Advantages and Disadvantages of Huffman Codes

It turns out that Huffman codes indeed have optimal code length.

Theorem: Optimality of Huffman codes

Let P_X be a source, and let C^* be the associated Huffman code. For any other uniquely decodable code C' with the same alphabet,

$$\ell_{C^*}(P_X) \leq \ell_{C'}(P_X).$$

Proof

see Cover/Thomas, Section 5.8

As we have seen, the average codeword length of Huffman codes is theoretically optimal. However, Huffman codes (and symbol codes in general) still have a number of disadvantages:

- When compressing, for example, an English text symbol-by-symbol, the probability distribution for each position may depend on the string of text that precedes it: for example, the letter ***n*** is a lot more likely than the letter ***a*** if it comes after the string ***informatio***. Given this change of distribution, the Huffman code may not produce the shortest possible code. This can be resolved by recomputing the Huffman code after every symbol, but this results in a lot of overhead.
- The average codeword length is upper bounded by $H(X) + 1$. This additive cost of 1 bit is fine when $H(X)$ is very large, but can be a significant overhead when $H(X)$ is small itself.

created: 2019-10-21