Third Lecture
September 20, 2022

Variable length source coding (cont.)

We stated and proved the Main Theorem in variable length coding:

Theorem 6 *Let us have an information source emitting symbol $x^{(i)} \in \mathcal{X}$ with probability $p(x^{(i)}) = p_i, (i = 1, \ldots, r)$. For any s-ary UD code $f : \mathcal{X} \to \mathcal{Y}^*$ of this source we have expected codeword length*

$$L(f) = \sum_{i=1}^{r} p_i |f(x^{(i)})| \geq H_s(P) = \frac{1}{\log s} H(P) = \frac{1}{\log s}\left(-\sum_{i=1}^{r} p_i \log p_i\right) = -\sum_{i=1}^{r} p_i \log_s p_i,$$

where P stands for the distribution $(p_1, \ldots, p_r)$. Thus, for a UD code the average codeword length is bounded from below by the entropy of the distribution governing the system.

For proving the theorem, we used McMillan theorem and the corollary of Jensen's inequality.

Proof of Theorem 6. We know from the McMillan theorem, that $\sum_{i=1}^{r} s^{-|f(x^{(i)})|} \leq 1$. Set $b = \sum_{i=1}^{r} s^{-|f(x^{(i)})|}$ and $q_i = \frac{s^{-|f(x^{(i)})|}}{b} \geq s^{-|f(x^{(i)})|}$. Then

$$\sum_{i=1}^{r} p_i |f(x^{(i)})| = -\sum_{i=1}^{r} p_i \log_s(q_i b) \geq -\sum_{i=1}^{r} p_i \log_s q_i = -\frac{1}{\log s}\sum_{i=1}^{r} p_i \log q_i.$$

Observe that $\sum_{i=1}^{r} q_i = 1$ and $q_i \geq 0$ for every i (so $(q_1, \ldots, q_r)$ could be considered a probability distribution). Thus by Corollary 4 of Jensen's inequality, we have that $-\sum_{i=1}^{r} p_i \log q_i \geq -\sum_{i=1}^{r} p_i \log p_i$ and the statement follows. □

We can have equality iff the distribution of *s-adic*, ie. for all i $p_i = s^{-l_i}$.

Example:
- $s = 2$ case: the distribution $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$ is diadic, since the probabilities in the distribution are $2^{-1}, 2^{-2}, 2^{-3}, 2^{-3}$. We have seen that the entropy of this distribution is 1.75bits and also we have seen a perfix code for this distribution with expected codeword length 1.75bits.
- $s = 3$ case: the distribution $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9}, \frac{1}{27}, \frac{1}{27}, \frac{1}{27}\right)$ is triadic. We calculated the entropy, and checked that the expected codeword length of the code $(0, 1, 20, 21, 220, 221, 222)$ equals the entropy.

Thus in this special case, we can reach the lower bound, we can find a code f such that $L(f) = H_s(P)$. For other source distributions, there isn't such a code, but there exists a code with expected codeword length close to the lower bound.

Theorem 7 *Let us have an information source emitting symbol $x^{(i)} \in \mathcal{X}$ with probability $p(x^{(i)}) = p_i, (i = 1, \ldots, r)$. There exists an s-ary prefix code for this source with average codeword length less than $H_s(P) + 1 = \frac{H(P)}{\log s} + 1$.*

Proof of Theorem 7. Kraft's theorem implies that there is a prefix code with codeword lengths $\left\lceil \log_s \frac{1}{p_1} \right\rceil, \ldots, \left\lceil \log_s \frac{1}{p_r} \right\rceil$, since

$$1 = \sum_{i=1}^{r} p_i = \sum_{i=1}^{r} s^{\log_s p_i} = \sum_{i=1}^{r} s^{-\log_s(1/p_i)} \geq \sum_{i=1}^{r} s^{-\lceil \log_s(1/p_i) \rceil}.$$

Such a code has average length

$$\sum_{i=1}^{r} p_i \left\lceil \log_s \frac{1}{p_i} \right\rceil < \sum_{i=1}^{r} p_i (\log_s \frac{1}{p_i} + 1) \leq \sum_{i=1}^{r} p_i \log_s \frac{1}{p_i} + \sum p_i = \sum_{i=1}^{r} p_i \log_s \frac{1}{p_i} + 1.$$

Shannon-Fano code

Next we introduced a code construction, called the *Shannon-Fano code*:

We assume $p_1 \geq p_2 \geq \cdots \geq p_n > 0$. Let $w_1 = 0$ and for $j > 1$ let $w_j = \sum_{i=1}^{j-1} p_i$. Let the codeword $f(x^{(j)})$ be the s-ary representation of the number w_j (which is always in the $[0,1)$ interval) without the starting integer part digit 0, and with minimal such length that it is not a prefix of any other such codeword. The latter condition already ensures that the code is prefix.

This construction is very closely related to the one on which the proof of Theorem 7 was based. Nevertheless, below we give a second proof of Theorem 7 directly using the Shannon-Fano code construction.

The above definition (of Shannon-Fano code) implies that the first $|f(x^{(j)})| - 1$ digits of $f(x^{(j)})$ is a prefix of another codeword and thus it must be the prefix of a codeword coming from a closest number w_h, thus w_{j-1} or w_{j+1}. This implies

$$p_j = p(x^{(j)}) = w_{j+1} - w_j \leq s^{-(|f(x^{(j)})|-1)}$$

or

$$p_{j-1} = p(x^{(j-1)}) = w_j - w_{j-1} \leq s^{-(|f(x^{(j)})|-1)}.$$

By $p_{j-1} \geq p_j$ in either case the first of the above two inequalities holds. Thus $\log_s p_j \leq -|f(x^{(j)})| + 1$ implying

$$-p_j \log_s p_j \geq p_j(|f(x^{(j)})| - 1),$$

and thus

$$-\sum_{j=1}^{r} p_j \log_s p_j + 1 \geq \sum_{j=1}^{r} p_j |f(x^{(j)})|.$$

Then we constructed the Shannon-Fano code for three probability distributions.

Examples:

- $s = 2$, consider the distribution $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$
- $s = 3$, consider the distribution $\left(\frac{3}{8}, \frac{1}{6}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{12}\right)$
- $s = 4$, consider the distribution $(0.36, 0.17, 0.09, 0.09, 0.07, 0.04, 0.04, 0.04, 0.03, 0.03, 0.02)$

In order to get the s-ary representation of the w_is, we took the interval $[0,1]$ and partitioned it into s parts of the same length $\left(\frac{1}{s}\right)$. The w_is falling into the first partition get a first digit 0, the w_is falling into the second partition get a first digit 1, etc the w_is falling into the last partition get a first digit $s-1$. We go on partitioning further the subintervals having more than one w_i falling there. And the corresponding codewords get a new digit. We do this until there are no partition with more than one w_i in it.

Optimal codes

We have seen constructions giving average codeword length close to the lower bound $H_s(P)$, but nothing guaranteed that any of these codes would be best possible. So the question of how to find an optimal average length code comes up. This will be answered by constructing the so-called Huffman code. We will study this only for the binary case, i.e, when the size of the code alphabet is $s = 2$.

Def. A code f is *optimal* if $\mathbb{E}\,|f(X)| \leq \mathbb{E}\,|f'(X)|$ for all codes $f' : \mathcal{X} \to \mathcal{Y}^*$

We discussed that optimal code does exist, since there are finitely many possible codes, and there are more than one possibility for an optimal code, since eg. inverting the bits doesn't change the average codeword length but results in a different code. Similarly interchanging the codewords of the same length.

Assume $p_1 \geq \cdots \geq p_r > 0$, $p_i = p(x^{(i)})$ and having an optimal binary code $C = (f(x^{(1)}), \ldots, f(x^{(r)}))$, $l_i := |f(x^{(i)})|$. By the foregoing we can assume that the code is prefix. (Note that the $p_r > 0$ assumption is not a real restriction: if we have 0-probability events, they need not be encoded. Or they could even be encoded into long codewords, since their contribution to the average length will be zero anyway.)

Theorem 8 *(Properties of optimal code) If the prefix code $f : \mathcal{X} \to \{0,1\}^*$ is optimal then (there is a reordering of the source symbols and the codewords of the same length such that)*
(1) $l_1 \leq l_2 \leq \cdots \leq l_r$
(2) $l_r = l_{r-1}$
(3) the two longest codewords $f(x^{(r)})$ and $f(x^{(r-1)})$ differ only in the last bit.

We will prove this theorem next time, and also then will we learn about a construction of an optimal code, the Huffman code.