SUSTech

Huffman. Code.

Examples of construct a Huffman tree.

Proof of Lemma 5.8.1.

① We can prove that $p_j > p_k$ and $l_j > l_k$ lead to non-optimal prefix code.

~~Step 1:~~ Suppose there is a code C with. $p_j > p_k$ and $l_j > l_k$.

The expected length

$$L(C) = p_j l_j + p_k l_k + \sum_{i \neq j,k} p_i l_i$$

We can construct a new code C', which swaps the codeword j and k in C.

$$L(C') = p_j l_k + p_k l_j + \sum_{i \neq j,k} p_i l_i$$

ACT TODAY IMPACT TOMORROW

SUSTech

$$LCC(C_m) - LCC(C_m')$$

$$= p_j l_j + p_k l_k - p_j l_k - p_k l_j$$

$$= \underbrace{(p_j - p_k)}_{>0} \underbrace{(l_j - l_k)}_{>0} > 0$$

$$\Rightarrow LCC(C_m') < LCC(C_m).$$

C_m is not the optimal prefix code.

SUSTech

(3) We can prove that "if two longest codewords have different lengths the code is not optimal."

Suppose codewords j and k are the longest for in code C

$c(j)$: $b_1 b_2 \cdots b_m$

$c(k)$: $b_1' b_2' \cdots b_m' b_{m+1}' \cdots b_{m+n}'$

Because of prefix code, $b_1 b_2 \cdots b_m \neq b_1' b_2' \cdots b_m'$

Change codeword $c(k)$ to $c'(k) = b_1' b_2' \cdots b_m'$, we get a new code C'

$$\begin{cases} c'(i) = c(i) & i \neq k \\ c'(k) = b_1' b_2' \cdots b_m' \end{cases}$$

Obviously, $\begin{cases} c'(k) \neq c'(i) & i \neq k. \\ c(i)\ (i \neq k) \text{ is not prefix of } c'(k). \end{cases}$

Hence C' is still a prefix code.

Moreover $L(C') < L(C)$

ACT TODAY IMPACT TOMORROW

SUSTech

③ Suppose the longest codeword in ~~this~~ optimal prefix code C is m.

According to Property 2. there are at least two codewords with length m. say,

$$b_1\ b_2\ b_3\ \cdots\ b_m$$
$$b_1'\ b_2'\ b_3'\ \cdots\ b_m'$$

~~const~~
consider a codeword $b_1 b_2 b_3 \cdots \bar{b}_m$,
if $b_1 b_2 \cdots \bar{b}_m \in C$. C is the code we want.

If $b_1 b_2 \cdots \bar{b}_m \notin C$, we replace $b_1' b_2' \cdots b_m'$ with $b_1 b_2 \cdots \bar{b}_m$, and obtain a new code C'.

No codeword in C' is the prefix of $b_1 b_2 \cdots \bar{b}_m$
$\Rightarrow$ C' is a prefix code with optimal expected length.

ACT TODAY IMPACT TOMORROW

SUSTech

Based on Lemma 5.8.1, suppose there is a distribution with PMF $p_1 \geq p_2 \geq \cdots \geq p_m$. There exists an optimal prefix code with lengths $l_1 \leq l_2 \leq l_3 \cdots \leq l_{m-1} = l_m$ and codewords $C(x_{m-1})$ and $C(x_m)$ differ only in the last digit.

canonical code

SUSTech

Optimality Proof of Huffman Code:

W.L.O.G., we consider binary code only.

Mathematical Induction:

Step 1: Huffman code is obviously optimal for a distribution with $|\mathcal{X}| = 2$.

Step 2: Suppose Huffman code is optimal for $|\mathcal{X}| = m \quad m \geq 2$.

Step 3: We should prove it is optimal for $|\mathcal{X}| = m+1$

Consider a distribution $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_{m+1}$,
After one step iteration, we get a m-ary distribution $(p_1, p_2, p_3, \cdots, p_{m-2}, p_{m-1}, p_m + p_{m+1})$

SUSTech

Define the following codes:

C_{m+1} : Huffman code for $(p_1, p_2, \cdots, p_m, p_{m+1})$

C^*_{m+1} : optimal prefix code of canonical form for $(p_1, p_2, \cdots, p_m, p_{m+1})$

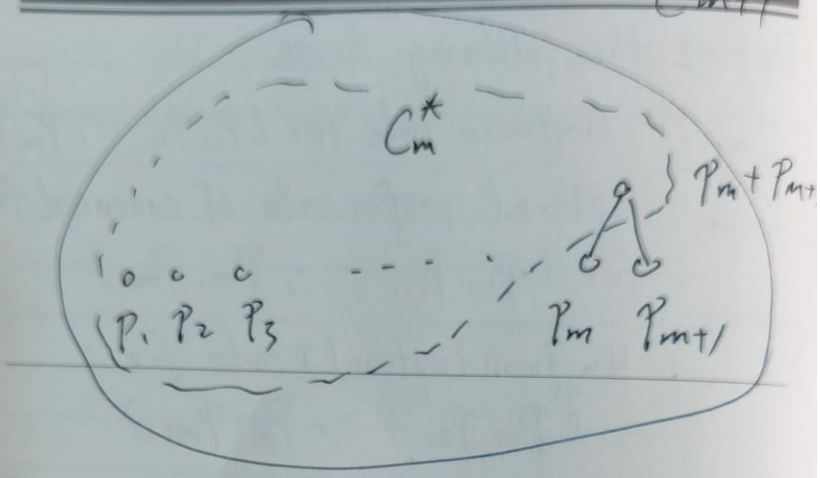
condense ↓

C^*_m : Huffman (optimal) code for $(p_1, p_2, \cdots, p_m + p_{m+1})$

C_m : prefix code condensed from C^*_{m+1} for $(p_1, p_2, \cdots, p_m + p_{m+1})$

	probability	codewords	
	⋮	⋮	
C^*_{m+1}	p_m	$b_1 b_2 \cdots b_{n-1} b_n$	
	p_{m+1}	$b_1 b_2 \cdots b_{n-1} b_n$	
	⋮	⋮	
C_m	⋮	same as C^*_{m+1}	C_m is a prefix code
	$p_m + p_{m+1}$	$b_1 b_2 \cdots b_{n-1}$	

ACT TODAY IMPACT TOMORROW



C_{m+1}		C_m^*
p_1	l_1	l_1
p_2	l_2	l_2
⋮		⋮
p_{m-1}	l_{m-1}	l_{m-1}
p_m, p_{m+1}	l_m	l_{m-1}

$$L(C_{m+1}) = \sum_{i=1}^{m-1} p_i l_i + (p_m + p_{m+1}) l_m$$

$$L(C_m^*) = \sum_{i=1}^{m-1} p_i l_i + (p_m + p_{m+1})(l_{m-1})$$

$$L(C_m^*) + P_m + P_{m+1} = L(C_{m+1})$$

$$L(C_m) + P_m + P_{m+1} = L(C_{m+1}^*)$$

$$\Rightarrow \underbrace{L(C_m^*) - L(C_m)}_{\leq 0} = \underbrace{L(C_{m+1}) - L(C_{m+1}^*)}_{\geq 0}$$

$$\Rightarrow L(C_m^*) - L(C_m) = L(C_{m+1}) - L(C_{m+1}^*) = 0$$

$$\Rightarrow L(C_{m+1}) = L(C_{m+1}^*)$$