Week 5: Entropie et Compression (Solutions)

1 Calculation of Entropy

The frequencies of appearance of the letters are as following:

A	space	S, T, B	H, L, V, I , Y, !
5/20	3/20	2/20	1/20

The entropy is therefore:

$$H = \frac{5}{20}\log_2\left(\frac{20}{5}\right) + \frac{3}{20}\log_2\left(\frac{20}{3}\right) + \frac{6}{20}\log_2\left(\frac{20}{2}\right) + \frac{6}{20}\log_2(20) = \frac{17}{10} - \frac{3}{20}\log_2(3) + \frac{3}{4}\log_2(5) \quad \simeq 3.2037$$

(Hint: to "remove" the $\log_2(20)$, notice $\log_2(20) = \log_2(4 \times 5) = \log_2(4) + \log_2(5)$.)

2 The entropy of a chess game

a. Here is an example of an optimal strategy (there are several equivalent ones):

Q1: Is this an empty box? If yes, we know that the box is empty (with 1 question asked). If not, we continue:

Q2: Is the piece black? Q3: Is the piece a pawn?

With these 2 additional questions, if the answer to question 3 is yes, we know that the square contains a pawn (white or black), with thus 3 questions asked in all. If not, we continue:

Q4: Is the piece a knight or a bishop?

If yes: Q5: Is the piece a knight? With the answers to these last 2 questions, we can identify the piece (thus within 5 questions in all).

If not: Q5: Is the piece a tower? If yes, we can identify the square again with 5 questions in all. If not, we must ask one more question: Q6: Is the piece a queen? And we obtain in this case the answer in 6 questions.

In total, taking into account the probabilities of the appearance of each piece on the chessboard, we find that the average number of questions to ask is:

$$\frac{32}{64} \times 1 + \frac{16}{64} \times 3 + \frac{12}{64} \times 5 + \frac{4}{64} \times 6 = \frac{1}{2} + \frac{3}{4} + \frac{15}{16} + \frac{3}{8} = \frac{41}{16} = 2.5625$$

b. In the diagram on the left, the number and nature of the parts have not changed from the original situation, and so the entropy remains the same, in the same way, that the order of the letters in a sequence of letters does not influence the entropy of the sequence: only the probabilities of occurrence matter in the calculation of the entropy.

In the diagram on the right, a white piece has disappeared. One could do a long and complicated calculation here to recalculate the entropy of the game with the formula of the enumeration, but a simple and intuitive argument allows us to conclude that the entropy decreases: with one less pawn on the board, the number of different positions that can be represented is less than the one that can be represented at the beginning. If you are not convinced, think now of the situation of an endgame where there are only the two kings on the board and one pawn, for example. In this case, the entropy of the game is smaller than at the beginning, because the number of different positions that can be formed with these 3 elements on the board is clearly much smaller than with all the pieces.

3 Entropy Comparisons

a. The word with the highest entropy is shown below in **bold**:

1. EPFL and EEPPFFLL: The entropies are equal (and worth 2) because in both cases, the 4 letters E, P, F and L all appear with probability 1/4.

2. **MEDITERRANNEE** and MEDETERRENNEE: The entropy of the first word is clearly greater because the number of letters in each word is the same, but all vowels are replaced by E's in the second word.

3. AAAH and **HAHA**: The entropy of the first word is $(3/4)\log_2(4/3) + (1/4)\log_2(4) \simeq 0.81$, while that of the second word is 1. We can also see that there are fewer possibilities to form different words with the 4 letters AAAH than with the 4 letters HAHA, so less entropy in the first word.

4. ABB and **ABBA**: The entropy of the first word is $(2/3)\log_2(3/2) + (1/3)\log_2(3) \simeq 0.91$, while that of the second is 1.

5. **ACD** and ACDC: The entropy of the first word is $\log_2(3) \simeq 1.58$, while that of the second is 1.5.

6. ABR and **ABRI**: The entropy of the first word is $\log_2(3)$, while that of the second is $\log_2(4)$.

7. CALC and **CALCUL**: The entropy of the first word is 1.5, while that of the second is $(2/3)\log_2(3) + (1/3)\log_2(6) \simeq 1.92$.

b. By simply counting the number of appearances of the letters, the following order for the entropies can easily be derived:

$$1. < 2. = 4. < 3.$$

The place of the 5^{th} sequence is more difficult to determine. By calculating the entropies explicitly (with the help of a calculating machine or of `http://www.shannonentropy.netmark.pl/`), we find that

$$1. < 2. = 4. < 5. < 3.$$

Alternatively, to compare the entropies of Sequences 3 and 5, we can focus on the entropy computation for the differences. The numbers of appearance of the letters are (in decreasing order):

in Sequence 3: 3 2 2 1 1 1 1 1 1 1 1 1

in Sequence 5: 4 2 1 1 1 1 1 1 1 1 1 1

The only difference between these two sequences is that in Sequence 3, we have 2 letters (the 1st and the 3rd), which appear respectively 3 and 2 times, while in Sequence 5, we have two letters which appear respectively 4 and 1 times. Given the definition of entropy, we need to compare only the two expressions (1) and (2) below. (Note that $1.5 = \frac{3}{2} < \log_2 3 = 1.5849 < \frac{5}{3} = 1.\dot{6}$.)

$$\text{(1) Part of Entropy of Seq. 3: } \frac{3}{16}\cdot\log_2\frac{16}{3} + \frac{2}{16}\cdot\log_2\frac{16}{2} = \frac{3}{16}\cdot\log_2(16) - \frac{3}{16}\cdot\log_2(3) + 0.375 =$$

$$= 1.125 - \frac{3}{16}\cdot\log_2(3) > 1.125 - \frac{3}{16}\cdot\frac{5}{3} = 0.8125$$

$$\text{(2) Part of Entropy of Seq. 5: } \frac{4}{16}\cdot\log_2\frac{16}{4} + \frac{1}{16}\cdot\log_2\frac{16}{1} = 0.5 + 0.25 = 0.75$$

Since expression (1) is larger than expression (2), the entropy of sequence 3 is larger than the entropy of sequence 5.

4 Does entropy increase or decrease?

a. The answer is no. It may be that entropy increases (for example, with ABB and ABBA) or decreases (for example, with ACD and ACDC).

b. Here the answer is yes (for example, with ABR and ABRI). The formal proof of this fact is beyond the scope of this course, but a simple intuition is as follows: adding a letter that is not part of the sequence brings something new to it and thus increases the number of possibilities of forming different words, which implies an increase in entropy.

5 Minimize the number of weighings

a. Observe first the similarity between this problem and the question game seen in class. Here, instead of an "oracle" that gives us "yes" or "no" answers, we now have a scale that tells us "tilt left", "tilt right", or "remain steady". Each weighing, therefore, allows us to divide the set of possibilities by 3 (and not by 2). It is a question of carrying out the right weighing to be the most effective possible.

Let's number the 9 pieces by letters: ABCDEFGHI. One of them is lighter: so there are 9 possibilities in all. To divide the set of possibilities by 3, we perform the following weighing:

ABC-DEF

(meaning that we place ABC on the left and DEF on the right of the scale, with the other pieces GHI remaining on the table).

- If the scale tilts to the left (which is noted as ABC>DEF: the weight of ABC is greater than that of DEF), then we know that the lighter piece is in DEF.

- If the scale tilts to the right (which we note ABC<DEF), then we know that the lighter piece is in ABC.

- If the balance remains stable (which we note ABC=DEF), then we know that the lighter piece is in GHI.

Thus, we have reduced the set of possibilities by 3. It is then enough to repeat the operation with the three remaining parts XYZ. More precisely, we carry out the weighing:

X-Y

If X>Y, the lighter piece is Y; if X<Y, the lighter piece is X; if X=Y, then necessarily, Z is the lighter piece, all the others having the same weight.

In conclusion, 2 weighings are enough to find the lighter piece. Note that $2 = \log_3(9)$. This is the definition of ternary entropy (whereas in the course and the previous exercises we were dealing with binary entropy).

b. For this problem, let's name the 4 pieces by letters: ABCD and let's call X the piece that we have in our pocket and that we know has the right weight. We have again 9 possibilities:

- either one of the 4 pieces is lighter than the others;

- or one of the 4 pieces is heavier than the others;

- or all the pieces have the same weight.

To divide this set of 9 possibilities into 3 equal parts, here is the first weight to be performed:

AB-CX

(with D remaining on the table).

1) If AB=CX, then we are left with 3 possibilities: either Piece D is heavier, or it is lighter, or all the pieces have the same weight. We, therefore, carry out a second weighing of D-X which gives us the answer.

2) If AB<CX, then there are also 3 possibilities: either one of the pieces A or B is lighter or Piece C is heavier. We then carry out the A-B weighing: if A=B, it is C which is heavier; if A<B, it is A which is lighter; if A>B, it is B which is lighter.

3) If AB>CX, then there are also 3 possibilities: either one of the pieces A or B is heavier, or Piece C is lighter. We then carry out the A-B weighing: if A=B, it is C which is heavier; if A<B, it is B which is heavier; if A>B, it is A which is heavier.

Thus, in total, 2 weights are enough to answer the questions of the statement. Again, it is no accident that $2 = \log_3(9)$.

6 Shannon-Fano algorithm

a. Using the question game, here is the code that we get (when impossible to have the same number of appearances on the left and on the right, having fewer on the left side):

Letter	#Appearances	#Questions	Code word
I	4	3	111
N	4	3	110
O	4	3	101
C	4	3	100
M	3	3	011
A	3	4	0101
T	3	4	0100
L	2	4	0011
U	2	4	0010
F	1	4	0001
R	1	5	00001
E	1	5	00000

b. In total, we therefore need $19 \times 3 + 11 \times 4 + 2 \times 5 = 111$ bits, i.e. $L(C) = \frac{111}{32} = 3.468975$ bits per letter.

c. The entropy of the sequence is equal to $H(X) = 3.42923$. So we see that $H(X) \leq L(C) \leq H(X) + 1$, we see in fact that $L(C)$ is much closer to $H(X)$ than to $H(X) + 1$.

d. The sequence contains 12 different letters: if we want to use the same number of bits per letter, we need at least 4 bits per letter (because $2^4 = 16$, which is the closest power of 2 above 12), and so $32 \times 4 = 128$ bits in total. So we use 17 bits more than with the Shannon-Fano code.

e. The sequence starting with DIDON... has a very large number of D's (11 to be more precise): we, therefore, expect a lower entropy.

f. Using the question game again, here is the code we get:

g. The total number of bits used is this time $17 \times 2 + 12 \times 3 + 1 \times 4 + 2 \times 5 = 84$ bits, i.e. $L(C) = \frac{84}{32} = 2.625$ bits per letter. The entropy of the sequence is $H(X) = 2.56475$. Again, we can see that $H(X) \leq L(C) \leq H(X)+1$, and that $L(C)$ is closer to $H(X)$ than to $H(X)+1$, even if the probabilities of the letters are quite irregular in the present case.

Letter	#Appearances	#Questions	Code word
D	11	2	11
N	6	2	10
O	5	3	011
I	4	3	010
U	3	3	001
A	1	4	0001
T	1	5	00001
S	1	5	00000

h. The sequence contains 8 different letters: if we want to use the same number of bits per letter, we need at least 3 bits per letter (because $2^3 = 8$) and therefore $32 \times 3 = 96$ bits in total. We, therefore, use 12 bits more than with the Shannon-Fano code.