

# What is Information?

Take a moment to consider this.

Related concepts:

- data,
- instruction,
- knowledge,
- meaning,
- mental stimulus,
- pattern,
- perception,
- representation.

Information is always *about* something. Information is *context-specific*.

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

**Information**  
Information  
Theory  
History

This Course  
Goal

Teaching  
Approach  
The Book  
Study Suggestion

# Information: Example

## Example

You walk into your office one day, and on your desk, there is a post-it message carrying a piece of data →



Is this information? What does it tell us?

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

**Information**  
Information  
Theory  
History

This Course  
Goal

Teaching  
Approach  
The Book  
Study Suggestion

# Information: Example

## Example

You walk into your office one day, and on your desk, there is a post-it message carrying a piece of data →



Is this information? What does it tell us?

- What does it say? (interpretation)
- Who sent it? (source)
- Is the source reliable? (trust)
- Is it relevant/valuable to us? (quality)

*Data* becomes *information* when we *learn* something from it.  
Also, data is *useless* until it becomes information.

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

**Information**  
Information  
Theory  
History

This Course

Goal  
Teaching  
Approach  
The Book  
Study Suggestion

So, what *is* information?

**Definition:** Context-specific data (generated by some *source*).

**Meaning:** Depends on the context.

**Subjectivity alert:** This conception of information is *beyond the scope of science*. It is not rigorous. Immeasurable.

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

**Information**  
Information  
Theory  
History

This Course

Goal  
Teaching  
Approach  
The Book  
Study Suggestion

# What is Information Theory?

- Branch of Applied Mathematics.
- Main concern: quantification of *information*.

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

Information  
**Information  
Theory**  
History

This Course

Goal  
Teaching  
Approach  
The Book  
Study Suggestion

# What is Information Theory?

- Branch of Applied Mathematics.
- Main concern: quantification of *information*.

**Problem:** But wait, we just said that information **cannot be quantified**.

**Solution:** *define information rigorously* to be that which is *quantifiable* in a piece of data.

- Care nothing about *qualitative* properties of data.

## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

Information  
**Information  
Theory**  
History

This Course  
Goal

Teaching  
Approach  
The Book

Study Suggestion

About Me

Who am I?

About  
Information  
TheoryInformation  
Theory  
History

This Course

Goal  
Teaching  
Approach  
The Book  
Study Suggestion

Wait,

**Q:** We can't just *re-define information*!

**A:** Yes we *can*; we are mathematicians.<sup>2</sup> It is common to associate subjective words with theoretical concepts when the word somewhat relates to our concept.

- Derivative
- Group, Ring
- Natural numbers

We could have called this concept whatever we wanted<sup>3</sup>. The name for our concept does not matter; it is how it is defined that does.

---

<sup>2</sup>We are not really re-naming information as we know it. We are associating the name “information” to the concept we are interested in.

<sup>3</sup>For instance, “Snickersnack”, “Shawarma”, or, more appropriately, “*uncertainty*”.

## Back on topic

The founder of this field had a very specific purpose in mind:

*Find the fundamental limits of compressing and reliably communicating data.*<sup>4</sup>

He devised a model suitable for his purpose.

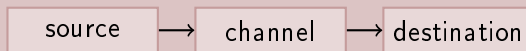


Figure: Communication System

He **defines** that the source has a finite set  $S$  of types of data (symbols) to send, and at any given time, has a certain probability to send symbol  $s$ ,  $\forall s \in S$ , that these probabilities do not change, and are independent on prior  $s$  sent.

---

<sup>4</sup>Eliminate redundancy.



## Definition (Information (informal))

*Information* in  $s$  sent by source  $S$  is the uncertainty involved in whether  $S$  would send  $s$  next.

- This *is* measurable; uncertainty is a function of probability<sup>5</sup>! This model is **objective** as we have completely disregarded *qualitative* properties of  $s$ .
- We shall see later that low uncertainty (predictability) implies redundancy.
- *This theory aims to minimise redundancy!*

So, when I say “information”, think along the lines of *randomness, redundancy, uncertainty, and surprise*.

---

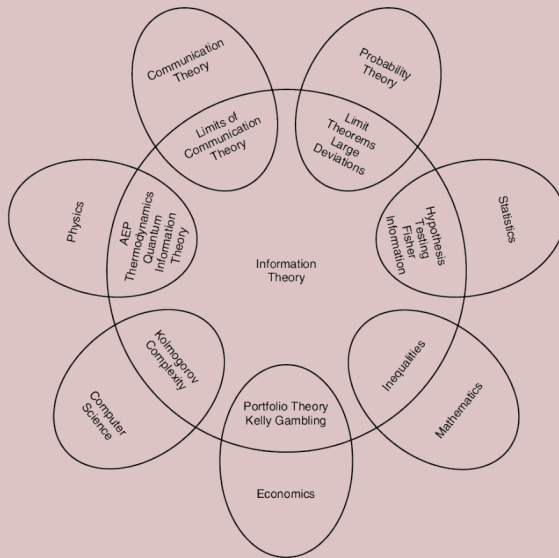
<sup>5</sup>We will see in a week or two what this function is (it is specially tailored for our needs. We’re mathematicians; we can do that).

About Me  
Who am I?

About  
Information  
Theory  
Information  
Theory  
History

This Course  
Goal  
Teaching  
Approach  
The Book  
Study Suggestion

# Relation to Other Fields



## Course Introduction

About Me  
Who am I?

About  
Information  
Theory

Information  
Information  
Theory  
**History**

This Course

Goal  
Teaching  
Approach  
The Book  
Study Suggestion

## Definition

Recall our prior, informal, definition of information (course introduction).

### Definition (Information (informal))

*Information* in  $s$  sent by source  $\mathcal{S}$  is the uncertainty involved in whether  $\mathcal{S}$  would send  $s$  next.

<sup>1</sup>Uncertainty decreases as  $p_i$  increases. There is no uncertainty in the certain event.

<sup>2</sup>Symbol emission of  $\mathcal{S}$  are iid; receiving  $s_i$  tells us *nothing* about  $s_j$  (this would change if iid is revoked).

# Definition

Recall our prior, informal, definition of information (course introduction).

## Definition (Information (informal))

*Information* in  $s$  sent by source  $\mathcal{S}$  is the uncertainty involved in whether  $\mathcal{S}$  would send  $s$  next.

Let's define a function,  $I : S^* \rightarrow [0, +\infty)$ , to express this. We require:

- (1)  $I$  is a decreasing function of the probability of  $p_i$  of  $s_i$ , with  $I(s_i) = 0$  if  $p_i = 1^1$ .
- (2)  $I(s_i s_j) = I(s_i) + I(s_j)^2$ .

Note, we have  $\Pr(s_i s_j) = \Pr(s_i) * \Pr(s_j) = p_i p_j, \forall i, j$ , since symbol emission in  $\mathcal{S}$  is iid.

<sup>1</sup>Uncertainty decreases as  $p_i$  increases. There is no uncertainty in the certain event.

<sup>2</sup>Symbol emission of  $\mathcal{S}$  are iid; receiving  $s_i$  tells us *nothing* about  $s_j$  (this would change if iid is revoked).

## Definition

Recall our prior, informal, definition of information (course introduction).

### Definition (Information (informal))

*Information* in  $s$  sent by source  $\mathcal{S}$  is the uncertainty involved in whether  $\mathcal{S}$  would send  $s$  next.

Let's define a function,  $I : \mathcal{S}^* \rightarrow [0, +\infty)$ , to express this. We require:

- (1)  $I$  is a decreasing function of the probability of  $p_i$  of  $s_i$ , with  $I(s_i) = 0$  if  $p_i = 1^1$ .
- (2)  $I(s_i s_j) = I(s_i) + I(s_j)^2$ .

Note, we have  $\Pr(s_i s_j) = \Pr(s_i) * \Pr(s_j) = p_i p_j$ ,  $\forall i, j$ , since symbol emission in  $\mathcal{S}$  is iid. Conditions (1) and (2) are satisfied if we define

### Definition (Information Measure $I(s)$ )

The *information* conveyed by symbol  $s_i \in \mathcal{S}$ , written  $I(s_i)$ , is defined

$$I(s_i) = -\log p_i = \log \frac{1}{p_i}.$$

<sup>1</sup>Uncertainty decreases as  $p_i$  increases. There is no uncertainty in the certain event.

<sup>2</sup>Symbol emission of  $\mathcal{S}$  are iid; receiving  $s_i$  tells us *nothing* about  $s_j$  (this would change if iid is revoked).

## Let us Verify

The function

$$I(s_i) = -\log p_i = \log \frac{1}{p_i}$$

is

Faithful to (1), since  $\frac{1}{p_i}$  decreases as  $p_i$  increases.

## Lecture 3: A Measure of Information

Optimality of Huffman Codes
$L(\mathcal{C})$ of Huffman Codes
Huffman Codes are Optimal
Source Extension
Information
$I(s)$ measure
$I(s)$ is uniquely defined
Entropy
$H(S)$ measure
Properties of $H(S)$
Epilogue

---

<sup>3</sup>Fortunately,  $s_i$  is never sent.

## Let us Verify

The function

$$I(s_i) = -\log p_i = \log \frac{1}{p_i}$$

is

Faithful to (1), since  $\frac{1}{p_i}$  decreases as  $p_i$  increases.

Faithful to (2), since

$$I(s_i s_j) = \log \frac{1}{p_i p_j} = \log \frac{1}{p_i} + \log \frac{1}{p_j} = I(s_i) + I(s_j).$$

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information

$I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy

$H(S)$  measure  
Properties of  $H(S)$

Epilogue

---

<sup>3</sup>Fortunately,  $s_i$  is never sent.

## Let us Verify

The function

$$I(s_i) = -\log p_i = \log \frac{1}{p_i}$$

is

Faithful to (1), since  $\frac{1}{p_i}$  decreases as  $p_i$  increases.

Faithful to (2), since

$$I(s_i s_j) = \log \frac{1}{p_i p_j} = \log \frac{1}{p_i} + \log \frac{1}{p_j} = I(s_i) + I(s_j).$$

Also, as  $I(s_i) \rightarrow +\infty$  as  $p_i \rightarrow 0$ , we say  $I(s_i) = +\infty$  if  $p_i = 0^3$ .

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information

$I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy

$H(S)$  measure  
Properties of  $H(S)$

Epilogue

---

<sup>3</sup>Fortunately,  $s_i$  is never sent.



## Let us Verify

The function

$$I(s_i) = -\log p_i = \log \frac{1}{p_i}$$

is

Faithful to (1), since  $\frac{1}{p_i}$  decreases as  $p_i$  increases.

Faithful to (2), since

$$I(s_i s_j) = \log \frac{1}{p_i p_j} = \log \frac{1}{p_i} + \log \frac{1}{p_j} = I(s_i) + I(s_j).$$

Also, as  $I(s_i) \rightarrow +\infty$  as  $p_i \rightarrow 0$ , we say  $I(s_i) = +\infty$  if  $p_i = 0^3$ .

## Example (Biased die)

For  $S = (6, 5, 4, 3, 2, 1)$  and  $P = (\frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ , we have

$$I(6) = \log \frac{1}{\frac{1}{2}} = \log 2 = 1 \quad I(3) = \log \frac{1}{\frac{1}{10}} = \log 10 \approx 2.885$$

Getting a 3 is more “surprising” than getting a 6.

<sup>3</sup>Fortunately,  $s_i$  is never sent.

# Could We Have Defined $I$ Differently?

## Theorem (Uniqueness of $I$ (Exercise 3.7))

Let  $f$  be

- (1) a strictly decreasing function  $(0, 1] \rightarrow \mathbb{R}$  s.t.
- (2)  $f(ab) = f(a) + f(b)$ ,  $\forall a, b \in (0, 1]$ .

Then  $f$  is defined as  $f(x) = -\log_r x$  for some  $r > 1$ , thus justifying  $I$ .

## Proof

We check how  $f$  reacts on specific input, and show that because of its behaviour, it must have the above-stated form.

# Could We Have Defined $I$ Differently?

## Theorem (Uniqueness of $I$ (Exercise 3.7))

Let  $f$  be

- (1) a strictly decreasing function  $(0, 1] \rightarrow \mathbb{R}$  s.t.
- (2)  $f(ab) = f(a) + f(b)$ ,  $\forall a, b \in (0, 1]$ .

Then  $f$  is defined as  $f(x) = -\log_r x$  for some  $r > 1$ , thus justifying  $I$ .

## Proof

We check how  $f$  reacts on specific input, and show that because of its behaviour, it must have the above-stated form.

Let  $g = f(e^{-x})$ . This function is strictly increasing on  $[0, +\infty)$ , with

$$\begin{aligned} g(x+y) &= f(e^{-(x+y)}) = f\left(\frac{1}{e^{x+y}}\right) = f\left(\frac{1}{e^x} * \frac{1}{e^y}\right) \\ &= f\left(\frac{1}{e^x}\right) + f\left(\frac{1}{e^y}\right) = f(e^{-x}) + f(e^{-y}) \\ &= g(x) + g(y). \end{aligned}$$

Cont.

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information  
 $I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy  
 $H(S)$  measure  
Properties of  $H(S)$

Epilogue

## Proof

**Note:**  $g = f(e^{-x})$ , with  $g(x+y) = g(x) + g(y)$ .

With  $x = y = 0$ , we have  $g(0+0) = \underline{g(0) = 0}$ . Let  $\underline{c = g(1)}$ . We now aim to show that  $g(x) = cx$ ,  $\forall x \geq 0$ . Why?

*Because then, since  $g(x) = f(e^{-x}) = cx$ , we must have*  
 $f(x) = -\ln x = -\log_r x$  as required, with  $r = e^{\frac{1}{c}} > 0$ .

So here we go:

### Lecture 3: A Measure of Information

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information

$I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy

$H(S)$  measure  
Properties of  $H(S)$

Epilogue

## Proof

**Note:**  $g = f(e^{-x})$ , with  $g(x+y) = g(x) + g(y)$ .

With  $x = y = 0$ , we have  $g(0+0) = \underline{g(0) = 0}$ . Let  $\underline{c = g(1)}$ . We now aim to show that  $g(x) = cx$ ,  $\forall x \geq 0$ . Why?

*Because then, since  $g(x) = f(e^{-x}) = cx$ , we must have  $f(x) = -\ln x = -\log_r x$  as required, with  $r = e^{\frac{1}{c}} > 0$ .*

So here we go: For  $n \geq 0$ ,

$$g(2^n) = g(\underbrace{1 + \dots + 1}_{2^n \text{ times}}) = \underbrace{g(1) + \dots + g(1)}_{2^n \text{ times}} = c2^n.$$

$$g(1) = g\left(2^n * \frac{1}{2^n}\right) = \underbrace{g\left(\frac{1}{2^n}\right) + \dots + g\left(\frac{1}{2^n}\right)}_{2^n \text{ times}}, \text{ so } g\left(\frac{1}{2^n}\right) = \frac{c}{2^n}.$$

### Lecture 3: A Measure of Information

Optimality of Huffman Codes  
 $L(\mathcal{C})$  of Huffman Codes  
 Huffman Codes are Optimal  
 Source Extension

#### Information

$I(s)$  measure  
 $I(s)$  is uniquely defined

#### Entropy

$H(S)$  measure  
 Properties of  $H(S)$

#### Epilogue

## Proof

**Note:**  $g = f(e^{-x})$ , with  $g(x+y) = g(x) + g(y)$ .

With  $x = y = 0$ , we have  $g(0+0) = \underline{g(0) = 0}$ . Let  $\underline{c = g(1)}$ . We now aim to show that  $g(x) = cx$ ,  $\forall x \geq 0$ . Why?

*Because then, since  $g(x) = f(e^{-x}) = cx$ , we must have  $f(x) = -\ln x = -\log_r x$  as required, with  $r = e^{\frac{1}{c}} > 0$ .*

So here we go: For  $n \geq 0$ ,

$$g(2^n) = g(\underbrace{1 + \dots + 1}_{2^n \text{ times}}) = \underbrace{g(1) + \dots + g(1)}_{2^n \text{ times}} = c2^n.$$

$$g(1) = g\left(2^n * \frac{1}{2^n}\right) = g\left(\underbrace{\frac{1}{2^n} + \dots + \frac{1}{2^n}}_{2^n \text{ times}}\right), \text{ so } g\left(\frac{1}{2^n}\right) = \frac{c}{2^n}.$$

As such,  $g(2^n) = c2^n$ ,  $\forall n$ . Almost done. How about all other  $x$ ?

### Lecture 3: A Measure of Information

Optimality of Huffman Codes  
 $L(\mathcal{C})$  of Huffman Codes  
 Huffman Codes are Optimal  
 Source Extension

#### Information

$I(s)$  measure  
 $I(s)$  is uniquely defined

#### Entropy

$H(S)$  measure  
 Properties of  $H(S)$

#### Epilogue

## Proof

**Note:**  $g = f(e^{-x})$ , with  $g(x+y) = g(x) + g(y)$ .

With  $x = y = 0$ , we have  $g(0+0) = \underline{g(0) = 0}$ . Let  $\underline{c = g(1)}$ . We now aim to show that  $g(x) = cx$ ,  $\forall x \geq 0$ . Why?

*Because then, since  $g(x) = f(e^{-x}) = cx$ , we must have  $f(x) = -\ln x = -\log_r x$  as required, with  $r = e^{\frac{1}{c}} > 0$ .*

So here we go: For  $n \geq 0$ ,

$$g(2^n) = g(\underbrace{1 + \dots + 1}_{2^n \text{ times}}) = \underbrace{g(1) + \dots + g(1)}_{2^n \text{ times}} = c2^n.$$

$$g(1) = g\left(2^n * \frac{1}{2^n}\right) = \underbrace{g\left(\frac{1}{2^n}\right) + \dots + g\left(\frac{1}{2^n}\right)}_{2^n \text{ times}}, \text{ so } g\left(\frac{1}{2^n}\right) = \frac{c}{2^n}.$$

As such,  $g(2^n) = c2^n$ ,  $\forall n$ . Almost done. How about all other  $x$ ? Well, any  $x$  can be written as a sum of 2s in powers. For  $a_i \in \{0, 1\}$ ,

$$x = \sum_{n=0}^N a_n 2^n, \text{ for } x \in \mathbb{Z}^+ \qquad x = \sum_{n=-\infty}^N a_n 2^n, \text{ for } x \in \mathbb{R}^+ \quad (1)$$

Cont.

### Lecture 3: A Measure of Information

Optimality of Huffman Codes  
 $L(\mathcal{C})$  of Huffman Codes  
 Huffman Codes are Optimal  
 Source Extension

#### Information

$I(s)$  measure  
 $I(s)$  is uniquely defined

#### Entropy

$H(S)$  measure  
 Properties of  $H(S)$

#### Epilogue

## Proof, cont.

**Note:**  $g = f(e^{-x})$ ,  $g(x+y) = g(x) + g(y)$ ,  $g(2^n) = c2^n$ ,  $\forall n$ .

Use this to prove  $g(x) = cx$ ,  $\forall x \geq 0$ .

For all  $M \leq N$ , we have

$$\sum_{n=M}^N a_n 2^n \leq x \leq \sum_{n=M}^N a_n 2^n + 2^M.$$

Apply  $g$ . From linearity of  $g$ , we get

$$\sum_{n=M}^N c a_n 2^n \leq g(x) \leq \sum_{n=M}^N c a_n 2^n + c 2^M.$$

Divide by  $c$ . We obtain

$$\sum_{n=M}^N a_n 2^n \leq \frac{g(x)}{c} \leq \sum_{n=M}^N a_n 2^n + 2^M.$$

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information

$I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy

$H(S)$  measure  
Properties of  $H(S)$

Epilogue



## Proof, cont.

**Note:**  $g = f(e^{-x})$ ,  $g(x+y) = g(x) + g(y)$ ,  $g(2^n) = c2^n$ ,  $\forall n$ .

Use this to prove  $g(x) = cx$ ,  $\forall x \geq 0$ .

For all  $M \leq N$ , we have

$$\sum_{n=M}^N a_n 2^n \leq x \leq \sum_{n=M}^N a_n 2^n + 2^M.$$

Apply  $g$ . From linearity of  $g$ , we get

$$\sum_{n=M}^N c a_n 2^n \leq g(x) \leq \sum_{n=M}^N c a_n 2^n + c 2^M.$$

Divide by  $c$ . We obtain

$$\sum_{n=M}^N a_n 2^n \leq \frac{g(x)}{c} \leq \sum_{n=M}^N a_n 2^n + 2^M.$$

“Sandwich” tightens for  $M \rightarrow -\infty$ ; *equality at limits*. So,  $\frac{g(x)}{c} = x$ .  
Thus,  $g(x) = cx$ . Result follows. □

**Note:** In this proof,  $r = e^{\frac{1}{c}}$ . And  $r$  is the base of our log.  $r$  still represents the arity of our code. To obtain a *desired*  $r$ , we pick a  $c$  such that  $e^{\frac{1}{c}}$  yields our desired  $r$ . Thus, we allow ourselves to use any  $r$ .

Optimality of  
Huffman Codes  
 $L(\mathcal{C})$  of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension  
Information  
 $I(s)$  measure  
 $I(s)$  is uniquely  
defined  
Entropy  
 $H(S)$  measure  
Properties of  $H(S)$   
Epilogue

# Information Entropy

- We have seen how to measure the information in a message:  $I(s)$ .
- How much information does  $\mathcal{S}$  convey on average?

## Definition (Expected (average) Value of a Random Variable)

Let  $X$  be a discrete random variable with image  $\mathcal{X} = \{x_1, \dots\}$  and probability mass function  $p: \mathcal{X} \rightarrow [0, 1]$ . The *expected value* of  $X$  is defined as

$$E[X] = \sum_{k=1}^{\infty} x_k p(x_k).$$

Denote by  $X$  the value  $\mathcal{S}$  might emit at some point.

## Definition (Information Entropy)

The average amount of information conveyed by  $\mathcal{S}$ ,  $H$ , called the *information entropy* of  $\mathcal{S}$ , is defined

$$H(X) = E[I(X)].$$

**Note:** We will abuse notation and write  $H(\mathcal{S})$  instead of  $H(X)$ .

# Explicit Definition

We have

$$H(X) = E[I(X)].$$

Let

$$I_r(s_i) = -\log_r p_i.^4$$

Since  $I_r(X)$  is *itself* a random variable,  $H_r$  becomes

$$H_r(S) = \sum_{i=1}^q p_i I_r(s_i) = \sum_{i=1}^q p_i \log_r \frac{1}{p_i} = - \sum_{i=1}^q p_i \log_r p_i.$$

Changing the base of  $I_r$  changes the base of  $H_r$ .

By adopting the convention  $p \log \frac{1}{p} \rightarrow 0$  as  $p \rightarrow 0$ , we say  $p \log \frac{1}{p} = 0$  when  $p = 0$ . This was the only potential discontinuity point, so now  $H(S)$  is continuous.

Optimality of  
Huffman Codes  
L(C) of Huffman  
Codes  
Huffman Codes  
are Optimal  
Source Extension

Information

$I(s)$  measure  
 $I(s)$  is uniquely  
defined

Entropy

$H(S)$  measure  
Properties of  $H(S)$

Epilogue

<sup>4</sup>Recall we could pick the base of the logarithm of  $I$  as we wished.

# Example

## Example (Unbiased and biased die)

Recall the unbiased die  $\mathcal{S}_u = (S, P_u)$  and biased die  $\mathcal{S}_b = (S, P_b)$ .

$$S = (6, 5, 4, 3, 2, 1),$$

$$P_u = \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right),$$

$$P_b = \left( \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10} \right).$$

We then have that

$$H_2(\mathcal{S}_b) = - \sum_{i=1}^6 p_{b_i} \log_2 p_{b_i} \approx 2.16096$$

$$H_2(\mathcal{S}_u) = \sum_{i=1}^6 p_{u_i} \log_2 \frac{1}{p_{u_i}} = 6 * \frac{1}{6} * \log_2 6 \approx 2.58496$$

**Spoiler:**  $\mathcal{S}$  approaches equiprobability  $\implies H(S)$  increases.

## Lower Bound

Theorem (Lower bound of  $H_r(S)$ )

$H_r(S) \geq 0$ , with equality iff  $p_i = 1$  for some  $i$ <sup>a</sup>

<sup>a</sup>Then  $p_j = 0$  for  $j \neq i$ .

## Proof.

Recall

$$H_r(S) = \sum_{i=1}^q p_i \log_r \frac{1}{p_i}.$$

Each term in this sum  $p * \log_r \frac{1}{p} \geq 0$ , with equality iff  $p \in \{0, 1\}$ . □

## Lecture 3: A Measure of Information

Optimality of Huffman Codes

$L(\mathcal{C})$  of Huffman Codes

Huffman Codes are Optimal

Source Extension

Information

$I(s)$  measure

$I(s)$  is uniquely defined

Entropy

$H(S)$  measure

Properties of  $H(S)$

Epilogue

## Upper Bound

## Lemma

For all  $x > 0$  we have  $\ln x \leq x - 1$ , w. equality iff  $x = 1$ .

## Corollary

Let  $x_i \geq 0$  and  $y_i > 0$  for  $1 \leq i \leq q$  and let  $\sum_i x_i = \sum_i y_i = 1^a$ . Then

$$\sum_{i=1}^q x_i \log_b \frac{1}{x_i} \leq \sum_{i=1}^q x_i \log_b \frac{1}{y_i},$$

(that is,  $\sum_i x_i \log_r \frac{y_i}{x_i} \leq 0$ ), with equality iff  $x_i = y_i, \forall i$ .

<sup>a</sup> $(x_1, \dots, x_q)$  and  $(y_1, \dots, y_q)$  are probability distributions

## Proof

Assuming each  $x_i > 0$ ,

$$\begin{aligned} \text{LHS} - \text{RHS} &= \sum_{i=1}^q x_i \log_r \frac{y_i}{x_i} = \frac{1}{\ln r} \sum_{i=1}^q x_i \ln \frac{y_i}{x_i} \\ &\leq \frac{1}{\ln r} \sum_{i=1}^q x_i \left( \frac{y_i}{x_i} - 1 \right) = \frac{1}{\ln r} \left( \sum_{i=1}^q y_i - \sum_{i=1}^q x_i \right) = 0 \end{aligned}$$

Cont.

# Lecture 3: A Measure of Information

Optimality of Huffman Codes

$L(\mathcal{C})$  of Huffman Codes

Huffman Codes are Optimal

Source Extension

Information

$I(s)$  measure

$I(s)$  is uniquely defined

Entropy

$H(S)$  measure

Properties of  $H(S)$

Epilogue

## Proof, cont.

So,  $\text{LHS} \leq \text{RHS}$  with equality iff each  $\frac{y_i}{x_i} = 1$ . If some  $x_i = 0$ , the terms in the original sums with  $x_i$  as factors are 0. We drop them from the sum, and the result follows.  $\square$

## Theorem

*If a source  $\mathcal{S}$  has  $q$  symbols, then  $H_r(\mathcal{S}) \leq \log_r q$ , with equality iff the symbols are equiprobable.*

## Proof.

By insertion to the corollary above, with  $x_i = p_i$  and  $y_i = \frac{1}{q}$ , we satisfy the condition there. So

$$H_r(\mathcal{S}) = \sum_{i=1}^q p_i \log_r \frac{1}{p_i} \leq \sum_{i=1}^q p_i \log_r q = \log_r q \sum_{i=1}^q p_i = \log_r q.$$

 $\square$ 

Thus,  $H(\mathcal{S})$  is “sandwiched”:

$$0 \leq H_r(\mathcal{S}) \leq \log_r q$$

## Lecture 3: A Measure of Information

Optimality of Huffman Codes  
 $L(\mathcal{C})$  of Huffman Codes  
 Huffman Codes are Optimal  
 Source Extension

## Information

$I(\mathcal{S})$  measure  
 $I(\mathcal{S})$  is uniquely defined

## Entropy

$H(\mathcal{S})$  measure  
 Properties of  $H(\mathcal{S})$

## Epilogue

$$H_r(\mathcal{S}) \leq L(\mathcal{C})$$

### Theorem

If  $\mathcal{C}$  is a uniquely decodable  $r$ -ary code for source  $\mathcal{S}$ , then  $H_r(\mathcal{S}) \leq L(\mathcal{C})$ .

### Proof.

Let

$$K = \sum_{i=1}^q r^{-l_i}.$$

By “McMillan’s Inequality”,  $K \leq 1$ . Let  $x_i = p_i$  and  $y_i = \frac{r^{-l_i}}{K}$ .

$$\begin{aligned} \underline{H_r(\mathcal{S})} &= \sum_{i=1}^q p_i \log_r \left( \frac{1}{p_i} \right) \leq \sum_{i=1}^q p_i \log_r \left( \frac{1}{y_i} \right) = \sum_{i=1}^q p_i \log_r (r^{l_i} K) \\ &= \sum_{i=1}^q p_i (l_i + \log_r K) = \sum_{i=1}^q p_i l_i + \log_r K \sum_{i=1}^q p_i \\ &= L(\mathcal{C}) + \log_r K \leq \underline{L(\mathcal{C})} \end{aligned}$$

( $K \leq 1$ ; thus  $\log_r K \leq 0$ )



<sup>a</sup> $y_i > 0$  and  $\sum_i y_i = 1$ . Recall:  $\sum_{i=1}^q x_i \log_b \frac{1}{x_i} \leq \sum_{i=1}^q x_i \log_b \frac{1}{y_i}$  (yesterday)



With luck,  $L(\mathcal{C}) = H_r(S)$

Corollary (Equality when  $p_i = r^{e_i}$ ;  $e_i \in \mathbb{Z}$ ;  $e_i \leq 0$ )

Let  $S$  be a source w. probabilities  $p_i$ . There exists a uniquely decodable  $r$ -ary code  $\mathcal{C}$  for  $S$  with  $L(\mathcal{C}) = H_r(S)$  iff  $\log_r p_i \in \mathbb{Z}$ ,  $\forall i$ .<sup>a</sup>

<sup>a</sup>That is,  $p_i = r^{e_i}$ , for some integer  $e_i \leq 0$

Proof.

$\Rightarrow$ : If  $L(\mathcal{C}) = H_r(S)$  in “lower bound” theorem, both “ $\leq$ ” must be “ $=$ ”.  
Then  $p_i = y_i = \frac{r^{-l_i}}{K}$  and  $\log_r K = 0$ . Thus  $K = 1$  and  $p_i = r^{-l_i}$ .

## Lecture 4: Shannon's Source Coding Theorem

Shannon-Fano  
Coding

$$H_r(S) \leq L(\mathcal{C})$$

$$L(\mathcal{C}) \leq 1 + H_r(S)$$

Shannon's Source  
Coding Theorem

$H_r(S)$  and Source  
Extension  
Theorem

Code Generation  
Story

Epilogue

The Big Picture  
Next Week

With luck,  $L(\mathcal{C}) = H_r(S)$

Corollary (Equality when  $p_i = r^{e_i}$ ;  $e_i \in \mathbb{Z}$ ;  $e_i \leq 0$ )

Let  $S$  be a source w. probabilities  $p_i$ . There exists a uniquely decodable  $r$ -ary code  $\mathcal{C}$  for  $S$  with  $L(\mathcal{C}) = H_r(S)$  iff  $\log_r p_i \in \mathbb{Z}$ ,  $\forall i$ .<sup>a</sup>

<sup>a</sup>That is,  $p_i = r^{e_i}$ , for some integer  $e_i \leq 0$

Proof.

$\Rightarrow$  : If  $L(\mathcal{C}) = H_r(S)$  in “lower bound” theorem, both “ $\leq$ ” must be “ $=$ ”.

Then  $p_i = y_i = \frac{r^{-l_i}}{K}$  and  $\log_r K = 0$ . Thus  $K = 1$  and  $p_i = r^{-l_i}$ .

$\Leftarrow$  : Assume  $\log_r p_i \in \mathbb{Z}$ . Thus  $p_i = \frac{1}{r^{l_i}}$ , with  $l_i \in \mathbb{Z}^+$ . Since

$$\sum_{i=1}^q \frac{1}{r^{l_i}} = \sum_{i=1}^q p_i = 1,$$

by “McMillan”, there exists a uniquely decodable  $r$ -ary code  $\mathcal{C}$  for  $S$  with word-lengths  $l_i$ . And,

$$L(\mathcal{C}) = \sum_{i=1}^q p_i l_i = \sum_{i=1}^q p_i \log_r \frac{1}{p_i} = H_r(S).$$



Hitting  $H_r(S)$  is hard.

### Example

Let  $S$  have  $q = 3$ , and let  $P_1 = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ . Here,

$$H_2(S) = \frac{2}{3} \log_2 3 + \frac{1}{3} \log_2 6 + \frac{1}{3} \log_2 6 \approx 1.2516,$$

### Lecture 4: Shannon's Source Coding Theorem

Shannon-Fano  
Coding

$$H_r(S) \leq L(C)$$

$$L(C) \leq 1 + H_r(S)$$

Shannon's Source  
Coding Theorem

$H_r(S)$  and Source  
Extension  
Theorem

Code Generation  
Story

Epilogue

The Big Picture  
Next Week

Hitting  $H_r(\mathcal{S})$  is hard.

### Example

Let  $\mathcal{S}$  have  $q = 3$ , and let  $P_1 = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ . Here,

$$H_2(\mathcal{S}) = \frac{2}{3} \log_2 3 + \frac{1}{3} \log_2 6 + \frac{1}{3} \log_2 6 \approx 1.2516,$$

while a Huffman code for  $\mathcal{S}$ ,  $\mathcal{C}_1 = (0, 10, 11)$ , has

$$L(\mathcal{C}_1) = \sum_{i=1}^q p_i l_i = \frac{2}{3} + \frac{1}{6} * 2 + \frac{1}{6} * 2 = 1 + \frac{1}{3} > \underline{H_2(\mathcal{S})}.$$

Hitting  $H_r(\mathcal{S})$  is hard.

### Example

Let  $\mathcal{S}$  have  $q = 3$ , and let  $P_1 = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ . Here,

$$H_2(\mathcal{S}) = \frac{2}{3} \log_2 3 + \frac{1}{3} \log_2 6 + \frac{1}{3} \log_2 6 \approx 1.2516,$$

while a Huffman code for  $\mathcal{S}$ ,  $\mathcal{C}_1 = (0, 10, 11)$ , has

$$L(\mathcal{C}_1) = \sum_{i=1}^q p_i l_i = \frac{2}{3} + \frac{1}{6} * 2 + \frac{1}{6} * 2 = 1 + \frac{1}{3} > \underline{H_2(\mathcal{S})}.$$

However, for  $P_2 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ ,

$$H_2(\mathcal{S}) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 = 1 + \frac{1}{2},$$

### Lecture 4: Shannon's Source Coding Theorem

Shannon-Fano  
Coding

$H_r(\mathcal{S}) \leq L(\mathcal{C})$   
 $L(\mathcal{C}) \leq 1 + H_r(\mathcal{S})$

Shannon's Source  
Coding Theorem

$H_r(\mathcal{S})$  and Source  
Extension  
Theorem

Code Generation  
Story

Epilogue

The Big Picture  
Next Week

Hitting  $H_r(\mathcal{S})$  is hard.

### Example

Let  $\mathcal{S}$  have  $q = 3$ , and let  $P_1 = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ . Here,

$$H_2(\mathcal{S}) = \frac{2}{3} \log_2 3 + \frac{1}{3} \log_2 6 + \frac{1}{3} \log_2 6 \approx 1.2516,$$

while a Huffman code for  $\mathcal{S}$ ,  $\mathcal{C}_1 = (0, 10, 11)$ , has

$$L(\mathcal{C}_1) = \sum_{i=1}^q p_i l_i = \frac{2}{3} + \frac{1}{6} * 2 + \frac{1}{6} * 2 = 1 + \frac{1}{3} > \underline{H_2(\mathcal{S})}.$$

However, for  $P_2 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ ,

$$H_2(\mathcal{S}) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 = 1 + \frac{1}{2},$$

while the same  $\mathcal{C}$  (denoted  $\mathcal{C}_2$  here) has

$$L(\mathcal{C}_2) = \sum_{i=1}^q p_i l_i = \frac{1}{2} + \frac{1}{4} * 2 + \frac{1}{4} * 2 = 1 + \frac{1}{2} = \underline{H_2(\mathcal{S})}.$$

Getting close to  $H_r(\mathcal{S})$ 

## Definition (Code Efficiency)

If  $\mathcal{C}$  is an  $r$ -ary code for  $\mathcal{S}$ , the *efficiency* of  $\mathcal{C}$  is given by

$$\eta = \frac{H_r(\mathcal{S})}{L(\mathcal{C})}$$

and its *redundancy* is  $\bar{\eta} = 1 - \eta$ .

For every uniquely decodable codes, we have  $0 \leq \eta \leq 1$ . We want  $\eta = 1$ . That is, we want  $L(\mathcal{C}) = H_r(\mathcal{S})$

Getting close to  $H_r(\mathcal{S})$ 

## Definition (Code Efficiency)

If  $\mathcal{C}$  is an  $r$ -ary code for  $\mathcal{S}$ , the *efficiency* of  $\mathcal{C}$  is given by

$$\eta = \frac{H_r(\mathcal{S})}{L(\mathcal{C})}$$

and its *redundancy* is  $\bar{\eta} = 1 - \eta$ .

For every uniquely decodable codes, we have  $0 \leq \eta \leq 1$ . We want  $\eta = 1$ . That is, we want  $L(\mathcal{C}) = H_r(\mathcal{S})$

**Q:** Why was  $L(\mathcal{C}_1) > H_r(\mathcal{S})$  but  $L(\mathcal{C}_2) = H_r(\mathcal{S})$ ?

**A:** Because  $r = 2$  and  $P_2 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = (\frac{1}{r^1}, \frac{1}{r^2}, \frac{1}{r^2})$ .

*Each  $p_i$  is  $r$  in a (negative) integer power.*

*Each  $w_i \in \mathcal{C}$  was attached to  $p_i \in P$  s.t.*

$$|w_i| = l_i = \log_r \frac{1}{p_i}.$$

We need to be very fortunate to satisfy this condition.

$$H_r(\mathcal{S}) \leq L(\mathcal{C})$$

$$L(\mathcal{C}) \leq 1 + H_r(\mathcal{S})$$