CSE 421
Algorithms

Huffman Codes:
An Optimal Data Compression Method
Compression Example

100k file, 6 letter alphabet:

File Size:
   ASCII, 8 bits/char:  800kbits
   $2^3 > 6; \ 3 \text{ bits/char: } 300\text{kbits}$

Why?
   Storage, transmission vs 5 Ghz cpu
Compression Example

100k file, 6 letter alphabet:

File Size:

- ASCII, 8 bits/char: 800kbits
- $2^3 > 6$; 3 bits/char: 300kbits
- better: 2.52 bits/char $74\% \times 2 + 26\% \times 4$: 252kbits
- Optimal?

E.g.:

Why not:

<table>
<thead>
<tr>
<th></th>
<th>ASCII</th>
<th>Why not:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>00</td>
<td>00</td>
</tr>
<tr>
<td>b</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>c</td>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>d</td>
<td>101</td>
<td>1101</td>
</tr>
<tr>
<td>e</td>
<td>110</td>
<td>1101</td>
</tr>
<tr>
<td>f</td>
<td>111</td>
<td>1110</td>
</tr>
</tbody>
</table>

1101110 = cf or ec?
Data Compression

Binary character code ("code")
  each k-bit source string maps to unique code word
  (e.g. k=8)
  “compression” alg: concatenate code words for
  successive k-bit “characters” of source

Fixed/variable length codes
  all code words equal length?

Prefix codes
  no code word is prefix of another (unique decoding)
Prefix Codes = Trees

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>45%</td>
</tr>
<tr>
<td>b</td>
<td>13%</td>
</tr>
<tr>
<td>c</td>
<td>12%</td>
</tr>
<tr>
<td>d</td>
<td>16%</td>
</tr>
<tr>
<td>e</td>
<td>9%</td>
</tr>
<tr>
<td>f</td>
<td>5%</td>
</tr>
</tbody>
</table>

```
1 0 1 0 0 0 0 0 1
f a b
```

```
1 1 0 0 0 0 1 0 1
f a b
```
Greedy Idea #1

Put **most frequent**
under root, then recurse …
Greedy Idea #1

Top down: Put *most* frequent under root, then recurse.

**Too greedy:**
unbalanced tree

\[0.45 \times 1 + 0.16 \times 2 + 0.13 \times 3 \ldots = 2.34\]
not too bad, but imagine if all freqs were \(~1/6\):
\[\frac{1+2+3+4+5+5}{6}=3.33\]
Greedy Idea #2

Top down: Divide letters into 2 groups, with ~50% weight in each; recurse (Shannon-Fano code)

Again, not terrible
2*0.5+3*0.5 = 2.5

But this tree can easily be improved! (How?)
Greedy idea #3

Bottom up: Group
least frequent letters
near bottom

<table>
<thead>
<tr>
<th>Letter</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>45%</td>
</tr>
<tr>
<td>b</td>
<td>13%</td>
</tr>
<tr>
<td>c</td>
<td>12%</td>
</tr>
<tr>
<td>d</td>
<td>16%</td>
</tr>
<tr>
<td>e</td>
<td>9%</td>
</tr>
<tr>
<td>f</td>
<td>5%</td>
</tr>
</tbody>
</table>

Diagram:

- Top node: 100
  - Child nodes: 25, 14
    - 25: c:12, b:13
    - 14: f:5, e:9
.45 * 1 + .41 * 3 + .14 * 4 = 2.24 bits per char
Huffman’s Algorithm (1952)

Algorithm:

insert node for each letter into priority queue by freq
while queue length > 1 do
    remove smallest 2; call them x, y
    make new node z from them, with f(z) = f(x) + f(y)
    insert z into queue

Analysis: \( O(n) \) heap ops: \( O(n \log n) \)

Goal: Minimize \( \text{Cost}(T) = \sum_{c \in C} \text{freq}(c) \times \text{depth}(c) \)

Correctness: ???
Correctness Strategy

Optimal solution may not be unique, so cannot prove that greedy gives the only possible answer.

Instead, show greedy’s solution is as good as any.

How: an exchange argument

Identify inversions: node-pairs whose swap improves tree

To compare trees T (arbitrary) to H (Huffman): run Huff alg, tracking subtrees in common to T & H; discrepancies flag inversions; swapping them incrementally xforms T to H
Defn: A pair of leaves $x, y$ is an inversion if
\[
\text{depth}(x) \geq \text{depth}(y)
\]
and
\[
\text{freq}(x) \geq \text{freq}(y)
\]

Claim: If we flip an inversion, cost never increases.

Why? All other things being equal, better to give more frequent letter the shorter code.

\[
(d(x) \times f(x) + d(y) \times f(y)) - (d(x) \times f(y) + d(y) \times f(x)) =
\]
\[
(d(x) - d(y)) \times (f(x) - f(y)) \geq 0
\]

I.e., non-negative cost savings.
General Inversions

Define the frequency of an *internal* node to be the sum of the frequencies of the leaves in that subtree (as shown in the example trees above). Given that, the definition of inversion on slide 13 easily generalizes to an arbitrary pair of nodes, and the associated claim still holds: exchanging an inverted pair of nodes (& associated subtrees) cannot raise the cost of a tree.

Proof: Homework
The following slide is heavily animated, which doesn’t show too well in print. The point is to illustrate the Lemma on slide 17. Idea is to run Huffman alg on the example above and compare successive subtrees it builds to subtrees in an arbitrary tree T. While they agree (marked by yellow), repeat; when they first differ (in this case, when Huffman builds node 30), identify an inversion in T whose removal would allow them to agree for at least one more step, i.e., T’ is more like H than T, but costs no more. Slide 16 is an example; slide 17 sketches the proof in general.
In short, where \( T \) first differs from \( H \) flags an inversion in \( T \) vsn: yellow only at roots. Conceptually clearer? Nice animation, but printed pdf overlays them.
Lemma: Any prefix code tree $T$ can be converted to a Huffman tree $H$ via inversion-exchanges

Pf Idea: Run Huffman alg; “color” $T$’s nodes to track matching subtrees between $T$, $H$. Inductively: yellow nodes in $T$ match subtrees of $H$ in Huffman’s heap at that stage in the alg. & yellow nodes partition leaves. Initially: leaves yellow, rest white.

At each step, Huffman extracts $A$, $B$, the 2 min heap items; both yellow in $T$.

Case 1: $A$, $B$ match siblings in $T$. Then their newly created parent node in $H$ corresponds to their parent in $T$; paint it yellow, $A$ & $B$ revert to white.

Case 2: $A$, $B$ not sibs in $T$. WLOG, in $T$, $\text{depth}(A) \geq \text{depth}(B)$ & $A$ is $C$’s sib. Note $B$ can’t overlap $C$ ($B = C \Rightarrow$ case 1; $B$ subtree of $C$ contradicts depth; $B$ contains $C$ contradicts partition). In $T$, the freq of $C$’s root $\geq$ freqs of all yellow nodes init ($\neq \emptyset$ since …).

Huff’s picks ($A$ & $B$) were min, so $\text{freq}(C) \geq \text{freq}(B)$. $\therefore$ $B:C$ is an inversion–$B$ is no deeper/no more frequent than $C$.

Swapping gives $T'$ more like $H$; repeating $\leq n$ times converts $T$ to $H$. 

---

$T$  \quad \text{↔} \quad T'$

$A$  \quad $B$  \quad $C$
Theorem: Huffman is optimal

Pf: Apply the above lemma to any optimal tree $T = T_1$. The lemma only exchanges inversions, which never increase cost, so, cost of successive trees is monotonically non-increasing, and the last tree is $H$: $\text{cost}(T_1) \geq \text{cost}(T_2) \geq \text{cost}(T_3) \geq \ldots \geq \text{cost}(H)$.

Corr: can convert any tree to $H$ by inversion-exchanges (general exchanges, not just leaf exchanges)
Data Compression

Huffman is optimal.

BUT still might do better!

- Huffman encodes fixed length blocks. What if we vary them?

- Huffman uses one encoding throughout a file. What if characteristics change?

- What if data has structure? E.g. raster images, video,…

- Huffman is lossless. Necessary?

LZW, MPEG, …
David A. Huffman, 1925-1999