Roberto De Prisco, MIT

"On the Redundancy and Data Expansion of Huffman Codes"

The problem of providing bounds on the redundancy of an optimal code for a discrete memoryless source in terms of the probability distribution of the source, has been extensively studied in the literature. The attention has mainly focused on binary codes for the case when the most or the least likely source letter probabilities are known.

In this paper we analyze the relationships among tight lower bounds on the redundancy r. Let r $\leq \phi_{D,i}(x)$ be the tight lower bound on r for D-ary codes in terms of the value x of the i-th most likely source letter probability. We prove that $\phi_{D,i-1}(x) \leq \phi_{D,i}(x)$ for all possible x and i. As a consequence, we can bound the redundancy when only the value of a probability (but not its rank) is known. Another consequence is a shorter and simpler proof of a known bound. We also provide some other properties of tight lower bounds. Finally, we determine an achievable lower bound on r in terms of the least likely source letter probability for $D \geq 3$, generalizing the known bound for the case D=2.

We also study the data expansion of Huffman codes. While compressing a file with a Huffman code, it is possible that the size of the file grows temporarily. This happens when the source letters with low frequencies (to which long codewords are assigned) are encoded first. The maximum data expansion is the average growth in bits per source letter resulting from the encoding of a source letter with a long codeword. It is a measure of the worst case temporary growth of the file. In this paper we provide some new properties of the maximum data expansion d of Huffman codes and using these properties we prove that d ¡ 1.256, improving on the previous best known upper bound.