Entropy Coding

Marek Rychlik

Department of Mathematics
University of Arizona

February 25, 2009

Kraft inequality (extended variant)

Theorem

If A is countable and $\mathcal{C} : A \to \{0,1\}^+$ is a lossless symbol code then

$$\sum_{a\in A} 2^{-D(a)} \leq 1.$$

where $D(a) = \ell(\mathcal{C}(a))$ is the length of the codeword assigned to a symbol a.

What does Kraft Inequality tell us?

- If a symbol code is lossless then there is a numerical restriction on the lengths of *codewords* $\ell(\mathcal{C}(a))$
- We cannot have short codewords for all symbols: if some symbols have short codewords, there must be symbols with long ones.

Shannon source coding theorem

Theorem

(Shannon, 1948) If a binary symbol code $\mathcal{C} : A \to \{0,1\}^+$ is lossless then the expected value of the length of the codeword satisfies the inequality:

$$\mathbb{E}(\ell \circ \mathcal{C}) \geq H(P)$$

where $H(P)$ is the Shannon entropy *of the distribution P:*

$$H(P) = \sum_{a \in A} P(a)(-\log_2 P(a))$$

The quantity $I(a) = -\log_2 P(a)$ is interpreted as the amount of information contained in one occurrence of symbol a.

What does Shannon Theorem tell us?

- It imposes a mathematical limit on the efficiency of coding completely random messages.
- The limit is formulated in terms of one number, the entropy of the distribution.

The existence of nearly optimal codes

Theorem

(Shannon-Fano, 1948) For every alphabet A and a distribution function $P : A \to (0,1]$ there exists a binary code $\mathcal{C} : A \to \{0,1\}^+$ such that:

$$H(P) \leq \mathbb{E}(\ell \circ \mathcal{C}) \leq H(P) + 1.$$

- Huffman codes (Huffman, MIT dissertation, 1954). Suboptimal if probabilities of symbols are not powers of 2.
- Arithmetic codes (Rissanen, IBM Research, 1976). Nearly Shannon-optimal. Hard to implement efficiently on a standard computer.
- Golomb-Rice codes (Golomb, 1966).

Shannon-Fano coding

- We order the alphabet A and the probabilities P.
- We form partial sums of the probabilities ($i = 0, 1, \ldots, n$):

$$q_i = \sum_{j<i} p_i$$

  (Note: $q_0 = 0$ is defined.)
- The Shannon-Fano code $\mathcal{C}$ maps the $i - th$ symbol a_i of the alphabet to the bits of the *binary rational number* in the interval $[q_{i-1}, q_i)$.
- The rational number with the fewest bits is selected.
- Shannon-Fano codes satisfy the upper bound of the Shannon-Fano theorem.

An example of a Shannon-Fano code

Example

Let $A = \{a, b\}$ and $P = \{2/3, 1/3\}$. Thus $q_1 = 2/3$ and $q_2 = 1$. The interval assignment is:

$$\begin{aligned} a &\rightarrow [0, 2/3) \\ b &\rightarrow [2/3, 1) \end{aligned}$$

The binary rational number with the fewest number of bits in the interval $[0, 2/3)$ is 0. Similarly, in the interval $[2/3, 1)$ we pick $3/4$ which has the binary expansion 0.11. This results in the symbol code:

$$\begin{aligned} a &\rightarrow 0 \\ b &\rightarrow 11 \end{aligned}$$

The codewords are the digits of the fractional parts of the rational numbers selected. The Shannon-Fano code is a prefix code.

Block codes that are nearly optimal

- A *block code* is a mapping $\mathcal{C}_{block} : A^r \to B$ of blocks of symbols of length r to some alphabet B.
- A block code is extended to A^+ by concatenation if message length L is a multiple of r: $L = k \cdot r$.

$$\begin{aligned} s_1 s_2 \ldots s_L \quad \longrightarrow \quad & \mathcal{C}(s_1 s_2 \ldots s_r)\mathcal{C}(s_{r+1} s_{r+2} \ldots s_{2r}) \ldots \\ & \mathcal{C}(s_{(k-1)r+1} s_{(k-1)r+2} \ldots s_L) \end{aligned}$$

Block codes that are nearly optimal

- A block code can be used to code messages whose length is a multiple of s by concatenation.
- Given that one has symbol codes for which $\mathbb{E}(\ell \circ \mathcal{C}) \leq H(P) + 1$, we can define a $\mathcal{C}_{block} : A^r \to B$ for which
$$\mathbb{E}(code(M)) \leq H(P) + \frac{1}{r}$$
  for every message M whose length is a multiple of block length r.
- As $r \to \infty$, the block code becomes optimal.

Arithmetic coding

- A practical realization of the Shannon-Fano idea is *arithmetic coding*. IBM research developed arithmetic coding in the 1970's and has held a number of patents in this area.
- An entire message (sometimes billions of symbols) are encoded as a single binary rational number, whose digits become the code to be stored or transmitted.
- The crux of the algorithm is to perform coding in time proportional to the message length, using only final precision arithmetic available in modern computers.

What is next?

- We can decrease entropy using statistical forecasting, digital filtering, and other clever ideas.
- Lower entropy leads to higher compression ratio.

“What we do here is decrease entropy.” — A Qbit physicist and engineer.