Huffman Heights

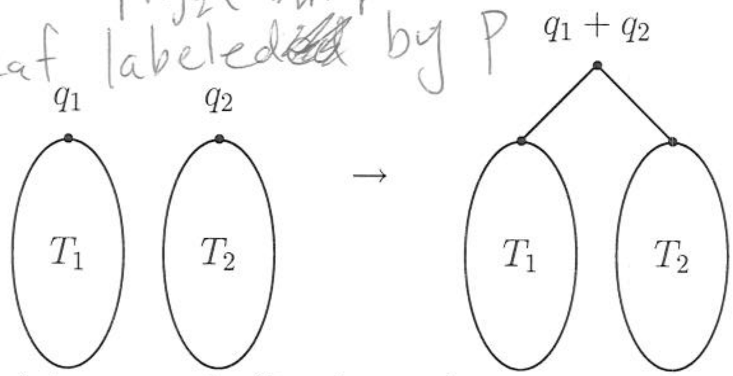
Theorem 3 *A letter that occurs with probability p will be represented by a leaf with height $h \leq \lceil \log_2 1/p \rceil$ in the Huffman tree.*

Huffman is better (or equal) to the tree from heights. (without trimming)

Proof (by induction) Assume that the height of a particular leaf in T_1 is $h \leq \lceil \log_2 q_1/p \rceil = \lceil \log_2 (1/(p/q_1)) \rceil$

Pick a leaf labeled by P in q_1

q_2 is next smaller tree $q_1 \leq q_2$



so The height of that same leaf in the resulting tree is exactly $h + 1$, which is bounded above by:

$$h + 1 \leq \left\lceil \log_2 \frac{q_1}{p} + 1 \right\rceil = \left\lceil \log_2 \frac{2q_1}{p} \right\rceil \leq \left\lceil \log_2 \frac{q_1 + q_2}{p} \right\rceil$$

$$= \left\lceil \log_2 q_1 \tfrac{1}{p} \right\rceil + 1$$

Note that $q_2 \geq q_1$, as assumed in the construction of the Huffman code.

at the last step

height of that leaf $\leq \left\lceil \log_2 \frac{1}{p} \right\rceil$

Expected Code Length

Theorem 4 *The Huffman Code yields expected code length within 1 of the entropy, H.*

Proof. Assume that $h_i = \lceil \log_2 1/p_i \rceil$.

$$\sum_{i=1}^{k} \frac{1}{2^{h_i}} = \sum_{i=1}^{k} \frac{1}{2^{\lceil \log_2 1/p_i \rceil}} \leq \sum_{i=1}^{k} \frac{1}{2^{\log_2 1/p_i}} = \sum_{i=1}^{k} \frac{1}{1/p_i} = 1$$

Therefore the sequence $h_1, h_2, \ldots h_k$ corresponds to a binary tree $\sum_{i=1}^{k} p_i \log_2(1/p_i)$

$/\!/$

$$\text{Expected code length (ECL)} = \sum_{i=1}^{k} p_i h_i = \sum_{i=1}^{k} p_i \lceil \log_2 1/p_i \rceil$$

$\leq \sum_{i=1}^{k} p_i(\log_2 1/p_i + 1)$

$$H = \sum_{i=1}^{k} p_i \log_2 1/p_i \leq ECL \leq \sum_{i=1}^{k} p_i(\log_2 1/p_i + 1) = H + 1$$