Compression

6.897
L13.1
Apr. 9, 2003

Huffman code [Huffman 1952]

Given: p_i's = prob. of letter i occurring (stochastic)

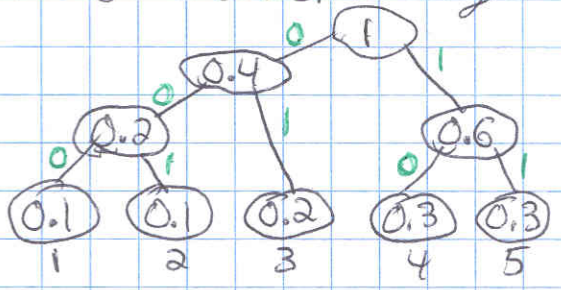
Want: prefix code – code word w_i $\forall i$ s.t. none are prefixes of another
⇒ after seeing w_i, know you saw i

Goal: minimize average codeword length $E[|w_i|] = \sum_{i=1}^{|\Sigma|} p_i |w_i|$

- Can view (binary) prefix codes as root-to-leaf paths in binary tree
- similar to optimal BST but without order & just leaves

Huffman code: greedily construct binary tree by merging lowest-prob. pair

1 – 0.1
2 – 0.1
3 – 0.2
4 – 0.3
5 – 0.3



Theorem: Huffman minimizes E[codeword length] = E[depth of leaf] among all (binary) prefix codes

Proof: greedy exchange argument

Theorem: $E[\text{Huffman codeword length}] \in [H, H+1)$

Proof: Claim you can pick node ℓ_i in complete binary tree $\forall i$ such that $\text{depth}(\ell_i) = \lceil \lg 1/p_i \rceil$ and no ℓ_i is ancestor of ℓ_j.

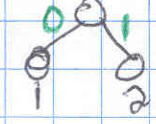
⇒ take tree on those leaves & $E[\text{depth}] = \sum p_i \lceil \lg 1/p_i \rceil < 1 + H$.

Suppose $p_1 \geq p_2 \geq \cdots \geq p_n$ & pick $\ell_1, \ell_2, \ldots$ in increasing depth order

After picking ℓ_1 arbitrarily, there are $\lceil \lceil 1/p_1 \rceil \rceil - 1$ nodes at that depth remaining

Enough to support $1 - p_1 \leq 1 - \lfloor \lfloor p_1 \rfloor \rfloor$ remaining probability. □

- can be generalized to nonbinary codes, k-grams instead of letters
- but far from entropy: $p_1 = \frac{1}{1024}$ $p_2 = 1 - \frac{1}{1024}$ $H \approx 0.01$

code:  ⇒ E[codelength] = 1