3/4/21

Huffman's encoding

prefix codes binary

ASCII [ ] 8 bits
A = 64

Englisih
E, A > 10%
X, Z, Q 1%

Z: 215
E: 74809
A: 48,000

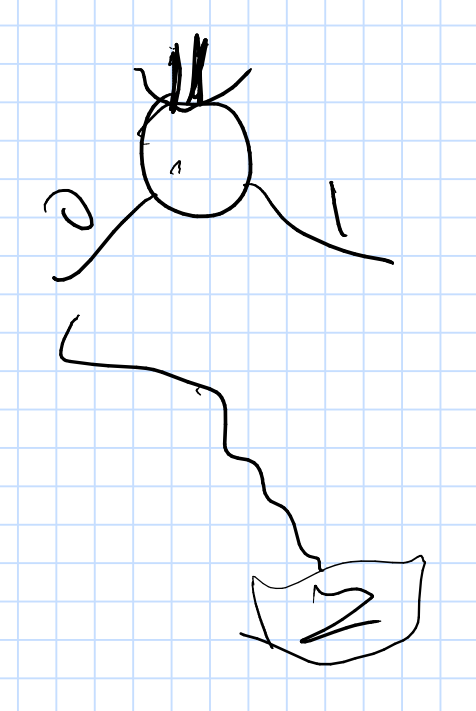
E 011
Z 111101000

ABRA CADABRA

[ ][ ] . . . - - - [ ] encoded

a B
011,0111 10111 ~~~
e

prefix code w_i: code word
1, 2, . . . , n
$w_1, w_2, \ldots, w_n \in \{0,1\}^*$
There is node code word that is prefix of another codword.] prefix free code



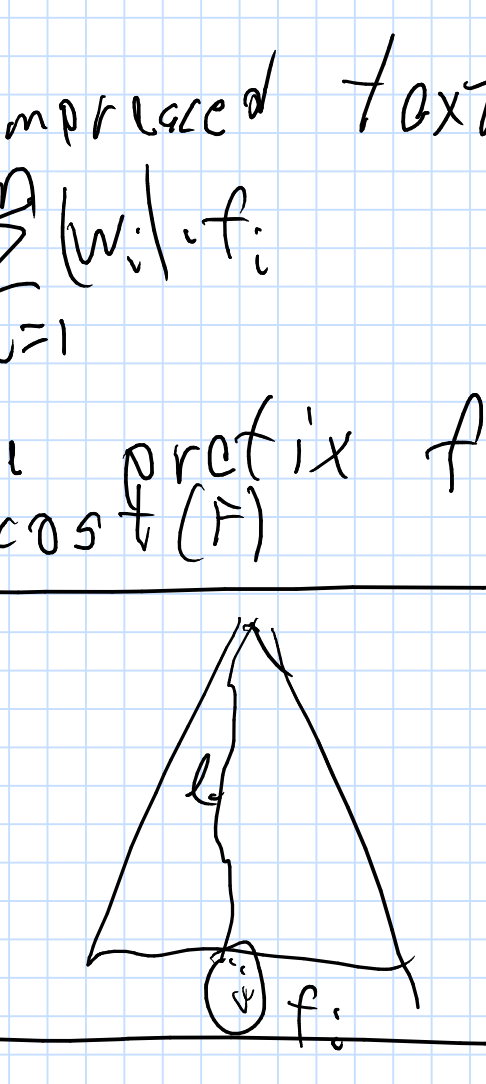
0111 0 . . .
Z

$f[i]$ = frequency of letter i.

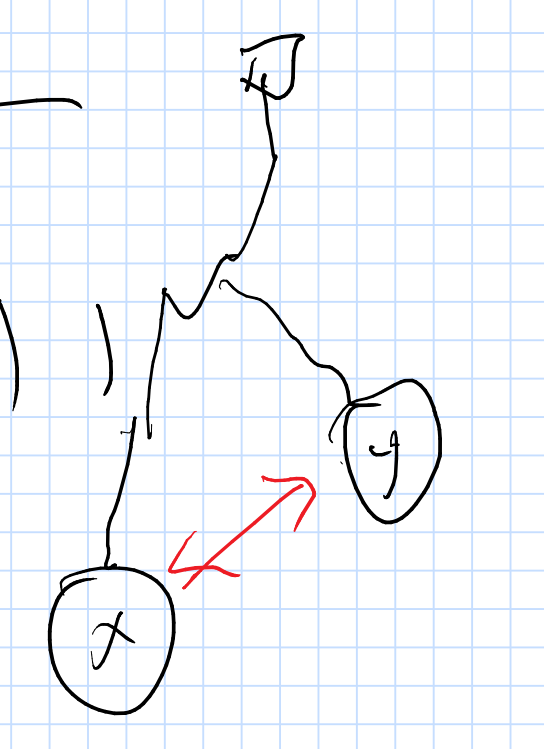
C: code $\begin{cases} w_1 & l_1 = |w_1| \text{ in bits} \\ w_2 & l_2 \\ \vdots & \vdots \\ w_n & l_n \end{cases}$ $F = \{f_1, f_2, \ldots, f_n\}$

size compressed text $(F) \equiv cost(F)$
$= \sum_{i=1}^{n} |w_i| \cdot f_i$

kompute prefix free code that min cost(F)



property 1



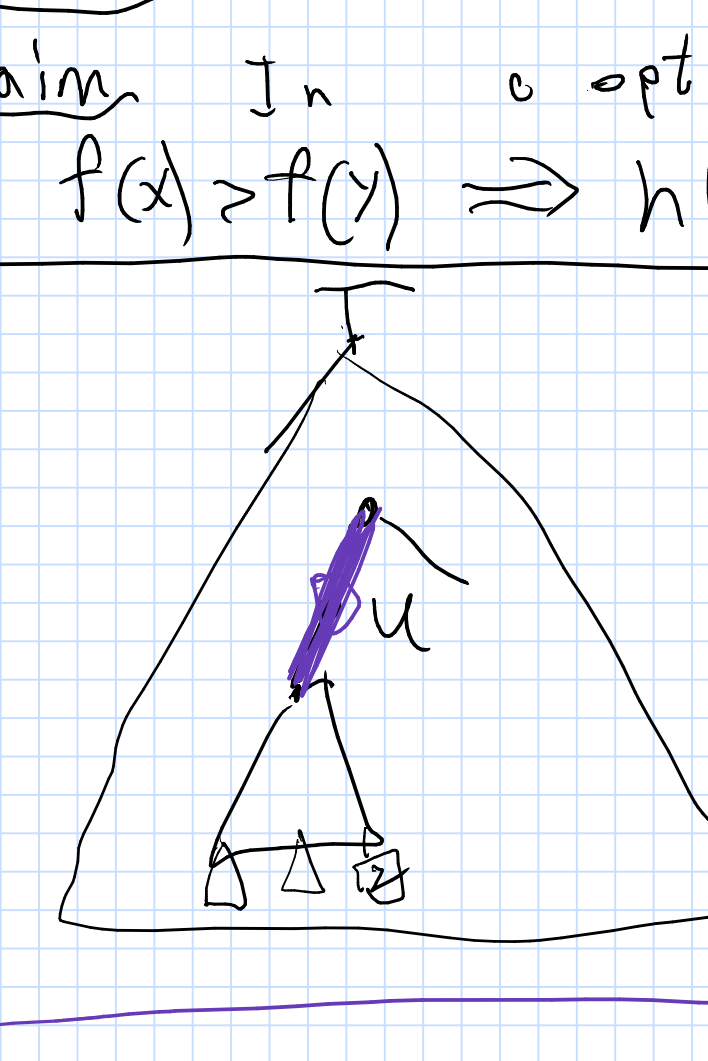
$h(x) > h(y)$ optimal code tree.

$\Rightarrow f(x) < f(y)$
If false $f(x) > f(y)$.

$f(x) \cdot h(x) + f(y) \cdot h(y)$ $h(x) > h(y)$
$f(x) h(y) + f(y) h(x)$ $f(x) > f(y)$

proved:

claim In a optimal
if $f(x) > f(y) \Rightarrow h(x) \leq h(y)$.



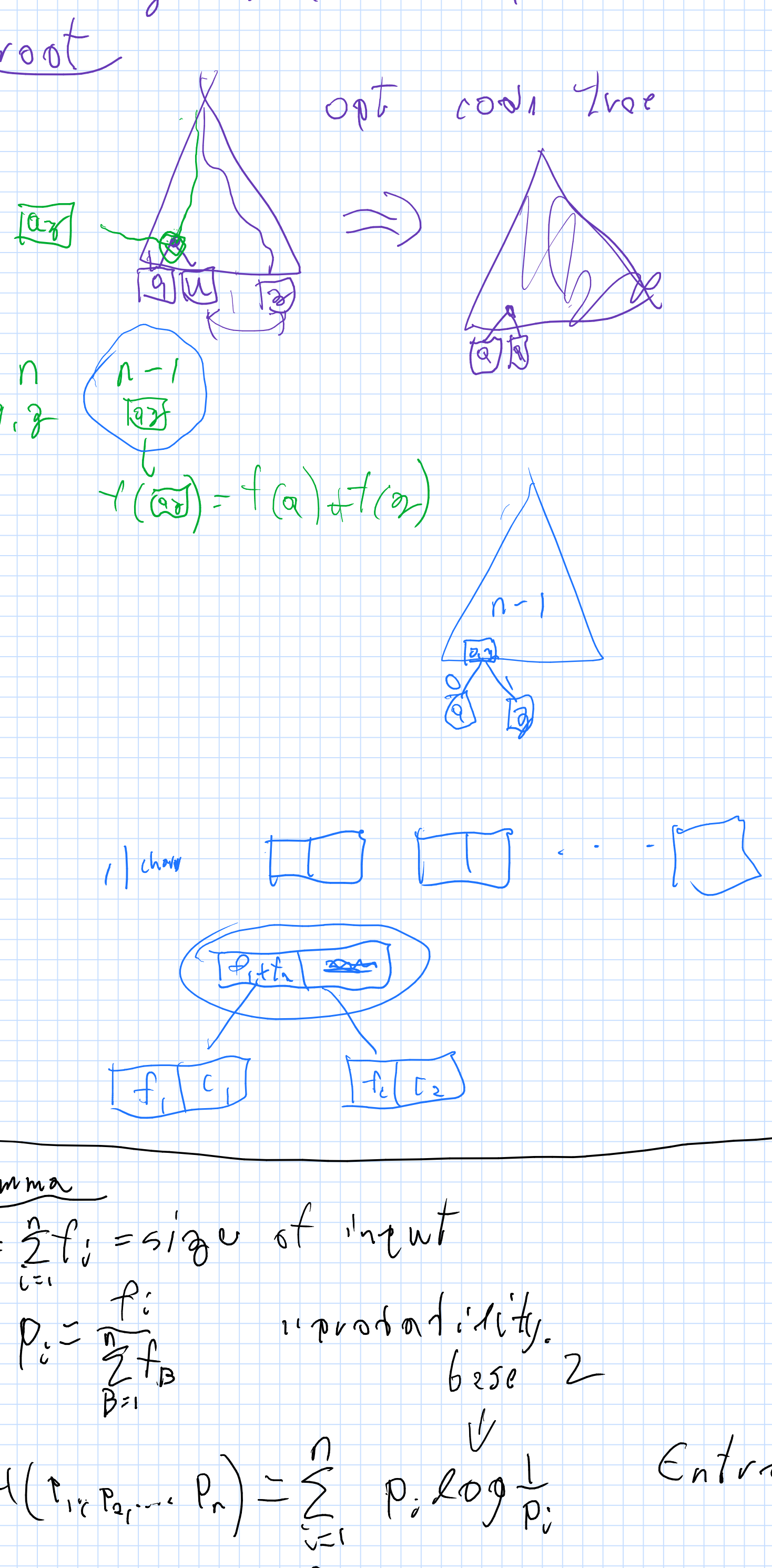
delete u

Lemma Let q and z be the two characters such that $f(q) \leq f(z) \leq$ frequency of any other character

$\Rightarrow \exists$ optimal code tree T such that
- q and z are siblings
- q is the deepost leaf.
- z is also deepost leaf.

proof

opt code tree



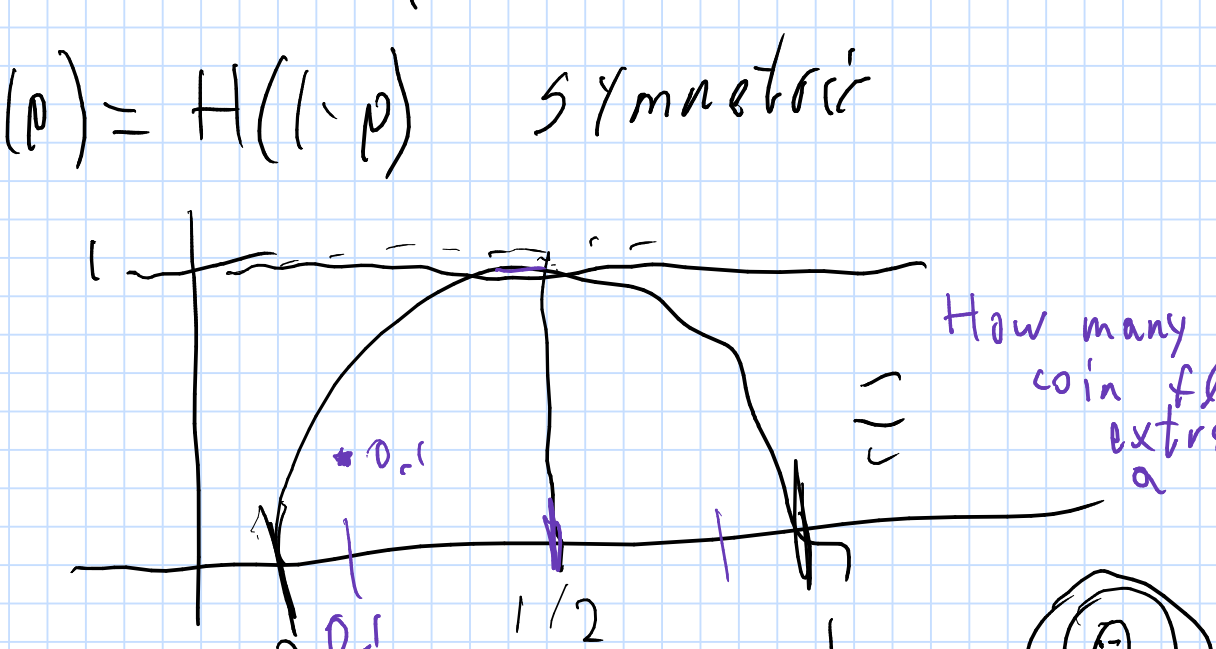
Lemma
$\alpha = \sum_{i=1}^{n} f_i$ = size of input
$P_i = \frac{f_i}{\sum_{B=1}^{n} f_B}$ probability.

base 2

$H(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$ Entropy
$\sum_{i=1}^{n} - p_i \log p_i$

$H(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ binary entropy

$H(p) = H(1-p)$ symmetric



10% probability heads

= How many unbiased coin flips can be extracted from a single flip of the biased coin

$p \equiv$ probability for head

$p(1-p)^9 \binom{10}{1} = 0.1 \cdot 0.9^9 \cdot 10 \neq \frac{1}{2}$

0	01
1	10
repeat	00
	11

$0.1 \cdot 0.9 = 0.09$

0.18

$2 \cdot \frac{1}{0.18} \geq 10 \ldots$

Lemma
Given n characters with probabilities $p_1, p_2, \ldots, p_n$ such that
p_i is a power of 2
$p_i = 1/2^{t_i}$

Then, the average cost of encoding the input using Hoffman encoding is the entropy:

$E\left[\begin{matrix}\text{expected} \\ \text{len code} \\ \text{rd}\end{matrix}\right] = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$ $p_i = \frac{1}{2^{t_i}}$ t_i

$\Leftrightarrow$ the ith character with probability $p_i = \frac{1}{2^{t_i}}$ has a word of length t_i.

proof
By induction:
n=2 $p_1 = p_2 = 1/2$

$t_1 = t_2 = 1$

n>2 Let 1 and 2 be the two characters of least probability.

Then $p_1 = p_2 \Rightarrow$

0.0 1
0.00 . . . 0 1
0.001
0.1
0.001