

What Is Information?

Uncertainty, Entropy and Mutual Information

Balaji Dhayabaran

Kanithakalam · kanithakalam.com/foundations-of-information-theory

© 2026 Balaji Dhayabaran · Licensed under CC BY-NC-ND 4.0

A question to begin with

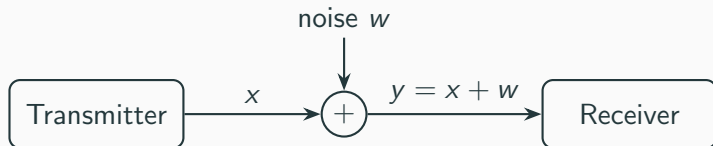
A noisy signal reaches your phone.

How much information actually got through?

A question to begin with

A noisy signal reaches your phone.

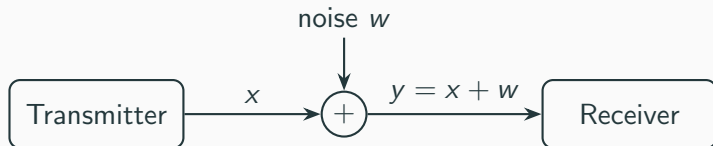
How much information actually got through?



A question to begin with

A noisy signal reaches your phone.

How much information actually got through?



By the end of this lecture, we will **calculate** that number for a real channel.

Our plan

1. What do we mean by *information*?

Our plan

1. What do we mean by *information*?
2. Two small channels, and a puzzle

Our plan

1. What do we mean by *information*?
2. Two small channels, and a puzzle
3. Measuring uncertainty: surprise and **entropy**

Our plan

1. What do we mean by *information*?
2. Two small channels, and a puzzle
3. Measuring uncertainty: surprise and **entropy**
4. Solving the puzzle: **conditional entropy**

Our plan

1. What do we mean by *information*?
2. Two small channels, and a puzzle
3. Measuring uncertainty: surprise and **entropy**
4. Solving the puzzle: **conditional entropy**
5. **Mutual information**, and a look ahead to capacity

Our plan

1. What do we mean by *information*?
2. Two small channels, and a puzzle
3. Measuring uncertainty: surprise and **entropy**
4. Solving the puzzle: **conditional entropy**
5. **Mutual information**, and a look ahead to capacity

One channel, followed from start to finish.

What Is Information?

Before and after

- **Before** transmission: the receiver does not know which bit will be sent. It only knows how likely each bit is.

Before and after

- **Before** transmission: the receiver does not know which bit will be sent. It only knows how likely each bit is.
- **During** transmission: it observes a *noisy* version, y .

Before and after

- **Before** transmission: the receiver does not know which bit will be sent. It only knows how likely each bit is.
- **During** transmission: it observes a *noisy* version, y .
- **After**: the noisy observation still gives clues, so its doubt is *reduced*.

Before and after

- **Before** transmission: the receiver does not know which bit will be sent. It only knows how likely each bit is.
- **During** transmission: it observes a *noisy* version, y .
- **After**: the noisy observation still gives clues, so its doubt is *reduced*.

$$\begin{aligned} &\text{Information Transferred} \\ &= \text{Initial Uncertainty} - \text{Remaining Uncertainty} \end{aligned}$$

Before and after

- **Before** transmission: the receiver does not know which bit will be sent. It only knows how likely each bit is.
- **During** transmission: it observes a *noisy* version, y .
- **After**: the noisy observation still gives clues, so its doubt is *reduced*.

$$\begin{aligned} &\text{Information Transferred} \\ &= \text{Initial Uncertainty} - \text{Remaining Uncertainty} \end{aligned}$$

Everything in this lecture is about making this one equation precise.

The model

$$y = x + w$$

The model

$$y = x + w$$

- $x \in \{0, 1\}$ is the transmitted bit, with $P(x = 0) = P(x = 1) = \frac{1}{2}$

The model

$$y = x + w$$

- $x \in \{0, 1\}$ is the transmitted bit, with $P(x = 0) = P(x = 1) = \frac{1}{2}$
- w is additive noise, independent of x

The model

$$y = x + w$$

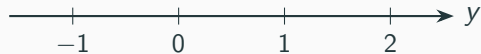
- $x \in \{0, 1\}$ is the transmitted bit, with $P(x = 0) = P(x = 1) = \frac{1}{2}$
- w is additive noise, independent of x
- Real noise is Gaussian (AWGN). Today, to see the ideas clearly, the noise takes only two values.

The model

$$y = x + w$$

- $x \in \{0, 1\}$ is the transmitted bit, with $P(x = 0) = P(x = 1) = \frac{1}{2}$
- w is additive noise, independent of x
- Real noise is Gaussian (AWGN). Today, to see the ideas clearly, the noise takes only two values.
- The receiver knows the probabilities, but not which bit was actually sent.

Example 1: noise $w \in \{-1, +1\}$

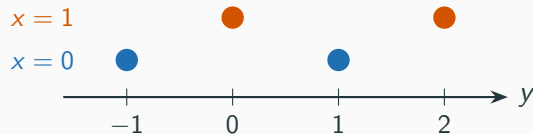


Example 1: noise $w \in \{-1, +1\}$



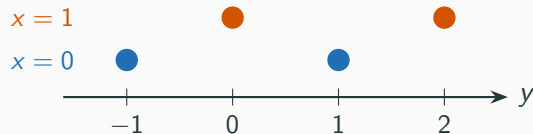
- If $x = 0$: $y \in \{-1, 1\}$

Example 1: noise $w \in \{-1, +1\}$



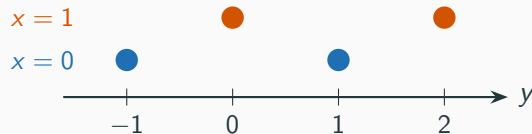
- If $x = 0$: $y \in \{-1, 1\}$
- If $x = 1$: $y \in \{0, 2\}$

Example 1: noise $w \in \{-1, +1\}$



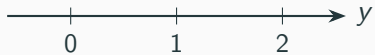
- If $x = 0$: $y \in \{-1, 1\}$
- If $x = 1$: $y \in \{0, 2\}$
- The two sets **never overlap**: every y points to exactly one x .

Example 1: noise $w \in \{-1, +1\}$



- If $x = 0$: $y \in \{-1, 1\}$
- If $x = 1$: $y \in \{0, 2\}$
- The two sets **never overlap**: every y points to exactly one x .
- All uncertainty is removed: **1 bit** is transferred per transmission.

Example 2: noise $w \in \{0, 1\}$



Example 2: noise $w \in \{0, 1\}$



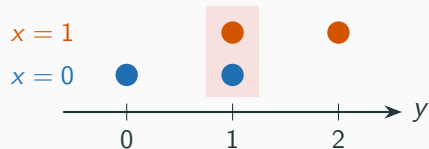
- If $x = 0$: $y \in \{0, 1\}$

Example 2: noise $w \in \{0, 1\}$



- If $x = 0$: $y \in \{0, 1\}$
- If $x = 1$: $y \in \{1, 2\}$

Example 2: noise $w \in \{0, 1\}$



- If $x = 0$: $y \in \{0, 1\}$
- If $x = 1$: $y \in \{1, 2\}$
- $y = 0$ or $y = 2$: the receiver is certain. $y = 1$: it could be either!

In Example 2, how much information gets through
per transmission?

In Example 2, how much information gets through per transmission?

(A) 1 bit

(B) 0 bits

(C) somewhere in between

The puzzle

In Example 2, how much information gets through per transmission?

(A) 1 bit

(B) 0 bits

(C) somewhere in between

Keep your answer. We will come back and calculate it exactly.

Measuring Uncertainty

Rare events carry more doubt

- An event with probability close to 1: we expected it. Little surprise.

Rare events carry more doubt

- An event with probability close to 1: we expected it. Little surprise.
- An event with probability close to 0: very surprising.

Rare events carry more doubt

- An event with probability close to 1: we expected it. Little surprise.
- An event with probability close to 0: very surprising.
- So our measure should decrease as p increases.

Rare events carry more doubt

- An event with probability close to 1: we expected it. Little surprise.
- An event with probability close to 0: very surprising.
- So our measure should decrease as p increases.

First guess: $\frac{1}{p}$

$$p = 0.01 \Rightarrow 100$$

$$p = 0.99 \Rightarrow 1.01$$

The first guess fails

Send two independent bits.

The first guess fails

Send two independent bits.

- Probabilities **multiply**: $p_{\text{joint}} = p_1 \times p_2$

The first guess fails

Send **two independent bits**.

- Probabilities **multiply**: $p_{\text{joint}} = p_1 \times p_2$
- So $1/p$ also multiplies: $\frac{1}{p_{\text{joint}}} = \frac{1}{p_1} \times \frac{1}{p_2}$

The first guess fails

Send **two independent bits**.

- Probabilities **multiply**: $p_{\text{joint}} = p_1 \times p_2$
- So $1/p$ also multiplies: $\frac{1}{p_{\text{joint}}} = \frac{1}{p_1} \times \frac{1}{p_2}$
- But two bits should carry $1 + 1 = 2$ bits of information: information should **add**.

The first guess fails

Send **two independent bits**.

- Probabilities **multiply**: $p_{\text{joint}} = p_1 \times p_2$
- So $1/p$ also multiplies: $\frac{1}{p_{\text{joint}}} = \frac{1}{p_1} \times \frac{1}{p_2}$
- But two bits should carry $1 + 1 = 2$ bits of information: information should **add**.

We need a function that turns **products** into **sums**.

The first guess fails

Send **two independent bits**.

- Probabilities **multiply**: $p_{\text{joint}} = p_1 \times p_2$
- So $1/p$ also multiplies: $\frac{1}{p_{\text{joint}}} = \frac{1}{p_1} \times \frac{1}{p_2}$
- But two bits should carry $1 + 1 = 2$ bits of information: information should **add**.

We need a function that turns **products** into **sums**.

The only well-behaved such function is the **logarithm**.

Surprise

$$S(p) = \log_2 \frac{1}{p} \quad (\text{measured in bits})$$

Surprise

$$S(p) = \log_2 \frac{1}{p} \quad (\text{measured in bits})$$

p	$S(p)$
1/2	1 bit
1/4	2 bits
1/8	3 bits
0.01	6.64 bits

Surprise

$$S(p) = \log_2 \frac{1}{p} \quad (\text{measured in bits})$$

p	$S(p)$
1/2	1 bit
1/4	2 bits
1/8	3 bits
0.01	6.64 bits

Two independent events: $S(p_1 p_2) = S(p_1) + S(p_2)$ ✓

Surprise

$$S(p) = \log_2 \frac{1}{p} \quad (\text{measured in bits})$$

p	$S(p)$
1/2	1 bit
1/4	2 bits
1/8	3 bits
0.01	6.64 bits

Two independent events: $S(p_1 p_2) = S(p_1) + S(p_2)$ ✓

This argument goes back to Claude Shannon, 1948.

From one outcome to the whole source

- Surprise belongs to **one outcome**.

From one outcome to the whole source

- Surprise belongs to **one outcome**.
- The uncertainty of a **source** should describe all its outcomes together.

From one outcome to the whole source

- Surprise belongs to **one outcome**.
- The uncertainty of a **source** should describe all its outcomes together.
- So we take the **average surprise**, weighting each outcome by how often it occurs:

From one outcome to the whole source

- Surprise belongs to **one outcome**.
- The uncertainty of a **source** should describe all its outcomes together.
- So we take the **average surprise**, weighting each outcome by how often it occurs:

$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i}$$

This is the **entropy** of the source.

Entropy of a binary source

- Fair bit, $q = \frac{1}{2}$:
 $H = 1$ bit

Entropy of a binary source

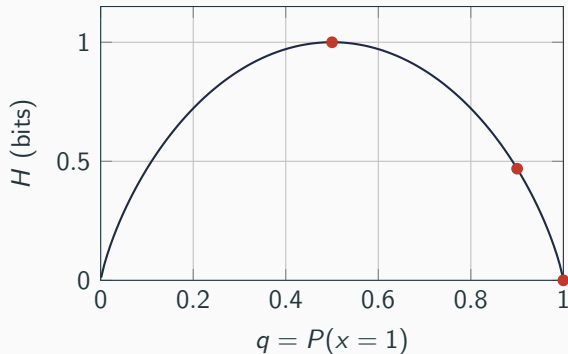
- Fair bit, $q = \frac{1}{2}$:
 $H = 1$ bit
- Biased bit, $q = 0.9$:
 $H \approx 0.469$ bits

Entropy of a binary source

- Fair bit, $q = \frac{1}{2}$:
 $H = 1$ bit
- Biased bit, $q = 0.9$:
 $H \approx 0.469$ bits
- Stuck transmitter, $q = 1$:
 $H = 0$ bits

Entropy of a binary source

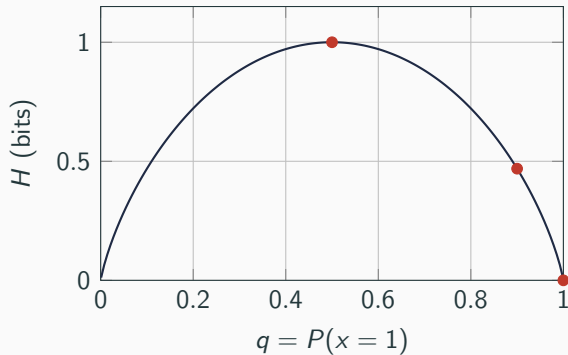
- Fair bit, $q = \frac{1}{2}$:
 $H = 1$ bit
- Biased bit, $q = 0.9$:
 $H \approx 0.469$ bits
- Stuck transmitter, $q = 1$:
 $H = 0$ bits



Entropy of a binary source

- Fair bit, $q = \frac{1}{2}$:
 $H = 1$ bit
- Biased bit, $q = 0.9$:
 $H \approx 0.469$ bits
- Stuck transmitter, $q = 1$:
 $H = 0$ bits

A binary digit carries **at most** 1 bit of information.



Why not just average $1/p$?

Two three-symbol sources:

$$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \quad (0.98, 0.01, 0.01)$$

Why not just average $1/p$?

Two three-symbol sources:

	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(0.98, 0.01, 0.01)$
Average of $1/p$	3	3

Why not just average $1/p$?

Two three-symbol sources:

	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(0.98, 0.01, 0.01)$
Average of $1/p$	3	3
Entropy H	1.585 bits	0.161 bits

Why not just average $1/p$?

Two three-symbol sources:

	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(0.98, 0.01, 0.01)$
Average of $1/p$	3	3
Entropy H	1.585 bits	0.161 bits

- Each term $p \cdot \frac{1}{p} = 1$: the average of $1/p$ just counts the outcomes.

Why not just average $1/p$?

Two three-symbol sources:

	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(0.98, 0.01, 0.01)$
Average of $1/p$	3	3
Entropy H	1.585 bits	0.161 bits

- Each term $p \cdot \frac{1}{p} = 1$: the average of $1/p$ just counts the outcomes.
- It cannot tell a fair source from a predictable one. Entropy can.

Solving the Puzzle

Back to Example 2

$$y = x + w, \quad x \in \{0, 1\}, \quad w \in \{0, 1\}$$

Back to Example 2

$$y = x + w, \quad x \in \{0, 1\}, \quad w \in \{0, 1\}$$

- $P(x = 0) = P(x = 1) = \frac{1}{2}$, so the initial uncertainty is $H(X) = 1$ bit

Back to Example 2

$$y = x + w, \quad x \in \{0, 1\}, \quad w \in \{0, 1\}$$

- $P(x = 0) = P(x = 1) = \frac{1}{2}$, so the initial uncertainty is $H(X) = 1$ bit
- $P(w = 0) = P(w = 1) = \frac{1}{2}$, and w is independent of x

Back to Example 2

$$y = x + w, \quad x \in \{0, 1\}, \quad w \in \{0, 1\}$$

- $P(x = 0) = P(x = 1) = \frac{1}{2}$, so the initial uncertainty is $H(X) = 1$ bit
- $P(w = 0) = P(w = 1) = \frac{1}{2}$, and w is independent of x

Initial uncertainty is known. What is the remaining uncertainty?

How often does each y occur?

$$x \quad w \quad y = x + w \quad P(x, w)$$

How often does each y occur?

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$

Independence: $P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

How often does each y occur?

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$

Independence: $P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

How often does each y occur?

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$
1	0	1	$\frac{1}{4}$

Independence: $P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

How often does each y occur?

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$
1	0	1	$\frac{1}{4}$
1	1	2	$\frac{1}{4}$

Independence: $P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

How often does each y occur?

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$
1	0	1	$\frac{1}{4}$
1	1	2	$\frac{1}{4}$

Independence: $P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$$P(y = 0) = \frac{1}{4}, \quad P(y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad P(y = 2) = \frac{1}{4}$$

Imagine 1000 transmissions

- $(x=0, w=0) \rightarrow y=0$: about 250
- $(x=0, w=1) \rightarrow y=1$: about 250
- $(x=1, w=0) \rightarrow y=1$: about 250
- $(x=1, w=1) \rightarrow y=2$: about 250

Imagine 1000 transmissions

- $(x=0, w=0) \rightarrow y=0$: about 250
- $(x=0, w=1) \rightarrow y=1$: about 250
- $(x=1, w=0) \rightarrow y=1$: about 250
- $(x=1, w=1) \rightarrow y=2$: about 250

The receiver sees $y = 1$.

Keep **only** those cases: about 500.

Imagine 1000 transmissions

- $(x=0, w=0) \rightarrow y=0$: about 250
- $(x=0, w=1) \rightarrow y=1$: about 250
- $(x=1, w=0) \rightarrow y=1$: about 250
- $(x=1, w=1) \rightarrow y=2$: about 250

The receiver sees $y = 1$.

Keep **only** those cases: about 500. Of

these 500, how many had $x = 0$?

About 250.

Imagine 1000 transmissions

- $(x=0, w=0) \rightarrow y=0$: about 250
- $(x=0, w=1) \rightarrow y=1$: about 250
- $(x=1, w=0) \rightarrow y=1$: about 250
- $(x=1, w=1) \rightarrow y=2$: about 250

The receiver sees $y = 1$.

Keep **only** those cases: about 500. Of

these 500, how many had $x = 0$?

About 250.

$$P(x = 0 \mid y = 1) = \frac{250}{500} = \frac{1/4}{1/2} = \frac{1}{2}$$

Imagine 1000 transmissions

- $(x=0, w=0) \rightarrow y=0$: about 250
- $(x=0, w=1) \rightarrow y=1$: about 250
- $(x=1, w=0) \rightarrow y=1$: about 250
- $(x=1, w=1) \rightarrow y=2$: about 250

The receiver sees $y = 1$.

Keep **only** those cases: about 500. Of

these 500, how many had $x = 0$?

About 250.

$$P(x = 0 \mid y = 1) = \frac{250}{500} = \frac{1/4}{1/2} = \frac{1}{2}$$

Seeing $y = 1$ leaves the receiver **exactly where it started**: 50/50.

What the receiver believes after seeing y

Observed y	$P(x = 0 \mid y)$	$P(x = 1 \mid y)$
0	1	0
1	$\frac{1}{2}$	$\frac{1}{2}$
2	0	1

What the receiver believes after seeing y

Observed y	$P(x = 0 \mid y)$	$P(x = 1 \mid y)$
0	1	0
1	$\frac{1}{2}$	$\frac{1}{2}$
2	0	1

In every case, we divided:

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

What the receiver believes after seeing y

Observed y	$P(x = 0 \mid y)$	$P(x = 1 \mid y)$
0	1	0
1	$\frac{1}{2}$	$\frac{1}{2}$
2	0	1

In every case, we divided:

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

This is **Bayes' rule**, and we got it just by counting.

Remaining uncertainty for each observation

After seeing y , the receiver has a *new* distribution for x . Its uncertainty is its entropy:

$$H(X \mid Y = y) = \sum_x P(x \mid y) \log_2 \frac{1}{P(x \mid y)}$$

Remaining uncertainty for each observation

After seeing y , the receiver has a *new* distribution for x . Its uncertainty is its entropy:

$$H(X | Y = y) = \sum_x P(x | y) \log_2 \frac{1}{P(x | y)}$$

$$y = 0 : (1, 0)$$

$$H = 0 \text{ bits}$$

Remaining uncertainty for each observation

After seeing y , the receiver has a *new* distribution for x . Its uncertainty is its entropy:

$$H(X | Y = y) = \sum_x P(x | y) \log_2 \frac{1}{P(x | y)}$$

$$y = 0 : (1, 0)$$

$$H = 0 \text{ bits}$$

$$y = 1 : (\frac{1}{2}, \frac{1}{2})$$

$$H = 1 \text{ bit}$$

Remaining uncertainty for each observation

After seeing y , the receiver has a *new* distribution for x . Its uncertainty is its entropy:

$$H(X | Y = y) = \sum_x P(x | y) \log_2 \frac{1}{P(x | y)}$$

$$y = 0 : (1, 0) \qquad H = 0 \text{ bits}$$

$$y = 1 : \left(\frac{1}{2}, \frac{1}{2}\right) \qquad H = 1 \text{ bit}$$

$$y = 2 : (0, 1) \qquad H = 0 \text{ bits}$$

Remaining uncertainty for each observation

After seeing y , the receiver has a *new* distribution for x . Its uncertainty is its entropy:

$$H(X | Y = y) = \sum_x P(x | y) \log_2 \frac{1}{P(x | y)}$$

$$y = 0 : (1, 0) \qquad H = 0 \text{ bits}$$

$$y = 1 : \left(\frac{1}{2}, \frac{1}{2}\right) \qquad H = 1 \text{ bit}$$

$$y = 2 : (0, 1) \qquad H = 0 \text{ bits}$$

Exactly as our intuition said: $y = 1$ leaves all the doubt.

Average over what the receiver might see

The receiver does not choose y . As with entropy, we **average**:

Average over what the receiver might see

The receiver does not choose y . As with entropy, we **average**:

$$H(X | Y) = \sum_y P(y) H(X | Y = y)$$

Average over what the receiver might see

The receiver does not choose y . As with entropy, we **average**:

$$H(X | Y) = \sum_y P(y) H(X | Y = y)$$

$$= \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 0 = \frac{1}{2} \text{ bit}$$

Average over what the receiver might see

The receiver does not choose y . As with entropy, we **average**:

$$H(X | Y) = \sum_y P(y) H(X | Y = y)$$

$$= \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 0 = \frac{1}{2} \text{ bit}$$

This average is the **conditional entropy**:
the uncertainty that remains, on average, after observing y .

The same structure as entropy

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

The same structure as entropy

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

$$H(X | Y) = \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x | y)}$$

The same structure as entropy

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

$$H(X | Y) = \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x | y)}$$

- Both are an **average of surprise**.

The same structure as entropy

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

$$H(X | Y) = \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x | y)}$$

- Both are an **average of surprise**.
- Entropy: surprise **before** observing y .

The same structure as entropy

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

$$H(X | Y) = \sum_{x,y} P(x,y) \log_2 \frac{1}{P(x | y)}$$

- Both are an **average of surprise**.
- Entropy: surprise **before** observing y .
- Conditional entropy: surprise **after** observing y , averaged over everything that can happen.

The answer to the puzzle

$$\begin{aligned} &\text{Information Transferred} \\ &= \text{Initial Uncertainty} - \text{Remaining Uncertainty} \end{aligned}$$

The answer to the puzzle

Information Transferred

= Initial Uncertainty – Remaining Uncertainty

$$= H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit per transmission}$$

The answer to the puzzle

Information Transferred

= Initial Uncertainty – Remaining Uncertainty

$$= H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit per transmission}$$

The answer was **(C)**: somewhere in between.

The answer to the puzzle

$$\begin{aligned} & \text{Information Transferred} \\ &= \text{Initial Uncertainty} - \text{Remaining Uncertainty} \end{aligned}$$

$$= H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit per transmission}$$

The answer was **(C)**: somewhere in between.

- Half the time ($y = 0$ or 2), the receiver learns x exactly.

The answer to the puzzle

$$\begin{aligned} &\text{Information Transferred} \\ &= \text{Initial Uncertainty} - \text{Remaining Uncertainty} \end{aligned}$$

$$= H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit per transmission}$$

The answer was **(C)**: somewhere in between.

- Half the time ($y = 0$ or 2), the receiver learns x exactly.
- Half the time ($y = 1$), it learns nothing.

Mutual Information

The reverse question

How much does knowing x reduce the uncertainty about y ?

The reverse question

How much does knowing x reduce the uncertainty about y ?

- Before: $y \in \{0, 1, 2\}$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \Rightarrow H(Y) = 1.5$ bits

The reverse question

How much does knowing x reduce the uncertainty about y ?

- Before: $y \in \{0, 1, 2\}$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \Rightarrow H(Y) = 1.5$ bits
- Once x is known, only the noise is left $\Rightarrow H(Y | X) = H(W) = 1$ bit

The reverse question

How much does knowing x reduce the uncertainty about y ?

- Before: $y \in \{0, 1, 2\}$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \Rightarrow H(Y) = 1.5$ bits
- Once x is known, only the noise is left $\Rightarrow H(Y | X) = H(W) = 1$ bit

$$H(Y) - H(Y | X) = 1.5 - 1 = \frac{1}{2} \text{ bit}$$

The reverse question

How much does knowing x reduce the uncertainty about y ?

- Before: $y \in \{0, 1, 2\}$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \Rightarrow H(Y) = 1.5$ bits
- Once x is known, only the noise is left $\Rightarrow H(Y | X) = H(W) = 1$ bit

$$H(Y) - H(Y | X) = 1.5 - 1 = \frac{1}{2} \text{ bit}$$

The **same** answer!

The reverse question

How much does knowing x reduce the uncertainty about y ?

- Before: $y \in \{0, 1, 2\}$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \Rightarrow H(Y) = 1.5$ bits
- Once x is known, only the noise is left $\Rightarrow H(Y | X) = H(W) = 1$ bit

$$H(Y) - H(Y | X) = 1.5 - 1 = \frac{1}{2} \text{ bit}$$

The **same** answer!

y varies by 1.5 bits: 1 bit of it is noise, $\frac{1}{2}$ bit carries information about x .

Mutual information

The reduction is the same in both directions. It is **shared** by x and y :

Mutual information

The reduction is the same in both directions. It is **shared** by x and y :

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$

Mutual information

The reduction is the same in both directions. It is **shared** by x and y :

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$

In general, it can be written symmetrically:

$$I(X; Y) = \sum_{x,y} P(x, y) \log_2 \frac{P(x, y)}{P(x) P(y)}$$

Mutual information

The reduction is the same in both directions. It is **shared** by x and y :

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$

In general, it can be written symmetrically:

$$I(X; Y) = \sum_{x,y} P(x, y) \log_2 \frac{P(x, y)}{P(x) P(y)}$$

If x and y are independent, every term is $\log_2 1 = 0$, so $I(X; Y) = 0$.

Information needs dependence.

Check: Example 1

Noise $w \in \{-1, +1\}$, equally likely. The receiver sees $y \in \{-1, 0, 1, 2\}$.

Check: Example 1

Noise $w \in \{-1, +1\}$, equally likely. The receiver sees $y \in \{-1, 0, 1, 2\}$.

- Every y comes from exactly one $x \Rightarrow H(X | Y) = 0$

Check: Example 1

Noise $w \in \{-1, +1\}$, equally likely. The receiver sees $y \in \{-1, 0, 1, 2\}$.

- Every y comes from exactly one $x \Rightarrow H(X | Y) = 0$
- $I(X; Y) = 1 - 0 = 1$ bit

Check: Example 1

Noise $w \in \{-1, +1\}$, equally likely. The receiver sees $y \in \{-1, 0, 1, 2\}$.

- Every y comes from exactly one $x \Rightarrow H(X | Y) = 0$
- $I(X; Y) = 1 - 0 = 1$ bit
- Other side: $H(Y) - H(Y | X) = 2 - 1 = 1$ bit

Check: Example 1

Noise $w \in \{-1, +1\}$, equally likely. The receiver sees $y \in \{-1, 0, 1, 2\}$.

- Every y comes from exactly one $x \Rightarrow H(X | Y) = 0$
- $I(X; Y) = 1 - 0 = 1$ bit
- Other side: $H(Y) - H(Y | X) = 2 - 1 = 1$ bit

The formulas confirm what we saw at the start.

Looking Ahead

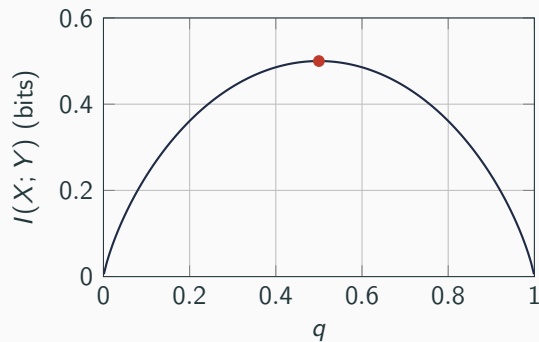
What if the transmitter changes q ?

- We assumed $P(x = 1) = q = \frac{1}{2}$.

What if the transmitter changes q ?

- We assumed $P(x = 1) = q = \frac{1}{2}$.
- For Example 2 with any q :

$$I(X; Y) = \frac{1}{2}H(q)$$



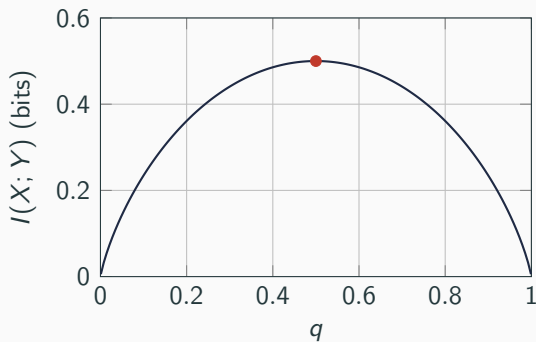
What if the transmitter changes q ?

- We assumed $P(x = 1) = q = \frac{1}{2}$.

- For Example 2 with any q :

$$I(X; Y) = \frac{1}{2} H(q)$$

- Largest at $q = \frac{1}{2}$: $\frac{1}{2}$ bit



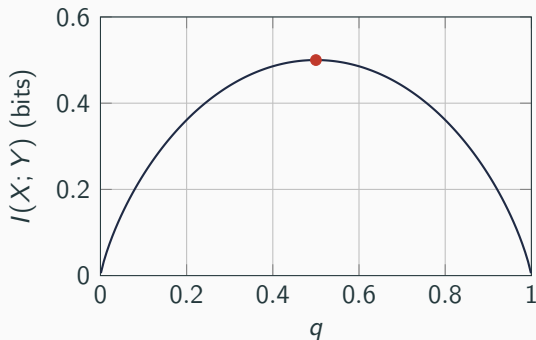
What if the transmitter changes q ?

- We assumed $P(x = 1) = q = \frac{1}{2}$.

- For Example 2 with any q :

$$I(X; Y) = \frac{1}{2} H(q)$$

- Largest at $q = \frac{1}{2}$: $\frac{1}{2}$ bit



What is the **most** information a channel can carry,
over all choices of the input? $\Rightarrow$ **Capacity**

Why this matters in a real system

- A 5G base station constantly asks: *how much can I send to this user right now?*

Why this matters in a real system

- A 5G base station constantly asks: *how much can I send to this user right now?*
- Link adaptation picks a modulation and coding scheme to match the channel.

Why this matters in a real system

- A 5G base station constantly asks: *how much can I send to this user right now?*
- Link adaptation picks a modulation and coding scheme to match the channel.
- Capacity sets the limit that no scheme can beat.

Why this matters in a real system

- A 5G base station constantly asks: *how much can I send to this user right now?*
- Link adaptation picks a modulation and coding scheme to match the channel.
- Capacity sets the limit that no scheme can beat.
- Mutual information is the quantity underneath all of it.

Why this matters in a real system

- A 5G base station constantly asks: *how much can I send to this user right now?*
- Link adaptation picks a modulation and coding scheme to match the channel.
- Capacity sets the limit that no scheme can beat.
- Mutual information is the quantity underneath all of it.

Next: capacity, and how close real systems come to it.

Summary

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	one outcome

Summary

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	one outcome
Entropy	$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$	uncertainty before observing

Summary

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	one outcome
Entropy	$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$	uncertainty before observing
Conditional entropy	$H(X Y) = \sum_y P(y) H(X Y=y)$	uncertainty left after observing

Summary

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	one outcome
Entropy	$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$	uncertainty before observing
Conditional entropy	$H(X Y) = \sum_y P(y) H(X Y=y)$	uncertainty left after observing
Mutual information	$I(X; Y) = H(X) - H(X Y)$	information transferred

Summary

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	one outcome
Entropy	$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$	uncertainty before observing
Conditional entropy	$H(X Y) = \sum_y P(y) H(X Y=y)$	uncertainty left after observing
Mutual information	$I(X; Y) = H(X) - H(X Y)$	information transferred

Example 1: 1 bit

Example 2: $\frac{1}{2}$ bit

Thank you

The full chapter (with worked problems), these slides and the video lecture:

kanithakalam.com/foundations-of-information-theory

Follow Kanithakalam on LinkedIn:

linkedin.com/company/kanithakalam

Questions and comments are welcome through the website or LinkedIn.

Balaji Dhayabaran · Kanithakalam

© 2026 Balaji Dhayabaran. Licensed under CC BY-NC-ND 4.0
(creativecommons.org/licenses/by-nc-nd/4.0).

You may share these slides for non-commercial purposes with attribution, without modification.