«Calcul formel avancé et application». Very brief lecture notes.

21.09.2023. Lecture 2.

1. Discussion of the homework.

The exercise about an optimal sorting for $k = 4$ and $k = 5$ objects remains unsolved.

2. The game guess a number revisited. We discussed the version of the game "guess a number," where the first player choses at random an integer number between 1 and k with (some fixed and known in advance) probabilities $p_1, \ldots, p_k$, and the second player should reveal this number by asking questions with answers *yes* or *no*, with the minimal *on average* number of questions.

Proposition 1. *For every random variable α distributed on a set of n values*

$$0 \le H(\alpha) \le \log n.$$

Moreover, $H(\alpha) = 0$ if and only if the distribution is concentrated at one point (one probability p_i is equal to 1, and the other p_j for $j \neq i$ are equal to 0), and $H(\alpha) = \log n$ if and only if the distribution is uniform ($p_1 = \ldots = p_n = \frac{1}{n}$).

Sketch of proof : We use the concavity of the function $\log x$ and Jensen's inequality for the concave functions.

Proposition 2. *For every random variable α and for every (deterministic) function F, Shannon's entropy of the random variable $\beta = F(\alpha)$ is not greater than Shannon's entropy of α.*

Sketch of proof : First of all, we observed that $H(\alpha) = H(\beta)$, if F is a bijection. Then, we proved that the entropy of a distribution decreases, when we merge together two points in this distribution ; in other words, $H(\alpha) \ge H(F(\alpha))$, if F merges together two points from the range of α and leaves distinct the other values of α. By iterating the basic "merging" operations, we prove the inequality $H(\alpha) \ge H(F(\alpha))$ for an arbitrary function F.

Given a pair of jointly distributed random variables (α, β) we can apply the definition of Shannon's entropy three times, with three protentially different distributions : we have Shannon's entropy of the entire distribution (denoted $H(\alpha, \beta)$) and the entropies of two marginals, $H(\alpha)$ and $H(\beta)$.

We have proved earlier that

Proposition 1. *In the game "guess a number," where the first player choses at random an integer number between 1 and k with (known in advance) probabilities $p_1, \ldots, p_k$, the average number of questions cannot be less than*

$$\sum_{i=1}^{k} p_i \log \frac{1}{p_i}$$

Now we proved an upper bound for the same game :

Proposition 2. *For the game "guess a number," where the first player choses at random an integer number between 1 and k with (known in advance) probabilities $p_1, \ldots, p_k$, there exists a strategy that requires on average less than*

$$\sum_{i=1}^{k} p_i \log \frac{1}{p_i} + 1$$

questions.

Sketch of the proof : W.l.o.g. we assume that $p_1 \ge p_2 \ge \ldots \ge p_n$. We define $\ell_i = \lceil \log_2 \frac{1}{p_i} \rceil$. Observe that $\sum 2^{-\ell_i} \le 1$. Then, we construct a binary tree with n leaves and branches of length $\ell_1, \ldots, \ell_n$.

On the first stage we choose the leftmost branch of length ℓ_1, then we choose the leftmost branch of ℓ_2 that is incompatible with the first branch, and so on. On the k-th step we choose the leftmost a branch of length ℓ_i that is not a continuation of any branch chosen on the stages $1, \ldots, (k-1)$. We show that this procedure can be repeated until stage n due to two key facts :

— the sum $\sum 2^{-\ell_i}$ is nit greater than 1,

— $\ell_1 \le \ell_2 \le \ldots \le \ell_n$.

End of proof.

We observed that strategies in the guessing number game are equivalent to prefix-free binary codes. Thus, we have shown that for every probability distribution $(p_1, \ldots, p_n)$ the minimal average length of a binary code $\sum p_i |c_i|$ is a number between $\sum\limits_{i=1}^{k} p_i \log \frac{1}{p_i}$ and $\sum\limits_{i=1}^{k} p_i \log \frac{1}{p_i} + 1$.

3. Huffman's encoding. We discussed the construction of Huffman's code and proved its optimality. For a detailed explanation see the textbook *Elements of information theory* by T. M. Cover and J. A. Thomas.

Exercise 2.1. Construct Huffman's code for the distribution of probabilities $(0.33, 0.34, 0.2, 0.1, 0.05)$ and find the average length of the codewords for this code.

4. Block coding. We discussed the problem of optimal compression for texts of length N over an alphabet $\{a_1, \ldots, a_k\}$ with known frequencies of letters $(p_1, \ldots, p_k)$. Using a counting (based on Stirling's formula) we showed that we need

$$\left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \right) \cdot N + o(N)$$

binary digits.