CSE 312

Foundations of Computing II

Lecture 16: Information Theory and Data Compression

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Stefano Tessaro

tessaro@cs.washington.edu

Announcements

- Office hours: I am available 1-3pm.
- Please make sure to read the instructions for the midterm.
- Practice midterm solutions posted in the afternoon.

Today

How much can we compress data?

How much information is really contained in data?

Central topic in **information theory**, a discipline based on probability which has been extremely useful across electrical engineering, computer science, statistics, physics, …

Claude Shannon, **“A Mathematical Theory of Communication”**, 1948

http://www.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Encoding Scheme

$$x \longrightarrow \boxed{\mathbf{enc}} \longrightarrow z = \mathbf{enc}(x) \longrightarrow \boxed{\mathbf{dec}} \longrightarrow x$$

$\mathbf{enc}: \mathcal{X} \to \{0,1\}^*$

$\mathbf{dec}: \mathcal{X} \to \{0,1\}^*$

Decodability. For all values $x \in \mathcal{X}$: $\mathbf{dec}(\mathbf{enc}(x)) = x$

Goal: Encoding should "<u>compress</u>"

[We will formalize this using the language of probability theory]

Encoding – Example

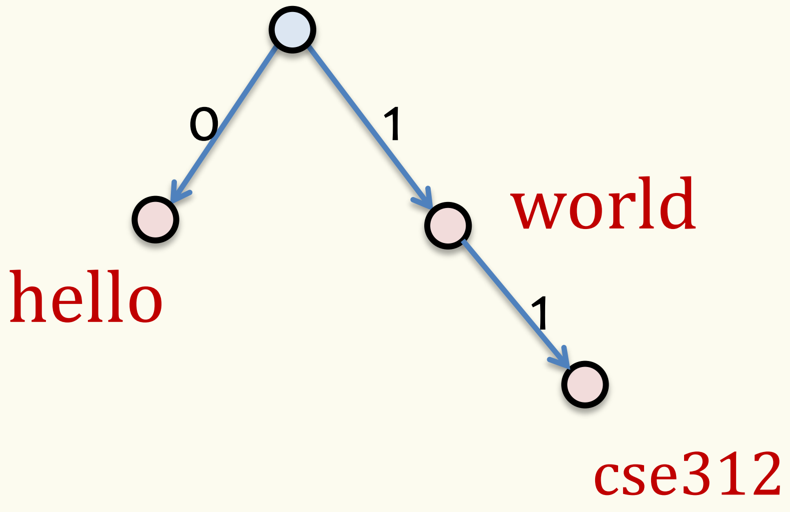
Say we need to encode a word from the set $\mathcal{X} = \{\text{hello}, \text{world}, \text{cse312}\}$

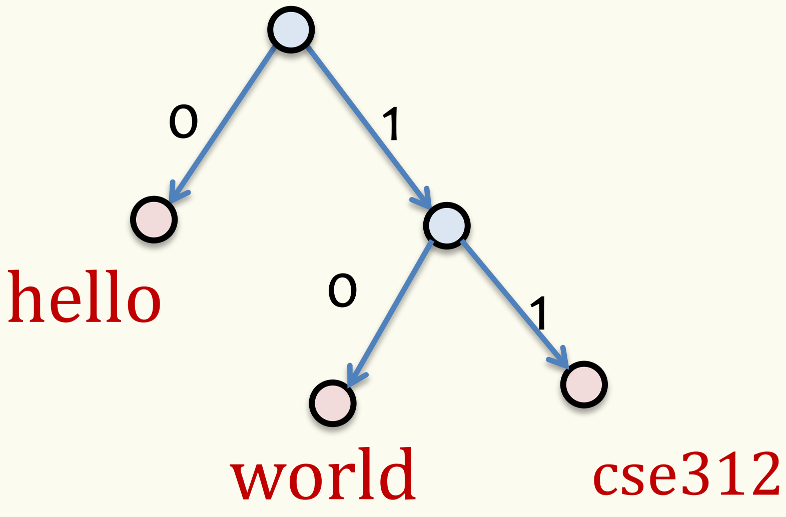
hello	$\longrightarrow$ 0		hello	$\longrightarrow$ 0		hello	$\longrightarrow$ 0
world	$\longrightarrow$ 1		world	$\longrightarrow$ 10		world	$\longrightarrow$ 11
cse312	$\longrightarrow$ 11		cse312	$\longrightarrow$ 11		cse312	$\longrightarrow$ 100000000
	$\mathbf{enc}$			$\mathbf{enc}$			$\mathbf{enc}$

Better Visualization – Trees

hello → 0
world → 1
cse312 → 11

hello → 0
world → 10
cse312 → 11

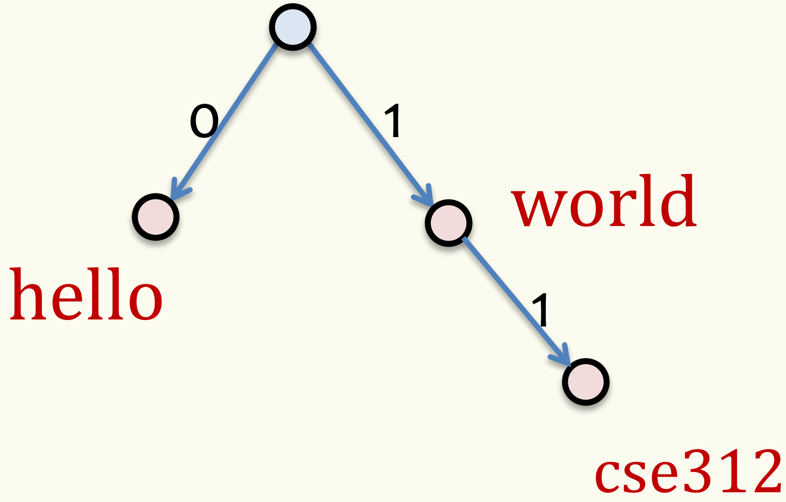




Focus – Prefix-free codes

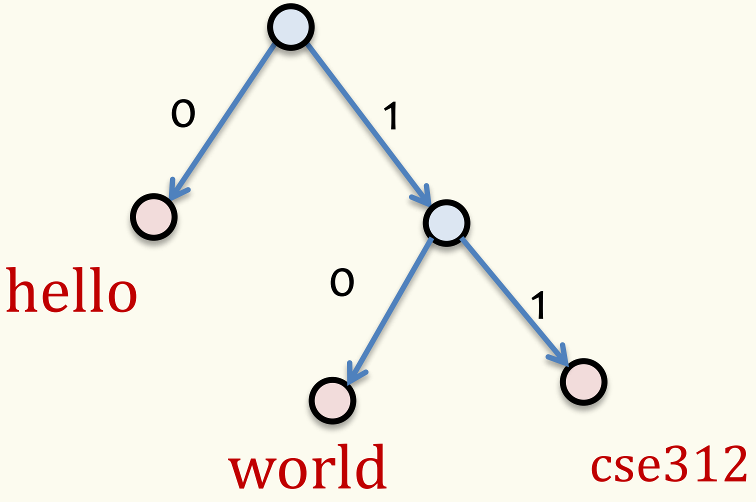
A code is **prefix-free** if no encoding is a **prefix** of another one.

i.e. every encoding is a leaf



Not prefix-free!

1 is a prefix of 11



Prefix-free!!

Random Variables – Arbitrary Values

We will consider random variables $X: \Omega \to \mathcal{X}$ taking values from a (finite) set $\mathcal{X}$. [We refer to these as a "random variable over the alphabet $\mathcal{X}$."]

Example: $\mathcal{X} = \{\text{hello}, \text{world}, \text{cse312}\}$

$\mathbb{p}_X(\text{hello}) = \frac{1}{2}$ $\mathbb{p}_X(\text{world}) = \frac{1}{4}$ $\mathbb{p}_X(\text{cse312}) = \frac{1}{4}$

The Data Compression Problem

Data = random variable X over alphabet $\mathcal{X}$

$$X \longrightarrow \boxed{\mathbf{enc}} \longrightarrow Z = \mathbf{enc}(X) \longrightarrow \boxed{\mathbf{dec}} \longrightarrow X$$

$\mathbf{enc}\colon \mathcal{X} \to \{0,1\}^*$

$\mathbf{dec}\colon \mathcal{X} \to \{0,1\}^*$

Two goals:

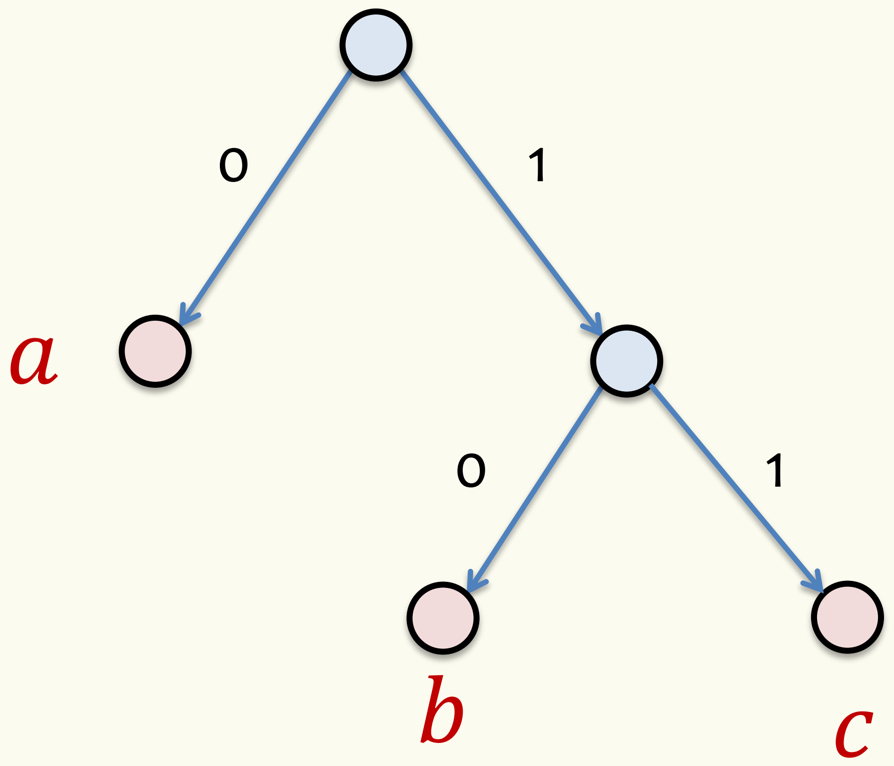
1. **Decodability.** For all values $x \in \mathcal{X}$: $\mathbf{dec}\big(\mathbf{enc}(x)\big) = x$
2. **Minimal length.** The <u>length</u> $|Z|$ of Z should be as small as possible

More formally: minimize $\mathbb{E}(|Z|)$

Expected Length – Example

$$\mathcal{X} = \{a, b, c\}$$

$$\mathbb{p}_X(a) = \frac{1}{2} \qquad \mathbb{p}_X(b) = \frac{1}{4} \qquad \mathbb{p}_X(c) = \frac{1}{4}$$



$$\mathbb{p}_Z(0) = \frac{1}{2}$$

$$\mathbb{p}_Z(10) = \frac{1}{4}$$

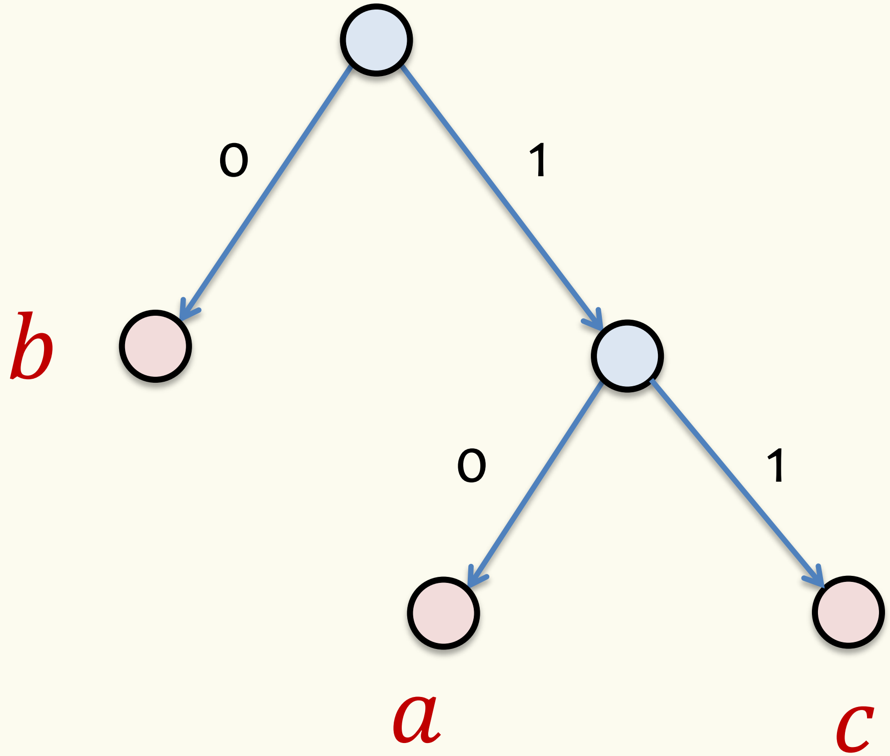
$$\mathbb{p}_Z(11) = \frac{1}{4}$$

$$\mathbb{E}(|Z|) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}$$

Expected Length – Example

$\mathcal{X} = \{a, b, c\}$

$\mathbb{p}_X(a) = \frac{1}{2}$ $\quad \mathbb{p}_X(b) = \frac{1}{4}$ $\quad \mathbb{p}_X(c) = \frac{1}{4}$



$$\mathbb{p}_Z(0) = \frac{1}{4}$$

$$\mathbb{p}_Z(10) = \frac{1}{2}$$

$$\mathbb{p}_Z(11) = \frac{1}{4}$$

$$\mathbb{E}(|Z|) = \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{7}{4}$$

What is the shortest encoding?

Problem. Given a random variable X, find optimal $(\mathbf{enc}, \mathbf{dec})$, i.e., $\mathbb{E}(|\mathbf{enc}(X)|)$ is a small as possible.

Next: There is an inherent limit on how short the encoding can be (in expectation).

Random Variables – Arbitrary Values

Assume you are given a random variable X with the following PMF:

x	a	b	c	d
$\mathbb{p}_X(x)$	$\frac{15}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$

You learn $X = a$; surprised? $s(a) = \log_2 16/15 \approx 0.09$

You learn $X = d$; surprised? $s(d) = 6$

Definition. The **surprise** of outcome x is $s(x) = \log_2\left(\frac{1}{\mathbb{p}_X(x)}\right)$

Entropy = Expected Surprise

Definition. The **entropy** of a discrete RV X over alphabet $\mathcal{X}$ is

$$\mathbb{H}(X) = \mathbb{E}\big(s(X)\big) = \sum_{x \in \mathcal{X}} \mathbb{p}_X(x) \cdot \log_2\left(\frac{1}{\mathbb{p}_X(x)}\right)$$

Weird convention: $0 \log_2 1/0 = 0$

Intuitively: Captures how surprising outcome of random variable is.

Entropy = Expected Surprise

Definition The entropy of a discrete RV X over alphabet $\mathcal{X}$ is

$$\mathbb{H}(X) = \mathbb{E}\big(s(X)\big) = \sum_{x \in \mathcal{X}} \mathbb{p}_X(x) \cdot \log_2 \left(\frac{1}{\mathbb{p}_X(x)} \right)$$

x	a	b	c	d
$\mathbb{p}_X(x)$	$\frac{15}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$

$$\mathbb{H}(X) = \frac{15}{16} \cdot \log_2 \frac{16}{15} + \frac{1}{32} \cdot 5 + \frac{1}{64} \cdot 6 + \frac{1}{64} \cdot 6$$

$$= \frac{15}{16} \log_2 \frac{16}{15} + \frac{11}{32} \approx 0.431 \ldots$$

Entropy = Expected Surprise

Definition. The **entropy** of a discrete RV X over alphabet $\mathcal{X}$ is

$$\mathbb{H}(X) = \mathbb{E}\big(s(X)\big) = \sum_{x \in \mathcal{X}} \mathbb{p}_X(x) \cdot \log_2\left(\frac{1}{\mathbb{p}_X(x)}\right)$$

x	a	b	c	d
$\mathbb{p}_X(x)$	1	0	0	0

x	a	b	c	d
$\mathbb{p}_X(x)$	$1/4$	$1/4$	$1/4$	$1/4$

$$\mathbb{H}(X) = 1 \cdot 0 + 3 \cdot 0 \log_2 \frac{1}{0} = 0$$

$$\mathbb{H}(X) = 4 \cdot \frac{1}{4} \log_2(4) = 2$$

Entropy = Expected Surprise

Definition The **entropy** of a discrete RV X over alphabet $\mathcal{X}$ is

$$\mathbb{H}(X) = \mathbb{E}\big(s(X)\big) = \sum_{x \in \mathcal{X}} \mathbb{p}_X(x) \cdot \log_2\left(\frac{1}{\mathbb{p}_X(x)}\right)$$

Proposition. $0 \le \mathbb{H}(X) \le \log_2 |\mathcal{X}|$

Uniform distribution

Takes one value with prob 1

Shannon's Source Coding Theorem

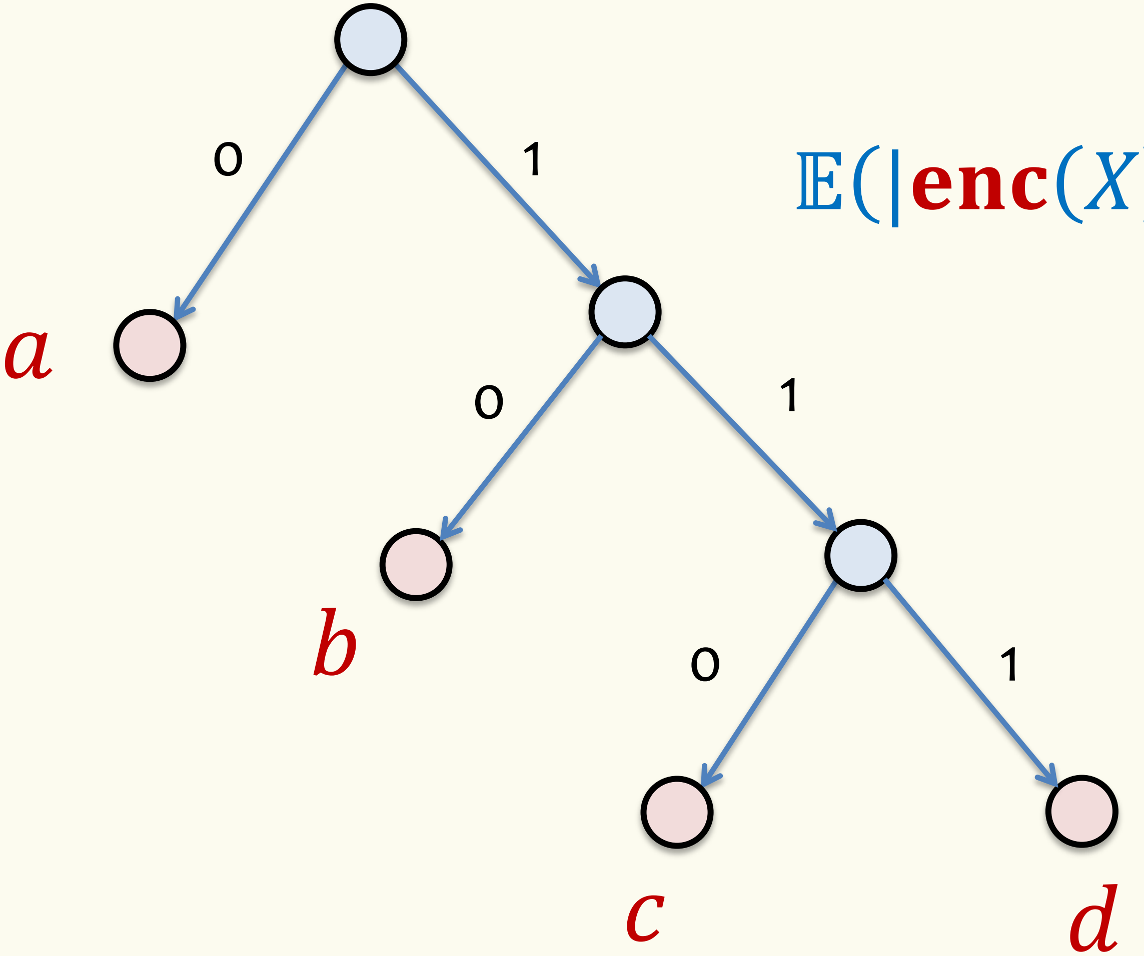
Theorem. (Source Coding Theorem) Let $(\mathbf{enc}, \mathbf{dec})$ be an optimal prefix-free encoding scheme for a RV X, then

$$\mathbb{H}(X) \leq \mathbb{E}(|\mathbf{enc}(X)|) \leq \mathbb{H}(X) + 1$$

- We cannot compress beyond the entropy
  - Corollary: "uniform" data cannot be compressed
- We can get within one bit of it.
- Example of optimal code: Huffman Code (CSE 143?)
- Result can be extended to uniquely decodable codes. (E.g., suffix free)

Example

x	a	b	c	d
$\mathbb{p}_X(x)$	$\frac{15}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$



$$\mathbb{E}(|\mathbf{enc}(X)|) = \frac{15}{16} \cdot 1 + \frac{1}{32} \cdot 2 + 2 \cdot \frac{1}{64} \cdot 3$$

$$= \frac{15}{16} + \frac{10}{64} = \frac{70}{64} \leq \mathbb{H}(X) + 1$$

Data Compression in the Real World

Main issue: we do not know the distribution of X

- Universal compression: Lempel/Ziv/Welch
  - See http://web.mit.edu/6.02/www/f2011/handouts/3.pdf
  - Used in GIF, UNIX compress.
  - General idea: Assume data is sequence of symbols generated from a random process to be "estimated".
- Whole area of computer science dedicated to the topic.
- This is lossless compression, very different from "lossy compression" used in images, videos, audio etc.
  - Assumes humans can be "fooled" with some loss of data