Exercise Sheet 4

for Lecture *Quantum-Information, -Computing, and -Sensing*

Stefan Scheel, Friedemann Reinhard, Clemens Cap, Arman Kashef

5th June 2024

Exercises have to be handed in before the exercise session, which takes place on June 11, 11:15 in the office of Clemens Cap, Zuse Haus, Albert Einstein Strasse 22, Room 355

1. Coding and Compression

Goal: We want to understand the effect of recodings of an information source. In this task, we will also employ something which is known as HUFFMAN. In part 2 of the lecture we shall meet this example again as an application of the Shannon source coding theorem.

An information source $A = \{a, b, c, d\}$, $p\colon (a, b, c, d) \mapsto (0.1, 0.2, 0.3, 0.4)$ shall be recoded over the binary alphabet $\{\texttt{0}, \texttt{1}\}$. In this recoding, more than one symbol of the binary alphabet may be used. The straight-forward encoding would read

a	`00`
b	`01`
c	`10`
d	`11`

Table 1: Encoding for our information source.

Task: Find an enocding, which is more efficient in terms of the average code word length!

Hint: Use more binary digits for the less frequent symbols and less binary digits for the more frequent symbols.

2. Entropy and German Language

Goal: We want to understand how modeling assumptions help us approximate the 'true' entropy found in concrete objects, such as the German language. This provides us with a better 'feeling', how external constraints can affect entropy.

The German language consists of words over the alphabet $\mathcal{A} := \{A, B, \ldots, Z\}$. The probabilities of the individual letters are given as in the table at `https://de.wikipedia.org/wiki/Buchstabenh%C3%A4ufigkeit` (We will use the upper table in the reference, not employing Umlauts or spaces, but recognizing ß).

Task 1: Determine the entropy of a single letter of German text.

Task 2: How far is this entropy below the entropy which we would get in an optimal information source with 27 characters.

Task 3: Determine the entropy of a six-letter word, assuming independence of the letter distributions at the individual positions and using the letter probabilities as given above.

Note: The average length of a German word in the corpus of the Duden is 5.99 letters. This calculation also respects the more and the less frequent use of words in a corpus `https://www.duden.de/sprachwissen/sprachratgeber/Durchschnittliche-Lange-eines-deutschen-Wortes`.

Task 4: Determine the entropy of a single word of the German language assuming that the German language has a total of 500.000 words and furthermore assuming equidistribution of all the words. `https://www.statistik-bw.de/Service/Veroeff/Monatshefte/20100911`

Task 5: Now determine the entropy of a single word of the German language using ZIPFs law. This law suggests that in a human language with n words the relative frequencies of the words follow a harmonic series

$$1/1 \quad 1/2 \quad 1/3 \quad 1/4 \ldots$$

In this sense the second most frequent word occurs half so often than the most frequent word. The third most frequent word occurs one third so often than the most frequent word.

3. A Mysterious Big Black Box

Goal: One task of information modeling is to understand how we can reduce the observed behavior of a complex object to more simple sub-systems. This exercise refers to the products and compounds which we studied in the lecture.

You have a big black box with which you can do experiments. You realize that in every round of the experiment you get one of four outcomes. For simplicity, we denote these outcomes with the letters of the set $A = \{a, b, c, d\}$. You have no idea

about the inner workings of the black box, so you do long phases of observation. The observations suggest that the individual outcomes of the experiment are independent from each other: There is no obvious correlation between an earlier experiment and a later experiment and the black box does not seem to have an inner state or memory. It looks like the machine performs a (fresh) random choice in every run. You determine the probabilities of the outcome seem to be as follows: $p\colon (a, b, c, d) \mapsto (1, 2, 4, 8)/15$.

Question: Is it reasonable to assume that the black box consists of two subsystems? If yes, provide such a partitioning into subsystems.

Hint: Technically spoken: Can you find a representation of the output set A as a product $R \times S$ of two sets R and S and probabilities $r\colon R \to [0, 1]$ on R and $s\colon S \to [0, 1]$ on S, which allow you to model the big black box as product of (R, r) and (S, s).