

Part 1: Theory

L03: Basics of entropy (1/7)

[measures of information, intuition behind entropy]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa26)

<https://northeastern-datalab.github.io/cs7840/fa26/>

9/17/2026

Let's gain some intuition for "measures of information"

The following numeric examples with hats and 4 balls are based on Chapter 1.1 from [Moser'18] Information Theory (lecture notes, 6th ed). https://moser-isi.ethz.ch/cgi-bin/request_script.cgi?script=it

Let's gain some intuition: What is information?

What is information? Let's look at some sentences with "information":

1. "It will rain tomorrow."
2. "It will snow tomorrow."
3. "(in 2024) The name of the next US president will be..."
 - a. ... Donald."
 - b. ... Donald Duck."
4. "Our university is called Northeastern University."



Let's gain some intuition: What is information?

What is information? Let's look at some sentences with "information":

1. "It will rain tomorrow."
2. "It will snow tomorrow."
3. "(in 2024) The name of the next US president will be..."
 - a. ... Donald."
 - b. ... Donald Duck."
4. "Our university is called Northeastern University."

⇒ Information (in a sentence) is linked to surprise (which is the delta of knowledge before and after seeing the sentence).

Let's next try to quantify "information" 😊

Let's try to quantify "information"

EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- The "pure" message U_1 that we care about in our abstraction is ...



Let's try to quantify "information"

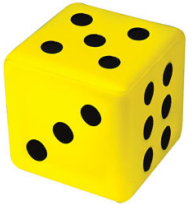
EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$

EXAMPLE 2: A gambler throws a fair die with **6** sides {**A**, **B**, **C**, **D**, **E**, **F**}.

- "Side **C** comes up."
- message $U_2 = \text{"C"}$



what has changed ?

Let's try to quantify "information"

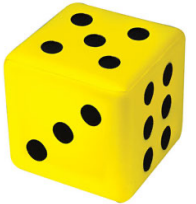
EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$
- There are **4** possible outcomes, each has a probability of $\frac{1}{4}$.

EXAMPLE 2: A gambler throws a fair die with 6 sides {**A**, **B**, **C**, **D**, **E**, **F**}.

- "Side **C** comes up."
- message $U_2 = \text{"C"}$
- There are **6** possible outcomes, each has a probability of $\frac{1}{6}$.



⇒ 1) The number of possible outcomes should be linked to "information"
(we need more space to encode a message)

Let's try to quantify "information"

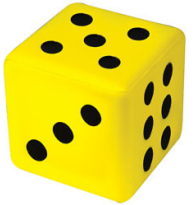
EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$, or in above binary encoding $U_1 = \text{"10"}$
- There are 4 possible outcomes, each has a probability of $\frac{1}{4}$.

EXAMPLE 2: A gambler throws a fair die with 6 sides {**A**, **B**, **C**, **D**, **E**, **F**}.

- "Side **C** comes up."
- message $U_2 = \text{"C"}$, or in above binary encoding $U_2 = \text{"010"}$
- There are 6 possible outcomes, each has a probability of $\frac{1}{6}$.



⇒ 1) The number of possible outcomes should be linked to "information"
(we need more space to encode a message)

Let's try to quantify "information"

EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$
- There are 4 possible outcomes, each has a probability of $\frac{1}{4}$.

EXAMPLE 3: The gambler throws the 4-sided die **three times**.

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message $U_3 = \text{"CBD"}$.



How many outcomes do we have now

?

Notice "**BCD**" is not the same as "**CBD**"

Let's try to quantify "information"

EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$
- There are **4** possible outcomes, each has a probability of $\frac{1}{4}$.

EXAMPLE 3: The gambler throws the 4-sided die **three times**.

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message $U_3 = \text{"CBD"}$.
- Now we had **64** = $4 \cdot 4 \cdot 4 = 4^3$ possible outcomes.



16 times more!

How much more information did we learn in situation 3?



Let's try to quantify "information"

EXAMPLE 1: A gambler throws a fair die with 4 sides {**A**, **B**, **C**, **D**}.



- "Side **C** comes up."
- message $U_1 = \text{"C"}$
- There are 4 possible outcomes, each has a probability of $\frac{1}{4}$.

EXAMPLE 3: The gambler throws the 4-sided die **three times**.

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message $U_3 = \text{"CBD"}$.
- Now we had $64 = 4 \cdot 4 \cdot 4 = 4^3$ possible outcomes.



We have 3 independent throws, the message U is 3 times as long, despite 4^3 possible total outcomes. Our information is 3 times as much.

⇒ 2) Information is additive in some sense

Hartley's measure of information [1928]



1 roll has 4 outcomes.



3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes.

$$\log_4(4) = 1$$

$$\log_4(64) = 3$$

Hartley's insight: use the **logarithm of the number of possible outcomes r** to measure the amount of information in an outcome.

Hartley's measure
of information

$$H_0(U) = \log_b(n)$$

n = number of outcomes



Hartley's measure of information [1928]



1 roll has 4 outcomes.



3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes.

$$\log_4(4) = 1$$

$$\log_4(64) = 3$$

Hartley's insight: use the **logarithm of the number of possible outcomes r** to measure the amount of information in an outcome.

Hartley's measure
of information

$$H_0(U) = \log_b(n)$$

n = number of outcomes



The basis b of the logarithm is not really important.
(just unit of information, like 1 km = 1000 m)

$$\log_2(c) = 1.443 \cdot \log_e(c)$$

$$2^{1.443} = e \Leftrightarrow 1.443 = \log_2(e)$$

We will
use: $\lg(c)$

$$e^z = (2^{1.443})^z = 2^{1.443 \cdot z}$$

Hartley's measure of information [1928]



1 roll has 4 outcomes.



3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes.

$$\log_4(4) = 1$$

$$\log_4(64) = 3$$

Hartley's insight: use the **logarithm of the number of possible outcomes r** to measure the amount of information in an outcome.

Hartley's measure
of information

$$H_0(U) = \log_b(n)$$

n = number of outcomes



For k independent trials,
the amount of information is:

$$\log_b(n^k) = ?$$

Hartley's measure of information [1928]



1 roll has 4 outcomes.



3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes.

$$\log_4(4) = 1$$

$$\log_4(64) = 3$$

Hartley's insight: use the **logarithm of the number of possible outcomes r** to measure the amount of information in an outcome.

Hartley's measure
of information

$$H_0(U) = \log_b(n)$$

n = number of outcomes



For k independent trials,
the amount of information is:

$$\log_b(n^k) = k \cdot \log_b(n)$$

the power of the **logarithm** 😊

Let's practice

EXAMPLE 4: A country has 1 million telephones. How long does the country's telephone numbers need to be?



Let's practice

EXAMPLE 4: A country has 1 million telephones. How long does the country's telephone numbers need to be?

$$\log_{10}(1,000,000) = 6$$

With 6 digits (like "123 456") we can represent 10^6 different telephones.

EXAMPLE 5: The current world population is 8,174,891,806 (as of Sat, September 7, 2024). How long must a binary telephone number be to connect to every person?

A tip: $2^{32} = 4,294, \dots, \dots$



Let's practice

EXAMPLE 4: A country has 1 million telephones. How long does the country's telephone numbers need to be?

$$\log_{10}(1,000,000) = 6$$

With 6 digits (like "123 456") we can represent 10^6 different telephones.

EXAMPLE 5: The current world population is 8,174,891,806 (as of Sat, September 7, 2024). How long must a binary telephone number be to connect to every person?

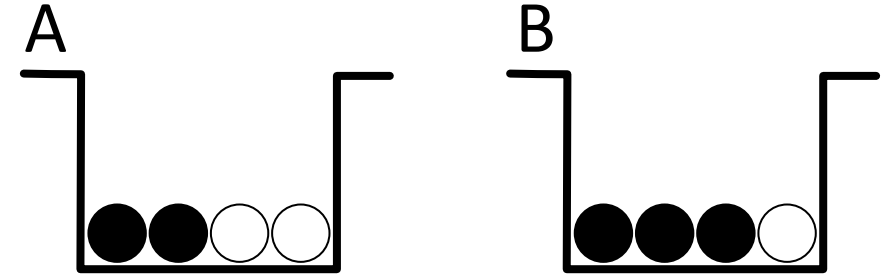
A tip: $2^{32} = 4,294, \dots, \dots$

$$\log_2(8,174,891,806) \approx 32.93$$

With 33 bits we can uniquely identify every person on the planet (today).

A problem with Hartley's information measure

EXAMPLE 6: we have two hats with indistinguishable black and white balls. There are 4 balls total in each hat.



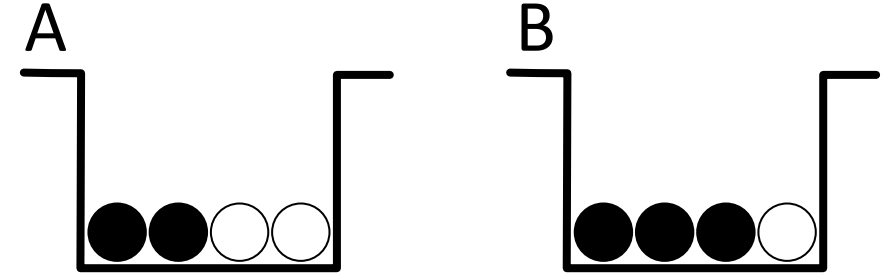
We randomly draw a ball from both hats. Let U_A , U_B be the color of the ball.

What does Hartley's information measure tell us
(maybe let's start with U_A)

?

A problem with Hartley's information measure

EXAMPLE 6: we have two hats with indistinguishable black and white balls. There are 4 balls total in each hat.



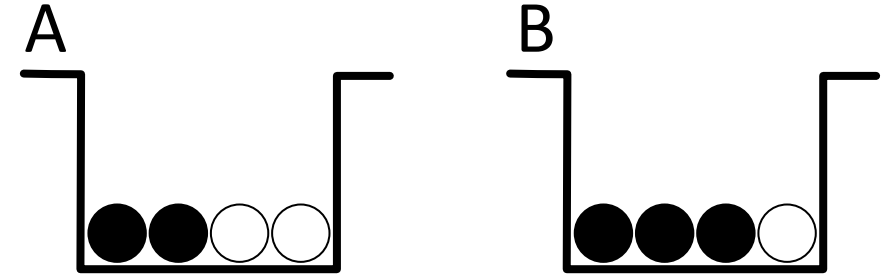
We randomly draw a ball from both hats. Let U_A , U_B be the color of the ball.

$$H_0(U_A) = \lg(2) = 1 \text{ bit} \quad (\text{we have 2 equally likely colors})$$

$$H_0(U_B) = ?$$

A problem with Hartley's information measure

EXAMPLE 6: we have two hats with indistinguishable black and white balls. There are 4 balls total in each hat.



We randomly draw a ball from both hats. Let U_A , U_B be the color of the ball.

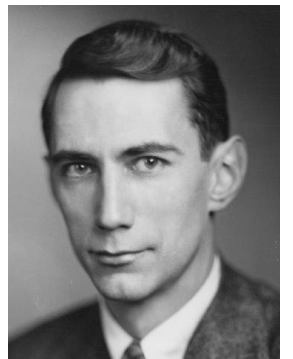
$$H_0(U_A) = \lg(2) = 1 \text{ bit}$$

$$H_0(U_B) = \lg(2) = 1 \text{ bit}$$

Problem: if $U = \text{black}$, then we get less information from U_B than from U_A (since we somehow expected that outcome)

$\Rightarrow$ 3) A proper measure of information should take into account the (possibly different) probabilities of the various outcomes.

This was the key insight of Claude Shannon [1948]

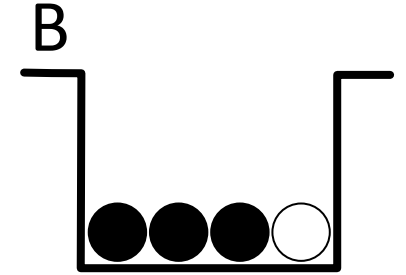


"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

What does Hartley tell us about the information we get after learning $U_B = \text{white}$



"Fixing" Hartley's information measure

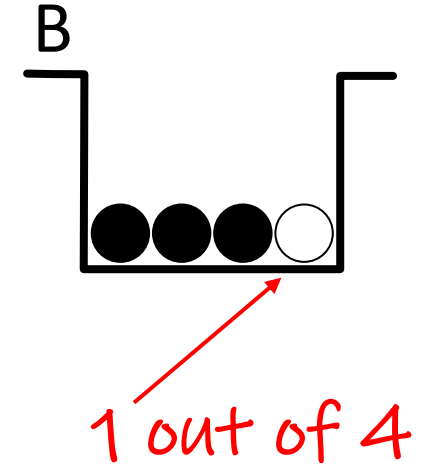
Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = ??? \quad ?$$



"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

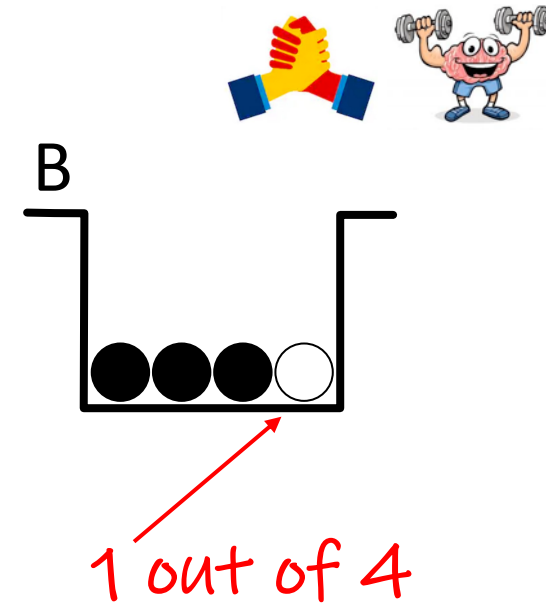
There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

Hartley does not work directly.
What can we do?



"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

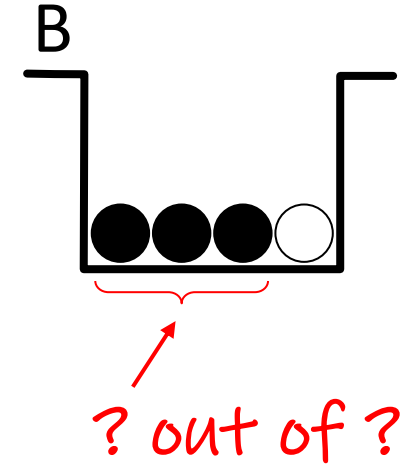
There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

What is our chance p to draw a black ball? ?



"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

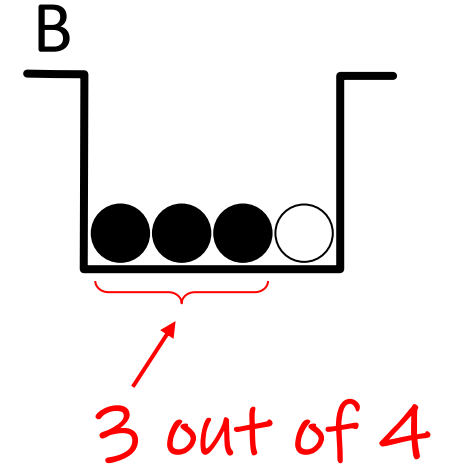
That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

There is a $p = 3/4$ chance to draw a black ball.

What do we do with the $3/4$? ?



"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

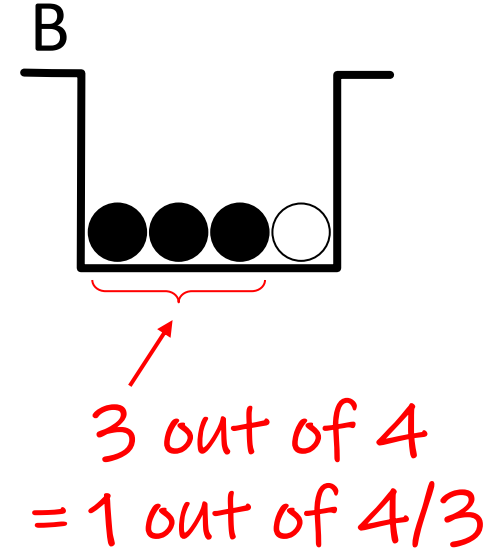
$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \text{?}$$



For Hartley, we need to have 1 black ball (and have "1 out of r outcomes"). We get this by normalizing, i.e. dividing by 3...

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$:

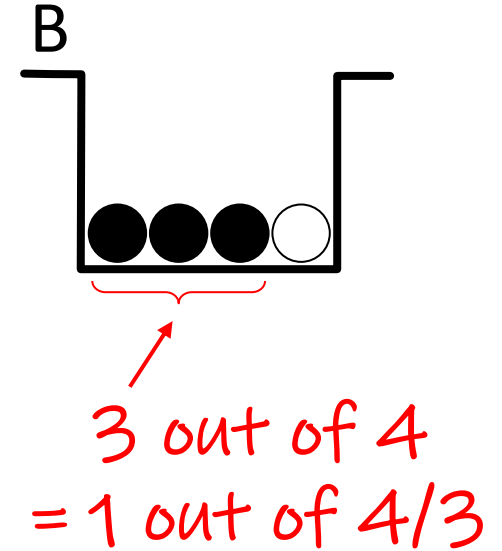
$$\lg\left(\frac{1}{p}\right)$$

#total balls /
#black balls

There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$



How do we combine these two possible outcomes to get one measure

?

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

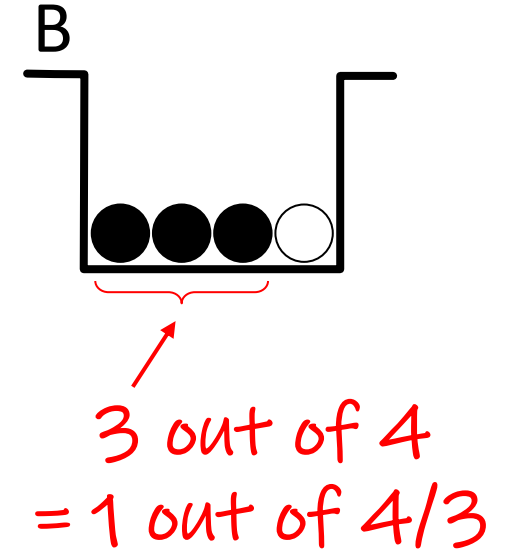
There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

Let's do "in expectation" 😊

$$\mathbb{E}[H_0(U_B)] = \frac{1}{4} \cdot \dots + \frac{3}{4} \cdot \dots$$



"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B :

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

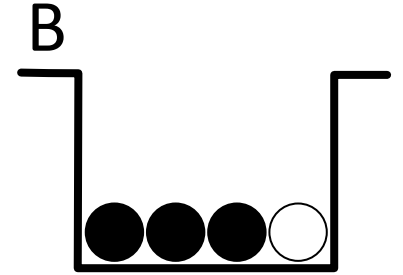
There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

Let's do "in expectation":

$$\mathbb{E}[H_0(U_B)] = \frac{1}{4} \cdot 2 \text{ bits} + \frac{3}{4} \cdot 0.415 \text{ bits} = 0.811 \text{ bits}$$



That's our expected amount of information we learn.

"Fixing" Hartley's information measure

Let's analyze the possible outcomes:

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

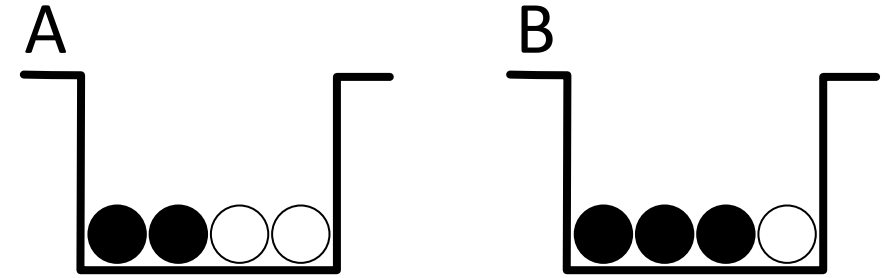
There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

Let's do "in expectation":

$$\mathbb{E}[H_0(U_B)] = \frac{1}{4} \cdot 2 \text{ bits} + \frac{3}{4} \cdot 0.415 \text{ bits} = 0.811 \text{ bits}$$



What would we get for
hat A instead of hat B ?

"Fixing" Hartley's information measure

Let's analyze the possible outcomes:

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

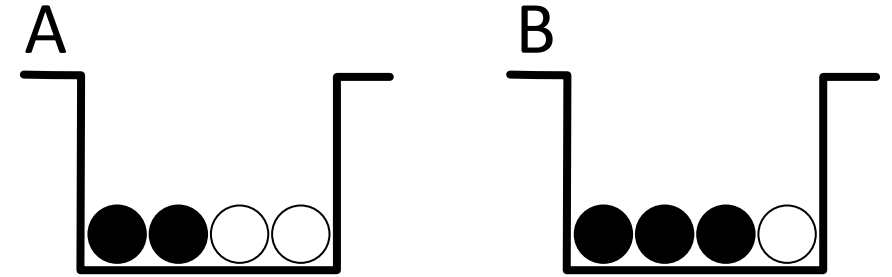
Let's do "in expectation":

$$\mathbb{E}[H_0(U_B)] = \frac{1}{4} \cdot 2 \text{ bits} + \frac{3}{4} \cdot 0.415 \text{ bits} = 0.811 \text{ bits}$$

1 bit for hat A

0.415 bits for hat B

Notice that 1 bit was the min unit of information for the Hartley measure. Expectation allowed us to go lower!



"Fixing" Hartley's information measure

Let's analyze the possible outcomes:

$U_B = \text{white}$:

There is a $p = 1/4$ chance to draw a white ball.

That's the result of 1 out of $n = 4$ possible outcomes.

$$H_0(U_B) = \lg(4) = 2 \text{ bits}$$

$U_B = \text{black}$: $\lg\left(\frac{1}{p}\right)$

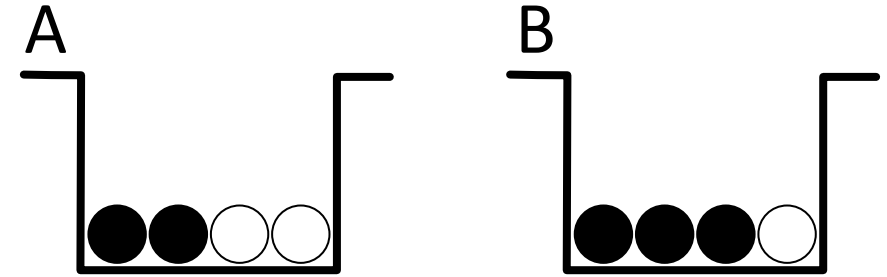
There is a $p = 3/4$ chance to draw a black ball.

That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_B) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

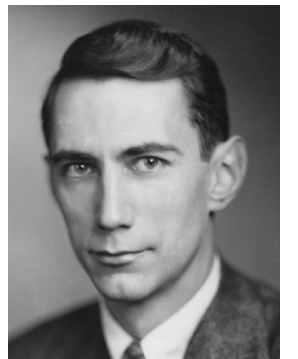
Let's do "in expectation":

$$\mathbb{E}[H_0(U_B)] = \frac{1}{4} \cdot \lg(4) + \frac{3}{4} \cdot \lg\left(\frac{4}{3}\right)$$



*This is Claude Shannon's
measure of information*

1 bit for hat A
= 0.811 bits hat B



Shannon's information content = Hartley information

$$\underbrace{\lg\left(\frac{1}{p_i}\right)}_{H_0(U)}$$

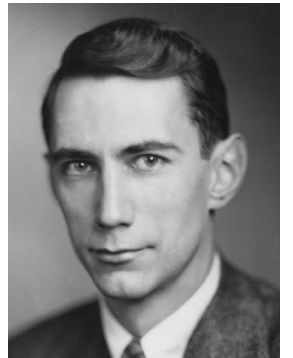
p_i = probability of the i -th possible outcome

Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

Shannon information content, quantifies level of surprise of a particular outcome with probability $p(U)$

Also, the amount of information you gain by observing outcome i

Then, entropy is the total expected surprise



Shannon's entropy

Shannon's measure of information as expected Hartley information (averaged over all possible outcomes)

$$H(\mathbf{p}) = \sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right) = - \sum_{i=1}^r p_i \cdot \lg(p_i) = \mathbb{E} \left[\lg\left(\frac{1}{p_i}\right) \right]$$

$H_0(U)$

p_i = probability of the i -th possible outcome

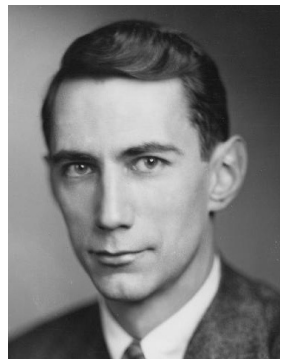
Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

1948:

A Mathematical Theory of Communication

By C. E. SHANNON

$$H = -K \sum_{i=1}^n p_i \log p_i$$



Shannon's entropy

Shannon's measure of information as expected Hartley information (averaged over all possible outcomes)

$$H(\mathbf{p}) = \sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right) = - \sum_{i=1}^r p_i \cdot \lg(p_i) = \mathbb{E} \left[\lg\left(\frac{1}{p_i}\right) \right]$$

$H_0(U)$

p_i = probability of the i -th possible outcome

Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

1928:

Transmission of Information

By R. V. L. HARTLEY

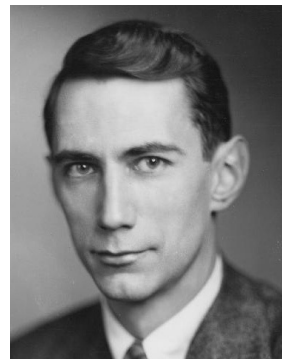
$$H = Kn,$$
$$H = n \log s$$

1948:

A Mathematical Theory of Communication

By C. E. SHANNON

$$H = -K \sum_{i=1}^n p_i \log p_i$$



Ralph Hartley. Transmission of information, The Bell System Technical Journal, 1928. <https://doi.org/10.1002/j.1538-7305.1928.tb01236.x>

Claude Shannon. A Mathematical Theory of Communication, The Bell System Technical Journal, 1948. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>

Gatterbauer. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/fa25/>

Shannon's entropy

Shannon's measure of information as expected Hartley information (averaged over all possible outcomes)

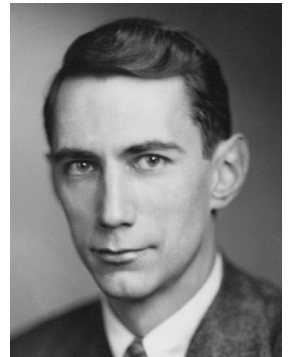
$$H(\mathbf{p}) = \sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right) = - \sum_{i=1}^r p_i \cdot \lg(p_i) = \mathbb{E} \left[\lg\left(\frac{1}{p_i}\right) \right]$$

$H_0(U)$

p_i = probability of the i -th possible outcome

Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

- 1) The **number of possible outcomes** should be linked to "information" H_0
- 2) Information is **additive** in some sense H_0
- 3) A proper measure of information should take into account the **different probabilities of the outcomes.** H



Ralph Hartley. Transmission of information, The Bell System Technical Journal, 1928. <https://doi.org/10.1002/j.1538-7305.1928.tb01236.x>

Claude Shannon. A Mathematical Theory of Communication, The Bell System Technical Journal, 1948. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>

Gatterbauer. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/fa25/>

- I am thinking of an integer from 0 to 255 (=256 or 2^8 choices).
 - You can ask me binary questions ("Is it < 30?")
 - How many questions do you have to maximally ask me to get the number?
 - You can always halve the range each time. Assume I choose 3:
 - Is it < 128? yes
 - Is it < 64? yes
 - Is it < 32? yes
 - Is it < 16? yes
 - Is it < 8? yes
 - Is it < 4? yes
 - Is it < 2? no
 - Is it < 3? no -> must be 3
 - 3 = 00000011 in binary (the answer to each question corresponds to 1 bit)

Shannon Information Content: smallest number of yes/no questions to identify that integer

Part 1: Theory

L04: Basics of entropy (2/7)

[measures of information, intuition behind entropy]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa26)

<https://northeastern-datalab.github.io/cs7840/fa26/>

9/21/2026

Pre-class conversations

- Last class recapitulation
- Changes slide posting policy (unedited version right after class on Piazza, updated version day before next class)
- Check out: Online Python notebook, feedback **very** welcome:
 - maybe creating those are only a learning experience for those doing it
 - also possibly useful for your own scribes
- Today:
 - Keep pen & paper ready for hands-on calculus, logarithms, lots of math
 - also see Schneider's "Information Theory Primer, With an Appendix on Logarithms"
 - information theory is a bit like probability theory, but in log space
 - Intuition behind entropy (and variants)

The value of both experimenting and formal training

The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka” but “That’s funny...”

—*Isaac Asimov (1920–1992)*

See also: <https://www.americanscientist.org/article/thats-funny>



“In the fields of observation chance favors only the prepared mind.”

— **Louis Pasteur**

tags: **chance**, **preparation**

See also: <https://www.goodreads.com/quotes/827857-in-the-fields-of-observation-chance-favors-only-the-prepared>

Focus on clear communication

"So if the field as we know it is to survive, mathematicians need to figure out how to adapt. One idea is to give less credit for solving a problem and more credit for work that communicates the ideas behind a solution in a way that can benefit the broader mathematical community. Yet it's hard to say exactly how that works in practice. Unlike solving problems, evaluating how well a mathematician has communicated with their peers has no clear metrics."

Properties of information (entropy) by example

Shannon entropy for unbiased outcomes

EXAMPLE 1: What is the entropy in a roll of an unbiased 8-sided die?



$$H(\mathbf{p}) = \sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right)$$

Shannon entropy for unbiased outcomes

EXAMPLE 1: What is the entropy in a roll of an unbiased 8-sided die?



$$H(\mathbf{p}) = \boxed{\sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right)} = \underbrace{\left(\sum_{i=1}^r p_i\right)}_1 \cdot \lg\left(\frac{1}{p_i}\right) = \lg\left(\frac{1}{p_i}\right) \quad ?$$

Shannon entropy for unbiased outcomes = Hartley measure

EXAMPLE 1: What is the entropy in a roll of an unbiased 8-sided die?



$$H(\mathbf{p}) = \boxed{\sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right)} = \underbrace{\left(\sum_{i=1}^r p_i\right)}_1 \cdot \lg\left(\frac{1}{p_i}\right) = \lg\left(\frac{1}{p_i}\right) = H_0\left(\frac{1}{p_i}\right)$$

= ?

number of outcomes

For unbiased outcomes, **Entropy** is exactly the **Hartley** information measure 😊

Shannon entropy for unbiased outcomes = Hartley measure

EXAMPLE 1: What is the entropy in a roll of an unbiased 8-sided die?



$$H(\mathbf{p}) = \sum_{i=1}^r p_i \cdot \lg\left(\frac{1}{p_i}\right) = \underbrace{\left(\sum_{i=1}^r p_i\right)}_1 \cdot \lg\left(\frac{1}{p_i}\right) = \lg\left(\frac{1}{p_i}\right) = H_0\left(\frac{1}{p_i}\right)$$

$= \lg(8) = 3$

number of outcomes

For unbiased outcomes, **Entropy** is exactly the **Hartley** information measure 😊

Characterization of the Hartley information measure

Shannon entropy for uniform sampling from n choices.

$$H_0(r) = H_0\left(\frac{1}{p_i}\right) = \lg(n)$$

two independent uniformly distributed RVs,
with alphabet size m and n

The Hartley function only depends on the number of elements in a set, and hence can be viewed as a function on natural numbers. Rényi showed that the Hartley function in base 2 is the only function mapping natural numbers to real numbers that satisfies

1. $H_0(mn) = H_0(m) + H_0(n)$ (additivity) $\lg(m \cdot n) = \lg(m) + \lg(n)$
2. $H_0(m) \leq H_0(m + 1)$ (monotonicity)
3. $H_0(2) = 1$ (normalization)

Condition 1 says that the uncertainty of the Cartesian product of two finite sets A and B is the sum of uncertainties of A and B . Condition 2 says that a larger set has larger uncertainty.

Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even.

How much information did we learn? ?

Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even.
How much information did we learn?

- **Before** the message:
- **After** the message:



Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even. How much information did we learn?

- Before the message: There are 8 choices: $\{1, 2, 3, 4, 5, 6, 7, 8\}$ $H_0(8) = 3$ bits
- After the message: There are 4 choices: $\{2, 4, 6, 8\}$ $H_0(4) = 2$ bits

How much uncertainty did we have before?

How much uncertainty did we have after



Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even. How much information did we learn?

- Before the message: There are 8 choices: $\{1, 2, 3, 4, 5, 6, 7, 8\}$ $H_0(8) = 3$ bits
- After the message: There are 4 choices: $\{2, 4, 6, 8\}$ $H_0(4) = 2$ bits

Let's think about encodings *(binary encoding with atypical 1-indexing)*

Before: $\{1, 2, 3, 4, 5, 6, 7, 8\}$
 ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
 000 001 010 011 100 101 110 111

After: $\{\text{1}, \text{2}, \text{3}, \text{4}, \text{5}, \text{6}, \text{7}, \text{8}\}$
 ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
 000 001 010 011 100 101 110 111

Do you notice something



Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even. How much information did we learn?

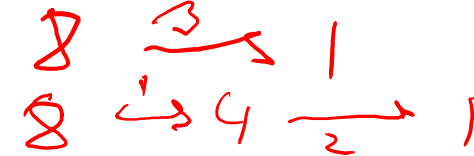
- Before the message: There are 8 choices: $\{1, 2, 3, 4, 5, 6, 7, 8\}$ $H_0(8) = 3$ bits
- After the message: There are 4 choices: $\{2, 4, 6, 8\}$ $H_0(4) = 2$ bits

Let's think about encodings

Before: $\{1, 2, 3, 4, 5, 6, 7, 8\}$
 ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
 000 001 010 011 100 101 110 111

After: $\{1, 2, 3, 4, 5, 6, 7, 8\}$
 ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
 000 00**1** 010 01**1** 100 10**1** 110 11**1**

We have learned 1 bit! ??**1**



"Grouping rule": Dividing the outcomes into two (last bit is 0 or 1), randomly choose one group (e.g. 1), and then randomly pick an element from that group (e.g. 111), gives same entropy as picking 111 from the start.

Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even. How much information did we learn?

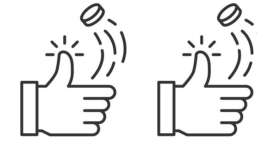
- Before the message: There are 8 choices: $\{1, 2, 3, 4, 5, 6, 7, 8\}$ $H_0(8) = 3$ bits
- After the message: There are 4 choices: $\{2, 4, 6, 8\}$ $H_0(4) = 2$ bits

Recall: **information is additive**:

1 flip of a 2-sided coin has 2 outcomes.



2 flips have $2^2 = 4$ outcomes.



3 flips have $2^3 = 8$ outcomes.



$$\lg(2) = 1$$

$$\lg(4) = 2$$

$$\lg(8) = 3$$

+1 bit

+1 bit

Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get a message with the information that the outcome of a roll is even. How much information did we learn?

- Before the message: There are 8 choices: $\{1, 2, 3, 4, 5, 6, 7, 8\}$ $H_0(8) = 3$ bits
- After the message: There are 4 choices: $\{2, 4, 6, 8\}$ $H_0(4) = 2$ bits

The power of the logarithm: transform multiplication into addition

Uncertainty before – Uncertainty after

$$\underbrace{\lg(8) - \lg(4)}$$

$$\lg\left(\frac{8}{4}\right) = \lg(2) = 1 \text{ bit}$$

$$H(U) = \lg\left(\frac{n}{m}\right)$$

Information content in a message U that reduces the number of unbiased outcomes from n to m

Learning partial information



EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.

We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?



$$H(U_1) = ?$$

$$H(U_2|U_1) = ?$$
 These are called "conditional entropies"!

$$H(U_3|U_{1,2}) = ?$$

$$H(U_4|U_{1-3}) = ?$$

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: $U_1 =$ "The outcome of the roll is not **1**", $U_2 =$ "... not **3**", $U_3 =$ "... not **5**", $U_4 =$ "... not **7**".

How much information do we learn from each individual message?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \text{?} \quad \text{will this be bigger or smaller}$$

$$H(U_3|U_{1,2}) = \text{?}$$

$$H(U_4|U_{1-3}) = \text{?}$$

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?

... and all of them together?



$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$

$$H(U_3|U_{1,2}) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$

$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?
... and all of them together?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$

$$H(U_3|U_{1,2}) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$

$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$

$$H(\{U_1, U_2, U_3, U_4\}) = \mathbf{1 \text{ bit}}$$

*How come that the **SUM** of these numbers turns out to be soooo nice?*



Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?
... and all of them together?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$

$$H(U_3|U_{1,2}) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$

$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$

$$H(\{U_1, U_2, U_3, U_4\}) = 1 \text{ bit}$$

$$H(\{U_1, U_2, U_3, U_4\})$$

$$= H(U_1) + H(U_2|U_1) + H(U_3|U_{1,2}) + H(U_4|U_{1-3})$$

*This is called the "chain rule"
(recall the chain rule we saw for probabilities)*

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?
... and all of them together?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$

$$H(U_3|U_{1,2}) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$

$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$

$$H(\{U_1, U_2, U_3, U_4\}) = 1 \text{ bit}$$

$$H(\{U_1, U_2, U_3, U_4\})$$

$$= H(U_1) + H(U_2|U_1) + H(U_3|U_{1,2}) + H(U_4|U_{1-3})$$

$$= \lg\left(\frac{8}{7}\right) + \lg\left(\frac{7}{6}\right) + \lg\left(\frac{6}{5}\right) + \lg\left(\frac{5}{4}\right)$$

$$= \text{?}$$

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides $\{1, 2, \dots, 8\}$.



We then get 4 messages, one after the other: U_1 = "The outcome of the roll is not **1**", U_2 = "... not **3**", U_3 = "... not **5**", U_4 = "... not **7**".

How much information do we learn from each individual message?

... and all of them together?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$

$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$

$$H(U_3|U_{1,2}) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$

$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$

$$H(\{U_1, U_2, U_3, U_4\}) = 1 \text{ bit}$$

$$H(\{U_1, U_2, U_3, U_4\})$$

$$= H(U_1) + H(U_2|U_1) + H(U_3|U_{1,2}) + H(U_4|U_{1-3})$$

$$= \lg\left(\frac{8}{7}\right) + \lg\left(\frac{7}{6}\right) + \lg\left(\frac{6}{5}\right) + \lg\left(\frac{5}{4}\right)$$

$$= \lg\left(\frac{8}{\cancel{7}} \cdot \frac{\cancel{7}}{\cancel{6}} \cdot \frac{\cancel{6}}{\cancel{5}} \cdot \frac{\cancel{5}}{4}\right) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

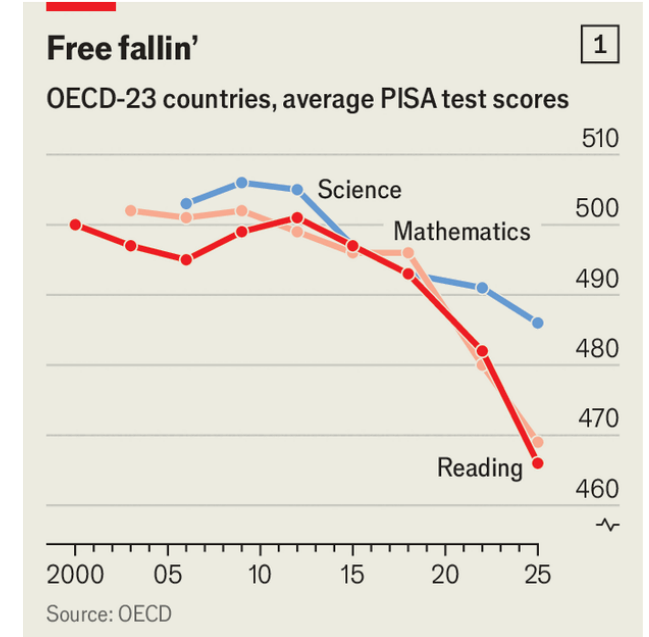
Again, the logarithm 😊

Why do we do so much basic logarithm math?

"Test scores are not the full measure of a child. But they are a good predictor of desirable outcomes and a bulwark against many of the worst ones. Kids who do well on tests tend to earn more, live longer and healthier lives, and commit fewer crimes. By one estimate, pushing up PISA scores by a mere quarter of a standard deviation might increase annual growth in rich countries by half a percentage point.

Perhaps, though, as artificial intelligence improves, the fact that children struggle intellectually will not matter so much? The adults of the future will be able to outsource their thinking to clever machines, just as they use machines to perform innumerable physical tasks. So why sweat about swotting?

This argument could not be more wrong. AI will not make it pointless to train human brains, any more than the invention of cars eliminated the need for physical exercise. On the contrary, adapting to a world of rapid, AI-driven change will require more mental agility, not less. And precisely because AI will increase the temptation for people to outsource difficult cognitive tasks, schools have an extra duty to help pupils exercise their brains and acquire the hard habits of thinking. Failure to furnish them with the basic building blocks of a well-developed mind, such as literacy and numeracy, is inexcusable."



Maximum Entropy distributions (with a bit of Calculus)

Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

Biased coin flip:



$$H_B(p) = ?$$

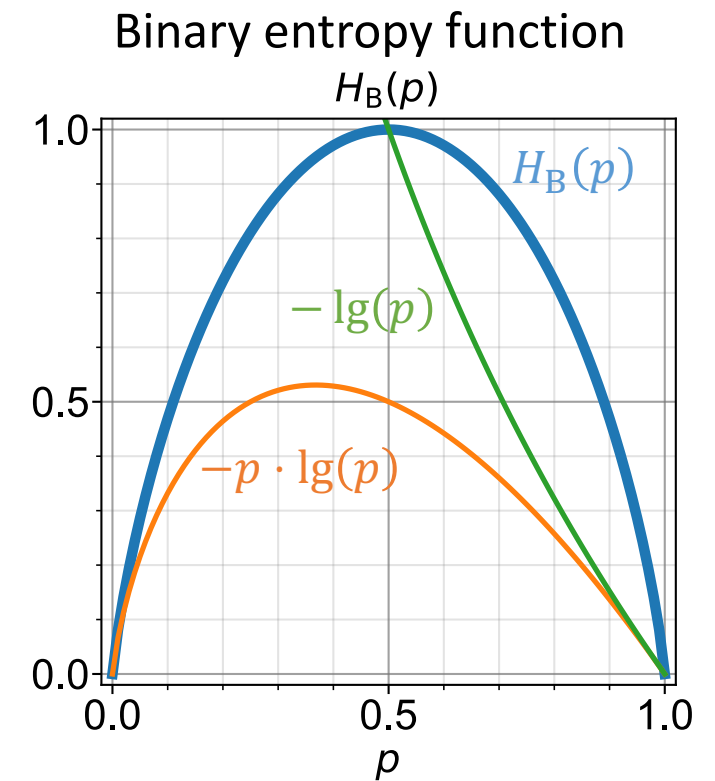
Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$



$$H_B(p) = -p \cdot \lg(p) - (1 - p) \cdot \lg(1 - p)$$

How to choose p in order to maximize entropy ?



Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

$$H_B(p) = -p \cdot \lg(p) - (1 - p) \cdot \lg(1 - p)$$

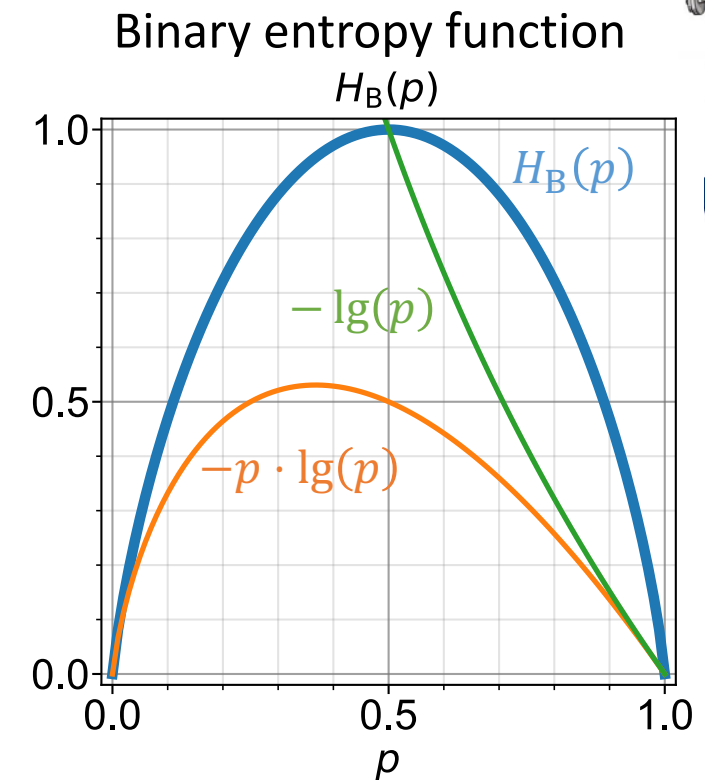
How to choose p in order to maximize entropy?

$\frac{dH}{dp} =$? We do the exercise together, but I first want you to create a logarithm cheat sheet

Understanding "change of basis"

$$\lg(x) = \log_2(x) = \frac{\ln(x)}{\ln(2)}$$

How do you derive that ?



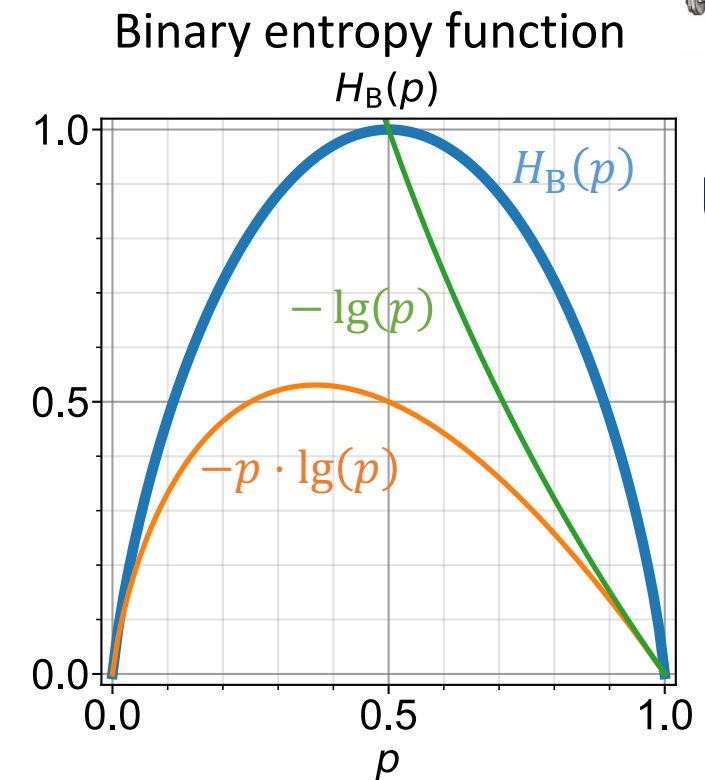
Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

$$H_B(p) = -p \cdot \lg(p) - (1 - p) \cdot \lg(1 - p)$$

How to choose p in order to maximize entropy?

$\frac{dH}{dp} =$  We do the exercise together, but I first want you to create a logarithm cheat sheet



Understanding "change of basis"

$$\lg(x) = \log_2(x) = \frac{\ln(x)}{\ln(2)} \quad \text{definition}$$

$$2^{\log_2(x)} = x \quad \text{apply } \ln(\dots) \text{ on both sides}$$

$$\ln(2^{\log_2(x)}) = \ln(x) \quad \log(a^b) = b \cdot \log(a)$$

$$\log_2(x) \cdot \ln(2) = \ln(x)$$

Calculus
cheat
sheet

$$\ln(x)' =$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)} \right)' =$$

$$(x \cdot \lg(x))' =$$

$$\lg(1 - x)' =$$

$$((1 - x) \cdot \lg(1 - x))' =$$



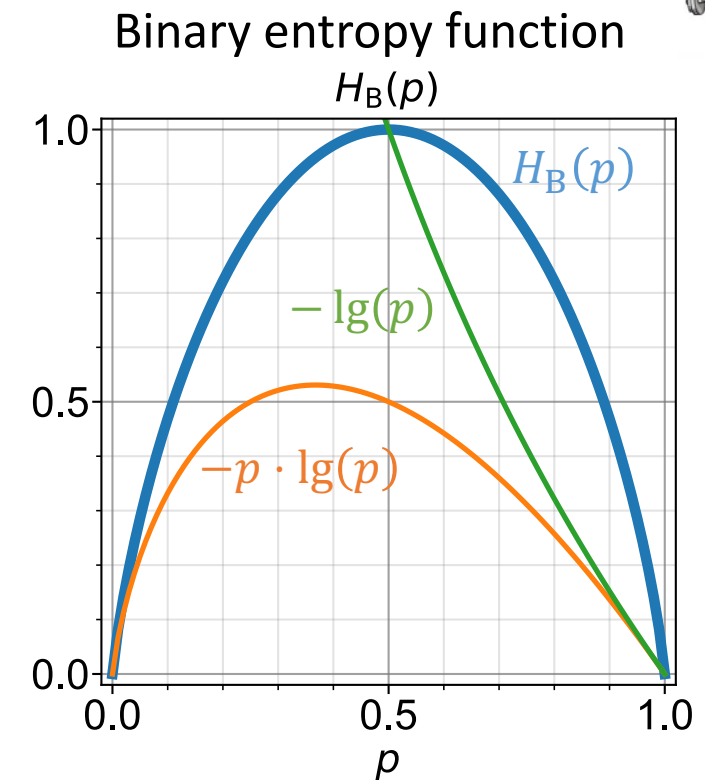
Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

$$H_B(p) = -p \cdot \lg(p) - (1 - p) \cdot \lg(1 - p)$$

How to choose p in order to maximize entropy?

$\frac{dH}{dp} =$? We do the exercise together, but I first want you to create a logarithm cheat sheet



Understanding "change of basis"

$$\lg(x) = \log_2(x) = \frac{\ln(x)}{\ln(2)} \quad \text{definition}$$

$$2^{\log_2(x)} = x \quad \text{apply } \ln(\dots) \text{ on both sides}$$

$$\ln(2^{\log_2(x)}) = \ln(x) \quad \log(a^b) = b \cdot \log(a)$$

$$\log_2(x) \cdot \ln(2) = \ln(x)$$

Calculus
cheat
sheet

$$\ln(x)' = \frac{1}{x}$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)} \right)' = \frac{1}{x \cdot \ln(2)}$$

product rule $(x \cdot \lg(x))' = \cancel{x} \frac{1}{x \ln(2)} + \lg(x)$

chain rule $\lg(1 - x)' = -\frac{1}{(1-x) \cdot \ln(2)}$

$$((1 - x) \cdot \lg(1 - x))' = -\frac{1}{\ln(2)} - \lg(1 - x)$$

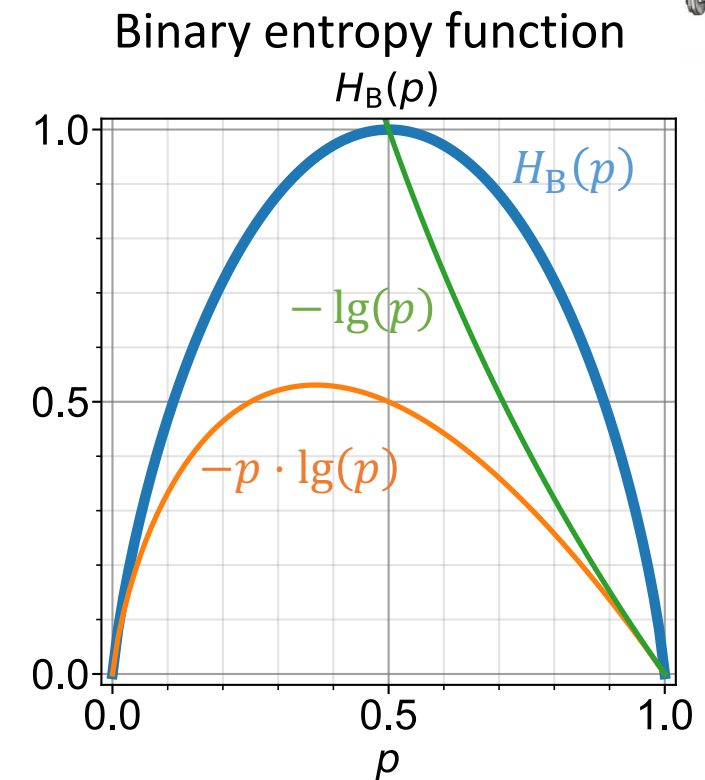
Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

$$H_B(p) = \underbrace{-p \cdot \lg(p)}_{\text{term 1}} - \underbrace{(1 - p) \cdot \lg(1 - p)}_{\text{term 2}}$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = \underbrace{-\cancel{\frac{1}{\ln(2)}} - \lg(p)}_{\text{derivative of term 1}} + \underbrace{\cancel{\frac{1}{\ln(2)}} + \lg(1 - p)}_{\text{derivative of term 2}}$$



Calculus
cheat
sheet

$$\ln(x)' = \frac{1}{x}$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)}\right)' = \frac{1}{x \cdot \ln(2)}$$

$$(x \cdot \lg(x))' = \cancel{x} \frac{1}{\cancel{x} \ln(2)} + \lg(x)$$

$$\lg(1 - x)' = -\frac{1}{(1 - x) \cdot \ln(2)}$$

$$((1 - x) \cdot \lg(1 - x))' = -\frac{1}{\ln(2)} - \lg(1 - x)$$

Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

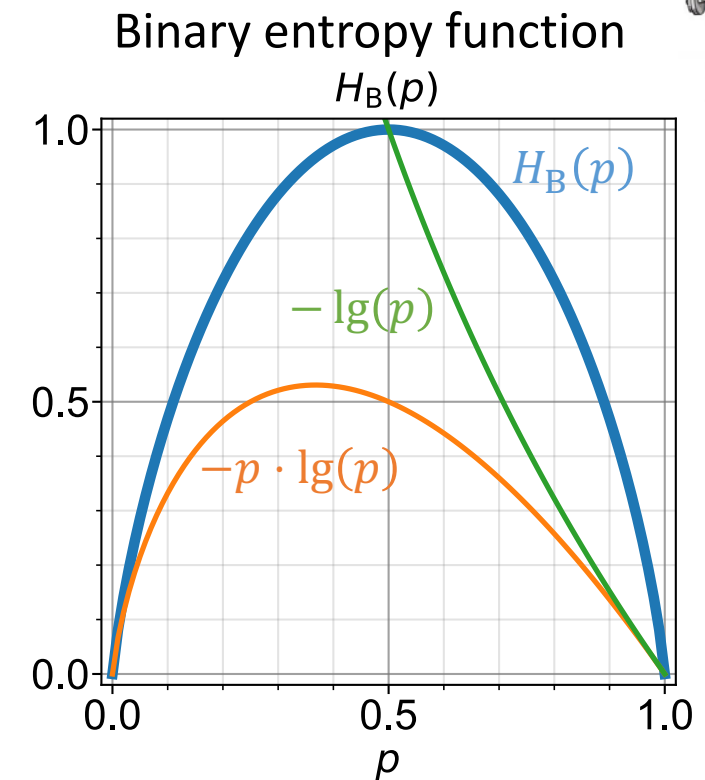
$$H_B(p) = \underbrace{-p \cdot \lg(p)}_{\text{term 1}} - \underbrace{(1 - p) \cdot \lg(1 - p)}_{\text{term 2}}$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = \underbrace{-\cancel{\frac{1}{\ln(2)}} - \lg(p)}_{\text{term 1}} + \underbrace{\cancel{\frac{1}{\ln(2)}} + \lg(1 - p)}_{\text{term 2}} = 0$$

$$\Leftrightarrow \lg\left(\frac{1-p}{p}\right) = 0 \Leftrightarrow \left(\frac{1-p}{p}\right) = 1 \Leftrightarrow \boxed{p = \frac{1}{2}}$$

$$\frac{d^2H}{dp^2} = ?$$



Calculus
cheat
sheet

$$\ln(x)' = \frac{1}{x}$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)}\right)' = \frac{1}{x \cdot \ln(2)}$$

$$(x \cdot \lg(x))' = \cancel{x} \frac{1}{x \ln(2)} + \lg(x)$$

$$\lg(1 - x)' = -\frac{1}{(1-x) \cdot \ln(2)}$$

$$((1 - x) \cdot \lg(1 - x))' = -\frac{1}{\ln(2)} - \lg(1 - x)$$

Binary Entropy Function

X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

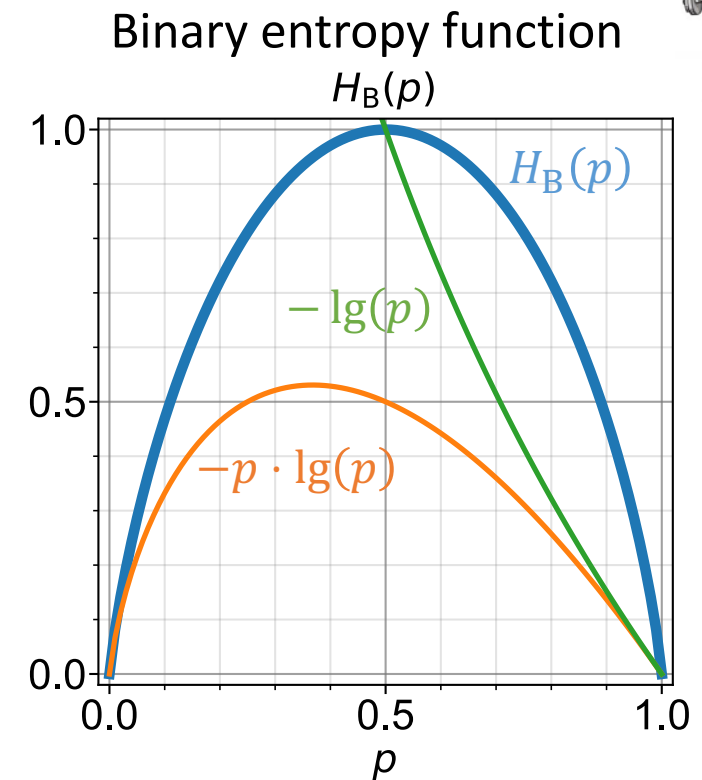
$$H_B(p) = \underbrace{-p \cdot \lg(p)}_{\text{red bracket}} - \underbrace{(1 - p) \cdot \lg(1 - p)}_{\text{red bracket}}$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = \underbrace{-\cancel{\frac{1}{\ln(2)}} - \lg(p)}_{\text{red bracket}} + \underbrace{\cancel{\frac{1}{\ln(2)}} + \lg(1 - p)}_{\text{red bracket}} = 0$$

$$\Leftrightarrow \lg\left(\frac{1-p}{p}\right) = 0 \Leftrightarrow \left(\frac{1-p}{p}\right) = 1 \Leftrightarrow \boxed{p = \frac{1}{2}}$$

$$\frac{d^2H}{dp^2} = -\frac{1}{p \cdot \ln(2)} - \frac{1}{(1-p) \cdot \ln(2)} < 0 \text{ concave}$$



Calculus
cheat
sheet

$$\ln(x)' = \frac{1}{x}$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)}\right)' = \frac{1}{x \cdot \ln(2)}$$

$$(x \cdot \lg(x))' = \cancel{x} \frac{1}{\cancel{x} \ln(2)} + \lg(x)$$

$$\lg(1 - x)' = -\frac{1}{(1-x) \cdot \ln(2)}$$

$$((1 - x) \cdot \lg(1 - x))' = -\frac{1}{\ln(2)} - \lg(1 - x)$$

Asking ChatGPT for help

Can you calculate step-by-step the derivative for p of:

$$H(p) = -p \cdot \lg(p) - (1-p) \cdot \lg[(1-p)]$$

Repeat but assume log is to the basis 2 instead of e

Sure! Let's differentiate the function

$$H(p) = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p),$$

where $\log_2$ denotes the logarithm base 2. The process is similar to before, so let's go through it step-by-step.

Step 1: Rewrite the Function

The function is already clear:

$$H(p) = -p \log_2(p) - (1-p) \log_2(1-p).$$

Step 2: Differentiate $H(p)$

We will differentiate each term separately.

Term 1: Differentiate $-p \log_2(p)$

Using the product rule:

$$\frac{d}{dp}[-p \log_2(p)] = -\left(\log_2(p) + p \cdot \frac{1}{p \ln(2)}\right) = -\log_2(p) - \frac{1}{\ln(2)}.$$

Term 2: Differentiate $-(1-p) \log_2(1-p)$

Again, using the product rule:

$$\frac{d}{dp}[-(1-p) \log_2(1-p)] = -\left(-\log_2(1-p) + (1-p) \cdot \frac{-1}{(1-p) \ln(2)}\right) = \log_2(1-p) - \frac{1}{\ln(2)}.$$

Step 3: Combine the Derivatives

Now we can combine the derivatives of both terms:

$$H'(p) = -\log_2(p) - \frac{1}{\ln(2)} + \log_2(1-p) - \frac{1}{\ln(2)}.$$

Step 4: Simplify

Combining the constant terms and rearranging gives:

$$H'(p) = \log_2(1-p) - \log_2(p) - \frac{2}{\ln(2)}.$$

This can also be expressed using properties of logarithms:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right) - \frac{2}{\ln(2)}.$$

Final Result

Thus, the derivative of $H(p)$ with respect to p is:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right) - \frac{2}{\ln(2)}.$$



ChatGPT made the *same* mistake as me!

Asking ChatGPT for help

Can you calculate step-by-step the derivative for p of:

$$H(p) = -p \cdot \lg(p) - (1-p) \cdot \lg[(1-p)]$$

Repeat but assume log is to the basis 2 instead of e

Sure! Let's differentiate the function

$$H(p) = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p),$$

where $\log_2$ denotes the logarithm base 2. The process is similar to before, so let's go through it step-by-step.

Step 1: Rewrite the Function

The function is already clear:

$$H(p) = -p \log_2(p) - (1-p) \log_2(1-p).$$

Step 2: Differentiate $H(p)$

We will differentiate each term separately.

Term 1: Differentiate $-p \log_2(p)$

Using the product rule:

$$\frac{d}{dp}[-p \log_2(p)] = -\left(\log_2(p) + p \cdot \frac{1}{p \ln(2)}\right) = -\log_2(p) - \frac{1}{\ln(2)}.$$

Term 2: Differentiate $-(1-p) \log_2(1-p)$

Again, using the product rule:

$$\frac{d}{dp}[-(1-p) \log_2(1-p)] = -\left(-\log_2(1-p) + (1-p) \cdot \frac{-1}{(1-p) \ln(2)}\right) = \log_2(1-p) - \frac{1}{\ln(2)}.$$

Step 3: Combine the Derivatives

Now we can combine the derivatives of both terms:

$$H'(p) = -\log_2(p) - \frac{1}{\ln(2)} + \log_2(1-p) - \frac{1}{\ln(2)}.$$

Step 4: Simplify

Combining the constant terms and rearranging gives:

$$H'(p) = \log_2(1-p) - \log_2(p) - \frac{2}{\ln(2)}.$$

This can also be expressed using properties of logarithms:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right) - \frac{2}{\ln(2)}.$$

Final Result

Thus, the derivative of $H(p)$ with respect to p is:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right) - \frac{2}{\ln(2)}.$$

You had some error in earlier step 2 term 2. Please check the minuses

Let's go through the differentiation of the term $-(1-p) \log_2(1-p)$ again carefully to ensure we get the signs correct.

Step 1: Differentiate $-(1-p) \log_2(1-p)$

Using the product rule, we have:

$$\frac{d}{dp}[-(1-p) \log_2(1-p)] = -\left(\frac{d}{dp}(1-p) \cdot \log_2(1-p) + (1-p) \cdot \frac{d}{dp}[\log_2(1-p)]\right).$$

1. The derivative of $1-p$ is -1 .

2. The derivative of $\log_2(1-p)$ is:

$$\frac{d}{dp}[\log_2(1-p)] = -\frac{1}{(1-p) \ln(2)}.$$

Step 2: Apply the Product Rule

Putting it all together:

$$\frac{d}{dp}[-(1-p) \log_2(1-p)] = -\left(-\log_2(1-p) + (1-p) \cdot \left(-\frac{1}{(1-p) \ln(2)}\right)\right).$$

Step 3: Simplify the Expression

This simplifies to:

$$\log_2(1-p) + \frac{1}{\ln(2)}.$$

Final Result for the Derivative of $H(p)$

Now, let's combine everything correctly for the overall derivative $H'(p)$:

1. The derivative of the first term $-p \log_2(p)$:

$$-\log_2(p) - \frac{1}{\ln(2)}.$$

2. The corrected derivative of the second term $-(1-p) \log_2(1-p)$:

$$\log_2(1-p) + \frac{1}{\ln(2)}.$$

Combining these:

$$H'(p) = \left(-\log_2(p) - \frac{1}{\ln(2)}\right) + \left(\log_2(1-p) + \frac{1}{\ln(2)}\right).$$

The $-\frac{1}{\ln(2)}$ and $+\frac{1}{\ln(2)}$ cancel out, leading to:

$$H'(p) = \log_2(1-p) - \log_2(p).$$

Using properties of logarithms:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right).$$

Final Result

Thus, the corrected derivative of $H(p)$ is:

$$H'(p) = \log_2\left(\frac{1-p}{p}\right).$$

Thank you for your patience, and I apologize for the confusion earlier!

Maximum Entropy distribution for nonbinary discrete RV

Entropy of a discrete distribution with n outcomes $H(\mathbf{p}) = -\sum_{i=1}^n p_i \cdot \lg(p_i)$

Which choice of $\mathbf{p} = (p_1, \dots, p_n)$ maximizes the entropy?

?

Maximum Entropy distribution for nonbinary discrete RV

Entropy of a discrete distribution with n outcomes $H(\mathbf{p}) = -\sum_{i=1}^n p_i \cdot \lg(p_i)$

Which choice of $\mathbf{p} = (p_1, \dots, p_n)$ maximizes the entropy?

Can be solved with constrained optimization:

$$\max_{\mathbf{p}} [H(\mathbf{p})] \quad \text{s.t.} \quad \sum_{i=1}^n p_i = 1$$

Form the Lagrangian:

?

Maximum Entropy distribution for nonbinary discrete RV

Entropy of a discrete distribution with n outcomes $H(\mathbf{p}) = - \sum_{i=1}^n p_i \cdot \lg(p_i)$

Which choice of $\mathbf{p} = (p_1, \dots, p_n)$ maximizes the entropy?

Can be solved with constrained optimization: $\max_{\mathbf{p}} [H(\mathbf{p})] \text{ s.t. } \sum_{i=1}^n p_i = 1$

Form the Lagrangian:

$$J(\mathbf{p}, \lambda) = - \sum_{i=1}^n p_i \cdot \lg(p_i) + \lambda \left(\sum_{i=1}^n p_i - 1 \right)$$

$$\frac{\partial J}{\partial p_i} = ?$$

Calculus	$\ln(x)' = \frac{1}{x}$
exercise	$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)} \right)' = \frac{1}{x \cdot \ln(2)}$
	$(x \cdot \lg(x))' = \frac{1}{\ln(2)} + \lg(x)$

Maximum Entropy distribution for nonbinary discrete RV

Entropy of a discrete distribution with n outcomes $H(\mathbf{p}) = -\sum_{i=1}^n p_i \cdot \lg(p_i)$

Which choice of $\mathbf{p} = (p_1, \dots, p_n)$ maximizes the entropy?

Can be solved with constrained optimization: $\max_{\mathbf{p}} [H(\mathbf{p})] \text{ s.t. } \sum_{i=1}^n p_i = 1$

Form the Lagrangian:

$$J(\mathbf{p}, \lambda) = -\sum_{i=1}^n p_i \cdot \lg(p_i) + \lambda \left(\sum_{i=1}^n p_i - 1 \right)$$

$$\frac{\partial J}{\partial \lambda} = \sum p_i - 1 = 0$$

$$\frac{\partial J}{\partial p_i} = -\frac{1}{\ln(2)} - \lg(p_i) + \lambda = 0$$

$$\Leftrightarrow \lg(p_i) = \lambda - \frac{1}{\ln(2)} \Leftrightarrow p_i = 2^{\lambda - \frac{1}{\ln(2)}}$$

What next?



Calculus
exercise

$$\ln(x)' = \frac{1}{x}$$

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)} \right)' = \frac{1}{x \cdot \ln(2)}$$

$$(x \cdot \lg(x))' = \frac{1}{\ln(2)} + \lg(x)$$

Maximum Entropy distribution for nonbinary discrete RV

Entropy of a discrete distribution with n outcomes $H(\mathbf{p}) = - \sum_{i=1}^n p_i \cdot \lg(p_i)$

Which choice of $\mathbf{p} = (p_1, \dots, p_n)$ maximizes the entropy?

Can be solved with constrained optimization:

$$\max_{\mathbf{p}} [H(\mathbf{p})] \quad \text{s.t.} \quad \sum_{i=1}^n p_i = 1$$

Form the Lagrangian: ~~$\partial p_1 \cdot \lg p_1 + p_2 \cdot \lg p_2 + \dots$~~

$$J(\mathbf{p}, \lambda) = - \sum_{i=1}^n p_i \cdot \lg(p_i) + \lambda \left(\sum_{i=1}^n p_i - 1 \right)$$

$$\frac{\partial J}{\partial p_i} = -\frac{1}{\ln(2)} - \lg(p_i) + \lambda = 0$$

$$\Leftrightarrow \lg(p_i) = \lambda - \frac{1}{\ln(2)} \Leftrightarrow p_i = 2^{\lambda - \frac{1}{\ln(2)}} =: C$$

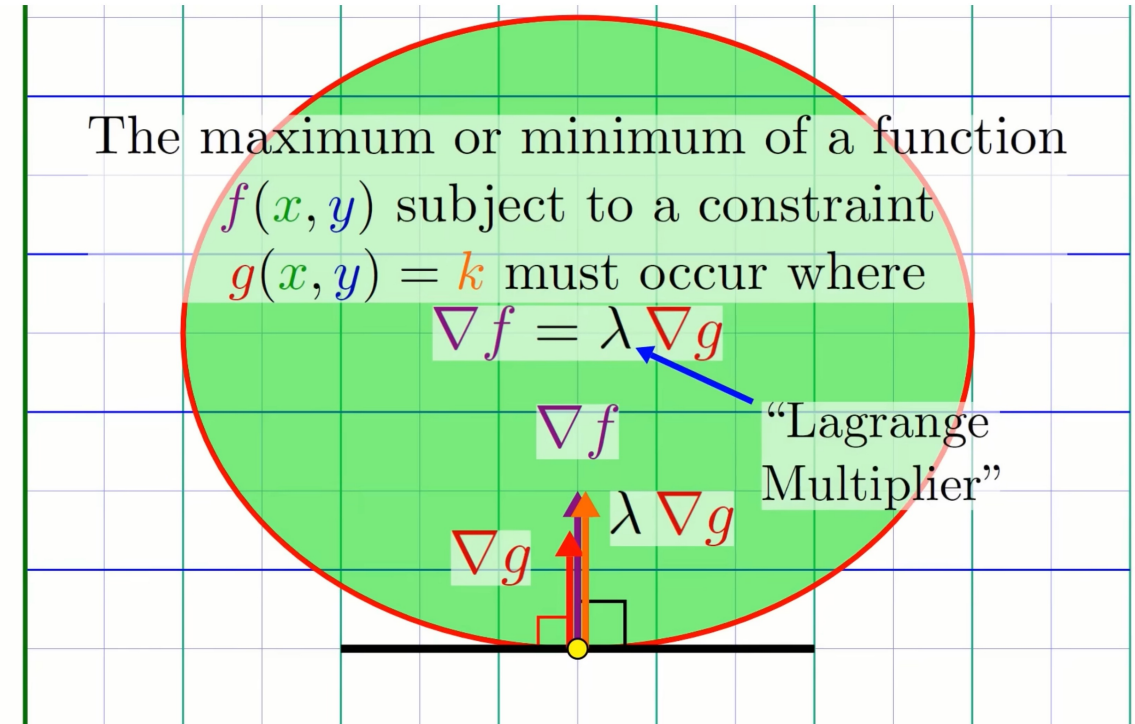
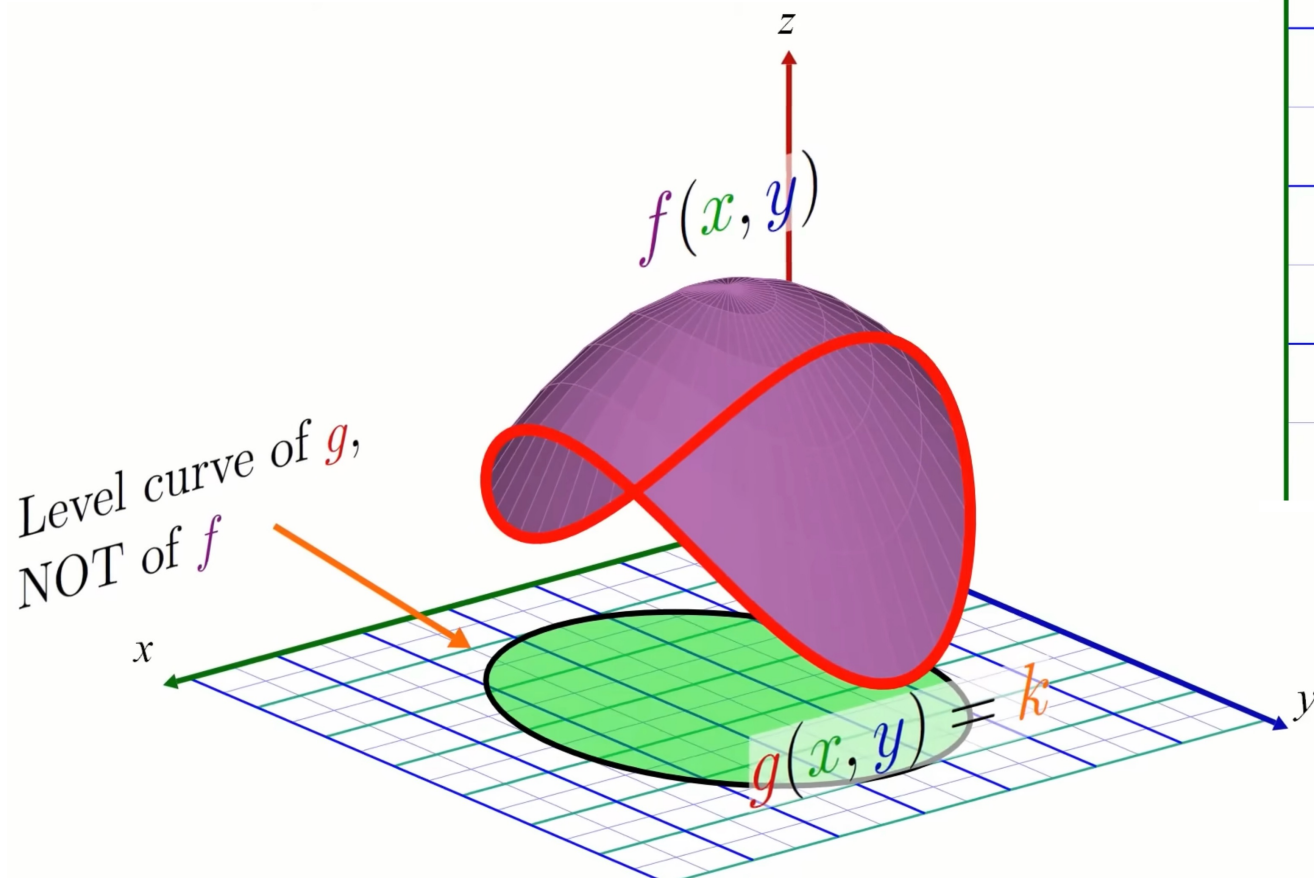
we are done ☺, all p_i are identical!

$$\sum_{i=1}^n p_i = 1 \Leftrightarrow \sum_{i=1}^n C = 1 \Leftrightarrow C = \frac{1}{n}$$

Calculus
exercise

$$\begin{aligned} \ln(x)' &= \frac{1}{x} \\ \lg(x)' &= \left(\frac{\ln(x)}{\ln(2)} \right)' = \frac{1}{x \cdot \ln(2)} \\ (x \cdot \lg(x))' &= \frac{1}{\ln(2)} + \lg(x) \end{aligned}$$

Nice video on Optimization w/ Lagrangian Multipliers



Properties of information (entropy) by example (continued)

Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . **How much information did we learn after each message?**

$$H(U_1) = ?$$

$$H(U_2) = ?$$

$$H(U_2|U_1) = ?$$

?

Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots,8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . How much information did we learn after each message?

$$\begin{array}{l} H(U_1) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit} \\ H(U_2) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit} \\ H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit} \end{array} \left. \vphantom{\begin{array}{l} H(U_1) \\ H(U_2) \\ H(U_2|U_1) \end{array}} \right\} \swarrow H(U_2|U_1) \overset{?}{=} H(U_2) = 1$$

Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots,8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . How much information did we learn after each message?

$$H(U_1) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit}$$

messages are independent

$$H(U_2|U_1) = H(U_2) = 1$$

1	2
3	4
5	6
7	8

*How do the messages
reduce the possible
outcomes?*



Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots,8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . How much information did we learn after each message?

$$H(U_1) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit}$$

messages are independent

$$H(U_2|U_1) = H(U_2) = 1$$

U_1	
	$\rightarrow$
1	2
3	4
5	6
7	8

Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots,8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . How much information did we learn after each message?

$$H(U_1) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit}$$

messages are independent

$$H(U_2|U_1) = H(U_2) = 1$$

the events are independent

$$p(U_2|U_1) = \underbrace{p(U_2)} = \frac{1}{2}$$

probability of the event $X \leq 4$

$\xrightarrow{U_1}$	
1	2
3	4
5	6
7	8
$\uparrow U_2$	

Learning partial information



EXAMPLE 4: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots,8\}$.

We get two messages: U_1 that the outcome of a roll is even, U_2 that the outcome of the same roll is ≤ 4 . How much information did we learn after each message?

$$H(U_1) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

$$H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit}$$

messages are independent

$$H(U_2|U_1) = H(U_2) = 1$$

the events are independent

$$p(U_2|U_1) = \underbrace{p(U_2)}_{\text{probability of the event } X \leq 4} = \frac{1}{2}$$

U_1		U_2
1	2	
3	4	
5	6	
7	8	

$$P(X_1, X_2) = P(X_1) \cdot P(X_2)$$

$$H(\{U_1, U_2\}) = H(U_1) + H(U_2|U_1)$$

$$= H(U_1) + H(U_2)$$

U_1 and U_2 are independent

{ 1, 2, 3, 4, 5, 6, 7, 8 }

000 001 010 011 100 101 110 111

We learned 2 bits independently

0?1

Joint entropy, Conditional entropy, Mutual information

Entropy

Given a discrete RV X with probability mass function (PMF) $p(x) = \mathbb{P}[X = x]$, for $x \in \mathcal{X}$. Entropy is defined as:

$$H(X) = \sum_x p(x) \cdot \lg\left(\frac{1}{p(x)}\right) = \mathbb{E}\left[\lg\left(\frac{1}{p(X)}\right)\right]$$

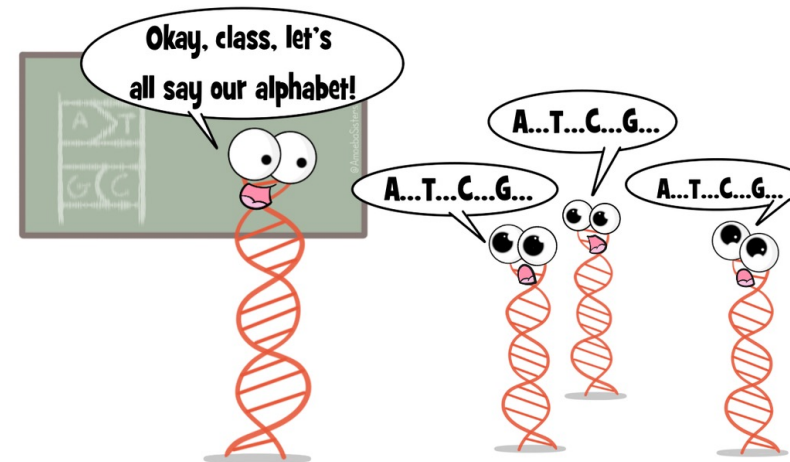
Alternative notation: $p(X) = p_X(x)$. Also: $\mathbb{E}_p[\dots]$ or $\mathbb{E}_X[\dots]$ or $\mathbb{E}_{X \sim p}[\dots]$ for the expected value operator w.r.t. the distribution p

Entropy is **label-invariant**, meaning that it depends only on the probability distribution and not on the actual values that the random variable X can take.

$$\mathcal{X} = \{1, 2, 3, 4\}$$



$$\mathcal{X} = \{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}$$



Joint Entropy

treat (X, Y) just like a single vector-valued RV $Z = \langle X, Y \rangle$

Given two RVs X and Y with PMF $p(X, Y)$, their **joint entropy** is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

Other notation: $p(X, Y) = p_{X,Y}(x, y)$.
Also: $\mathbb{E}_{X,Y \sim p}[\dots]$ or $\mathbb{E}_{X,Y \sim p}[\dots]$ or $\mathbb{E}_p[\dots]$

If X and Y are independent:

$$H(X, Y) = H(X) + H(Y)$$

How can we prove that? 

Joint Entropy

Given two RVs X and Y with PMF $p(X, Y)$, their **joint entropy** is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

Other notation: $p(X, Y) = p_{X,Y}(x, y)$.
Also: $\mathbb{E}_{X,Y \sim p}[\dots]$ or $\mathbb{E}_{X,Y \sim p}[\dots]$ or $\mathbb{E}_p[\dots]$

If X and Y are independent:

$$H(X, Y) = H(X) + H(Y)$$

$$\begin{aligned} H(X, Y) &= \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] \\ &= \mathbb{E} \left[\lg \left(\frac{1}{p(X) \cdot p(Y)} \right) \right] \\ &= \mathbb{E} \left[\lg \left(\frac{1}{p(X)} \right) + \lg \left(\frac{1}{p(Y)} \right) \right] \\ &= \mathbb{E} \left[\lg \left(\frac{1}{p(X)} \right) \right] + \mathbb{E} \left[\lg \left(\frac{1}{p(Y)} \right) \right] \\ &= H(X) + H(Y) \end{aligned}$$

Conditional Entropy, Chain rule of Entropy

Given two RVs X and Y with PMF $p(X, Y)$, their joint entropy is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

If X and Y are **not independent**:

$$H(X, Y) = \cancel{H(X) + H(Y)}$$

What do we
need to do? **?**

If X and Y are **not independent**, observing X might contain already some information about Y , so simply adding the information from each would **overcount**.

Conditional Entropy, Chain rule of Entropy

Given two RVs X and Y with PMF $p(X, Y)$, their joint entropy is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

If X and Y are not independent:

$$H(X, Y) = H(X) + H(Y|X)$$

Conditional entropy $H(Y|X)$: the expected amount of information needed to describe the outcome of RV Y given that the value of another RV X is known

$$H(Y|X) =$$

?

Conditional Entropy, Chain rule of Entropy

Given two RVs X and Y with PMF $p(X, Y)$, their joint entropy is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

If X and Y are not independent:

$$H(X, Y) = H(X) + H(Y|X)$$

$$H(X, Y) =$$

?

Conditional entropy $H(Y|X)$: the expected amount of information needed to describe the outcome of RV Y given that the value of another RV X is known

$$H(Y|X) = \sum_x p(x) \cdot H(Y|X = x)$$

$$= \mathbb{E}_{p(x)} [H(Y|X = x)]$$

DEFINITION of
conditional entropy

$$H(Y|X) = \sum_{x,y} p(x, y) \cdot \lg \left(\frac{1}{p(y|x)} \right)$$

Conditional Entropy, Chain rule of Entropy

Given two RVs X and Y with PMF $p(X, Y)$, their joint entropy is:

$$H(X, Y) = \mathbb{E} \left[\lg \left(\frac{1}{p(X, Y)} \right) \right] = \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right)$$

If X and Y are not independent:

$$H(X, Y) = H(X) + H(Y|X)$$

Conditional entropy $H(Y|X)$: the expected amount of information needed to describe the outcome of RV Y given that the value of another RV X is known

$$H(Y|X) = \sum_x p(x) \cdot H(Y|X = x) = \mathbb{E}_{p(x)} [H(Y|X = x)]$$

$$\begin{aligned} H(X, Y) &= \sum_x \sum_y p(x, y) \cdot \lg \left(\frac{1}{p(x, y)} \right) \\ &= \sum_x \sum_y p(x) \cdot p(y|x) \cdot \lg \left(\frac{1}{p(x) \cdot p(y|x)} \right) \\ &= \sum_x \sum_y p(x) \cdot p(y|x) \cdot \lg \left(\frac{1}{p(x)} \right) + \sum_x \sum_y p(x) \cdot p(y|x) \cdot \lg \left(\frac{1}{p(y|x)} \right) \\ &= \underbrace{\sum_x p(x) \cdot \lg \left(\frac{1}{p(x)} \right)}_{H(X)} \cdot \underbrace{\sum_y p(y|x)}_1 + \sum_x p(x) \cdot \underbrace{\sum_y p(y|x) \cdot \lg \left(\frac{1}{p(y|x)} \right)}_{H(Y|X = x)} \end{aligned}$$

$$\text{DEFINITION of conditional entropy} \quad H(Y|X) = \sum_{x,y} p(x, y) \cdot \lg \left(\frac{1}{p(y|x)} \right)$$

Chain rule for entropy

$$H(X, Y, Z) = H(X) + H(Y|X) + H(Z|X, Y)$$

... obvious generalization to
a chain of (not necessarily
independent) **observations**

If X and Y are not independent:

$$H(X, Y) = H(X) + H(Y|X)$$

conditional entropy

*Also notice the similarity to our earlier
probability factorizations, and Bayes' law:*

$$\begin{aligned}\mathbb{P}(A, B) &= \mathbb{P}(B|A) \cdot \mathbb{P}(A) \\ &= \mathbb{P}(A|B) \cdot \mathbb{P}(B)\end{aligned}$$

Learning partial information

EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, **how much information did we learn?**

?

Learning partial information

EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, how much information did we learn?

- We still learn $3-2=1$ bit about the roll of the die X .
 - We also learn 1 bit about the election outcome.
- } We learned 2 bits (U contains 2 bits)

How do these numbers add up?

?

Learning partial information

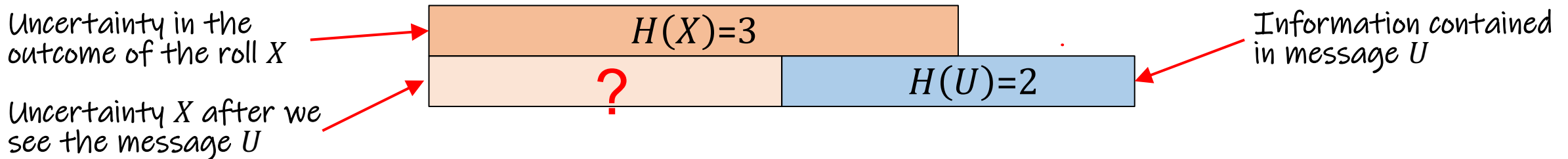
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, how much information did we learn?

- We still learn $3-2=1$ bit about the roll of the die X .
 - We also learn 1 bit about the election outcome.
- } We learned 2 bits (U contains 2 bits)

How do these numbers add up?



Learning partial information

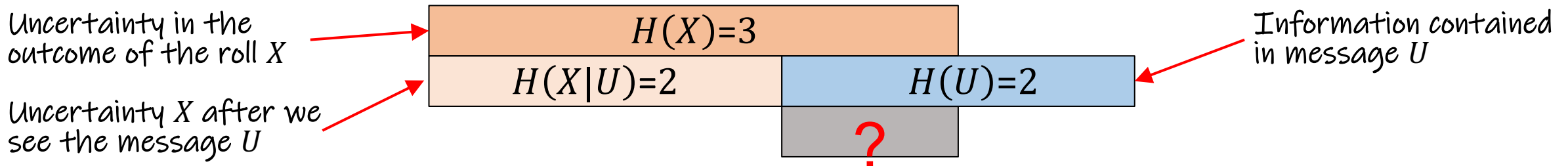
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, how much information did we learn?

- We still learn $3-2=1$ bit about the roll of the die X .
 - We also learn 1 bit about the election outcome.
- } We learned 2 bits (U contains 2 bits)

How do these numbers add up?



Learning partial information

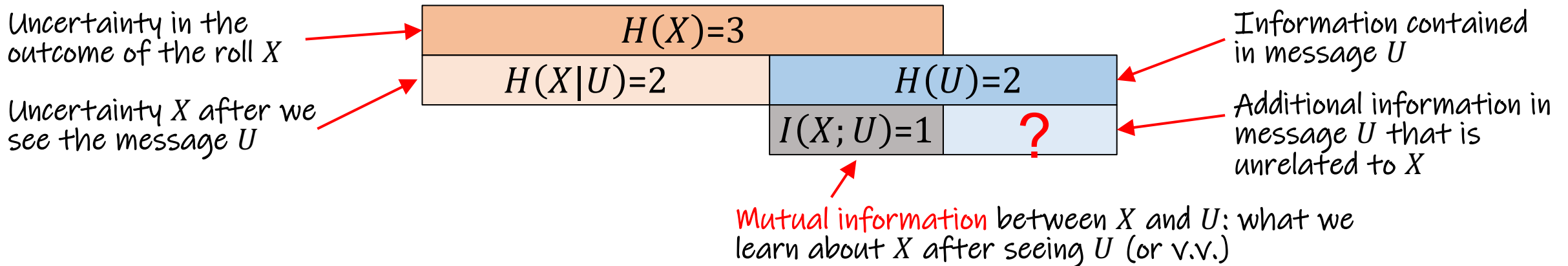
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, how much information did we learn?

- We still learn $3-2=1$ bit about the roll of the die X .
 - We also learn 1 bit about the election outcome.
- } We learned 2 bits (U contains 2 bits)

How do these numbers add up?



Learning partial information

EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}=\{1,2, \dots, 8\}$.



We then get a message U : "The outcome of the roll is even, and by the way, the next president of the US will be ...". Assuming two equally likely outcomes for the election, how much information did we learn?

- We still learn $3-2=1$ bit about the roll of the die X .
 - We also learn 1 bit about the election outcome.
- } We learned 2 bits (U contains 2 bits)

How do these numbers add up?

