Information Theory **DCC/ICEx/UFMG**
Prof. Mário S. Alvim **2017/02**

SOLUTION OF HOMEWORK

Symbol Codes
(MacKay - Chapter 5)

Necessary reading for this assignment:

- *Information Theory, Inference, and Learning Algorithms* (MacKay):
  - Chapter 5.1: *Symbol codes*
  - Chapter 5.2: *What limit is imposed by unique decodeability?*
  - Chapter 5.3: *What's the most compression we can hope for?*
  - Chapter 5.4: *How much can we compress?*
  - Chapter 5.5: *Optimal source coding with symbol codes: Huffman coding*
  - Chapter 5.6: *Disadvantages of the Huffman code*
  - Chapter 5.7: *Summary*

Note: The exercises are labeled according to their level of difficulty: `[Easy]`, `[Medium]` or `[Hard]`. This labeling, however, is subjective: different people may disagree on the perceived level of difficulty of any given exercise. Don't be discouraged when facing a hard exercise, you may find a solution that is simpler than the one the instructor had in mind!

Review questions.

1. Answer formally the following questions:

   (a) What is a symbol code for an ensemble? What is an extended code for an ensemble?

   Instructor's solution: A symbol code for an ensemble X is a function $C : \mathcal{A}_X \rightarrow \mathcal{D}^+$ mapping each symbol $a_i \in \mathcal{A}_X$ of the ensemble to a string on the encoding alphabet $\mathcal{D}$.
   If the encoding alphabet $\mathcal{D} = \{0, 1\}$ has only two elements, we have a binary symbol code.
   An extended code (with respect to a symbol code C) for an ensemble X is a function $C^+ : \mathcal{A}_X^+ \rightarrow \mathcal{D}^+$ mapping each string $x \in \mathcal{A}_X^+$ formed by symbols in the ensemble to a string formed on the encoding alphabet $\mathcal{D}$, such that $c^+(x_1, x_2, \ldots, x_N) = c(x_1)c(x_2)\cdots c(x_N)$.

   (b) When is a symbol code uniquely decodable? When is a symbol code prefix-free?

   Instructor's solution: A symbol code is uniquely decodable iff no two distinct strings formed by symbols in the ensemble are encoded into the same codeword. Formally, a code C is uniquely decodeable iff for all $x, y \in \mathcal{A}_X^+$ we have that $x \neq y \implies c^+(x) \neq c^+(y)$.
   A symbol code is prefix-free if no codeword is a prefix of any other codeword.

   (c) State Kraft's inequality and explain in what sense it is related to the notion of which prefix-free codes are actually possible.

Instructor's solution: Kraft's inequality states that for any instantaneous prefix-free code $C(X)$ over an alphabet having D symbols, the codelengths must satisfy $\sum_{i=1}^{I} D^{-l_i} \leq 1$, where $I = |\mathcal{A}_X|$.

Conversely, if a choice of codeword lengths satisfies the Kraft inequality, it is possible to construct a prefix-code with this choice of codeword lengths.

The importance of Kraft's inequality is the following: (i) if a code does not satisfy the Kraft inequality, it can't be prefix-free; and (ii) if a code satisfies the Kraft inequality, it can always be converted into a prefix-free code while maintaining the choice of codeword lengths.

(d) Explain what the Source coding theorem for symbol codes means in terms of the limits of compression of an ensemble.

Instructor's solution: The Source coding theorem for symbol codes states that for any ensemble X there exists a prefix code C with expected length satisfying $H(X) \leq L(C, X) \leq H(X) + 1$. The average length is equal to the entropy $H(X)$ only if the codelength for each outcome is equal to its Shannon information content.

The theorem gives a lower bound on the efficiency of prefix-codes for symbols. It means that no code that encodes individual symbols of a source can produce codewords with fewer than $H(X)$ bits per symbol in an average. Moreover, the theorem means that it is always possible to find a code that produces codewords with at most $H(X) + 1$ bits per symbol in an average.

Exercises.

2. (MacKay 5.19) `[Easy]`

Instructor's solution: The code $\{00, 11, 0101, 111, 1010, 100100, 0110\}$ is not uniquely decodable.

For instance, the string 111111 can be parsed as three consecutive uses of the codeword 11 or as two consecutive uses of the codeword 111.

3. (MacKay 5.20) `[Easy]`

Instructor's solution: The ternary code $\{00, 012, 0110, 0112, 100, 201, 212, 22\}$ is uniquely decodeable because it is prefix-free, as shown in Figure 1.

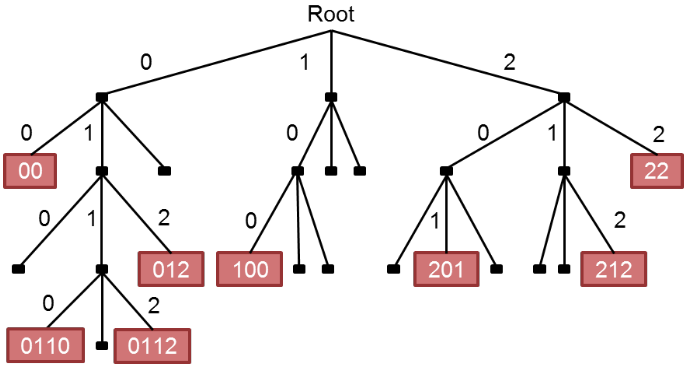

Figura 1: Exercise 5.20

4. (MacKay 5.21) `[Medium]` You only need to do:

(a) X^2 and X^3 when $\mathcal{P}_X = \{0.9, 0.1\}$; and

(b) X^2 when $\mathcal{P}_X = \{0.6, 0.4\}$.

Instructor's solution: Given in the textbook.

5. (MacKay 5.22) `[Medium]`

   Instructor's solution: Given in the textbook.

6. (MacKay 5.24) `[Easy]`

   Instructor's solution: If you wanna beat me on the 20-questions game, you better figure this one out yourself! ;-)