

## 15-853: Algorithms in the Real World

Data Compression: Lectures 1 and 2

15-853

Page 1

## Compression in the Real World

Generic File Compression

- **Files:** gzip (LZ77), bzip (Burrows-Wheeler), BOA (PPM)
- **Archivers:** ARC (LZW), PKZip (LZW+)
- **File systems:** NTFS

Communication

- **Fax:** ITU-T Group 3 (run-length + Huffman)
- **Modems:** V.42bis protocol (LZW), MNP5 (run-length+Huffman)
- **Virtual Connections**

15-853

Page 2

## Compression in the Real World

Multimedia

- **Images:** gif (LZW), jbig (context), jpeg-ls (residual), jpeg (transform+RL+arithmetic)
- **TV:** HDTV (mpeg-4)
- **Sound:** mp3

An example

Other structures

- **Indexes:** google, lycos
- **Meshes (for graphics):** edgebreaker
- **Graphs**
- **Databases:**

15-853

Page 3

## Compression Outline

➡ **Introduction:**

- Lossless vs. lossy
- Model and coder
- Benchmarks

**Information Theory:** Entropy, etc.

**Probability Coding:** Huffman + Arithmetic Coding

**Applications of Probability Coding:** PPM + others

**Lempel-Ziv Algorithms:** LZ77, gzip, compress, ...

**Other Lossless Algorithms:** Burrows-Wheeler

**Lossy algorithms for images:** JPEG, MPEG, ...

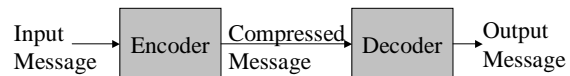
**Compressing graphs and meshes:** BBK

15-853

Page 4

## Encoding/Decoding

Will use "message" in generic sense to mean the data to be compressed



The encoder and decoder need to understand common compressed format.

15-853

Page 5

## Lossless vs. Lossy

**Lossless:** Input message = Output message

**Lossy:** Input message  $\approx$  Output message

Lossy does not necessarily mean loss of quality. In fact the output could be "better" than the input.

- Drop random noise in images (dust on lens)
- Drop background in music
- Fix spelling errors in text. Put into better form.

Writing is the art of lossy text compression.

15-853

Page 6

## How much can we compress?

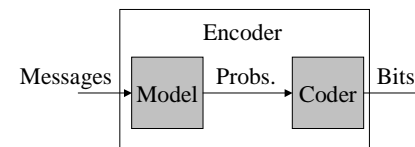
For lossless compression, assuming all input messages are valid, if even one string is compressed, some other must expand.

15-853

Page 7

## Model vs. Coder

To compress we need a bias on the probability of messages. The model determines this bias



Example models:

- Simple: Character counts, repeated strings
- Complex: Models of a human face

15-853

Page 8

## Quality of Compression

Runtime vs. Compression vs. Generality

Several standard corpuses to compare algorithms

### e.g. Calgary Corpus

2 books, 5 papers, 1 bibliography,  
1 collection of news articles, 3 programs,  
1 terminal session, 2 object files,  
1 geophysical data, 1 bitmap bw image

The **Archive Comparison Test** maintains a comparison of just about all algorithms publicly available

## Comparison of Algorithms

Program	Algorithm	Time	BPC	Score
RK	LZ + PPM	111+115	1.79	430
BOA	PPM Var.	94+97	1.91	407
PPMD	PPM	11+20	2.07	265
IMP	BW	10+3	2.14	254
BZIP	BW	20+6	2.19	273
GZIP	LZ77 Var.	19+5	2.59	318
LZ77	LZ77	?	3.94	?

## Compression Outline

**Introduction:** Lossy vs. Lossless, Benchmarks, ...

### ➔ **Information Theory:**

- Entropy
- Conditional Entropy
- Entropy of the English Language

**Probability Coding:** Huffman + Arithmetic Coding

**Applications of Probability Coding:** PPM + others

**Lempel-Ziv Algorithms:** LZ77, gzip, compress, ...

**Other Lossless Algorithms:** Burrows-Wheeler

**Lossy algorithms for images:** JPEG, MPEG, ...

**Compressing graphs and meshes:** BBK

## Information Theory

An interface between modeling and coding

### Entropy

- A measure of information content

### Conditional Entropy

- Information content based on a context

Entropy of the English Language

- How much information does each character in "typical" English text contain?

## Entropy (Shannon 1948)

For a set of messages  $S$  with probability  $p(s)$ ,  $s \in S$ , the **self information** of  $s$  is:

$$i(s) = \log \frac{1}{p(s)} = -\log p(s)$$

Measured in bits if the log is base 2.

The lower the probability, the higher the information **Entropy** is the weighted average of self information.

$$H(S) = \sum_{s \in S} p(s) \log \frac{1}{p(s)}$$

15-853

Page 13

## Entropy Example

$$p(S) = \{.25, .25, .25, .125, .125\}$$

$$H(S) = 3 \times .25 \log 4 + 2 \times .125 \log 8 = 2.25$$

$$p(S) = \{.5, .125, .125, .125, .125\}$$

$$H(S) = .5 \log 2 + 4 \times .125 \log 8 = 2$$

$$p(S) = \{.75, .0625, .0625, .0625, .0625\}$$

$$H(S) = .75 \log(4/3) + 4 \times .0625 \log 16 = 1.3$$

15-853

Page 14

## Conditional Entropy

The **conditional probability**  $p(s/c)$  is the probability of  $s$  in a context  $c$ . The **conditional self information** is

$$i(s|c) = -\log p(s|c)$$

The conditional information can be either more or less than the unconditional information.

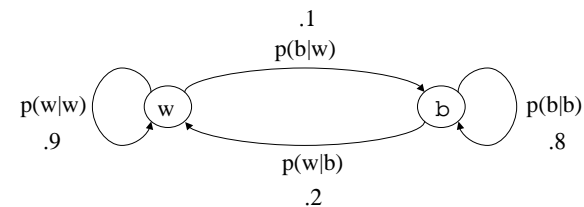
The **conditional entropy** is the weighted average of the conditional self information

$$H(S|C) = \sum_{c \in C} \left( p(c) \sum_{s \in S} p(s|c) \log \frac{1}{p(s|c)} \right)$$

15-853

Page 15

## Example of a Markov Chain



15-853

Page 16

## Entropy of the English Language

How can we measure the information per character?

ASCII code = 7

Entropy = 4.5 (based on character probabilities)

Huffman codes (average) = 4.7

Unix Compress = 3.5

Gzip = 2.5

RK = 1.79 (current close to best text compressor)

Must be less than 1.79.

15-853

Page 17

## Shannon's experiment

Asked humans to predict the next character given the whole previous text. He used these as conditional probabilities to estimate the entropy of the English Language.

The number of guesses required for right answer:

# of guesses	1	2	3	4	5	> 5
Probability	.79	.08	.03	.02	.02	.05

From the experiment he predicted

**H(English) = .6-1.3**

15-853

Page 18

## Compression Outline

**Introduction:** Lossy vs. Lossless, Benchmarks, ...

**Information Theory:** Entropy, etc.

➔ **Probability Coding:**

- Prefix codes and relationship to Entropy
- Huffman codes
- Arithmetic codes

**Applications of Probability Coding:** PPM + others

**Lempel-Ziv Algorithms:** LZ77, gzip, compress, ...

**Other Lossless Algorithms:** Burrows-Wheeler

**Lossy algorithms for images:** JPEG, MPEG, ...

**Compressing graphs and meshes:** BBK

15-853

Page 19

## Assumptions and Definitions

Communication (or a file) is broken up into pieces called **messages**.

Each message come from a **message set**  $S = \{s_1, \dots, s_n\}$  with a **probability distribution**  $p(s)$ .

Probabilities must sum to 1. Set can be infinite.

**Code  $C(s)$ :** A mapping from a message set to **codewords**, each of which is a string of bits

**Message sequence:** a sequence of messages

**Note:** Adjacent messages might be of a different types and come from a different probability distributions

15-853

Page 20

## Discrete or Blended

We will consider two types of coding:

**Discrete:** each message is a fixed set of bits

- Huffman coding, Shannon-Fano coding

01001 11 0001 011

message: 1 2 3 4

**Blended:** bits can be "shared" among messages

- Arithmetic coding

010010111010

message: 1,2,3, and 4

15-853

Page 21

## Uniquely Decodable Codes

A **variable length code** assigns a bit string (codeword) of variable length to every message value

e.g. a = 1, b = 01, c = 101, d = 011

What if you get the sequence of bits

1011?

Is it aba, ca, or, ad?

A **uniquely decodable code** is a variable length code in which bit strings can always be uniquely decomposed into its codewords.

15-853

Page 22

## Prefix Codes

A **prefix code** is a variable length code in which no codeword is a prefix of another word.

e.g., a = 0, b = 110, c = 111, d = 10

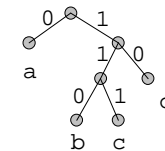
All prefix codes are uniquely decodable

15-853

Page 23

## Prefix Codes: as a tree

Can be viewed as a binary tree with message values at the leaves and 0s or 1s on the edges:



a = 0, b = 110, c = 111, d = 10

15-853

Page 24

## Some Prefix Codes for Integers

n	Binary	Unary	Gamma
1	..001	0	0
2	..010	10	10 0
3	..011	110	10 1
4	..100	1110	110 00
5	..101	11110	110 01
6	..110	111110	110 10

Many other fixed prefix codes:  
Golomb, phased-binary, subexponential, ...

15-853

Page 25

## Average Length

For a code  $\mathcal{C}$  with associated probabilities  $p(c)$  the **average length** is defined as

$$l_a(\mathcal{C}) = \sum_{c \in \mathcal{C}} p(c)l(c)$$

We say that a prefix code  $\mathcal{C}$  is **optimal** if for all prefix codes  $\mathcal{C}'$ ,  $l_a(\mathcal{C}) \leq l_a(\mathcal{C}')$

$l(c)$  = length of the codeword  $c$  (a positive integer)

15-853

Page 26

## Relationship to Entropy

**Theorem (lower bound):** For any probability distribution  $p(S)$  with associated uniquely decodable code  $\mathcal{C}$ ,

$$H(S) \leq l_a(\mathcal{C})$$

**Theorem (upper bound):** For any probability distribution  $p(S)$  with associated optimal prefix code  $\mathcal{C}$ ,

$$l_a(\mathcal{C}) \leq H(S) + 1$$

15-853

Page 27

## Kraft McMillan Inequality

**Theorem (Kraft-McMillan):** For any uniquely decodable code  $\mathcal{C}$ ,

$$\sum_{c \in \mathcal{C}} 2^{-l(c)} \leq 1$$

Also, for any set of lengths  $L$  such that

$$\sum_{l \in L} 2^{-l} \leq 1$$

there is a prefix code  $\mathcal{C}$  such that

$$l(c_i) = l_i (i = 1, \dots, |L|)$$

15-853

Page 28

## Proof of the Upper Bound (Part 1)

Assign each message a length:  $l(s) = \lceil \log(1/p(s)) \rceil$

We then have

$$\begin{aligned} \sum_{s \in S} 2^{-l(s)} &= \sum_{s \in S} 2^{-\lceil \log(1/p(s)) \rceil} \\ &\leq \sum_{s \in S} 2^{-\log(1/p(s))} \\ &= \sum_{s \in S} p(s) \\ &= 1 \end{aligned}$$

So by the Kraft-McMillan inequality there is a prefix code with lengths  $l(s)$ .

15-853

Page 29

## Proof of the Upper Bound (Part 2)

Now we can calculate the average length given  $l(s)$

$$\begin{aligned} l_a(S) &= \sum_{s \in S} p(s)l(s) \\ &= \sum_{s \in S} p(s) \cdot \lceil \log(1/p(s)) \rceil \\ &\leq \sum_{s \in S} p(s) \cdot (1 + \log(1/p(s))) \\ &= 1 + \sum_{s \in S} p(s) \log(1/p(s)) \\ &= 1 + H(S) \end{aligned}$$

And we are done.

15-853

Page 30

## Another property of optimal codes

**Theorem:** If  $C$  is an optimal prefix code for the probabilities  $\{p_1, \dots, p_n\}$  then  $p_i > p_j$  implies  $l(c_i) \leq l(c_j)$

**Proof:** (by contradiction)

Assume  $l(c_i) > l(c_j)$ . Consider switching codes  $c_i$  and  $c_j$ . If  $l_a$  is the average length of the original code, the length of the new code is

$$\begin{aligned} l'_a &= l_a + p_j(l(c_i) - l(c_j)) + p_i(l(c_j) - l(c_i)) \\ &= l_a + (p_j - p_i)(l(c_i) - l(c_j)) \\ &< l_a \end{aligned}$$

This is a contradiction since  $l_a$  is not optimal

15-853

Page 31

## Huffman Codes

Invented by Huffman as a class assignment in 1950. Used in many, if not most, compression algorithms gzip, bzip, jpeg (as option), fax compression,...

**Properties:**

- Generates optimal prefix codes
- Cheap to generate codes
- Cheap to encode and decode
- $l_a = H$  if probabilities are powers of 2

15-853

Page 32



## Huffman Codes

### Huffman Algorithm:

Start with a forest of trees each consisting of a single vertex corresponding to a message  $s$  and with weight  $p(s)$

**Repeat until one tree left:**

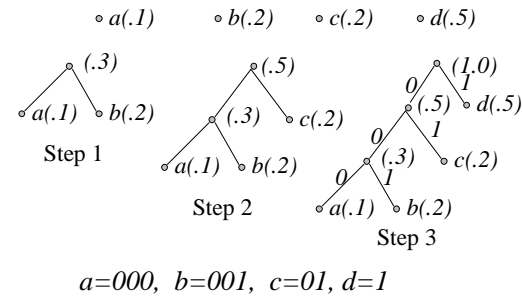
- Select two trees with minimum weight roots  $p_1$  and  $p_2$
- Join into single tree by adding root with weight  $p_1 + p_2$

15-853

Page 33

## Example

$$p(a) = .1, p(b) = .2, p(c) = .2, p(d) = .5$$



15-853

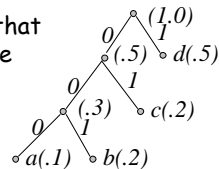
Page 34

## Encoding and Decoding

**Encoding:** Start at leaf of Huffman tree and follow path to the root. Reverse order of bits and send.

**Decoding:** Start at root of Huffman tree and take branch for each bit received. When at leaf can output message and return to root.

There are even faster methods that can process 8 or 32 bits at a time



15-853

Page 35

## Huffman codes are "optimal"

**Theorem:** The Huffman algorithm generates an optimal prefix code.

**Proof outline:**

Induction on the number of messages  $n$ .

Consider a message set  $S$  with  $n+1$  messages

1. Can make it so least probable messages of  $S$  are neighbors in the Huffman tree
2. Replace the two messages with one message with probability  $p(m_1) + p(m_2)$  making  $S'$
3. Show that if  $S'$  is optimal, then  $S$  is optimal
4.  $S'$  is optimal by induction

15-853

Page 36

## Problem with Huffman Coding

Consider a message with probability .999. The self information of this message is

$$-\log(.999) = .00144$$

If we were to send a 1000 such message we might hope to use  $1000 * .0014 = 1.44$  bits.

Using Huffman codes we require at least one bit per message, so we would require 1000 bits.

15-853

Page 37

## Arithmetic Coding: Introduction

Allows "blending" of bits in a message sequence.

Only requires 3 bits for the example

Can bound total bits required based on sum of self information:

$$l < 2 + \sum_{i=1}^n s_i$$

Used in PPM, JPEG/MPEG (as option), DMM

More expensive than Huffman coding, but integer implementation is not too bad.

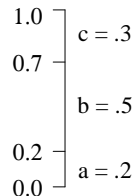
15-853

Page 38

## Arithmetic Coding: message intervals

Assign each probability distribution to an interval range from 0 (inclusive) to 1 (exclusive).

e.g.



$$f(i) = \sum_{j=1}^{i-1} p(j)$$

$$f(a) = .0, \quad f(b) = .2, \quad f(c) = .7$$

The interval for a particular message will be called the **message interval** (e.g for b the interval is [.2,.7))

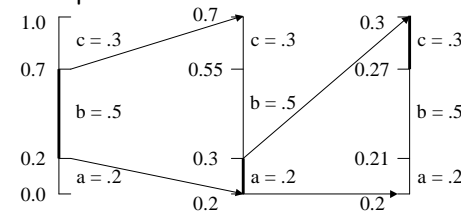
15-853

Page 39

## Arithmetic Coding: sequence intervals

Code a message sequence by composing intervals.

For example: **bac**



The final interval is [.27, .3)

We call this the **sequence interval**

15-853

Page 40

## Arithmetic Coding: sequence intervals

To code a sequence of messages with probabilities

$p_i (i = 1..n)$  use the following:

$$l_i = f_i \quad l_i = l_{i-1} + s_{i-1}f_i$$

$$s_1 = p_1 \quad s_i = s_{i-1}p_i$$

Each message narrows the interval by a factor of  $p_i$ .

Final interval size:  $s_n = \prod_{i=1}^n p_i$

## Warning

Three types of interval:

- **message interval** : interval for a single message
- **sequence interval** : composition of message intervals
- **code interval** : interval for a specific code used to represent a sequence interval (discussed later)

## Uniquely defining an interval

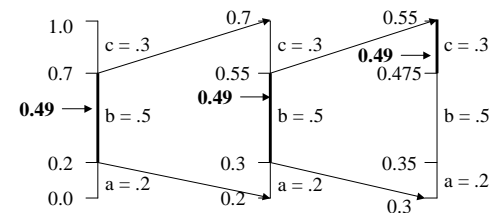
**Important property:** The sequence intervals for distinct message sequences of length  $n$  will never overlap

**Therefore:** specifying any number in the final interval uniquely determines the sequence.

Decoding is similar to encoding, but on each step need to determine what the message value is and then reduce interval

## Arithmetic Coding: Decoding Example

Decoding the number .49, knowing the message is of length 3:



The message is **bbc**.

## Representing Fractions

Binary fractional representation:

$$\begin{aligned} .75 &= .11 \\ 1/3 &= .010\bar{1} \\ 11/16 &= .1011 \end{aligned}$$

So how about just using the smallest binary fractional representation in the sequence interval.

e.g.  $[0, .33) = .01$   $[\.33, .66) = .1$   $[\.66, 1) = .11$

But what if you receive a 1?  
Should we wait for another 1?

15-853

Page 45

## Representing an Interval

Can view binary fractional numbers as intervals by considering all completions. e.g.

	min	max	interval
.11	$.11\bar{0}$	$.11\bar{1}$	$[\.75, 1.0)$
.101	$.101\bar{0}$	$.101\bar{1}$	$[\.625, .75)$

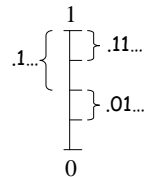
We will call this the **code interval**.

15-853

Page 46

## Code Intervals: example

$$[0, .33) = .01 \quad [\.33, .66) = .1 \quad [\.66, 1) = .11$$



Note that if code intervals overlap then one code is a prefix of the other.

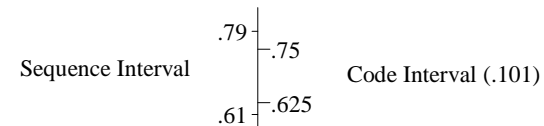
**Lemma:** *If a set of code intervals do not overlap then the corresponding codes form a prefix code.*

15-853

Page 47

## Selecting the Code Interval

To find a prefix code find a binary fractional number whose code interval is contained in the sequence interval.



Can use the fraction  $l + s/2$  truncated to

$$\lceil -\log(s/2) \rceil = 1 + \lceil -\log s \rceil$$

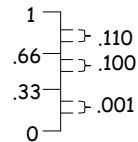
bits

15-853

Page 48

## Selecting a code interval: example

$$[0, .33) = .001 \quad [.33, .66) = .100 \quad [.66, 1) = .110$$



e.g: for  $[\text{.33...}, \text{.66...})$ ,  $l = \text{.33...}$ ,  $s = \text{.33...}$

$$l + s/2 = \text{.5} = \text{.1000...}$$

truncated to  $1 + \lceil -\log s \rceil = 1 + \lceil -\log(.33) \rceil = 3$  bits is  $\text{.100}$

Is this the best we can do for  $[0, .33)$ ?

## RealArith Encoding and Decoding

### RealArithEncode:

Determine  $l$  and  $s$  using original recurrences

Code using  $l + s/2$  truncated to  $1 + \lceil -\log s \rceil$  bits

### RealArithDecode:

Read bits as needed so code interval falls within a message interval, and then narrow sequence interval.

Repeat until  $n$  messages have been decoded .

## Bound on Length

**Theorem:** For  $n$  messages with self information  $\{s_1, \dots, s_n\}$  RealArithEncode will generate at most

$$2 + \sum_{i=1}^n s_i \text{ bits.}$$

$$\begin{aligned} \text{Proof: } 1 + \lceil -\log s \rceil &= 1 + \left\lceil -\log \left( \prod_{i=1}^n p_i \right) \right\rceil \\ &= 1 + \left\lceil \sum_{i=1}^n -\log p_i \right\rceil \\ &= 1 + \left\lceil \sum_{i=1}^n s_i \right\rceil \\ &< 2 + \sum_{i=1}^n s_i \end{aligned}$$

## Integer Arithmetic Coding

Problem with RealArithCode is that operations on arbitrary precision real numbers is expensive.

### Key Ideas of integer version:

Keep integers in range  $[0..R)$  where  $R=2^k$

Use rounding to generate integer sequence interval

Whenever sequence interval falls into top, bottom or middle half, expand the interval by factor of 2

This integer Algorithm is an approximation or the real algorithm.

## Integer Arithmetic Coding

The probability distribution as integers

Probabilities as counts:

e.g.  $c(a) = 11$ ,  $c(b) = 7$ ,  $c(c) = 30$

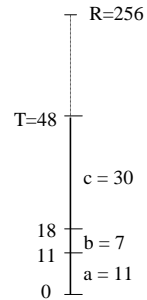
T is the sum of counts

e.g.  $48$  ( $11+7+30$ )

Partial sums  $f$  as before:

e.g.  $f(a) = 0$ ,  $f(b) = 11$ ,  $f(c) = 18$

Require that  $R > 4T$  so that probabilities do not get rounded to zero



15-853

Page 53

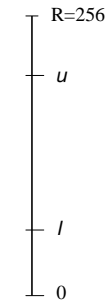
## Integer Arithmetic (contracting)

$$l_i = 0, \quad s_i = R$$

$$s_i = u_i - l_i + 1$$

$$u_i = l_i + \lfloor s_i \cdot (f_i + s_i) / T \rfloor - 1$$

$$l_i = l_i + \lfloor s_i \cdot f_i / T \rfloor$$



15-853

Page 54

## Integer Arithmetic (scaling)

If  $l \geq R/2$  then **(in top half)**

Output 1 followed by  $m$  0s

$m = 0$

Scale message interval by expanding by 2

If  $u < R/2$  then **(in bottom half)**

Output 0 followed by  $m$  1s

$m = 0$

Scale message interval by expanding by 2

If  $l \geq R/4$  and  $u < 3R/4$  then **(in middle half)**

Increment  $m$

Scale message interval by expanding by 2

15-853

Page 55