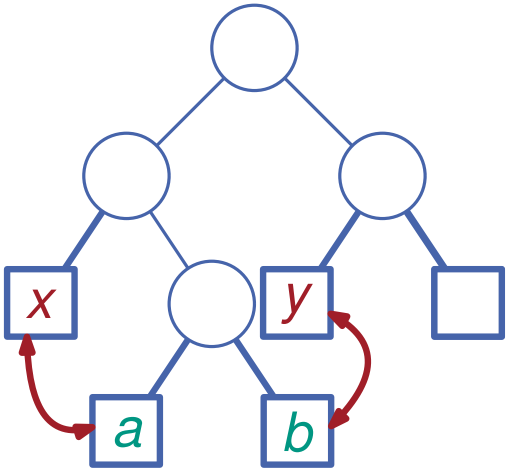
COSC 302, Spring 2018
Lecture 9: Greedy Algorithms, Huffman Codes

Prof. Darren Strash

Department of Computer Science
Colgate University



Greedy Algorithms: Huffman Codes

(Prefix) codes

Given: Characters $c_1, c_2, \ldots, c_n$.

And: Frequencies of occurrence $f_1, f_2, \ldots, f_n$.

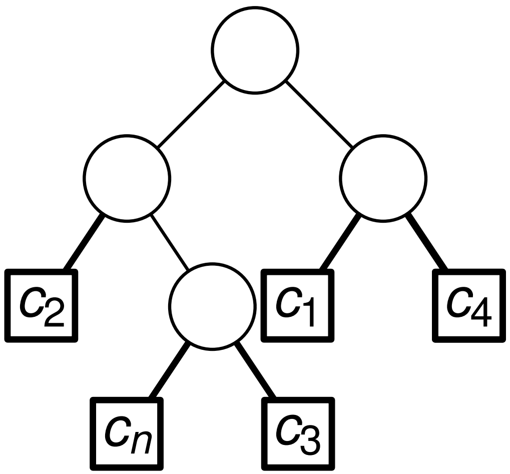
Goal: Compute a binary code of length l_i for each character c_i to minimize the total size of encoding a text.

Minimize:

$$\sum_{i=1}^{n} f_i \cdot l_i$$

and no code should be a prefix of another code.

$\rightarrow$ Binary tree representation



Huffman's Algorithm

Repeat until one character remains:

- Select two characters x, y with smallest frequency. Make them *sibling leaves*.
- Collapse x, y into a single 'meta' character z with frequency $f_z = f_x + f_y$

Huffman's Algorithm

Repeat until one character remains:

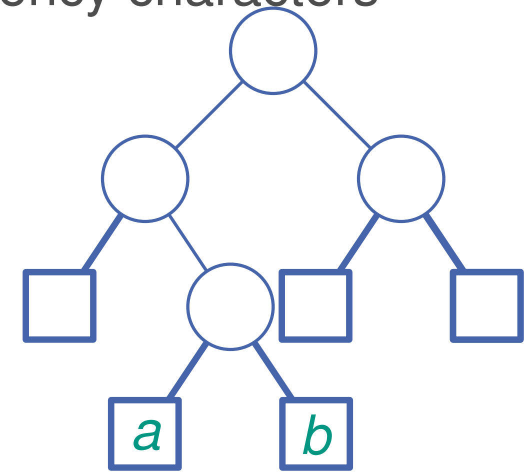
- Select two characters x, y with smallest frequency. Make them *sibling leaves*.
- Collapse x, y into a single 'meta' character z with frequency $f_z = f_x + f_y$

Locally optimal choice is globally optimal $\rightarrow$ Greedy algorithm

Greedy Choice Property

Why can we make lowest frequency characters into sibling leaves?

$\rightarrow$ There **is** an optimal code with lowest frequency characters as sibling leaves.



$\rightarrow$ Let T be an optimal code, such that $\sum_i f_i l_i$ is minimized.

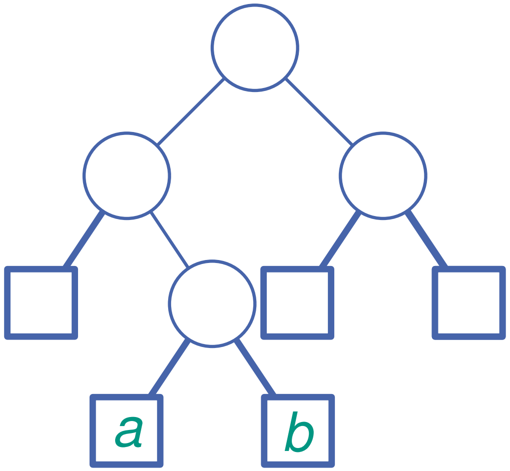
Let characters *a* and *b* be sibling leaves with largest depth in T.

Either they are the two smallest frequency characters or not.

If yes, we are done.

Greedy Choice Property

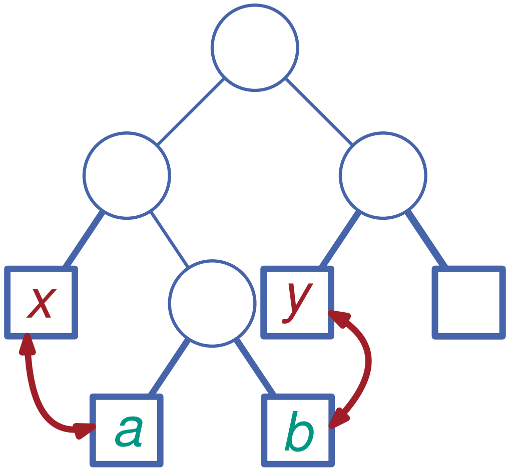
Otherwise, we show that characters with lowest frequency can be swapped with *a*, *b*.



Greedy Choice Property

Otherwise, we show that characters with lowest frequency can be swapped with *a*, *b*.

Let $x < y$ be two characters with smallest (and second smallest) frequency.

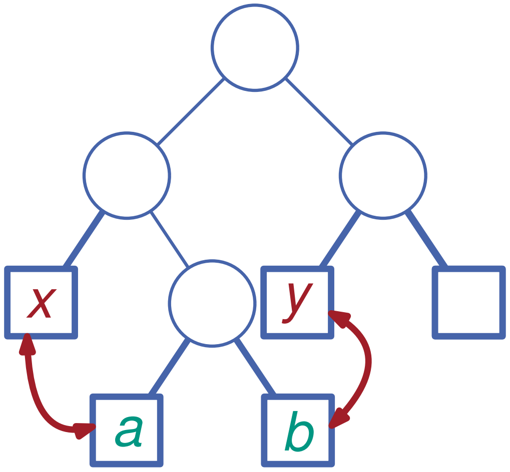


Greedy Choice Property

Otherwise, we show that characters with lowest frequency can be swapped with *a*, *b*.

Let $x < y$ be two characters with smallest (and second smallest) frequency.

Then the $\sum_i f_i l_i$ must remains the same!



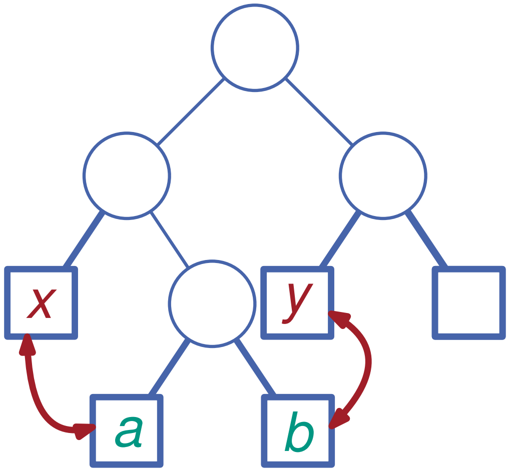
Greedy Choice Property

Otherwise, we show that characters with lowest frequency can be swapped with a, b.

Let $x < y$ be two characters with smallest (and second smallest) frequency.

Then the $\sum_i f_i l_i$ must remains the same!

New code length after swapping x and a?



$$\begin{aligned}
\sum_i f_i l_i \ &- f_a l_a + - f_x l_x + f_x l_a + f_a l_x \\
&= \sum_i f_i l_i \ + l_a(f_x - f_a) - l_x(f_x - f_a) \\
&= \sum_i f_i l_i \ + (l_a - l_x)(f_x - f_a) \\
&\leq \sum_i f_i l_i
\end{aligned}$$

Greedy Choice Property

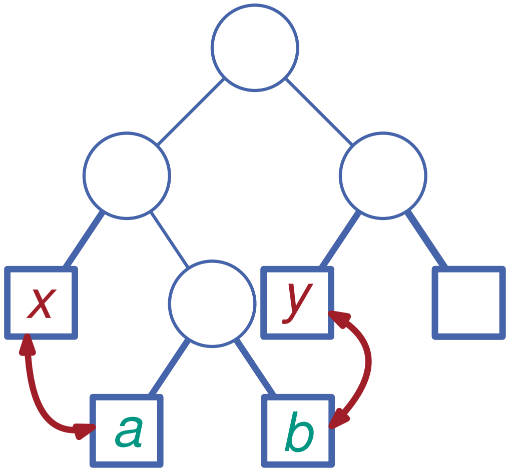
Otherwise, we show that characters with lowest frequency can be swapped with a, b.

Let $x < y$ be two characters with smallest (and second smallest) frequency.

Then the $\sum_i f_i l_i$ must remains the same!

New code length after swapping x and a?

$$\begin{aligned}
&\sum_i f_i l_i \;\; - f_a l_a + - f_x l_x + f_x l_a + f_a l_x \\
&\quad = \sum_i f_i l_i \;\; + l_a(f_x - f_a) - l_x(f_x - f_a) \\
&\quad = \sum_i f_i l_i \;\; + (l_a - l_x)(f_x - f_a) \\
&\quad \le \sum_i f_i l_i
\end{aligned}$$



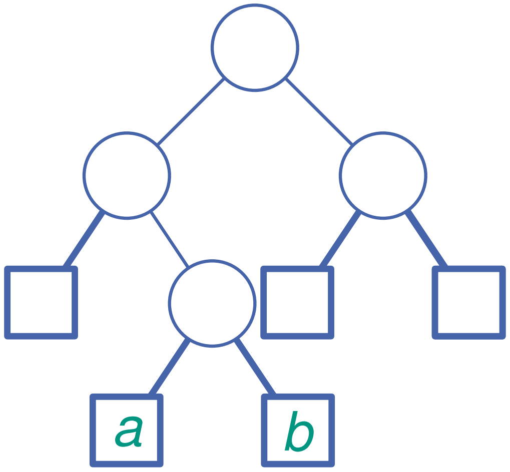
Repeat argument for b and y

Optimal Substructure

Why can we continue this procedure and get an optimal solution?

$\rightarrow$ Collapsing characters into one meta character gives us a subproblem.

$\rightarrow$ Must show that solving this subproblem optimally also optimally solves our original problem.



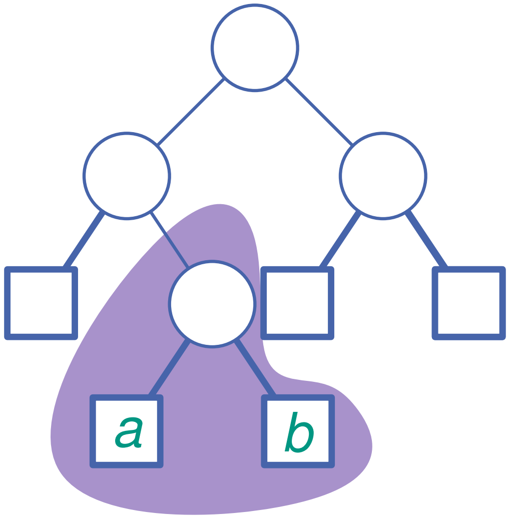
Let *a* and *b* be two characters with smallest (and second smallest) frequency.

Optimal Substructure

Why can we continue this procedure and get an optimal solution?

$\rightarrow$ Collapsing characters into one meta character gives us a subproblem.

$\rightarrow$ Must show that solving this subproblem optimally also optimally solves our original problem.



$z \quad f_z = f_a + f_b$

Let *a* and *b* be two characters with smallest (and second smallest) frequency.

Show that an optimal solution to total problem requires optimal solution to subproblem with z

Optimal Substructure

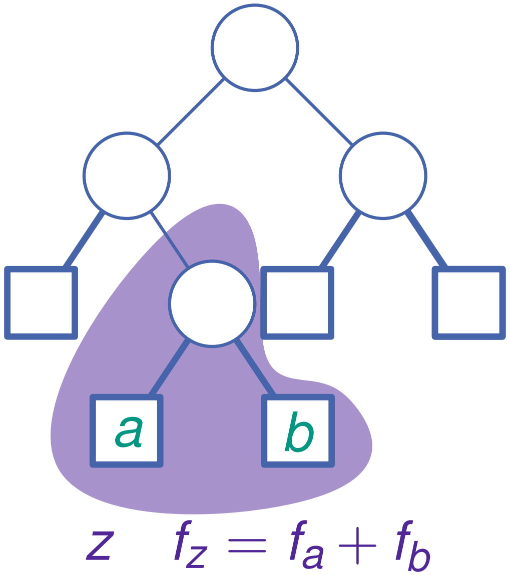
Let T be the minimum cost for the total problem (with a and b), and S be the cost for the subproblem (with z)

$$T = \sum_i f_i l_i = \sum_{i \neq a,b} f_i l_i + f_a l_a + f_b l_b$$

$$= \sum_{i \neq a,b} f_i l_i + (l_a - 1)(f_a + f_b) + f_a + f_b$$

$$= \sum_{i \neq a,b} f_i l_i + l_z(f_z) + f_a + f_b$$

$$= S + f_a + f_b$$



$\therefore$ if S is not minimized, then T is not minimized.