cdot 2 + 2 \cdot (1/8) \cdot 3 = 2.25$ bits also. This conforms with the Source Coding Theorem, which asserts that a code can be constructed whose average codeword length comes arbitrarily close to the entropy of the source (in the limit of a large enough alphabet).

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.

Solutions (similar to the different Exercise 3b!):

(*i*) The probability that the n^{th} flip is the first head, preceded by $n-1$ tails, is $\left(\frac{1}{2}\right)^{n-1} \times \frac{1}{2}$. Thus $P(X = n) = 2^{-n}$. Since the probability that the answer is n declines by one-half with each successive value of n, there is no more efficient series of questions to ask than: "Was it on the first flip? ...the second? ...the third? ... etc", because each of these questions in succession has a probability 0.5 of generating a "yes," and therefore each yes-no question extracts the maximum possible amount of information (i.e., one bit).

(*ii*) On average, the answer will be extracted after only two questions have been asked, because when we take the weighted average of each possible number n of necessary questions, weighted by the probability of having had to ask so many, we get the following series with its well-known convergence limit:

$$\lim_{N\to\infty} \sum_{n=1}^{N} n \left(\frac{1}{2}\right)^n = 2$$

(*iii*) The sequence of questions is equivalent to a variable-length prefix code for X, with each "yes" or "no" response corresponding to a 1 or 0. Although X can approach infinity, its average codeword length in bits (or average number of questions needed) is only two.

(*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, $H(X)$?

Solution:

(*i*) Entropy of the source, $H(X)$, is 1 bit.

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

Solution:

(*ii*) Output probabilities are: $p(y = 0) = (0.5)(1 - \epsilon) + (0.5)\epsilon = 0.5$ and likewise $p(y = 1) = (0.5)(1 - \epsilon) + (0.5)\epsilon = 0.5$. Entropy of this distribution is $H(Y) = 1$ bit, just as for the entropy $H(X)$ of the input distribution.

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

Solution:

(*iii*) Joint probability distribution $p(X, Y)$ is:

$$\begin{pmatrix} 0.5(1-\epsilon) & 0.5\epsilon \\ 0.5\epsilon & 0.5(1-\epsilon) \end{pmatrix}$$

and the entropy of this joint distribution is $H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$

$$= -(1-\epsilon)\log(0.5(1-\epsilon)) - \epsilon\log(0.5\epsilon) = (1-\epsilon) - (1-\epsilon)\log(1-\epsilon) + \epsilon - \epsilon\log(\epsilon)$$

$$= 1 - \epsilon\log(\epsilon) - (1-\epsilon)\log(1-\epsilon)\ .$$

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

Solution:

(*iv*) The mutual information is $I(X;Y) = H(X) + H(Y) - H(X,Y)$, which we can evaluate from the quantities above as: $1 + \epsilon\log(\epsilon) + (1-\epsilon)\log(1-\epsilon)$.

(*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?

Solution:

(*v*) In the <u>two</u> cases of $\epsilon = 0$ and $\epsilon = 1$ (perfect transmission, and perfectly erroneous transmission), the mutual information reaches its maximum of 1 bit and this is also then the channel capacity.

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

Solution:

(*vi*) If $\epsilon = 0.5$, the channel capacity is minimal and equal to 0.

(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 $p(Y = 1)$.

Solution:

(i) The output probabilities are: $p(0) = (0.5)(1 - \alpha + \beta)$, $p(1) = (0.5)(1 + \alpha - \beta)$.

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

Solution:

(ii) The capacity of this channel would be maximised if we have either $(\alpha, \beta) = (0, 0)$ or if $(\alpha, \beta) = (1, 1)$. In either of those cases, the mutual information between input and output reaches its maximum of 1 bit, which is therefore the channel capacity.

(iii) Define all paired values of (α, β) that would minimise the capacity of this channel, and state what that capacity would then be.

Solution:

(iii) If $(\alpha, \beta) = (0.5, 0.5)$ or $(0, 1)$ or $(1, 0)$, then the capacity of this channel is minimised, and it is equal to 0 bits. Note that in the case $(\alpha, \beta) = (0.5, 0.5)$ this is really a symmetric binary channel and we have the same answer as for Question $4(b)(vi)$. In the case $(\alpha, \beta) = (0, 1)$, this channel is stuck on always emitting a 0 regardless of the input. In the case $(\alpha, \beta) = (1, 0)$, this channel is stuck on always emitting a 1 regardless of the input. More generally, it can be shown that whenever $\alpha + \beta = 1$, the capacity of this asymmetric channel is minimised to 0 bits.