Sixth Lecture

October 11, 2022

We started the class by discussing the solutions of the midterm problems. After that we turned to the topic of block coding.

Block coding

We know that for all UD codes $\frac{H(X)}{\log s} \leq L(f) = \mathbb{E}(|f(X)|)$, and there exists a (e.g. Shannon-Fano or achieving optimal average length a Huffman) code satisfying $L(f) = \mathbb{E}(|f(X)|) < \frac{H(X)}{\log s} + 1$. The overhead of at most 1 bit is due to the fact that $\log \frac{1}{p_i}$ is not always an integer. This overhead can be reduced by spreading it out over symbols. That's why it worth using so called block codes, when the source symbols are not encoded separately but a few consecutive symbols are regarded as a supersymbol and encoded together.

A function $f : \mathcal{X}^k \to \mathcal{Y}^*$ is a block code using block size k.

Obviously the lower and upper bounds above remain true for this type of code as well. For all UD codes

$$\frac{H(X_1, \ldots, X_k)}{\log s} \leq L(f) = \mathbb{E}(|f(X_1, \ldots, X_k)|)$$

and there exists a code satisfying

$$L(f) = \mathbb{E}(|f(X_1, \ldots, X_k)|) < \frac{H(X_1, \ldots, X_k)}{\log s} + 1.$$

If the random variables are independent and identically distributed then $H(X_1, X_2, \ldots, X_k) = kH(X_1)$ thus the expected codeword length *per letter* for this code is

$$\frac{1}{k} L(f) \leq \frac{1}{k}\left(\frac{H(X_1, \ldots, X_k)}{\log s} + 1\right) = \frac{H(X_1)}{\log s} + \frac{1}{k}$$

So for block codes the upper bound for $\frac{1}{k}L(f)$ can be much closer to the lower bound even if the consecutive symbols (letters) are independent.

Example:
Let $\mathcal{X} = \{a, b\}$, $\mathcal{Y} = \{0, 1\}$, suppose that the distribution of the source symbols is $\mathbb{P}(X = a) = \frac{1}{4}$, $\mathbb{P}(X = b) = \frac{3}{4}$ and that the consecutive random variables are independent.
Let the code $f_1 : \mathcal{X} \to \mathcal{Y}^*$ be the following: $f_1(a) = 0, f_1(b) = 1$. Then $L(f_1) = 1$ and obviously we cannot do better than that if we encode symbols separately.
Let the code $f_2 : \mathcal{X}^2 \to \mathcal{Y}^*$ be the following: $f_1(aa) = 111, f_1(ab) = 110, f_1(ba) = 10, f_1(bb) = 0$. Because of independence the probabilities of the blocks are $p(aa) = \frac{1}{16}$, $p(ab) = \frac{3}{16}$, $p(ba) = \frac{3}{16}$, $p(bb) = \frac{9}{16}$. Thus $L(f_2) = \frac{27}{16}$ and the expected codeword length per symbol is $\frac{1}{2}L(f_2) = \frac{27}{32}$ that is less than 1. The block code performs better.

Entropy rate of information sources

Def. An *information source* is a stochastic process, i.e. a sequence of indexed random variables, $\mathbb{X} = X_1, X_2, \ldots$.

Def. A source $\mathbb{X} = X_1, X_2, \ldots$ is *memoryless* if the X_i's are independent.

Def. A source is *stationary* if for every n and k $(X_1, \dots, X_k)$ and $(X_{n+1}, \dots, X_{n+k})$ has the same distribution.

Def. The *entropy rate* of a source emitting the sequence of random variables $X_1, X_2, \dots$ is

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \dots, X_n),$$

provided that this limit exists.

The above limit trivially exists for stationary memoryless sources defined above. Indeed, if the source is stationary and memoryless, then $H(X_1, X_2, \dots, X_n) = nH(X_1)$, since in that case the random variables are independent and identically distributed, so we have $\lim_{n\to\infty} \frac{1}{n} H(X_1, \dots, X_n) = \lim_{n\to\infty} \frac{1}{n} nH(X_1) = H(X_1)$.

In fact, once a source is stationary it always has an entropy, it need not be memoryless.

Theorem 10 *If a source* $\mathbb{X} = X_1, X_2, \dots$ *is stationary then its entropy rate exists and is equal to*

$$\lim_{n\to\infty} H(X_n | X_1, \dots, X_{n-1}).$$

Remark: Note that $\lim_{n\to\infty} H(X_n | X_1, \dots, X_{n-1})$ can be much smaller than $H(X_1)$. Think about a source with source alphabet $\{0, 1\}$ that emits the same symbol as the previous one with probability 9/10 and the opposite with probability 1/10. In the long run we have the same number of 0's and 1's, $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 0) = 1/2$, so $H(X_1) = 1$, while $\lim_{n\to\infty} H(X_n | X_1, \dots, X_{n-1}) = H(X_n | X_{n-1}) = h(0.9) < 1$.

Proof. By the source being stationary, we have

$$H(X_n | X_1, \dots, X_{n-1}) = H(X_{n+1} | X_2, \dots, X_n) \geq H(X_{n+1} | X_1, X_2, \dots, X_n).$$

Thus the sequence $H(X_i | X_1, \dots, X_{i-1})$ is non-increasing and since all its elements are non-negative, it has a limit.

From the Chain rule we can write

$$\frac{1}{n} H(X_1, \dots, X_n) = \frac{1}{n} \left(H(X_1) + \sum_{i=2}^{n} H(X_i | X_1, \dots, X_{i-1}) \right).$$

To complete the proof we refer to a lemma of Toeplitz that says that if $\{a_n\}_{n=1}^{\infty}$ is a convergent sequence of reals with $\lim_{n\to\infty} a_n = a$, then defining $b_n := \frac{1}{n} \sum_{i=1}^{n} a_i$, we have that $\{b_n\}_{n+1}^{\infty}$ is also convergent and $\lim_{n\to\infty} b_n = a$, too. Applying this to $a_n := H(X_n | X_1, \dots, X_{n-1})$ the statement follows. □

Note that the proof implies that the sequence $\frac{1}{n} H(X_1, \dots, X_n)$ is also non-increasing.

Theorem 11 *If the stationary source* $\mathbb{X}$ *is encoded with a uniquely decodable block code* $f : \mathcal{X}^k \to \mathcal{Y}^*$ *using block size* k *then for the expected codeword length per symbol*

$$\frac{H(\mathbb{X})}{\log s} \leq \frac{1}{k} L(f) = \frac{1}{k} \mathbb{E}(|f(X_1, \dots, X_k)|)$$

and if k *is large enough then there exists a code with* $\frac{1}{k} L(f)$ *arbitrarily close to* $\frac{H(\mathbb{X})}{\log s}$.

This follows from the upper and lower bounds for block codes, the definition of entropy rate and the remark above.

Markov chain, Markov source

Def. A stochastic process $\mathbb{Z} = Z_1, Z_2, \ldots$ is *Markov (or Markovian)* if for every n we have

$$\mathbb{P}(Z_n = z_n | Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}) = \mathbb{P}(Z_n = z_n | Z_{n-1} = z_{n-1})$$

We say that the variables $Z_1, Z_2, \ldots$ form a *Markov chain.*

Intuitively the above definition means that knowing just the previous Z_i tells us everything we could know about the next one even if we knew the complete past, i.e. given Z_{k-1} the random variable Z_k is conditionally independent of all preceding random variables. Such situations often occur.

Def. A Markov chain $\mathbb{Z}$ is *homogenous* or *time invariant* if $\mathbb{P}(Z_n = j | Z_{n-1} = i)$ is independent of n. The possible values of the random variables in a Markov chain are called *states.*

When the homogenous Markov chain has r *states* its behavior is described by an $r \times r$ stochastic matrix (each row is a probability distribution) Π defined by $\Pi[i, j] = \mathbb{P}(Z_2 = j | Z_1 = i)$.

Theorem 12 *The entropy rate of a homogenous stationary Markov chain $\mathbb{Z}$ is $H(\mathbb{Z}) = H(Z_2|Z_1)$*

This follows from Theorem 10 above, the Markov property and time invariance.

Def. A general *Markov source* or *Hidden Markov model* is a stochastic process $\mathbb{X}$, for which each X_i can be written as a function of two random variables, namely $X_i = F(Z_i, Y_i)$ where $\mathbb{Z}$ is a homogenous Markov chain and $\mathbb{Y}$ is a stationary and memoryless source that is independent of $\mathbb{Z}$.

A Markov source can model a situation where, for example, Z is a text or speech and Y is the noise.

We started to talk about the entropy rate of a Markov source, but this will come next time. We will also discuss some examples and exercises.