Eighth Lecture
October 25, 2022

We discussed the home works.

Exercise 1: This exercise is similar to the previous example, only the numbers differ. Let the 3×3 transition probability matrix Π of a Markov chain $\mathbb{Z}$ with three states A, B, C have first row: $7/8, 1/8, 0$, second row: $0, 7/8, 1/8$, third row: $1/3, 1/3, 1/3$. Determine the entropy of the source whose outcome is the actual state of this Markov chain.

Solution: We need to calculate the stationary distribution. Let a, b, c denote the stationary probabilities of the system being in state A, B, C, respectively. Then from the first column we have $a = \frac{7}{8}a + \frac{1}{3}c$ giving $a = \frac{8}{3}c$ and from the third column we have $c = \frac{1}{8}b + \frac{1}{3}c$ giving $b = \frac{16}{3}c$. Using $a + b + c = 1$ we obtain $\frac{8}{3}c + \frac{16}{3}c + c = 1$ which implies $c = \frac{3}{27} = \frac{1}{9}$. Thus $a = \frac{8}{27}, b = \frac{16}{27}$ and the requested entropy value is

$$H(X) = \frac{8}{27}H(X_n|X_{n-1} = a) + \frac{16}{27}H(X_n|X_{n-1} = b) + \frac{1}{9}H(X_n|X_{n-1} = c) =$$
$$\frac{8}{27}h(1/8) + \frac{16}{27}h(1/8) + \frac{1}{9}\log 3 = \frac{8}{9}h(1/8) + \frac{1}{9}\log 3.$$

$\diamondsuit$

Exercise 2: Let $X_1, X_2, \ldots$ be a Markov chain for which $\text{Prob}(X_1 = 0) = \text{Prob}(X_) = 1) = \frac{1}{2}$ and let the transition probabilities for $i \geq 1$ be given by $\text{Prob}(X_{i+1} = 0|X_i = 0) = \text{Prob}(X_{i+1} = 1|X_i = 0) = \frac{1}{2}$, while $\text{Prob}(X_{i+1} = 0|X_i = 1) = 0$ and $\text{Prob}(X_{i+1} = 1|X_i = 1) = 1$. Calculate the entropy of the source whose outcome is the resulting sequence of random variables $X_1, X_2, \ldots$.

Intuitively the solution is quite clear: This source emits some number (perhaps zero) 0's first, but after the first 1 it will emit only 1's. As i gets larger and larger, the probability of $X_i = 0$ is smaller and smaller (in fact it will be $\frac{1}{2^i}$), so if i is large, then X_i is almost certainly 1. Therefore the uncertainty about the value of X_i approaches zero, so the entropy of the source should be 0.

This intuition is easy to confirm by calculation: by Theorem 10

$$H(X) = \lim_{n\to\infty} H(X_n|X_1, \ldots, X_{n-1}) = \lim_{n\to\infty} H(X_n|X_{n-1}) =$$
$$\lim_{n\to\infty} \text{Prob}(X_{n-1} = 0)h(1/2) + \text{Prob}(X_{n-1} = 1)h(0) =$$
$$\lim_{n\to\infty} \frac{1}{2^{n-1}} + (1 - \frac{1}{2^{n-1}})0 = 0.$$

Note that the exercise can also be solved similarly as the previous one: realizing that the stationary distribution is concentrated on the value 1 we get that the entropy of the source is $0 \cdot h(1/2) + 1 \cdot h(1) = 0$. $\diamondsuit$

We considered the next version of the Lempel-Ziv algorithms.

Third version: LZW

This is the most popular version of the algorithm that is a modification of LZ78 as suggested by Welch. We now start with a codebook that already contains all the one-character sequences. (They have an index which serves as a codeword for them; we can think about their codeword as the s-ary, or simply binary representation of this index.) We now read the longest new part p of the text that can be found in the codebook and the next character, let it be a. Then the output is simply the index of p, we extend the codebook with the new sequence pa (that we obtain by simply putting a to the end of p) giving it the next index, and we consider the extra character a as the beginning of the not yet encoded part of the text.

dabbacdabbacdabbacdabbacdeecdeecdee

4, 1, 2, 2, 1, 3, 6, 8, 10, 12, 9, 11, 7, 16, 4, 5, 5, 11, 21, 23, 5

index	bejegyzés
1	*a*
2	*b*
3	*c*
4	*d*
5	*e*
6	*da*
7	*ab*
8	*bb*
9	*ba*
10	*ac*
11	*cd*
12	*dab*
13	*bba*

index	bejegyzés
14	*acd*
15	*dabb*
16	*bac*
17	*cda*
18	*abb*
19	*bacd*
20	*de*
21	*ee*
22	*ec*
23	*cde*
24	*eec*
25	*cdee*

Home-work

Let the source alphabet be $\mathcal{X} = \{a, b, c\}$ and the initial dictionary contain the letters a, b and c with their indexes (1, 2 and 3 respectively). Using the Lempel-Ziv-Welch algorithm
(a) encode the sequence *cabcbcbcb*
(b) decode the sequence $3, 4, 5, 6, 7, 1$

Quantization

In many practical situations the source variables are real numbers, thus have a continuum range. If we want to use digital communication we have to discretize, which means that some kind of "rounding" is necessary.

Def. Let $X = X_1, X_2, \ldots$ be a stationary source, where the X_i's are real-valued random variables. A (1-dimensional) *quantized* version of this source is a sequence of discrete random variables (another source) $Q(X_1), Q(X_2), \ldots$ obtained by a map $Q : R \to R$ where the range of the map is finite. The function $Q(.)$ is called the *quantizer.*

Goal: Quantize a source so that the caused distortion is small.

How can we measure the distortion? We will do it by using the quadratic distortion measure $D(Q)$ defined for n-length blocks as

$$D(Q) = \frac{1}{n} E\left(\sum_{i=1}^{n} (X_i - Q(X_i))^2\right),$$

where $E(.)$ means expected value.
Since our X_i's are identically distributed we have

$$D(Q) = E((X - Q(X))^2).$$

(Here X is meant to have the same distribution as all the X_i's.)

Let the range of $Q(.)$ be the set $\{x_1, \ldots, x_N\}$, where the x_i's are real numbers. $Q(.)$ is uniquely defined by the values $x_1, \ldots, x_N$ and the sets $B_i = \{x : Q(x) = x_i\}$. Once we fix $x_1, \ldots, x_N$, we will have the smallest distortion $D(Q)$ if every x is "quantized" to the closest x_i, i.e.,

$$B_i = \{x : |x - x_i| \leq |x - x_j| \ \forall j \neq i\}.$$

(Note that this rule puts some values into two neighboring B_i's (considering $x_1 < x_2 < \cdots < x_N$, we have $x = \frac{1}{2}(x_i + x_{i+1})$ in both B_i and B_{i+1}). This can easily be resolved by saying that all these values go to (say) the smaller indexed B_i.)

If now we consider the B_i's fixed then the smallest distortion $D(Q)$ is obtained if the x_i values lie in the barycenter of the B_i, which is $E(X|B_i) := E(X|X \in B_i) = \frac{\int_{B_i} x f(x) dx}{\int_{B_i} f(x) dx}$, where $f(x)$ is the density function of the random variable X. (We will always assume that $f(x)$ has all the "nice" properties needed for the existence of the integrals we mention.)

We proved the previous claim, i.e. smallest distortion is achieved for given quantization intervals B_i when $Q(x) = E(X|B_i)$ for $x \in B_i$. Here you can find a different proof for the statement:

This holds for all B_i separately, so it is enough to show it for one of them. By the linearity of expectation

$$E((X - c)^2) = E(X^2) - c(2E(X) - c),$$

and this is smallest when $c(2E(X) - c)$ is largest. Since the sum of c and $2E(X) - c$ does not depend on c, one can see simply from the inequality between the arithmetic and geometric mean ($\frac{a+b}{2} \geq \sqrt{ab}$ with equality iff $a = b$) that this product is largest when $c = E(X)$. (At least this is the case if we can assume that both c and $2E(X) - c$ are non-negative and so the inequality $\frac{a+b}{2} \geq \sqrt{ab}$ can be used. If this is not the case, we can still easily obtain that $c(2E(X) - c)$ is maximized by $c = E(X)$ by looking at the derivatives.)

Lloyd-Max algorithm

The above suggests an iterative algorithm to find a good quantizer: We fix some quantization levels $x_1 < \cdots < x_N$ first and optimize for them the B_i domains by defining them as above: let $y_i = \frac{x_i + x_{i+1}}{2}$ for $i = 1, \ldots, N-1$ and

$$B_1 := (-\infty, y_1], \quad B_i := (y_i, y_{i+1}],\ i = 2, \ldots, N-1, \quad B_N = (y_{N-1}, \infty).$$

Notice that in general there is no reason for the x_i's to be automatically the barycenters of the domains B_i obtained in the previous step. So now we can consider these domains B_i fixed and optimize the quantization levels with respect to them by redefining them as the corresponding barycenters:

$$x_i := \frac{\int_{B_i} x f(x) dx}{\int_{B_i} f(x) dx}.$$

Now we can consider again the so-obtained x_i's fixed and redefine the B_i's for them, and so on. After each step (or after each "odd" step when we optimize the B_i domains for the actual x_i's) we can check whether the current distortion is below a certain threshold. If yes we stop the algorithm, if no, then we continue with further iterations.

The distortion is non-increasing in each step, therefore it converges to somewhere since it is non-negative. The problem is that not necessarily to the global optimum, the limit might as well be a local optimum. To solve this the algorithm can be started from different initial quantization levels and then the one with smallest distortion is chosen.

It should be clear from the above that if either of the two steps above changes the x_i quantization levels or the B_i domains, then the quantizer before that step was not optimal. It is possible, that no such change is attainable already and the quantizer is still not optimal.

A quantizer is called a Lloyd-Max quantizer if the two steps of the Lloyd-Max algorithm have no effect on them.

Example
Let X be a random variable that takes its values on the finite set $\{1, 2, 3, 4\}$ with uniform distribution. (That is $P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = 1/4$.) Let $N = 2$ that is we are allowed to use two values for the quantization. There are three different quantizers for which neither of the above steps can cause any improvement, but only one of them is optimal.
These three quantizers can be described by

$$Q_1(1) = 1, Q_1(2) = Q_1(3) = Q_1(4) = 3;$$

$$Q_2(1) = Q_2(2) = 1.5, Q_2(3) = Q_2(4) = 3.5;$$

$$Q_3(1) = Q_3(2) = Q_3(3) = 2, Q_3(4) = 4.$$

Home-work
Calculate the distortion of the above three quantizers.