Greedy Algorithms

→ locally optimal, near-sighted, immediate benefit
→ sometimes intuitively clear, sometimes not.
→ correctness of the algorithm is most often not obvious, and thus requires a concrete argument

Classic Example: minimum spanning tree

Interval Scheduling: [Kleinberg Tardos Chapter 4]

Given a set of intervals (on real number line) find the largest subset of disjoint intervals.

Example: → (1, 5), (0, 3), (2, 4), (4, 6), (3, 4), (5, 7)

Application: resource allocation requests for fixed intervals of time

• resource can be allocated to only one request at a time
• no partial allocation
• cater to maximum number of requests.

Searching through all possible solutions ≈ \(2^n\) solutions
Greedy Strategy 1: keep choosing the smallest interval (and removing those which intersect with chosen ones)

Greedy Strategy 2: earliest starting time

Greedy Strategy 3: keep choosing the interval with smallest no. of overlaps.

Greedy Strategy 4: keep choosing the interval with earliest ending time
Intuitively, you are maximizing the remaining time, thus maximizing the possible number of requests catered afterwards.

Claim: Greedy strategy 4 always gives an optimal solution.

General Framework of argument:

1. Show that there exists an optimal solution that agrees with the first greedy step.
2. The rest of the argument will work inductively.

Consider an optimal solution $I_1, I_2, I_3, \ldots, I_k$.

Earliest ending interval $I_0$

\[ I_0 \quad I_1 \quad I_2 \quad \ldots \quad I_k \]

Swap $I_0$ with $I_1$.

Consider a new solution $I_0, I_2, I_3, \ldots, I_k$.

Claim: It is valid solution.

\[
\text{endtime}(I_0) < \text{endtime}(I_1) < \text{starttime}(I_2)
\]

\[
\Rightarrow I_0 \text{ does not overlap with } I_2, I_3, \ldots, I_k
\]

$I_0, I_2, I_3, \ldots, I_k$ is an optimal solution agreeing with first greedy step.
Inductive proof based on the number of intervals.

**Inductive Hypothesis:** Greedy algo 4 works for any instance with up to \( n-1 \) intervals.

**Inductive Step:** It works for all instances with \( n \) intervals.

**Base Case:** \( n = 1 \). Obvious.

**Input:** \( I = \{ I_1, I_2, I_3, \ldots, I_n \} \)

For convenience, assume end time \( (I_1) \leq \text{end time} \ (I_2) \leq \cdots \).

Algorithm chooses \( I_1 \), then removes all intervals overlapping with \( I_1 \).

\[ \rightarrow I' = I - \{ \text{intervals overlapping with } I_1 \} \]

Recursively applying the same algorithm on \( I' \).

By the inductive hypothesis, we know that the algorithm gives optimal solution for \( I' \).

That is, \( \text{Algo} (I') = \text{OPT} (I') \).

**Output:** \( I_1 + \text{Algo} (I') = I_1 + \text{OPT} (I') \).

**Claim:** \( (I_1 + \text{any optimal solution for } I') \)

is an optimal solution for \( I \).
Claim: $I_1 + \text{OPT}(X')$ is an optimal solution for $X$.

Proof: Recall that we showed that there exist an optimal solution for $X$ that contains $I_1$.

Let it be $I_1, I_1, I_2, \ldots, I_k$.

Clearly, $I_1, I_2, \ldots, I_k$ are disjoint from $I_1$.

Hence, $\{I_1, I_2, \ldots, I_k\}$ is a valid solution for $X'$.

\[\text{OPT}(X') \geq \{I_1, I_1, I_2, \ldots, I_k\}\]

\[\Rightarrow |I_1 + \text{OPT}(X')| \geq \{I_1, I_1, I_2, \ldots, I_k\}|\]

optimal for $X = |\text{OPT}(X)|$

Pseudocode

Input: $(s_1, f_1), (s_2, f_2), \ldots, (s_n, f_n)$

- Sort according to $f_j$.

$f = -\infty$ (finish time of latest interval selected so far)

for ($i = 1 \text{ to } n$)

- if ($s_i \geq f$) then

  select $(s_i, f_i)$

  $f \leftarrow f_i$
HW Assignments

deadlines $d_1, d_2, \ldots, d_n$
time required $l_1, l_2, \ldots, l_n$

\[ \text{Lateness of } i\text{-th assignment } = (t_i - d_i) \text{ if } t_i > d_i \]

Minimize maximum lateness over all assignments.

→ smallest length first?
→ earliest deadline first?
→ minimum $d_i - l_i$ first?

Example:

<table>
<thead>
<tr>
<th>A_1</th>
<th>A_2</th>
<th>A_3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A_1</th>
<th>A_2</th>
<th>A_3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lateness</th>
<th>Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$A_3 A_2 A_1$</td>
</tr>
<tr>
<td>1</td>
<td>$A_2 A_1 A_3$</td>
</tr>
<tr>
<td>0</td>
<td>$A_1 A_2 A_3$</td>
</tr>
</tbody>
</table>

Red shows maximum lateness.
Best order is $A_1 A_2 A_3$.

Interval scheduling: cater to all requests with using minimum number of servers.
(one server can cater to one request at a time)

Equivalently, finding the minimum number of platforms required for a set of trains stopping at a station.
Interval Scheduling: another variant

- select a set of disjoint intervals to maximize the total length of selected intervals.

For example, there might be a profit proportional to the duration of use.

Greedy Strategy 1: keep picking the longest length intervals

Greedy Strategy 2: earliest starting time
Dynamic Programming:

Recursion with better memory management.

**General Idea:** try to categorize the possible solutions into different types. For each type, find the optimal solution via recursion. Then compare the various types of optimal solutions with each other.

Interval scheduling with maximum total length:

\[ I = \{ I_1, I_2, I_3, \ldots, I_n \} \]

Two kinds of solutions:

- contain \( I_1 \)
- don't contain \( I_1 \)

Can we find optimal from both the kinds recursively?

\[ I' = I - \text{overlap}(I_1) \]

\[ \text{OPT}(I') + I_1 \]

\[ I'' = I - I_1 \]

\[ \text{OPT}(I'') \]
$$\text{OPT}(A) = \max \{ I_1 + \text{OPT}(I') , \text{OPT}(I'') \}$$

Example

$$I = \{ (0,5), (1,8), (5,7), (4,9), (3,4) \}$$

$$I_1 = (0,5)$$

$$I' = \{ (5,7) \}$$

$$I'' = \{ (1,8), (5,7), (4,9), (3,4) \}$$

$$I \rightarrow \{(1,8), (5,7), (4,9), (3,4)\}$$

$$\emptyset \rightarrow \{(5,7), (4,9), (3,4)\}$$

$$\{(3,4)\} \rightarrow \{(4,9), (3,4)\}$$

$$\{(3,4)\}$$

Nothing clever here. We are simply trying to go over all possible solutions recursively.

No. of recursive calls seems to be growing exponentially because each call makes two new recursive calls.

Efficient implementation possible if no. of total distinct recursive calls is small.

$$I = \{ (1,3), (11,13), (21,23), (31,33), (2,4), (12,14), (22,24), (32,34) \}$$

$$I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8$$
HW work out the recursion tree and figure out whether the number of recursive calls is growing exponentially or polynomially.

\[ \mathcal{I} = \left\{ \begin{array}{c}
(1,3) \quad (11,13) \quad (21,23) \quad (31,33) \\
\mathcal{I}_1 \quad \mathcal{I}_2 \quad \mathcal{I}_3 \quad \mathcal{I}_4 \\
(2,4) \quad (12,14) \quad (22,24) \quad (32,34) \\
\mathcal{I}_5 \quad \mathcal{I}_6 \quad \mathcal{I}_7 \quad \mathcal{I}_8
\end{array} \right\} \]

If we have \( 2n \) intervals, we will have at least \( 2^n \) distinct recursive calls or subproblems.
\[ \mathcal{I} = \{ (1,3), (2,4), (11,13), (12,14), (21,23), (22,24), (31,33), (32,34) \} \]

Only \(2n\) distinct recursive calls or subproblems.

It seems if the intervals are arranged in a particular order, the number of distinct subproblems will be small.

Possible orders:
1. Starting time
2. Ending time
sort the intervals in increasing order of starting time

$I_1, I_2, I_3, \ldots, I_n$

Claim: The input set of intervals for any recursive call will look like

$\{I_j, I_{j+1}, I_{j+2}, \ldots, I_n\}$

(as opposed to an arbitrary subset of intervals)

This is because when we remove intervals overlapping with $I_k$, the remaining set is simply

all intervals with starting time $> \text{end-time}(I_k)$.

No. of distinct subproblems $\leq n$. 
Recursive Implementation:

\[
\text{Opt} \leftarrow \text{array with all zeros.}
\]
\[
\text{Opt}[n] \leftarrow \text{length}(\text{In})
\]

Output \( \text{ALG}(i) \):

\[
\text{ALG}(j): \quad \text{// ALG}(j) \text{ computes optimal solution for } \{I_j, I_{j+1}, \ldots, I_n\}
\]

\[
\text{if } \text{Opt}[j] > 0 \text{ return } \text{Opt}[j]
\]

\quad \text{// means already solved and stored.}

\[
\text{else}
\]

\[
\text{Opt}[j] \leftarrow \max \begin{cases} 
\text{ALG}(j+1) \\
\text{length}(I_j) + \text{ALG}(p(j))
\end{cases}
\]

\[
\text{return } \text{Opt}[j]
\]

Iterative Implementation:

\[
\text{Opt} \leftarrow \text{array with all zeros.}
\]
\[
\text{Opt}[n] \leftarrow \text{length}(\text{In})
\]

for (j = n-1 to 1)

\[
\text{Opt}[j] \leftarrow \max \left\{ \begin{array}{l}
\text{Opt}[j+1] \\
\text{length}(I_j) + \text{Opt}[p(j)]
\end{array} \right. 
\]
Add code to compute the optimal set of intervals.

Conclusion: Order of processing the input is important.

Intervals: Order of starting time/ending time

Sequences: left to right

Try to ensure that no. of distinct subproblems is small.

Maximum travel in a day - \( d \).

Night stay prices - \( p_1, p_2, \ldots \).

Minimize total stay cost during the journey.

\[ O(n) \]

No. of distinct subproblems = \( n \).

\[ \begin{array}{c}
2 \ 5 \ 3 \ 1 \ 2 \ 5 \ 3 \ 1 \ 3 \ 1 \ 2 \ 4 \ 1 \\
\end{array} \]

\( A = b a c b c b a b c a c b a \)

\( B = b c b c a \)

Match \( B \) inside \( A \) with minimum cost

\[ O(mn) \]
Subset sum problem

Given set of integers \( A = \{a_1, a_2, a_3, \ldots, a_n\} \)

is there a subset with sum zero?

Example:
\[
2, -5, 1, 7, -4, -6
\]
\[
\{ 2, 7, -5, -4 \}
\]

Solutions

containing an

not containing an

is there a subset of \( \{2, -5, 1, 7, -4\} \)

with sum zero?

is there a subset of \( \{2, -5, 1, 7, -4\} \)

with sum six?

Subproblem \( A_j = \{a_1, \ldots, a_j\} \), number \( N \)

is there a subset of \( A_j \) with sum \( N \).

No. of distinct subproblems = \( n \times \sum_{i} |a_i| \)

\( \text{ALG}(j, N) : \)

\( \rightarrow \text{ALG}(j-1, N) \) OR \( \text{ALG}(j-1, N-a_j) \)

Pseudo polynomial time.

A polynomial time algorithm is supposed to take time \( \text{poly}(n, \text{no. of bits in } a_1, a_2, \ldots) \).
Subset Sum Pseudocode

\( S \leftarrow \) Boolean two-dim array of size \( n \times (\sum |a_i| + 1) \)

\( S[j, k] \) will denote whether there is a subset of first \( j \) numbers with sum = \( k \).

Say, range of \( j \) \( 1 \leq j \leq n \)

and range of \( k \) \( \text{neg} \leq k \leq \text{pos} \)

\( \downarrow \)

Sum of all negative numbers \( \rightarrow \) Sum of all positive numbers

Initialization:

\( S[1, k] \leftarrow \begin{cases} \text{True} & \text{for } k = a_1 \\ \text{False} & \text{for all other values of } k \end{cases} \)

for \( (j = 2 \text{ to } n) \)

for \( (\text{neg} \leq k \leq \text{pos}) \)

\( S[j, k] \leftarrow S[j-1, k] \text{ OR } S[j-1, k-a_j] \)

To construct the desired set

Assume \( S[n, 0] \) is true.

\( \text{sum} \leftarrow 0 \)

for \( (j = n \text{ to } 1) \)

if \( S[j-1, \text{sum}] = \text{true} \), don't take \( a_j \)

else if \( S[j-1, \text{sum} - a_j] = \text{true} \), take \( a_j \) and \( \text{sum} \leftarrow \text{sum} - a_j \)
Knapsack problem

n objects weights \( w_1, w_2, \ldots, w_n \)
Values \( v_1, v_2, \ldots, v_n \)

Select a subset \( S \) s.t.

\[
\sum_{i \in S} w_i \leq W
\]

and the total value \( \sum_{i \in S} v_i \) is maximized.

---------------

| Sep 6 |

Balanced margins

\[ W = 68 \quad \rightarrow \quad \text{Slack} \]

Suppose there are ten tourist guides, of which, six can speak French and six can speak German (two can speak both French and German). Everyone comes with their own charges. We want to select five.

\[
\begin{align*}
\text{average slack} & \quad 14/3 = 4.6 \\
0 & \quad 8 \\
6 & \quad 6
\end{align*}
\]

We are given a sequence of words of lengths \( 21 \)

\( w_1, w_2, w_3 \ldots w_n \).

Each line can have at most \( W \) characters.

\[
\overline{w_i}, \overline{w_{i+1}}, \ldots, \overline{w_j}
\]

If a line has from the \( i \)-th to \( j \)-th word then its slack is defined to be

\[
S = W - \left[ (w_i + 1) + (w_{i+1} + 1) + \cdots + (w_{j-1} + 1) + w_j \right]
\]
Variance = \( \text{avg}(l_1^2, l_2^2, \ldots, l_k^2) - (\text{avg}(l_1, l_2, \ldots, l_k))^2 \)

Arrange the words in lines to minimize the sum of squares of the slacks of all lines.

**Puzzle** find integers \( a, b, c \) such that
\[
a + b + c = 14
\]
and \( a^2 + b^2 + c^2 \) is minimized.

**Answer:**

What do you observe?

**Balanced Margins**

**Idea 1:** fit as many words as you can

\[ \rightarrow \] can be very unbalanced

**Greedy Idea:**

compute the average slack per line. Go line by line and try to keep the slack for each line as close as possible to the average slack.

doesn't give an optimal solution.

\[
\begin{array}{cccc}
13 & 15 & 5 & 2 \\
\downarrow & \downarrow & \downarrow & \downarrow \\
5 & 5 & 4 & 51, 45
\end{array}
\]

\( = \frac{11}{3} \approx 3.67 \)
Dynamic Programming.

Try to categorize the set of all possible solutions.

Each solution is a partition of words into $k$ lines:

$$ n = n_1 + n_2 + n_3 + \ldots + n_k $$

Categories:

- $n_k = 1$ last line has 1 word.
- $n_k = 2$ last line has 2 words.
- $n_k = n-1$ last line has $n-1$ words.

Assuming last line has words $w_{p+1}, \ldots, w_n$, (i.e. $n-p$ words) can you compute the optimal solution via a subproblem?

$$ \text{OPT}(w_1, w_2, \ldots, w_p) + \text{Slack}(w_{p+1}, w_{p+2}, \ldots, w_n)^2 $$

$$ \text{OPT}(n) = \min \begin{cases} 
\text{OPT}(n-1) + [W - w_n]^2 \\
\text{OPT}(n-2) + [W - w_n - w_{n-1}]^2 \\
\text{OPT}(n-3) + [W - w_n - w_{n-1} - w_{n-2}]^2 \\
\vdots \\
\text{OPT}(n) = \min \end{cases} $$

Running time $O(n^2)$
Pseudocode for computing the optimal value and the optimal solution.

\[ S \leftarrow \text{array of length } n+1 \]

\[ \text{// } S[j] \text{ denotes the minimum sum of squares of slacks for first } j \text{ words, for } 1 \leq j \leq n \]

\[ N \leftarrow \text{array of length } n+1 \]

\[ \text{// } N[j] \text{ denotes the index of the first word in the last line in the optimal arrangement of first } j \text{ words.} \]

\[ s[0] \leftarrow 0 \; ; \; s[j] \leftarrow \infty \text{ for } j > 0 \; ; \; N[0] \leftarrow 0 \]

for (j = 1 to n)

  for (r = j to 1)

    Slack \leftarrow W - (w_r + w_{r+1} + \ldots + w_j + j \cdot r)

    if (Slack \geq 0 \text{ and } S[j] > S[r-1] + (Slack)^2)

      \[ S[j] \leftarrow S[r-1] + (Slack)^2 \]

      \[ N[j] \leftarrow r \]

Optimal Arrangement:

Run Arrange(n).

Arrange(j) {

  Arrange(N(j)-1)

  print words from N(j) to j in a new line.

}
Optimal Binary Search Tree

\[ a_1 \leq a_2 \leq a_3 \leq a_4 \]

7x2
5x1
4x3
6x2
\underline{43}

Suppose we know how frequent are the search queries for each of the elements.

Say frequencies

\[
\begin{align*}
    a_1 & \rightarrow 7 \\
    a_2 & \rightarrow 5 \\
    a_3 & \rightarrow 4 \\
    a_4 & \rightarrow 6
\end{align*}
\]

First binary tree 43

Second binary tree 47

Compute the binary search tree with minimum total search cost \( \sum_i f_i \cdot h_i \)

where \( f_i \rightarrow \text{frequency} (a_i) \), \( h_i \rightarrow \text{depth} (a_i) \)

Greedy Idea 1:

maximum frequency element \( \rightarrow \) root

\[ a_i < a_i \rightarrow a_i > a_i \]
Dynamic Programming \( a_1 < a_2 < a_3 \ldots < a_n \)

Possible solutions: all valid binary search trees.

Categories:

- Root \( \leftarrow a_1 \)
- Root \( \leftarrow a_2 \)
- Root \( \leftarrow a_n \)

\[
\text{Tree}(a_1, \ldots, a_n) = \min_{1 \leq i \leq n} \left\{ \text{Tree}(a_1, \ldots, a_{i-1}) + \text{Tree}(a_{i+1}, \ldots, a_n) + \sum_{j=1}^{n} f_j \cdot 1 \right\}
\]

How many distinct subproblems = \( \binom{n}{2} \)

\[
\text{Tree}(i, j) = \min_{k} \left\{ \text{Tree}(i, k-1) + \text{Tree}(k+1, j) \right\} + \sum_{k=i}^{j} f_k
\]

\( j > i \)
\[ \text{OPT} \leftarrow 2 \text{ dim array.} \]

\[ \text{OPT} [i, i] \leftarrow f_i \text{ for each i} \]

for \((i = n-1 \text{ to } 1)\)

for \((j = i+1 \text{ to } n)\)

\[ \text{OPT} [i, j] = \sum_{k=1}^{i} f_k + \min \{ \text{OPT} [i, k-1] + \text{OPT} [k+1, j] \} \]

Optimal cost

for \((a_i, a_{i+1}, \ldots, a_j)\)

Add/modify code to compute the optimal solution.
Sequence Alignment

Computational biology, spell checking

Similarity between two strings

Needleman and Wunsch defined a notion of similarity

Example

\[
\begin{align*}
UGCTGACU & \quad \text{UGCTGACU} \\
\rightarrow \quad \text{GAATGCA} & \quad \text{GAATGCA}
\end{align*}
\]

Given

Gap Penalty \( \delta \)

Mismatch cost (for each pair) \( \alpha_{xy} \)

Total cost = sum of the gap and mismatch costs.

Find an alignment of the two strings with minimum total cost.

Input: \( x_1, x_2, \ldots, x_m \) and \( y_1, y_2, \ldots, y_n \)

\( \delta, \{ \alpha_{x_i,y_j} \} \)

Categories of solutions.

1. \[
\begin{array}{c}
\text{---} \\
\text{---}
\end{array}
\]

2. \[
\begin{array}{c}
x_{m-1} \\
x_m
\end{array}
\]

3. \[
\begin{array}{c}
x_m \\
y_n
\end{array}
\]

\[
\begin{array}{c}
y_{n-1} \\
y_n
\end{array}
\]
$$\text{OPT}(m, n) = \begin{cases} 
\alpha_{x_m, y_n} + \text{OPT}(m-1, n-1) \\
\min \{ \delta + \text{OPT}(m-1, n), \delta + \text{OPT}(m, n-1) \} 
\end{cases}$$

$$\text{OPT}(i, j) \quad \text{no. of distinct subproblems: m \times n}$$

Implementation

$$A \leftarrow 2D \text{ array } m \times n$$

// $A[i, j]$ denotes the minimum cost of alignment for $x_1x_2\ldots x_i$ and $y_1y_2\ldots y_j$

$$A[i, 0] \leftarrow \delta_i \text{ for each } i$$

$$A[0, j] \leftarrow \delta_j \text{ for each } j$$

for (i = 1 to n)

for (j = 1 to n)

$$A[i, j] = \min \left\{ \alpha_{x_i, y_j} + A[i-1, j-1], \delta + A[i-1, j], \delta + A[i, j-1] \right\}$$

**HW**

Space $O(mn)$

Can you get space $O(m+n)$ and time $O(mn)$?
Summarizing Greedy and Dynamic Programming

- Dividing the set of possible solutions into multiple categories.

 Greedy: there must be an optimal solution in a certain category $C^*$ (greedy choice)

DP: will take best of the optimal solutions from each category.

- To compute the optimal solution from a chosen category $\rightarrow$ smaller subproblem (Recursion)

DP

No. of distinct subproblems should be small.
Data Compression: Coding

Assign a fixed length 0-1 string to each character:

\[
\begin{align*}
a & \rightarrow 00001 \\
b & \rightarrow 00010 \\
\vdots
\end{align*}
\]

5-bit encoding can work for up to 32 characters.

Is a smaller length encoding possible?

With fixed length - not possible.

Variable length encoding

- can be more efficient when no of characters is not \(2^k\)
- can use smaller length codes for more frequent characters.

Example: Morse Code (dots and dashes and spaces)

- e
- t
- a

\[
\begin{align*}
ez & \rightarrow -\ldots \\
q & \rightarrow -\ldots
\end{align*}
\]

Problem: \cdot \ldots

\[
\begin{align*}
aa & \leftrightarrow \cdot \ldots \\
eta & \leftrightarrow \cdot \ldots \\
\text{eta} & \leftrightarrow \cdot \ldots
\end{align*}
\]

Solution: Gap after every character.
**Prefix Code**

**Def:** For any two different characters \( x \) and \( y \), \( c(x) \) should not be a prefix of \( c(y) \).

\[
\begin{align*}
\text{Ex} & \quad x \rightarrow 00 \quad \{ \text{Not a prefix code} \} \\
& \quad y \rightarrow 001
\end{align*}
\]

\[
\begin{align*}
\text{Ex} & \quad a \rightarrow 0 \\
& \quad b \rightarrow 10 \\
& \quad c \rightarrow 11 \\
& \quad \{ \text{Prefix code} \}
\end{align*}
\]

\[
\begin{align*}
abaca & \rightarrow 0100110
\end{align*}
\]

**Claim:** For a prefix code, any 0-1 string is unambiguously decodable.

Just scan left to right, as soon as the current substring matches one of the codewords, output the corresponding character.

---

**Optimal Prefix Codes**

<table>
<thead>
<tr>
<th>Freq</th>
<th>Code 1</th>
<th>Code 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>A 00</td>
<td>0</td>
</tr>
<tr>
<td>0.4</td>
<td>T 01</td>
<td>10</td>
</tr>
<tr>
<td>0.1</td>
<td>P 10</td>
<td>110</td>
</tr>
<tr>
<td>0.1</td>
<td>G 11</td>
<td>111</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Avg Bit length} & = \frac{1}{180} \\
\text{Chars} & = 100 \\
\text{Bits} & = 200
\end{align*}
\]

\[
\begin{align*}
\text{Avg Bit length} & = 0.4 \times 1 + 0.4 \times 2 + 0.1 \times 3 + 0.1 \times 3 \\
& = 1.8
\end{align*}
\]
Prob: Given frequencies $f_1, f_2, \ldots, f_n$ for $n$ characters find a prefix code that minimizes $\sum_i f_i \cdot l_i$.

Observation: Each prefix code corresponds to a binary tree. Codewords correspond to leaves of the tree.

Approach 1: assign 0 to highest frequency.

Approach 2: assign 0 to highest frequency if it is above some threshold.

Approach 3: Do a balanced division of frequencies into two parts.

0.32, 0.25, 0.18, 0.25
Observations:

- Low frequency → higher length
- High frequency → lower length

Approach 1: for highest frequency character, assign ‘0’ i.e. length one codeword.

\[(0.25, 0.25, 0.25, 0.25)\]
\[(0.28, 0.24, 0.24, 0.24)\]
\[(0.37, 0.21, 0.21, 0.21)\]
\[(0.35, 0.35, 0.15, 0.15)\]

Assign ‘0’ if frequency higher than certain threshold.

Doesn’t work:

Cannot decide \(f_1\) just by looking at \(f_1\).
Suppose there is some way to decide the encoding length for highest frequency character.

Say length 2. ‘00’

0.3, 0.25, 0.2, 0.15, 0.1

Can we reduce the rest of the encodings to a subproblem?

Not able to frame it as a smaller instance of the same problem.

→ Suppose we can fix the length for the least frequent character.
Shannon and Fano (1940’s)

Balanced Partition

Divide the list into two parts such that total frequency of each part is as close as possible to 0.5.

0.35, 0.35, 0.2, 0.1

0.35 \times 1 + 0.35 \times 2 + 0.2 \times 3 + 0.1 \times 3 = 1.95

By balanced partition approach we are getting average encoding length = 2. But there is a better solution with 1.95.

→ Try various possibilities for the partition?

Expontentially many possibilities...
Huffman [1952]

Obs 1: If you fix a binary tree, then there is a natural way to characters to leaves

0.35, 0.35, 0.2, 0.1

Lowest frequency character has the largest depth.

Obs 2: The leaf with the largest depth must have a sibling which is a leaf.

\[ f_1 > f_2 > \ldots > f_{n-1} > f_n \]

Claim: 
Lowest and second lowest frequency characters can be mapped to two largest depth siblings.
Now, the rest of the tree can be found recursively.

\[ f_1, f_2, f_3, f_4, f_5 \text{ (in decreasing order)} \]

\[
\begin{array}{l}
\text{Cost}_0: 2f_1 + 2f_2 + 2f_3 + 3f_4 + 3f_5 \\
\text{Cost}_1: 2f_1 + 2f_2 + 2f_3 + 2f_4 + 2f_5 \\
\text{Cost}_0 = \text{Cost}_1 + f_4 + f_5
\end{array}
\]

\[ \rightarrow \text{for any binary tree with} \ n \ \text{leaves where} \ f_{n-1} \ \text{and} \ f_n \ \text{are siblings, there is a corresponding binary tree with} \ n-1 \ \text{leaves with labelings} \]

\[ f_1, f_2, f_3, \ldots, f_{n-2}, f_{n-1} + f_n \]

Algorithm:

\[ \text{Input:} \ f_1 \geq f_2 \geq \cdots \geq f_{n-1} \geq f_n \ (n \ \text{char}) \]

Recursively compute optimal binary tree for

\[ f_1, f_2, \ldots, f_{n-2}, f_{n-1} + f_n \ (n-1 \ \text{char}) \]

For the leaf labeled \( f_{n-1} + f_n \), add two children with label \( f_{n-1} \), \( f_n \).
Example

0.3, 0.25, 0.18, 0.15, 0.12

0.3, 0.25, 0.18, 0.27

0.3, 0.27, 0.43

0.43, 0.57

\[ \sum f_i l_i \]
Proof of Correctness

1. There is an optimal solution where \( f_{n-1} \) and \( f_n \) are siblings.

2. There is a one-to-one correspondence between

   \[ A = \{ \text{full binary trees with } n \text{ leaves labeled } f_1, f_2, ..., f_n \} \]
   \[ B = \{ \text{full binary trees with } n-1 \text{ leaves labeled } f_1, f_2, f_3, ..., f_{n-2}, f_{n-1} + f_n \} \]

   Say, \( A = \{ T_1, T_2, T_3, ..., T_N \} \)
   \[ B = \{ R_1, R_2, R_3, ..., R_N \} \]

   Remove leaves labeled \( f_{n-1}, f_n \).
   Label their parent \( f_{n-1} + f_n \).

   For the leaf labeled \( f_{n-1} + f_n \),
   add two children labeled \( f_{n-1}, f_n \).

   Moreover \( \text{cost}(T_j) = \text{cost}(R_j) + f_{n-1} + f_n \)

   Thus, \( R_j \) is optimal in \( B \iff T_j \) is optimal in \( A \).
Huffman Codes:
Text over some large alphabet size → space efficient bit representation

- Que: Can we apply this technique when the data is already in 0/1 bit representation?

Suppose there is data where 0 is much more frequent than 1.

```
   a  g  a  a
001000110101000011100110000010
```

```
\[\text{(a)}\quad b\left\{\begin{array}{c}
\text{000} \quad 2/8 \\
\text{001} \quad 3/16
\end{array}\right.\rightarrow 00
\]
```

```
\[\text{(b)}\quad c\rightarrow 100
\]
```

```
\[\text{(c)}\quad h\rightarrow 1101
\]
```

- There are many other data compression techniques
  - Adaptive
  - Algebraic