Course «Information theory». Very brief lecture notes.

08.09.2020. Lecture 1.

1. The game guess a number : one players chooses an integer number between 0 and $n-1$, another player should find this number by asking questions with answers *yes* or *no*. There is a simple strategy that allows to find the chosen number in $\lceil \log n \rceil$ questions [1] (bisection). Moreover, there exists even a non-adaptive strategy with the same number of questions (the second player asks bits of the binary expansion of the chosen number).

These strategies are optimal : no strategy helps to reveal the chosen number in less than $\lceil \log n \rceil$ questions (in the worst case). The lower bound can be explained in terms of the *missing information* : every valid strategy for this game can be represented as a binary that contain at least n leaves (one leaf for every possible answer). Therefore, the height of this tree must be at least log of the number of possible answers.

2. Sorting algorithms. We are given n objects ("stones") and balance scales ; in one operation we can compare weights of two stones.
Claim. To sort n stones by their weights, we have to do in the worst case $\log(n!)$ pairwise comparisons (no algorithm can guarantee the right answer with less than $\log(n!)$ weighings). This is because every strategy (algorithm) can be represented by a binary tree with final answers written in the leaves, and there are $n!$ possible answers (different orderings of n elements). Thus, the depth of this tree cannot be less than $\log(n!)$.

There exist sorting algorithms that do the job with $O(n \log n)$ comparisons. Thus,

$$\log(n!) \leq [\text{the optimal number of weighings}] \leq O(n \log n).$$

We observed that the difference between the lower bound $\log(n!)$ and the upper bound $O(n \log n)$ is only a constant factor (see also the homework).

3. Fake coins problems. *(i) A light fake coin.* We are given $n = 25$ coins. One of them is fake, which is lighter than all other (identical) coins. We can use balance scales to compare weights of any two *groups* of coins. How many operations do we need to find the fake coin ?

In this case, every weighing strategy can be represented by a ternary rooted tree, and the maximal number of operations is the height of this tree. We proved that for $n = 25$ coins an optimal strategy requires 3 weighings. In general, to find a light fake coin in a heap of n coins we need $\lceil \log_3 n \rceil$ weighings.

(ii) A fake coin with unknown weight. We are given $n = 12$ coins. One of them is fake, and can be lighter or heavier than all other (identical) coins. Again, we can use balance scales to compare weights of any two groups of coins. We are not obliged to find out the relative weight of the fake coin (we ignore whether it is lighter or heavier than a genuine one). How many operations do we need to find the fake coin ?

1. In these notes, $\log n$ stands for the binary logfraithm $\log_2 n$.

We proved that an optimal strategy requires 3 weighings. For the same question with $n = 15$ we proved that 3 weighings are not enough.

4. Combinatorial definition of the information quantity. Following Ralph Hartley, we say that the *quantity of information* in a *finite set* X is defined as

$$\mathrm{Inf}(X) := \log |X|$$

(roughly speaking, $\mathrm{Inf}(X)$ is the number of *binary digits* needed to give a unique name to each element in X). For every finite set X, the value $\mathrm{Inf}(X)$ is a non-negative real number.

1. For a set $X \subset \mathbb{Z}^2$ we can define the "quantity of information" in X as well as in its projections on the first and the second coordinates (denoted $\pi_1[X]$ and $\pi_2[X]$ respectively). We have the following simple property :

$$\mathrm{Inf}(X) \leq \mathrm{Inf}(\pi_1[X]) + \mathrm{Inf}(\pi_2[X]),$$

the quantity of information in X is not greater than the sum of the information quantities in two its components. This statement is equivalent to the obvious inequality $X \subset \pi_1[X] \times \pi_2[X]$, and, therefore,

$$|X| \leq |\pi_1[X]| \cdot |\pi_2[X]|.$$

2. Similarly, for a finite set $X \subset \mathbb{Z}^3$ we have

$$\mathrm{Inf}(X) \leq \mathrm{Inf}(\pi_1[X]) + \mathrm{Inf}(\pi_2[X]) + \mathrm{Inf}(\pi_3[X]),$$

which corresponds to the trivial fact that for a 3-dimensional set X

$$X \subset \pi_1[X] \times \pi_2[X] \times \pi_3[X].$$

3. One more (much less trivial!) property is known to be true for Hartley's information : for a finite set $X \subset \mathbb{Z}^3$ we have

$$2 \cdot \mathrm{Inf}(X) \leq \mathrm{Inf}(\pi_{12}[X]) + \mathrm{Inf}(\pi_{23}[X]) + \mathrm{Inf}(\pi_{13}[X]),$$

where π_{ij} denotes the projection on the coordinates (i, j). For example, for a point $x = (a, b, c)$ we have $\pi_{13}(x) = (a, c)$. We will prove this inequality later.

15.09.2020. Lecture 2.

1. Discussion of the homework : We discussed an algorithm for sorting an array of n elements and proved by induction that it runs time $O(n \log n)$ in the worst case. We found optimal sorting algorithms for $n = 2, 3, 4$ elements. We proved a fake coin in heap of 14 coins can be found in at most 4 weighings (we do not known in advance whether the fake coin is lighter or heavier that the other ones).

2. The game guess a number. We discussed another version of the game "guess a number," where the first player choses at random an integer number between 1 and k with (known in advance) probabilities $p_1, \ldots, p_k$, and the second player should reveal this number by asking questions with answers *yes* or *no*, with the minimal *on average* number of questions. We discussed several specific examples and suggested a general scheme of "modified dichotomy." In this strategy, on each step the second player divides all numbers (that have not been excluded earlier) into two groups with balanced measures, i.e., the sums of probabilities in both groups must be as close to each other as possible. We discovered a plausible approximation : the average number of steps in this strategy is close to

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

though we did not prove it formally.

3. Shannon's entropy. For a random variable α with n possible values $a_1, \ldots, a_n$ such that $\mathrm{Prob}[\alpha = a_i] = p_i$, we define its Shannon's entropy as

$$H(\alpha) := \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$

(with the usual convention $0 \cdot \log \frac{1}{0} = 0$). We proved several properties of Shannon's entropy.

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, $H(\alpha) \geq H(F(\alpha))$, 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 $H(\alpha) \geq H(F(\alpha))$ 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 $H(\beta)$.

Proposition 3. *For every pair of jointly distributed random variables α and β*

$$H(\alpha, \beta) \leq H(\alpha) + H(\beta).$$

Sketch of proof : We used again the concavity of the function of logarithm and Jensen's inequality.

Proposition 4. *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*

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

Sketch of proof : Let have a strategy that permits to guess a number by asking questions with answers yes and no. We represent this strategy by a binary tree. Denote l_i the length of the branch from the root of this tree to the leaf marked i (i.e., the number of questions that we ask before we get the answer i). Observe that

$$1/2^{l_1} + \ldots + 1/2^{l_k} = 1$$

(this is true for every binary tree). We need to prove that

$$\sum_{i=1}^{k} p_i l_i \geq \sum_{i=1}^{n} p_i \log \frac{1}{p_i}.$$

This is inequality rewrites to the form

$$\sum_{i=1}^{k} p_i \log \left(\frac{2^{-l_i}}{p_i} \right) \leq 0.$$

And the last one easily follows for Jensen's inequality :

$$\sum_{i=1}^{k} p_i \log \left(\frac{2^{-l_i}}{p_i} \right) \leq \log \left(\sum_{i=1}^{k} p_i \cdot \frac{2^{-l_i}}{p_i} \right) = \log \left(\sum_{i=1}^{k} 2^{-l_i} \right) = \log 1 = 0.$$

22.09.2020. Lecture 3.

1. Discussion of the homework : We used Shannon's entropy to prove that we cannot find a fake coin among $n = 14$ coins (the fake coin can be heavier or lighter than the other ones) in less than 4 weighings.

2. Discussion of the homework : We proved that

$$H(\alpha, \beta) = H(\alpha) + H(\beta)$$

if and only if α and β are independent.

3. Discussion of the homework : The upper bound for the average number of questions in the game "guess a number". We use Shannon's entropy to estimate the average number of questions needed in the randomized version of the game "guess a number" (with a probability distribution on the set of possible integers).

Lemma 3.1. For integer numbers $l_1, \ldots, l_k$ such that

$$\sum_{i=1}^{k} \frac{1}{2^{l_i}} \leq 1$$

there exists a binary tree with k leaves such that the length of the path from the root to the i-th leaf is equal to l_i.

Theorem 3.2. In the game "guess a number" (with *yes or no* questions) where the number $i = 1, \ldots, k$ is chosen with probabilities $p_1, \ldots, p_k$, there exists a strategy that uses on average less than $\left(\sum\limits_{i=1}^{k} p_i \log \frac{1}{p_i} \right) + 1$ questions.

Sketch of proof : We define $l_i := \lceil \log \frac{1}{p_i} \rceil$, notice that $\sum 2^{-l_i} \leq 1$, and use Lemma 3.1 to construct a strategy where each i-th leaf is on the distance l_i form the root.

3. Kraft's inequality. A *prefix code* is a set of strings $\{c_1, \ldots, c_k\}$ where no codeword c_i is a prefix of any other code word c_j in this set. A *uniquely decodable code* is a set of strings $\{c_1, \ldots, c_k\}$ such that for every string x there exists at most one representation

$$x = c_{i_1} \circ c_{i_2} \circ \ldots \circ c_{i_n}$$

(where $\circ$ denotes concatenation). Every prefix code is uniquely decodable, but not vice-versa. We discussed the correspondence between strategies for the game *guess a number* and prefix codes.

In what follows we assume that all codes contain only binary codewords (words in the alphabet of two letters).

Lemma 3.3. [known as Kraft's inequality] For every uniquely decodable code $\{c_1, \ldots, c_k\}$ we have

$$\sum_{i=1}^{k} 2^{-|c_i|} \leq 1.$$

Lemma 3.4. [known as Kraft's inequality] For every list of integers $l_1, \ldots, l_k$ such that

$$\sum_{i=1}^{k} 2^{-l_i} \leq 1$$

there exists a prefix code $\{c_1, \ldots, c_k\}$ such that $|c_i| = l_i$ for each i.

Theorem 3.5. For every *uniquely decodable* code $\{c_1, \ldots, c_k\}$ there exists a *prefix code* $\{d_1, \ldots, d_k\}$ with the same lengths of the codewords, i.e., $|d_i| = |c_i|$ for each i.

29.09.2020. Lecture 4.

1. Huffman's coding. Let us have an alphabet $X = \{x_1, \ldots, x_k\}$ with a probability distribution $(p_1, \ldots, p_k)$. We are looking for a uniquely decodable code $\{c_1, \ldots, c_k\}$, that minimizes the average length of the codeword

$$p_1|c_1| + \ldots p_k|c_k| \to \min$$

In the class we discussed the recursive algorithm of Huffman that allows to construct for any given distribution of probabilities an optimal prefix code. The base of this recursive construction is trivial : if $k = 2$, then the two codewords are $c_1 = 0$ and $c_2 = 1$. The inductive step uses the idea that we can "merge" to minimal probabilities in the distribution and reduce the problem on a k-element distribution $(p_1, \ldots, p_k)$ to a similar problem on a $(k-1)$-element distribution $(p_1, \ldots, p_{k-2}, q = p_{k+1} + p_k)$.

The proof of optimality of Huffman's code was based on the following lemmas :

Lemma 1. If $p_1 \geq \ldots \geq p_k$, then in *every* optimal code for this distribution $|c_1| \leq \ldots \leq |c_k|$.

Lemma 2. In *every* optimal code there is no codeword that is strictly longer than all other codewords. In particular, if $|c_1| \leq \ldots \leq |c_k|$, then $|c_k| = |c_{k-1}|$.

Lemma 3. If $p_1 \geq \ldots \geq p_k$, then in *some* optimal code for this distribution the codewords c_k and c_k differ in only the last letter (these codewords correspond to a pair of "brothers" vertices in the binary tree). If a code satisfies this property, we call it *regular*.

Lemma 4. Let $p_1 \geq \ldots \geq p_k$ and let $c_1, \ldots, c_k$ be a regular code for this distribution. Denote by d the common prefix of c_k and c_{k-1} (in a regular code these codewords differ only in the last position). Then the code $c_1, c_2, \ldots, c_{k-1}, c_k$ is optimal for the distribution $(p_1, \ldots, p_k)$ if and only if the code $c_1, \ldots, c_{k-2}, d$ is optimal for the reduced distribution of probabilities $(p_1, \ldots, p_{k-2}, q = p_{k-1} + p_k)$.

2. Stirling's formula. It is known

$$N! = \sqrt{2\pi N} \left(\frac{N}{e}\right)^N \cdot (1 + o(1))$$

as $N \to \infty$. This formula is called Stirling's approximation. In the class we proved a weaker version of this theorem. We showed that for some constants

$c_1, c_2 > 0$ and for all natural N

$$c_1\sqrt{N}\left(\frac{N}{e}\right)^N \leq N! \leq c_2\sqrt{N}\left(\frac{N}{e}\right)^N$$

It follows, in particular, that $\log(N!) = N \cdot \log(N/e) + O(\log n)$.

The idea of the proof : we estimated the difference between the discrete sum $\ln(N!) = \ln 1 + \ln 2 + \ldots + \ln N$ and the integral $\int_1^N \ln x dx$.

3. Block coding : we discussed Exercise 1 from the previous homework.

4. Block coding for typical sequences. Let $(p_1, \ldots, p_k)$ be a probability distribution (each p_i is non negative and the sum of all p_i is equal to 1). We say that a word $x \in \{x_1, \ldots, x_k\}^*$ is *typical* is the frequencies of letters $x_1, \ldots, x_k$ in this word are exactly the numbers $p_1, \ldots, p_k$ respectively.

Theorem. For every $h > \sum p_i \log \frac{1}{p_i}$ and for all large enough n there is an injective mapping F_n that assigns to each typical n-letter word $x \in \{x_1, \ldots, x_k\}^n$ a string of bits in $\{0,1\}^{\lceil h \cdot n \rceil}$.

Sketch of the proof : We count the number of all typical words of length n. It is equal to

$$\frac{n!}{(p_1 n)! \cdot \ldots \cdot (p_k n)!}$$

To encode all these words, we need

$$\log \frac{n!}{(p_1 n)! \cdot \ldots \cdot (p_k n)!} = \left(\sum p_i \log \frac{1}{p_i}\right) \cdot n + O(\log n)$$

binary digits (the computation uses Stirling's approximation for the factorials). It remains to notice that $\left(\sum p_i \log \frac{1}{p_i}\right) \cdot n + O(\log n) < h \cdot n$ for all large enough integer numbers n.

06.10.2020. Lecture 5.

1. Mathematical expectation and variance of random variables.
Let α be random variables distribute on $\mathbb{R}$. In what follows we assume that the distribution is concentrated on a finite set of real numbers. Let $p_i = \text{Prob}[\alpha = x_i]$ for $i = 1, \dots, k$, and $\sum\limits_{i=1}^{k} p_i = 1$.

Definition. Expectation of α is defined as

$$E(\alpha) := \sum_{i=1}^{k} p_i x_i.$$

Simple properties of the expectation :

- $E(\alpha + c) = E(\alpha) + c$ for every constant c;
- $E(c \cdot \alpha) = c \cdot E(\alpha)$ for every constant c;
- $E(\alpha + \beta) = E(\alpha) + E(\beta)$ for every pair of jointly distributed α and β;
- $E(\alpha \cdot \beta) = E(\alpha) \cdot E(\beta)$ for all *independent* α and β;

Remark. The equality $E(\alpha \cdot \beta) = E(\alpha) \cdot E(\beta)$ is false for some correlated α, β. But, of course, there are examples of dependent pairs (α, β) such that the expectation of the product is still equal to the product of the expectations. Here is a simple example : the pairs (α, β) is uniformly distributed on the set of pairs $(1, 1)$, $(1, -1)$, $(-1, 1)$, $(-1, -1)$, $(2, 2)$, $(2, -2)$, $(-2, 2)$, $(-2, -2)$. In other words,

$$\text{Prob}[\alpha = a \text{ and } \beta = b] = 1/8$$

for $a = \pm 1$ & $b = \pm 1$ and for $a = \pm 2$ & $b = \pm 2$. Obviously, these α and β are not independent. However, it is easy to see that for this distribution $E(\alpha \cdot \beta) = 0 = E(\alpha) \cdot E(\beta)$.

Proposition 5.1. [Markov inequality] If the distribution of α is concentrated on only non-negative real numbers, then for every real number T

$$\text{Prob}[\alpha > T] < \frac{E(\alpha)}{T}.$$

Definition. Variance of α is defined as $\text{var}(\alpha) := E\left((\alpha - E(\alpha))^2\right)$.

Simple properties of the variance :

- $\text{var}(\alpha + c) = \text{var}(\alpha)$ for every constant c;
- $\text{var}(c \cdot \alpha) = c^2 \cdot \text{var}(\alpha)$ for every constant c;
- $\text{var}(\alpha) = E(\alpha^2) - (E(\alpha))^2$;
- $\text{var}(\alpha + \beta) = \text{var}(\alpha) + \text{var}(\beta)$ for every pair of *independent* α and β.

Proposition 5.2. [the Chebyshev inequality] For every real number T

$$\text{Prob}[|\alpha - E(\alpha)| > \delta] < \frac{\text{var}(\alpha)}{\delta^2}.$$

Example 1. Let α be a random variable such that $\text{Prob}[\alpha = 1] = p$ and $\text{Prob}[\alpha = 0] = 1 - p$. Then $E(\alpha) = p$ and $\text{var}(\alpha) = p(1-p)$.

Example 2. Let α_i, $i = 1, \ldots, n$ be a sequence of independent identically distributed random variable such that $\text{Prob}[\alpha_i = 1] = p$ and $\text{Prob}[\alpha_i = 0] = 1-p$ for every i. Then

$$E(\alpha_1 + \ldots + \alpha_n) = pn,$$

$$E\left(\frac{1}{n}(\alpha_1 + \ldots + \alpha_n)\right) = p$$

and

$$\text{var}(\alpha_1 + \ldots + \alpha_n) = p(1-p)n,$$

$$\text{var}\left(\frac{1}{n}(\alpha_1 + \ldots + \alpha_n)\right) = \frac{p(1-p)}{n}.$$

From the Chebyshev inequality we obtain

$$\text{Prob}\left[\left|\frac{\alpha_1 + \ldots + \alpha_n}{n} - p\right| \geq \delta\right] \leq \frac{p(1-p)}{\delta^2 n}.$$

2. Shannon's coding theorem for block coding.

Theorem 5.1. Let $\alpha_1, \ldots, \alpha_n$ be a sequence of independent identically distributed random variables, and let $h > H(\alpha_i)$. Denote A the range (the alphabet) of all α_i and $k(n) := \lceil hn \rceil$. Then there exists a sequence of functions (encoding and decoding)

$$\begin{array}{rcccc} C_n & : & A^n & \to & \{0,1\}^{k(n)}, \\ D_n & : & \{0,1\}^{k(n)} & \to & A^n \end{array}$$

such that probability of the decoding error

$$\epsilon_n := \text{Prob}_{(\alpha_1 \ldots \alpha_n)}\left[D_n(C_n(w_1 \ldots w_n)) \neq w_1 \ldots w_n\right]$$

tends to 0 as $n \to \infty$. (Here the n-letter words $w_1 \ldots w_n$ is a randomly value of the sequence of random variables $(\alpha_1, \ldots, \alpha_n)$. In other words, each letter w_j is chosen with the distribution α_j, independently of other letters.)

3. Discussion of the homework We showed that for every triple of non-negative real numbers h_1, h_2, h_3 there exists a pair of jointly distributed random variables (α, β) such that

$$\begin{array}{rcl} H(\alpha) & = & h_1 + h_2 \\ H(\beta) & = & h_1 + h_3 \\ H(\alpha, \beta) & = & h_1 + h_2 + h_3 \end{array}$$

4. Conditional entropy.

Definition. Let (α, β) be jointly distributed random variables, with

$$p_{ij} = \text{Prob}[\alpha = 1_i \ \& \ \beta = b_j].$$

For each value a_i we have a conditional distribution on the values of β with probabilities

$$p'_j = \text{Prob}[\beta = a_j \,|\, \alpha = a_i] = \frac{\text{Prob}[\alpha = a_i \ \& \ \beta = b_j]}{\text{Prob}[\alpha = a_i]}.$$

This conditional distribution has Shannon's entropy $\sum_j p'_j \log \frac{1}{p'_j}$; we denote it $H(\beta \,|\, \alpha = a_i)$.

Definition. We define the entropy of β conditional on α as the average

$$H(\beta \,|\, \alpha) := \sum_i \text{Prob}[\alpha = a_i] \cdot H(\beta \,|\, \alpha = a_i).$$

There are simple properties of *conditional entropy* :

(a) $H(\alpha, \beta) = H(\alpha \,|\, \beta) + H(\beta)$

(b) $H(\alpha \,|\, \beta) \leq H(\alpha)$

(c) $H(\alpha \,|\, \beta) = H(\alpha)$ if and only if α and β are independent

We proved (a) in the class ; this property follows directly from the definition. The property (b) rewrites to

$$H(\alpha, \beta) - H(\beta) \leq H(\alpha),$$

and we already proved this inequality (see Proposition 3 in lecture 2.) Moreover, we know that $H(\alpha, \beta) - H(\beta) = H(\alpha)$ if and only if α and β are independent ; this implies (c).

To be continued (in December).