5112

8-31

Section 4.8

Huffman Codes

Data Compression

Encoding Messages

"Everyone got an A"

$\longrightarrow$

0110001011

How do I encode my message in binary?

Encoding Messages

First Idea 26 English letters + space + , . ;

= 32 characters $32 = 2^5$

a 00000
b 00001
c 00010
d 00011
⋮

This is basically what ASCII does

Also similar to unicode.

A	01	N	10
B	1000	O	111
C	1010	P	0110
D	100	Q	1101
E	0	R	010
F	0010	S	000
G	110	T	1
H	0000	U	001
I	00	V	0001
J	0111	W	011
K	101	X	1001
L	0100	Y	1011
M	11	Z	1100

Morse Code

0101

c

aa

etet

eta

A	01	N	10
B	1000	O	111
C	1010	P	0110
D	100	Q	1101
E	0	R	010
F	0010	S	000
G	110	T	1
H	0000	U	001
I	00	V	0001
J	0111	W	011
K	101	X	1001
L	0100	Y	1011
M	11	Z	1100

Short encoding ⇝ more frequent

Long encoding ⇝ less frequent

Average bits per letter: ~2.53

Morse code •, —, pause

ternary encoding

A •—	J •———	S •••
B —•••	K —•—	T —
C —•—•	L •—••	U ••—
D —••	M ——	V •••—
E •	N —•	W •——
F ••—•	O ———	X —••—
G ——•	P •——•	Y —•——
H ••••	Q ——•—	Z ——••
I ••	R •—•	

Prefix Codes

Set of characters $S \xrightarrow{\gamma}$ Sequences of $\{0,1\}$

such that if $x, y \in S$, $\gamma(x)$ is not a prefix of $\gamma(y)$

Example $S = \{a, b, c, d, e\}$.

a	11
b	01
c	001
d	10
e	0001
f	0000

01 is a prefix of 011

011

11001000010

a c e d

How do we know it's not part of a longer sequence? Prefix code!

Prefix Codes

Frequencies: Each letter $x \in S$ has some frequency f_x.

fraction of characters in the text that are x.

$$\sum_{x \in S} f_x = 1$$

Given a prefix code γ and a message of length n,

$$\text{encoding length} = \sum_{x \in S} n f_x |\gamma(x)|$$

$ABL(\gamma)$ = average bits per letter $= \sum_{x \in S} f_x |\gamma(x)|$

Prefix Codes

γ_1

$a \longmapsto 11$
$b \longmapsto 01$
$c \longmapsto 001$
$d \longmapsto 10$
$e \longmapsto 000$

$f_a = .32$
$f_b = .25$
$f_c = .20$
$f_d = .18$
$f_e = .05$

$$ABL(\gamma_1) = 2 \cdot .32 + 2 \cdot .25 + 3 \cdot .20 + 2 \cdot .18 + 3 \cdot .05 = 2.25$$

γ_2

$a \longmapsto 11$
$b \longmapsto 10$
$c \longmapsto 01$
$d \longmapsto 001$
$e \longmapsto 000$

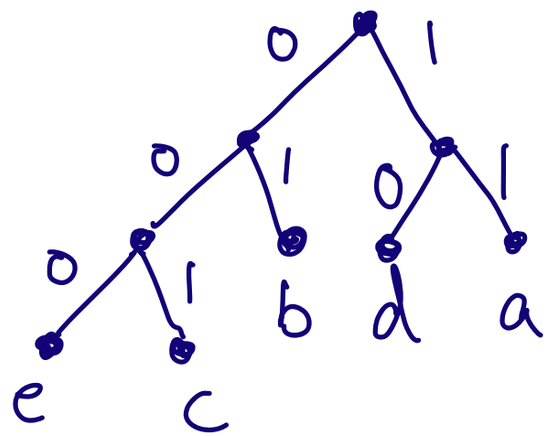
$$ABL(\gamma_2) = 2 \cdot .32 + 2 \cdot .25 + 2 \cdot .20 + 3 \cdot .18 + 3 \cdot .05 = 2.23$$

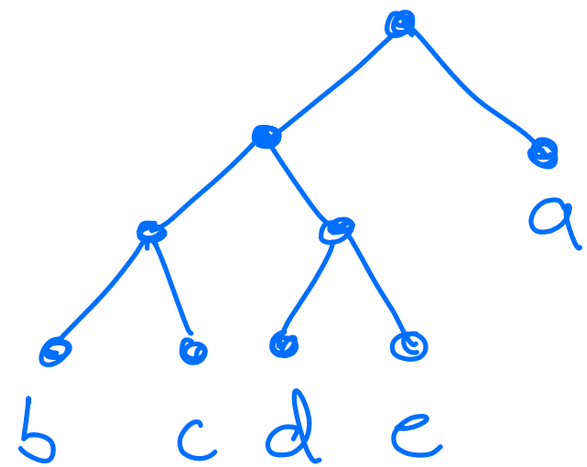
What is the optimal prefix code?

Prefix Codes and Binary Trees

γ_1

$a \longmapsto 11$
$b \longmapsto 01$
$c \longmapsto 001$
$d \longmapsto 10$
$e \longmapsto 000$





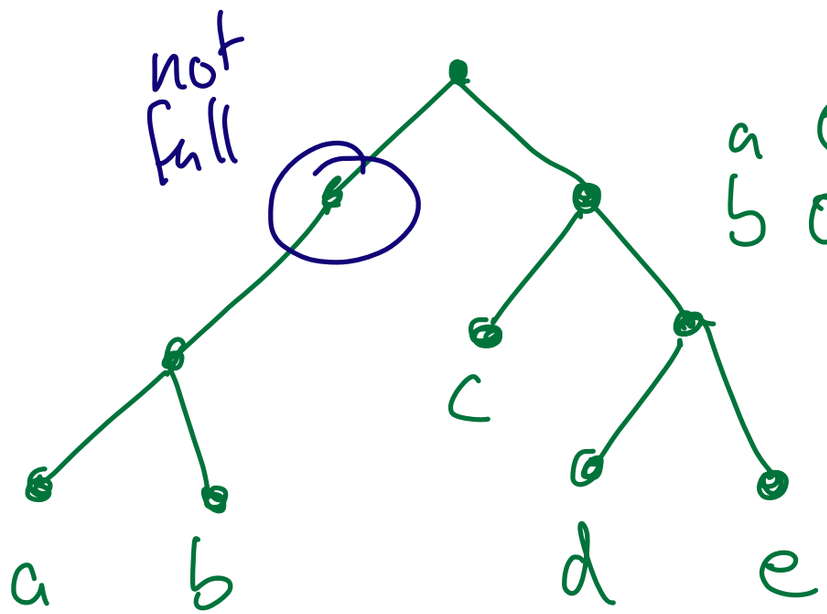
$a \longmapsto 1$
$b \longmapsto 000$
$c \longmapsto 001$
$d \longmapsto 010$
$e \longmapsto 011$

There is a 1-1 correspondence between prefix codes and binary trees with leaves labelled by S.

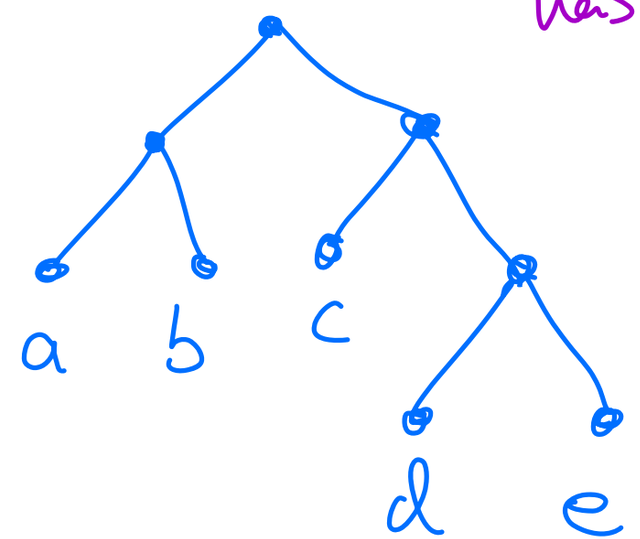
Optimal Prefix Codes

Lemma 1: The binary tree corresponding to an optimal prefix code is full.

every interior node has 2 children.



a 000
b 001

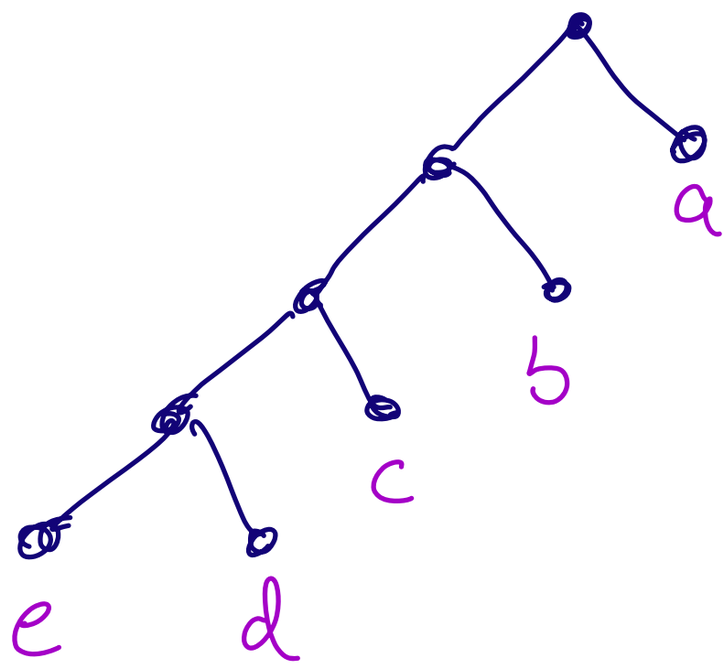


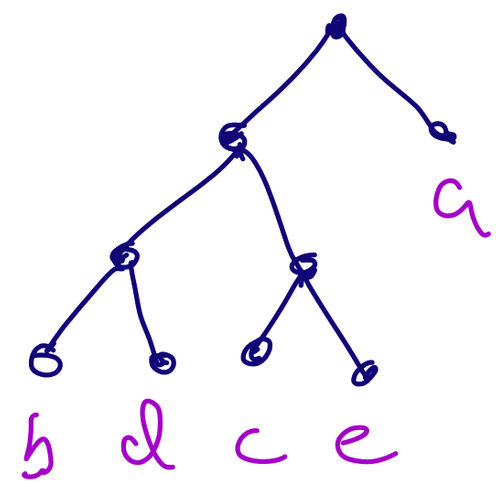
a 00
b 01

Optimal Prefix Codes

How would you label this tree?

$f_a = .32$
$f_b = .25$
$f_c = .20$
$f_d = .18$
$f_e = .05$

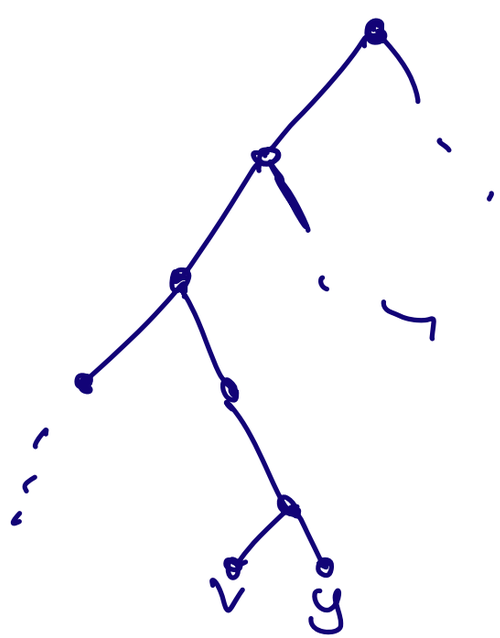




Huffman Codes

Start with the least frequent letter v.

Know v has maximum depth.
Know v has a sibling w. (Lemma 1)
Consider second least frequent letter y.
$\Rightarrow$ y has the same depth as v and w
$\Rightarrow$ we can relabel so that y and v are siblings without changing the ABL



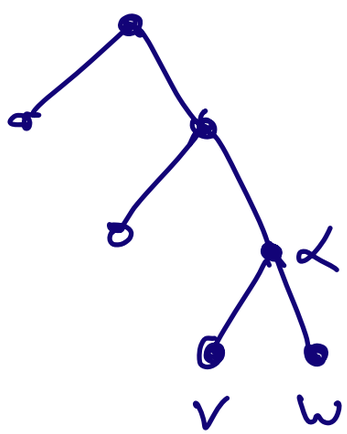
$\Rightarrow$ There exists an OPC where the two least frequent characters are siblings (of maximal depth)

Huffman Codes

Start with the two least frequent letters v and w

Create a new "letter" $\alpha = v|w$

$f_\alpha = f_v + f_w$

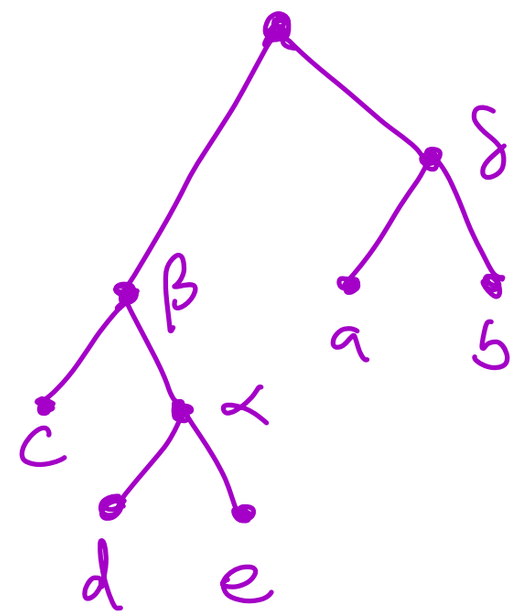


Recurse on $S' = S \setminus \{v, w\} \cup \{\alpha\}$.

Huffman Codes

$f_a = .32$
$f_b = .25$
$f_c = .20$
$f_d = .18$
$f_e = .05$

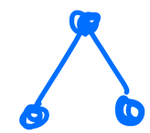
$f_\gamma = .57$
$f_\beta = .43$
$f_\alpha = .23$



Huffman Codes

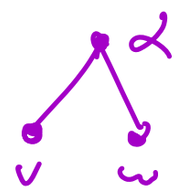
Theorem. Huffman code are optimal

Proof. Induction on $k = |S|$.

Base case $|S| = 2 \longrightarrow$  0 for one character, 1 for the other.

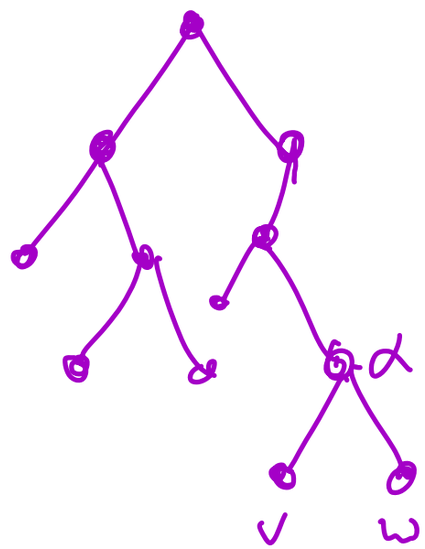
Assume that Huffman code for $|S| = k-1$ is optimal.

Do one step of Huffman code construction.

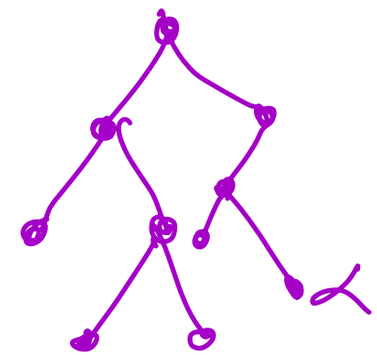


Encoding of $S' = S \setminus \{v, w\} \cup \{\alpha\}$ is optimal.

Let T be the tree for the encoding on S and T' be the tree for the encoding on S'.



encoding for S



encoding for S'