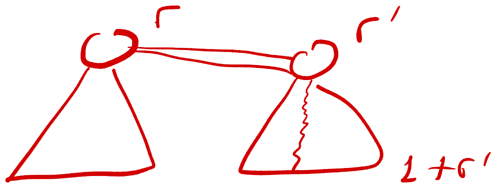
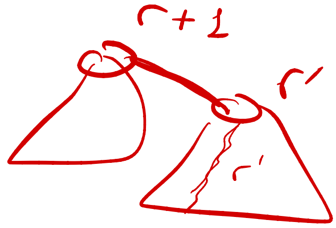
Data Compression



$r > r'$

$r \geq 1 + r'$



$r = r'$

if root of a tree has rank r
$\Rightarrow$ number of vertices in tree $\geq 2^r$

Minimum Connected Subgraph

Given $G = (V, E)$

undirected

connected

$w(u,v)$ for each $(u,v) \in E$

Want to find $S \subseteq E$

such that

- graph (V, S) is connected
- $\sum_{(a,b) \in S} w(a,b)$ minimized

Given

sequence of elements from a set V

$V = \{a, b, \ldots, z, \sqcup\}$

given "we hold these truths to be self."

---

use $\lceil \log_2 |V| \rceil$ bits for each symbol from V

---

file of n symbols $n \cdot \lceil \log_2 |V| \rceil$

---

want to use k bits for each symbol of V

2^k k-bit strings

$2^k \geq V$

$k \geq \lceil \log_2 V \rceil$

$V = \{a, b, c\}$

$k=1$ 0 1

Variable length encodings

a 0
b 000
c 010
d 011
e 1
f
⋮

c 010

aea 010

---

a → 01
b → 10
c → 001
d → 11
e → 000

prefix-free

1001110000001

An injective mapping

$$V \to \{0,1\}^K$$

is always prefix free

two different binary strings of same length

impossible one is prefix of other

$\left[\begin{array}{l} 0110 \\ 011011 \end{array}\right.$ not prefix-free

V fixed-length encoding

$\lceil \log_2 V \rceil$ per symbol

V sequence of length n

for each v

$f(v) = \#$ times that v appears in the sequence

suppose I find a variable-length encoding of V where each $v \in V$ is mapped to a bit string of length $\ell(v)$

Then sequence has encoding that uses $\sum_v f(v) \cdot \ell(v)$ bits

Given V, frequencies $f(v)$ for each $v \in V$ find a prefix-free encoding of V such that

$$\sum_v f(v)\ell(v)$$

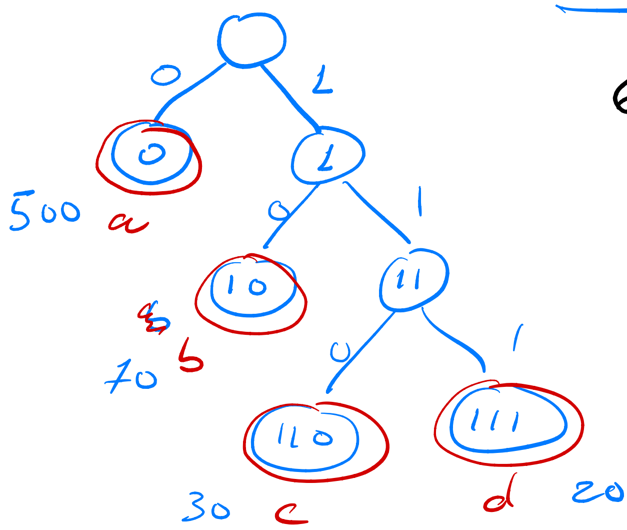
is minimized, where $\ell(v)$ is the number of bits in the encoding of v

V	f()	encoding	
a	500	0	00
b	70	10	01
c	30	110	10
d	20	111	11

bits in encoding

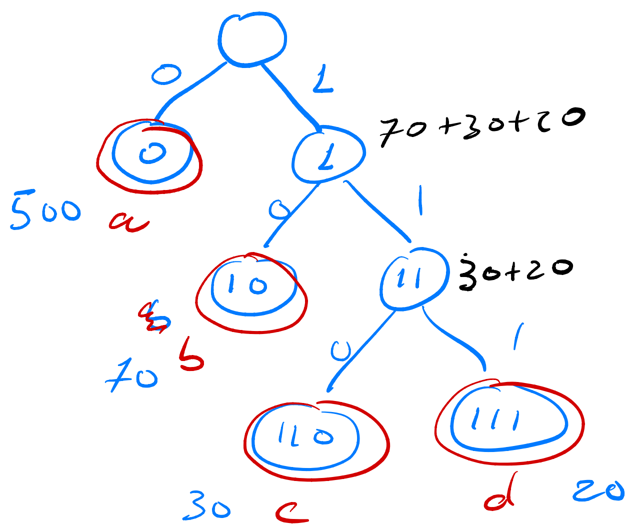
$$500 \cdot 1 + 70 \cdot 2 + 30 \cdot 3 + 20 \cdot 3 = 790$$

$$620 \cdot 2 = 1240$$



1 0 0 1 1 0 1 1 1 0 1 0 1 1 0 0

b a c d a b c d



$$
\begin{aligned}
&500\\
+\;&70+70\\
+\;&30+30+30\\
+\;&20+20+20
\end{aligned}
$$

$$500 + 120 + 70 + 50 + 30 + 20$$

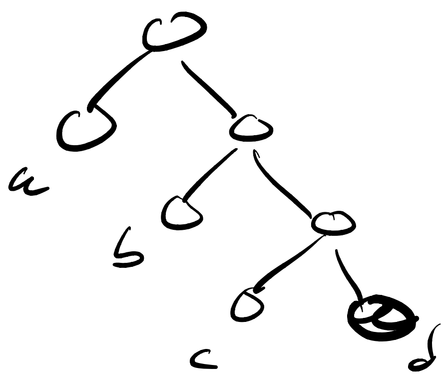
$$= 790$$

given $V = \{a, b, c, d\}$

f $\quad f(a) = 500 \quad f(b)$ --

---

want to construct a binary tree and associate elements of v to the leaves



associate to each node sum of $f(v)$ for all descendents of node

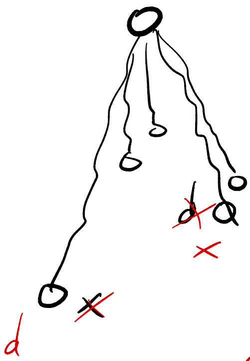
minimize sum of above over all non-root nodes a 500

b 70

c 30

d 20

An optimal tree exists in which d is a deepest leaf (has an encoding ≥ all other encodings)



$f(x) \geq f(d)$

$\ell(x) > \ell(d)$

switch x, d

before $\left(\sum_{v \neq x,d} f(v)\,\ell(v)\right) + f(x)\ell(x) + f(d)\ell(d)$

after $\left(\sum_{v \neq x,d} f(v)\,\ell(v)\right) + f(x)\ell(d) + f(d)\ell(x)$

before − after

$$= f(x)\,\ell(x) + f(d)\,\ell(d)$$

$$- f(x)\,\ell(d) - f(d)\,\ell(x)$$

$$= f(x)\big(\ell(x) - \ell(d)\big)$$

$$- f(d)\big(\ell(x) - \ell(d)\big)$$

$$= \underbrace{\big(f(x) - f(d)\big)}_{\geq 0}\underbrace{\big(\ell(x) - \ell(d)\big)}_{>0}$$

$$\geq 0$$

before $\geq$ after

V set of symbols

f frequencies

z is an element of V with smallest $f(\cdot)$

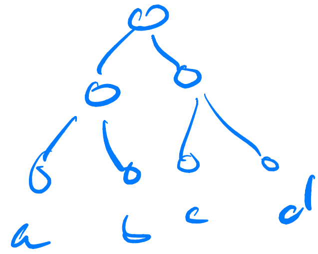
---

minimizes $\sum_{v \in V} f(v) \ell(v)$

Then there is a minimal prefix-free encoding of V such that z has a longest encoding, that is

$$\forall v \in V \quad \ell(z) \geq \ell(v)$$

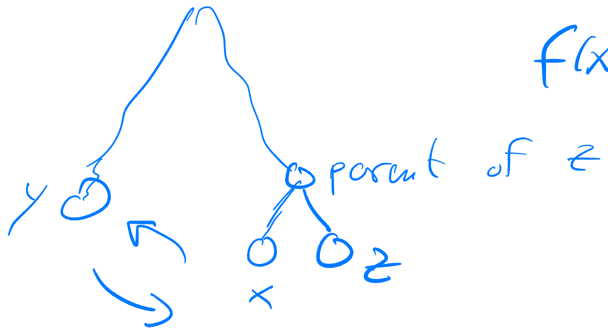
a	100
b	100
c	100
d	100



Suppose y, z two least frequent elements of V [$\forall x \in V - \{y, z\}$ $f(x) \geq f(y)$ and $f(x) \geq f(z)$]

There is an optimal solution in which y, z are siblings and longest

From before, there is optimal solution in which z deepest



$f(x) \geq f(y)$

$\ell(x) \geq \ell(y)$

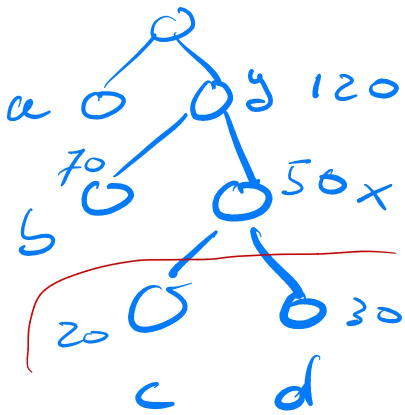
exchange x, y solution also optimal

$\geq$ as good

In optimal solution every non-leaf node has two children

"merge" a and b

a	500
b	70
c	30
d	20



a	500
b	70
x	50

a	500
y	120