19 Communication theory

19.2 Information content of natural language (4 units)

Background material for this project is contained in the Part IIB course Communication Theory.

Let I_m be a set of m *m*essages which may be transmitted with non-zero probability p_i, $i = 1, \ldots, m$. If successive messages are independent the source called *B*ernoulli — we do not assume this in general. Define the *s*ource entropy to be

$$h = -\sum_{i=1}^{n} p_i \log p_i.$$

The *H*uffman binary code for I_m is produced by the algorithm:

(i) Order the messages in I_m so that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m$.

(ii) Assign **0** to be the last character of the codeword for message $m-1$, and **1** for message m.

(iii) If $m > 2$, combine messages $m-1$ and m to form a *r*educed alphabet $I_{m-1} = \{1, 2, \ldots, m-2, (m-1, m)\}$ with respective probabilities $p_1, p_2, \ldots, p_{m-2}$ and $p_{m-1} + p_m$ and start again at step (i).

Whether or not a message source is Bernoulli, we can often improve the expected code-word length on a per-message basis by *s*egmentation, that is, grouping messages in blocks of n and regarding them as coming from the message set I_m^n.

The files, on the CATAM website, *dat*1902*x.dat*, where x is one of `A`, `B`, `C` or `D`, contain samples of English texts encoded by `A` $= 1, ...,$ `Z` $= 26$ with space $= 0$. Each file contains 401 records with 25 numbers per record, except the last, which contains a single negative number. (These files can be made available on other systems or on disc.)

Choose one of the data files to work with.

> **Question 1** Estimate the source entropy of English text, construct the corresponding Huffman code and find the expected codeword length. Compare with the Shannon–Fano code. Discuss how segmentation would improve the expected length if the source were assumed Bernoulli.
>
> **Question 2** Discuss the extent to which English text is not Bernoulli. Construct the Huffman code for pairs of letters. What effect does segmentation have in this case? Compare the effect of segmentation on English text with its effect on a Bernoulli source with the same distribution of letter frequencies as English.

References

[1] C.M. Goldie and R.G.E. Pinch, *C*ommunication theory, CUP, 1991.

