5.5 DATA COMPRESSION

- introduction
- run-length coding
- Huffman compression
- LZW compression

Data compression

Compression reduces the size of a file:
- To save space when storing it.
- To save time when transmitting it.
- Most files have lots of redundancy.

Who needs compression?
- Moore's law: # transistors on a chip doubles every 18–24 months.
- Parkinson's law: data expands to fill space available.
- Text, images, sound, video, ...

“Everyday, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone.” — IBM report on big data (2011)

Basic concepts ancient (1950s), best technology recently developed.

Applications

Generic file compression.
- Files: GZIP, BZIP, 7z.
- Archivers: PKZIP.
- File systems: NTFS, HFS+, ZFS.

Multimedia.
- Images: GIF, JPEG.
- Sound: MP3.
- Video: MPEG, DivX™, HDTV.

Communication.
- ITU-T T4 Group 3 Fax.
- V.42bis modem.
- Skype.

Databases. Google, Facebook, ....
Lossless compression and expansion

**Message.** Binary data $B$ we want to compress.
**Compress.** Generates a "compressed" representation $C(B)$.
**Expand.** Reconstructs original bitstream $B$.

![Diagram of data compression](chart.png)

Compress takes input $B$ and produces output $C(B)$.
Expand takes input $C(B)$ and produces output $B$.

**Compression ratio.** Bits in $C(B)$ / bits in $B$.

**Ex.** 50–75% or better compression ratio for natural language.

---

Food for thought

**Data compression has been omnipresent since antiquity:**
- Number systems.
- Natural languages.
- Mathematical notation.

has played a central role in communications technology,
- Grade 2 Braille.
- Morse code.
- Telephone system.

and is part of modern life.
- MP3.
- MPEG.

Q. What role will it play in the future?

---

Data representation: genomic code

**Genome.** String over the alphabet { A, C, T, G }.

**Goal.** Encode an $N$-character genome: ATAGATGCATAG...

**Standard ASCII encoding.**
- 8 bits per char.
- $8N$ bits.

<table>
<thead>
<tr>
<th>char</th>
<th>hex</th>
<th>binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41</td>
<td>01000001</td>
</tr>
<tr>
<td>C</td>
<td>43</td>
<td>01000011</td>
</tr>
<tr>
<td>T</td>
<td>54</td>
<td>01010100</td>
</tr>
<tr>
<td>G</td>
<td>47</td>
<td>01000111</td>
</tr>
</tbody>
</table>

**Two-bit encoding.**
- 2 bits per char.
- $2N$ bits.

<table>
<thead>
<tr>
<th>char</th>
<th>binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>00</td>
</tr>
<tr>
<td>C</td>
<td>01</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
</tr>
</tbody>
</table>

**Fixed-length code.** $k$-bit code supports alphabet of size $2^k$.
**Amazing but true.** Initial genomic databases in 1990s used ASCII.

---

Reading and writing binary data

**Binary standard input and standard output.** Libraries to read and write bits from standard input and to standard output.

```
public class BinaryStdIn
{
  boolean readBoolean(); // read 1 bit of data and return as a boolean value
  char readChar(); // read 8 bits of data and return as a char value
  char readChar(int r); // read r bits of data and return as a char value
  // [similar methods for byte (8 bits), short (16 bits), int (32 bits), long and double (64 bits)]
  boolean isSane(); // is the bitstream empty?
  void close(); // close the bitstream
}
```

```
public class BinaryStdOut
{
  void write:boolean b); // write the specified bit
  void write(char c); // write the specified 8-bit char
  void write(char c, int r); // write the r least significant bits of the specified char
  // [similar methods for byte (8 bits), short (16 bits), int (32 bits), long and double (64 bits)]
  void close(); // close the bitstream
}
```
Writing binary data

Date representation. Three different ways to represent 12/31/1999.

A character stream (StdOut)

```java
StdOut.print(month + "/" + day + "/" + year);
```

Three ints (BinaryStdOut)

```java
BinaryStdOut.write(month); BinaryStdOut.write(day); BinaryStdOut.write(year);
```

A 4-bit field, a 5-bit field, and a 12-bit field (BinaryStdOut)

```java
BinaryStdOut.write(month, 4); BinaryStdOut.write(day, 5); BinaryStdOut.write(year, 12);
```

Universal data compression

US Patent 5,533,051 on "Methods for Data Compression", which is capable of compression all files.

Slashdot reports of the Zero Space Tuner™ and BinaryAccelerator™.

"ZeoSync has announced a breakthrough in data compression that allows for 100:1 lossless compression of random data. If this is true, our bandwidth problems just got a lot smaller…. "

Binary dumps

Q. How to examine the contents of a bitstream?

Standard character stream

```text
% more abra.txt
ABRACADABRA!
```

Bitstream represented with hex digits

```java
% java BinaryDump 16 < abra.txt
41 42 52 41 43 41 44 41 42 52 41 21
```

Bitstream represented as pixels in a Picture

```java
% java PictureDump 16 6 < abra.txt
```

Universal data compression

**Proposition.** No algorithm can compress every bitstring.

**Pf 1.** [by contradiction]
- Suppose you have a universal data compression algorithm $U$ that can compress every bitstream.
- Given bitstring $B_i$, compress it to get smaller bitstring $B_1$.
- Compress $B_1$ to get a smaller bitstring $B_2$.
- Continue until reaching bitstring of size 0.
- Implication: all bitstrings can be compressed to 0 bits!

**Pf 2.** [by counting]
- Suppose your algorithm that can compress all 1,000-bit strings.
- $2^{1000}$ possible bitstrings with 1,000 bits.
- Only $1 + 2 + 4 + \ldots + 2^{998} + 2^{999}$ can be encoded with $\leq 999$ bits.
- Similarly, only 1 in $2^{500}$ bitstrings can be encoded with $\leq 500$ bits!
Undecidability

A difficult file to compress: one million (pseudo-) random bits

```java
public class RandomBits {
    public static void main(String[] args) {
        int x = 11111;
        for (int i = 0; i < 1000000; i++) {
            x = x * 314159 + 218281;
            BinaryStdOut.write(x > 0);
        }
        BinaryStdOut.close();
    }
}
```

Rudendcany in Enlgsh Inagugae

Q. How much redundancy is in the Enlgsh Inagugae?

“... randomising letters in the middle of words [has] little or no effect on the ability of skilled readers to understand the text. This is easy to demonstrate. In a publication of New Science you could randomise all the letters, keeping the first two and last two the same, and reibadality would hardly be affected. My analysis did not come to much because the theory at the time was for shape and sequence retention. Saberi’s work suggests we may have some powerful parallel processors at work. The reason for this is surely that identifying content by parallel processing speeds up recognition. We only need the first and last two letters to spot changes in meaning.” — Graham Rawlinson

A. Quite a bit.

Run-length encoding

Simple type of redundancy in a bitstream. Long runs of repeated bits.

```
00000000000011111111100000001111111111
```

40 bits

Representation. 4-bit counts to represent alternating runs of 0s and 1s:

```
15 11 11 11 10 11 11 11
```

16 bits (instead of 40)

Q. How many bits to store the counts?
A. We’ll use 8 (but 4 in the example above).

Q. What to do when run length exceeds max count?
A. If longer than 255, intersperse runs of length 0.

Applications. JPEG, ITU-T T4 Group 3 Fax, ...
### Run-length encoding: Java implementation

```java
public class RunLength {
    private final static int R = 256;
    private final static int lgR = 8;

    public static void compress() {
        // see textbook */
    }

    public static void expand() {
        boolean bit = false;
        while (!BinaryStdIn.isEmpty()) {
            int run = BinaryStdIn.readInt(lgR);
            for (int i = 0; i < run; i++) {
                BinaryStdOut.write(bit);
                bit = !bit;
            }
        }
        BinaryStdOut.close();
    }
}
```

---

### 5.5 Data Compression

- introduction
- run-length coding
- Huffman compression
- LZW compression

David Huffman

5.5  DATA COMPRESSION

### Variable-length codes

Use different number of bits to encode different chars.

**Ex.** Morse code: • • • – – • • •

**Issue.** Ambiguity.
- SOS?
- V7?
- IAMIE?
- EEWNI?

**In practice.** Use a medium gap to separate codewords.

- Codeword for S is a prefix of codeword for V

Q. How do we avoid ambiguity?
A. Ensure that no codeword is a prefix of another.

**Ex 1.** Fixed-length code.
**Ex 2.** Append special stop char to each codeword.
**Ex 3.** General prefix-free code.
Prefix-free codes: trie representation

Q. How to represent the prefix-free code?
A. A binary trie!
   - Chars in leaves.
   - Codeword is path from root to leaf.

Huffman trie node data type

Prefix-free codes: expansion

Compression.
- Method 1: start at leaf; follow path up to the root; print bits in reverse.
- Method 2: create ST of key-value pairs.

Expansion.
- Start at root.
- Go left if bit is 0; go right if 1.
- If leaf node, print char and return to root.

Prefix-free codes: expansion

Running time. Linear in input size $N$. 

public void expand() 
{
   Node root = readTrie();
   int N = BinaryStdIn.readInt();

   for (int i = 0; i < N; i++)
   {
      Node x = root;
      while (!x.isLeaf())
      {
         if (!BinaryStdIn.readBoolean())
            x = x.left;
         else
            x = x.right;
      }
      BinaryStdOut.write(x.ch, 8);
   }
   BinaryStdOut.close();
}
Prefix-free codes: how to transmit

Q. How to write the trie?
A. Write preorder traversal of trie; mark leaf and internal nodes with a bit.

private static void writeTrie(Node x)
{
    if (x.isLeaf())
    {
        BinaryStdOut.write(true);
        BinaryStdOut.write(x.ch, 8);
        return;
    }
    BinaryStdOut.write(false);
    writeTrie(x.left);
    writeTrie(x.right);
}

private static void writeTrie(Node x)
{
    if (x.isLeaf())
    {
        BinaryStdOut.write(true);
        BinaryStdOut.write(x.ch, 8);
        return;
    }
    BinaryStdOut.write(false);
    writeTrie(x.left);
    writeTrie(x.right);
}

Using preorder traversal to encode a trie as a bitstream

Note. If message is long, overhead of transmitting trie is small.

Shannon-Fano codes

Q. How to find best prefix-free code?

Shannon-Fano algorithm:
• Partition symbols \( S \) into two subsets \( S_0 \) and \( S_1 \) of (roughly) equal freq.
• Codewords for symbols in \( S_0 \) start with 0; for symbols in \( S_1 \) start with 1.
• Recur in \( S_0 \) and \( S_1 \).

Shannon-Fano algorithm demo

• Count frequency for each character in input.

 Huffman algorithm demo

<table>
<thead>
<tr>
<th>char</th>
<th>freq</th>
<th>encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1...</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0...</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1...</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1...</td>
</tr>
<tr>
<td>!</td>
<td>1</td>
<td>1...</td>
</tr>
</tbody>
</table>

input

A B R A C A D A B R A !

Problem 1. How to divide up symbols?
Problem 2. Not optimal!
Huffman algorithm demo

<table>
<thead>
<tr>
<th>char</th>
<th>freq</th>
<th>encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1 1 1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1 0 1 1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1 0 0</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1 1 0</td>
</tr>
<tr>
<td>!</td>
<td>1</td>
<td>1 0 1 0</td>
</tr>
</tbody>
</table>

Constructing a Huffman encoding trie: Java implementation

```java
private static Node buildTrie(int[] freq)
{
    MinPQ<Node> pq = new MinPQ<Node>();
    for (char i = 0; i < R; i++)
        if (freq[i] > 0)
            pq.insert(new Node(i, freq[i], null, null));

    while (pq.size() > 1)
    {
        Node x = pq.delMin();
        Node y = pq.delMin();
        Node parent = new Node(\'\0\', x.freq + y.freq, x, y);
        pq.insert(parent);
    }

    return pq.delMin();
}
```

Huffman codes

Q. How to find best prefix-free code?

Huffman algorithm:

- Count frequency freq[i] for each char i in input.
- Start with one node corresponding to each char i (with weight freq[i]).
- Repeat until single trie formed:
  - select two tries with min weight freq[i] and freq[j]
  - merge into single trie with weight freq[i] + freq[j]

Applications:

- JPEG
- PDF
- mp3
- DIVX
- g7lp

Huffman encoding summary

Proposition. [Huffman 1950s] Huffman algorithm produces an optimal
prefix-free code.

Pf. See textbook.

Implementation.

- Pass 1: tabulate char frequencies and build trie.
- Pass 2: encode file by traversing trie or lookup table.

Running time. Using a binary heap \( \Rightarrow N + R \log R \).

Q. Can we do better? [stay tuned]
5.5 DATA COMPRESSION

- introduction
- run-length coding
- Huffman compression
- LZW compression

Statistical methods

**Static model.** Same model for all texts.
- Fast.
- Not optimal: different texts have different statistical properties.
- Ex: ASCII, Morse code.

**Dynamic model.** Generate model based on text.
- Preliminary pass needed to generate model.
- Must transmit the model.
- Ex: Huffman code.

**Adaptive model.** Progressively learn and update model as you read text.
- More accurate modeling produces better compression.
- Decoding must start from beginning.
- Ex: LZW.

LZW compression example

<table>
<thead>
<tr>
<th>input</th>
<th>A</th>
<th>B</th>
<th>R</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>D</th>
<th>A</th>
<th>B</th>
<th>R</th>
<th>A</th>
<th>B</th>
<th>R</th>
<th>A</th>
<th>B</th>
<th>R</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>matches</td>
<td>A</td>
<td>B</td>
<td>R</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>B</td>
<td>R</td>
<td>A</td>
<td>B</td>
<td>R</td>
<td>A</td>
<td>B</td>
<td>R</td>
<td>A</td>
</tr>
<tr>
<td>value</td>
<td>41</td>
<td>42</td>
<td>52</td>
<td>41</td>
<td>43</td>
<td>41</td>
<td>44</td>
<td>81</td>
<td>83</td>
<td>82</td>
<td>88</td>
<td>41</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LZW compression for A B R A C A D A B R A B R A B R A A

<table>
<thead>
<tr>
<th>key</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>81</td>
</tr>
<tr>
<td>BR</td>
<td>82</td>
</tr>
<tr>
<td>RA</td>
<td>83</td>
</tr>
<tr>
<td>AC</td>
<td>84</td>
</tr>
<tr>
<td>CA</td>
<td>85</td>
</tr>
<tr>
<td>AD</td>
<td>86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>key</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>87</td>
</tr>
<tr>
<td>ABR</td>
<td>88</td>
</tr>
<tr>
<td>RAB</td>
<td>89</td>
</tr>
<tr>
<td>BRA</td>
<td>8A</td>
</tr>
<tr>
<td>ABRA</td>
<td>8B</td>
</tr>
</tbody>
</table>

Lempel-Ziv-Welch compression

**LZW compression.**
- Create ST associating W-bit codewords with string keys.
- Initialize ST with codewords for single-char keys.
- Find longest string \(s\) in ST that is a prefix of unscanned part of input.
- Write the W-bit codeword associated with \(s\).
- Add \(s + c\) to ST, where \(c\) is next char in the input.

Q. How to represent LZW compression code table?
A. A trie to support longest prefix match.
LZW compression: Java implementation

```java
public static void compress()
{
    String input = BinaryStdIn.readString();
    TST<Integer> st = new TST<Integer>();
    for (int i = 0; i < R; i++)
        st.put("", (char) i, i);
    int code = R+1;
    while (input.length() > 0)
    {
        String s = st.longestPrefixOf(input);
        BinaryStdOut.write(st.get(s), W);
        int t = s.length();
        if (t < input.length() && code < L)
            st.put(input.substring(0, t+1), code);
        input = input.substring(t);
    }
    BinaryStdOut.write(R, W);
    BinaryStdOut.close();
}
```

LZW expansion

LZW expansion:
- Create ST associating string values with W-bit keys.
- Initialize ST to contain single-char values.
- Read a W-bit key.
- Find associated string value in ST and write it out.
- Update ST.

Q. How to represent LZW expansion code table?
A. An array of size $2^W$. 

LZW expansion for 41 42 52 41 43 41 44 81 83 82 88 41 80

<table>
<thead>
<tr>
<th>key</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>42</td>
<td>B</td>
</tr>
<tr>
<td>43</td>
<td>C</td>
</tr>
<tr>
<td>44</td>
<td>D</td>
</tr>
<tr>
<td>81</td>
<td>AB</td>
</tr>
<tr>
<td>82</td>
<td>BR</td>
</tr>
<tr>
<td>83</td>
<td>RA</td>
</tr>
<tr>
<td>84</td>
<td>AC</td>
</tr>
<tr>
<td>85</td>
<td>CA</td>
</tr>
<tr>
<td>86</td>
<td>AD</td>
</tr>
<tr>
<td>87</td>
<td>DA</td>
</tr>
<tr>
<td>88</td>
<td>ABR</td>
</tr>
<tr>
<td>89</td>
<td>RAB</td>
</tr>
<tr>
<td>8A</td>
<td>BRA</td>
</tr>
<tr>
<td>8B</td>
<td>ABRA</td>
</tr>
</tbody>
</table>

LZW example: tricky case

```
input matches value
A B A B A B A A
41 42 81 83 80
```

LZW compression for ABABABA

<table>
<thead>
<tr>
<th>key</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41</td>
</tr>
<tr>
<td>B</td>
<td>42</td>
</tr>
<tr>
<td>C</td>
<td>43</td>
</tr>
<tr>
<td>D</td>
<td>44</td>
</tr>
<tr>
<td>AB</td>
<td>81</td>
</tr>
<tr>
<td>BA</td>
<td>82</td>
</tr>
<tr>
<td>ABA</td>
<td>83</td>
</tr>
<tr>
<td>ABRA</td>
<td></td>
</tr>
</tbody>
</table>

Codeword table
LZW example: tricky case

<table>
<thead>
<tr>
<th>key</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>42</td>
<td>B</td>
</tr>
<tr>
<td>81</td>
<td>B</td>
</tr>
<tr>
<td>83</td>
<td>A</td>
</tr>
<tr>
<td>80</td>
<td>A</td>
</tr>
</tbody>
</table>

LZW expansion for 41 42 81 83 80

need to know which key has value 83 before it is in ST!

codeword table

LZW in the real world

Lempel-Ziv and friends.

- LZ77.
- LZ78.
- LZW.
- Deflate / zlib = LZ77 variant + Huffman.

LZW in the real world

Unix compress, GIF, TIFF, V.42bis modem: LZW.
zip, 7zip, gzip, jar, png, pdf: deflate / zlib.
iPhone, Sony Playstation 3, Apache HTTP server: deflate / zlib.

LZW implementation details

How big to make ST?
- How long is message?
- Whole message similar model?
- [many other variations]

What to do when ST fills up?
- Throw away and start over. [GIF]
- Throw away when not effective. [Unix compress]
- [many other variations]

Why not put longer substrings in ST?
- [many variations have been developed]
Lossless data compression benchmarks

<table>
<thead>
<tr>
<th>year</th>
<th>scheme</th>
<th>bits / char</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967</td>
<td>ASCII</td>
<td>7.00</td>
</tr>
<tr>
<td>1950</td>
<td>Huffman</td>
<td>4.70</td>
</tr>
<tr>
<td>1977</td>
<td>LZ77</td>
<td>3.94</td>
</tr>
<tr>
<td>1984</td>
<td>LZMW</td>
<td>3.32</td>
</tr>
<tr>
<td>1987</td>
<td>LZH</td>
<td>3.30</td>
</tr>
<tr>
<td>1987</td>
<td>move-to-front</td>
<td>3.24</td>
</tr>
<tr>
<td>1987</td>
<td>LZB</td>
<td>3.18</td>
</tr>
<tr>
<td>1987</td>
<td>gzip</td>
<td>2.71</td>
</tr>
<tr>
<td>1988</td>
<td>PPMC</td>
<td>2.48</td>
</tr>
<tr>
<td>1994</td>
<td>SAKDC</td>
<td>2.47</td>
</tr>
<tr>
<td>1994</td>
<td>PPM</td>
<td>2.34</td>
</tr>
<tr>
<td>1995</td>
<td>Burrows-Wheeler</td>
<td>2.29</td>
</tr>
<tr>
<td>1997</td>
<td>BOA</td>
<td>1.99</td>
</tr>
<tr>
<td>1999</td>
<td>RK</td>
<td>1.89</td>
</tr>
</tbody>
</table>

Data compression summary

**Lossless compression.**
- Represent fixed-length symbols with variable-length codes. [Huffman]
- Represent variable-length symbols with fixed-length codes. [LZW]

**Lossy compression.** [not covered in this course]
- JPEG, MPEG, MP3, ...
- FFT, wavelets, fractals, ...

**Theoretical limits on compression.** Shannon entropy: \( H(X) = -\sum p(x_i) \log p(x_i) \)

**Practical compression.** Use extra knowledge whenever possible.