Information, Codes and Ciphers

Hao Ren

November 7, 2020

3 Compression Coding

3.1 Instantaneous and UD codes

3.1.1 Definition

a source S	with q source symbols	$s_1, s_2, \cdots, s_q$,
	with probabilities	$p_1, p_2, \cdots, p_q$,
encoded by a code C	with q codewords	$c_1, c_2, \cdots, c_q$,
	of lengths	$l_1, l_2, \cdots, l_q$.

- with a radix r codewords,
- variable length codes,
- not channel noise for source coding.

3.1.2 UD and I-code

A code C is

UD uniquely decodable codes if it can always be decoded unambiguously,

I-code instantaneous if no codeword is the prefix of others.

3.1.3 Comma Codes

The standard comma code of length n is

- a code which every codeword has length $\leq n$,
- a code which every codeword contains at most one 0,
- and if a codeword contains 0 then 0 must be the final symbol in the codeword.

3.1.4 Decision Trees

3.1.5 The Kraft-McMillan Theorem

Theorem 3.1 (The Kraft-McMillan Theorem)
A UD-code of radix r with q codewords $c_1, c_2, \cdots, c_q$ of lengths $l_1 \leq l_2 \leq \cdots \leq l_q$ exists
if and only if *an I-code with the same parameters exists*
if and only if

$$K = \sum_{i=1}^{q} \frac{1}{r^{l_i}} \leq 1.$$

3.1.6 Length and Variance

The expected or **average length** of codewords is given by

$$L = \sum_{i=1}^{q} p_i l_i$$

and the **variance** is given by

$$V = \sum_{i=1}^{q} p_i l_i^2 - L^2.$$

Our aim is to minimise L for a given source S and, if more than one code C gives this value, to minimise V.

Theorem 3.2 (Minimal UD-codes)
Let C be a UD-code with minimal expected length L for the given source S. Then, after permuting codewords of equally likely symbols if necessary,

- $l_1 \leq l_2 \leq \cdots \leq l_q$ *and*
- $l_{q-1} = l_1$.

Furthermore, if C is instantaneous, then

- c_{q-1} *and* c_q *differ only in their last place.*

If C is binary, then

- $K = \sum_{i=1}^{q} 2^{-l_i} = 1.$

3.2 Huffman's Algorithm