Updated 9/16/2025

Part 1: Theory
L03: Basics of entropy (1/6)
[measures of information, intuition behind entropy]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

https://northeastern-datalab.github.io/cs7840/fa25/

9/15/2025

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. "The name of the next president of the USA 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. "The name of the next president of the USA 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 $U_1 =$ "**C**"

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

- "Side **C** comes up."
- message $U_2 =$ "**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 $U_1 =$ "**C**"
- There are 4 possible outcomes, each has a probability of ¼.

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

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

⇒ 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**}.

A	**B**	**C**	**D**
00	01	10	11

- "Side **C** comes up."
- message U_1 = "**C**", or in above binary encoding U_1 = "**10**"
- There are 4 possible outcomes, each has a probability of ¼.

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

A	**B**	**C**	**D**	**E**	**F**
000	001	010	011	100	101

- "Side **C** comes up."
- message U_2 = "**C**", or in above binary encoding U_2 = "**010**"
- There are 6 possible outcomes, each has a probability of 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 = "**C**"
- There are 4 possible outcomes, each has a probability of ¼.

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

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message U_3 = "**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 = "**C**"
- There are 4 possible outcomes, each has a probability of ¼.

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

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message U_3 = "**CBD**".
- Now we had $64 = 4 \cdot \underbrace{4 \cdot 4}_{\text{16 times more!}} = 4^3$ possible outcomes.

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 $U_1 =$ "**C**"
- There are 4 possible outcomes, each has a probability of ¼.

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

- "The sequence of sides are: (**C**, **B**, **D**)"
- The message $U_3 =$ "**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. $\qquad \log_4(4) = 1$

3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes. $\qquad \log_4(64) = 3$

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

$$\boxed{H_0(U) = \log_b(n)}$$

n = number of outcomes

Ralph Hartley. Transmission of information, The Bell System Technical Journal, 1928. https://doi.org/10.1002/j.1538-7305.1928.tb01236.x
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Hartley's measure of information [1928]

1 roll has 4 outcomes. $\qquad \log_4(4) = 1$

3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes. $\qquad \log_4(64) = 3$

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 $\boxed{H_0(U) = \log_b(n)}$

n = number of outcomes

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

$\boxed{\log_2(c) = 1.443 \cdot \log_e(c)}$ $\qquad 2^{1.443} = e \Leftrightarrow 1.443 = \log_2(e)$

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. $\log_4(4) = 1$

3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes. $\log_4(64) = 3$

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

$$\boxed{H_0(U) = \log_b(n)}$$

n = number of outcomes

For k independent trials, the amount of information is: $\log_b(n^k) = ?$

Ralph Hartley. Transmission of information, The Bell System Technical Journal, 1928. https://doi.org/10.1002/j.1538-7305.1928.tb01236.x
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Hartley's measure of information [1928]

1 roll has 4 outcomes. $\quad \log_4(4) = 1$

3 rolls have $64 = 4 \cdot 4 \cdot 4 = 4^3$ outcomes. $\quad \log_4(64) = 3$

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

$$\boxed{H_0(U) = \log_b(n)}$$

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 ☺

Ralph Hartley. Transmission of information, The Bell System Technical Journal, 1928. https://doi.org/10.1002/j.1538-7305.1928.tb01236.x
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

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: $2^{32} = 4{,}294, \ldots, \ldots$

?

Sources: https://www.worldometers.info/world-population/

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: $2^{32} = 4{,}294, \ldots, \ldots$

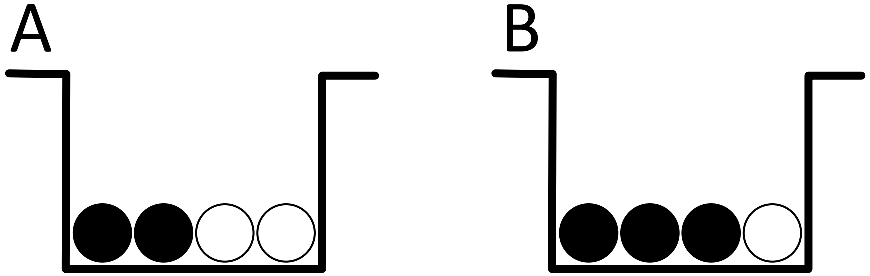
$$\log_2(8{,}174{,}891{,}806) \approx 32.93$$

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

Sources: https://www.worldometers.info/world-population/

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)

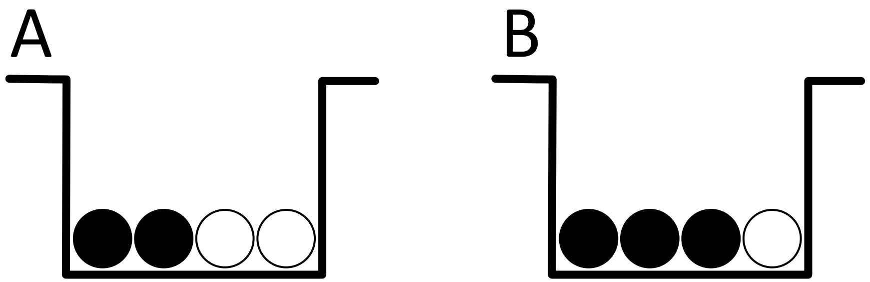
?

Numeric example from: Moser. Information Theory (lecture Notes, 6th ed). https://moser-isi.ethz.ch/cgi-bin/request_script.cgi?script=it

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

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.

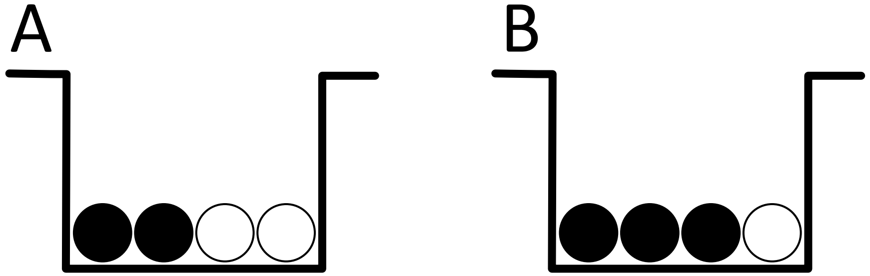
$H_0(U_A) = \lg(2) = 1$ bit (we have 2 equally likely colors)

$H_0(U_B) =$?

Numeric example from: Moser. Information Theory (lecture Notes, 6th ed). https://moser-isi.ethz.ch/cgi-bin/request_script.cgi?script=it
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

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.

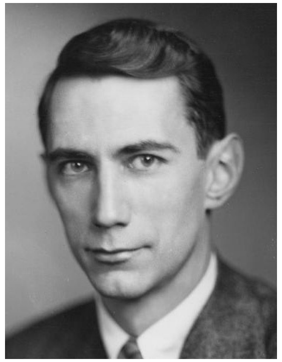


We randomly draw a ball from both hats. Let U_A, U_B be the color of the ball.

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

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

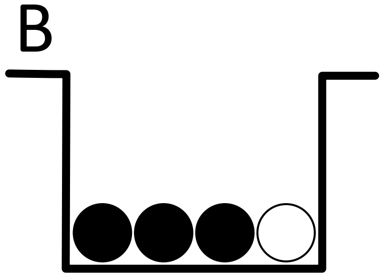
⇒ 3) A proper measure of information should take into account the (possibly different) probabilities of the various outcomes.



This was the key insight of Claude Shannon [1948]

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_{B}:

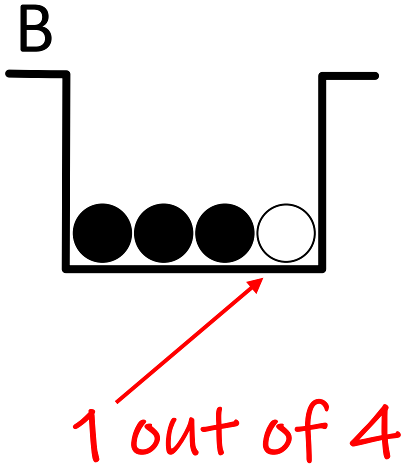


$\underline{U_B = \text{white}}$:

What does Hartley tell us about the information we get after learning U_B=white ?

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_{B}:



$\underline{U_B = \text{white}}$:

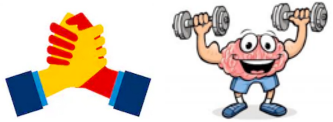
There is a $p = ¼$ chance to draw a white ball.

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

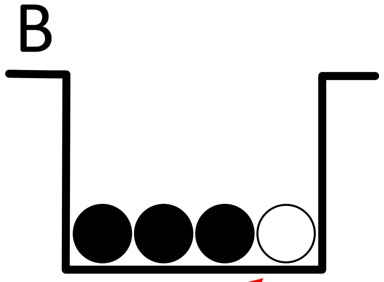
$$H_0(U_{\mathrm{B}}) = ???$$

?

"Fixing" Hartley's information measure



Let's analyze the possible outcomes for U_{B}:



1 out of 4

$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

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

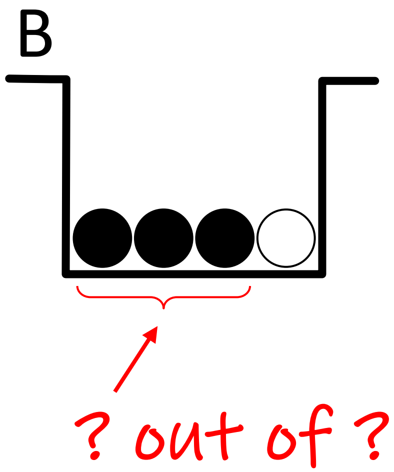
$\underline{U_B = \text{black}}$:

Hartley does not work directly.
What can we do?

?

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_{B}:



$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

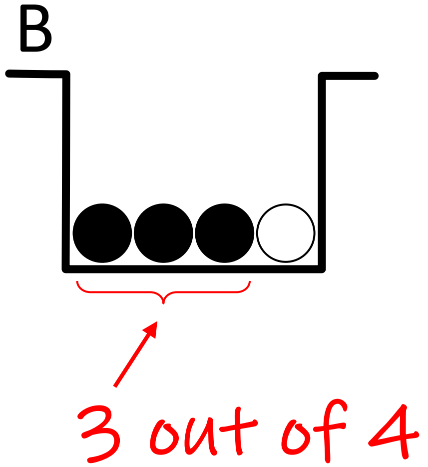
$$\lg\left(\frac{1}{p}\right)$$

$\underline{U_B = \text{black}}$:

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

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_B:



$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_\mathrm{B}) = \lg(4) = 2 \text{ bits}$$

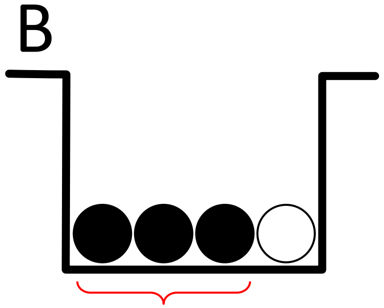
$\underline{U_B = \text{black}}$: $\quad \lg\left(\frac{1}{p}\right)$

There is a $p = ¾$ chance to draw a black ball.

What do we do with the ¾ ? ?

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_{B}:



3 out of 4
= 1 out of 4/3

$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

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

There is a $p = ¾$ chance to draw a black ball.

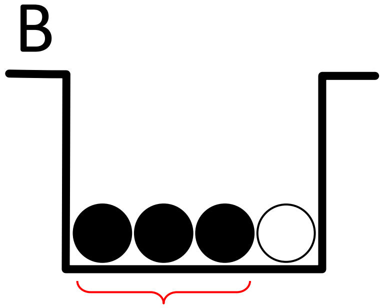
That's the result of 1 out of $n = 4/3$ possible outcomes.

$$H_0(U_{\mathrm{B}}) = \quad ?$$

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:



3 out of 4
= 1 out of 4/3

$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_\mathrm{B}) = \lg(4) = 2 \text{ bits}$$

$\underline{U_B = \text{black}}$:

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

#total balls / #black balls

There is a $p = ¾$ chance to draw a black ball.

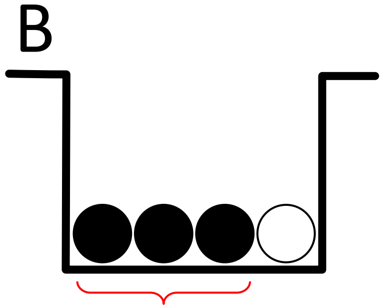
That's the result of 1 out of $n = 4/3$ possible outcomes.

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

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}:



3 out of 4
= 1 out of 4/3

$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

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

There is a $p = ¾$ chance to draw a black ball.

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

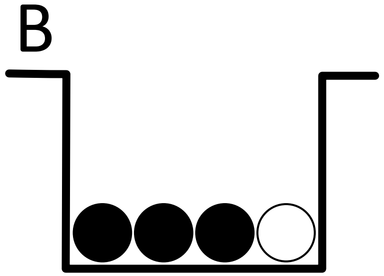
$$H_0(U_{\mathrm{B}}) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

Let's do "in expectation" ☺

$$\mathbb{E}[H_0(U_{\mathrm{B}})] = \frac{1}{4} \cdot \quad \ldots \quad + \frac{3}{4} \cdot \quad \ldots$$

"Fixing" Hartley's information measure

Let's analyze the possible outcomes for U_{B}:



$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

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

There is a $p = ¾$ chance to draw a black ball.

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

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

That's our expected amount of information we learn.

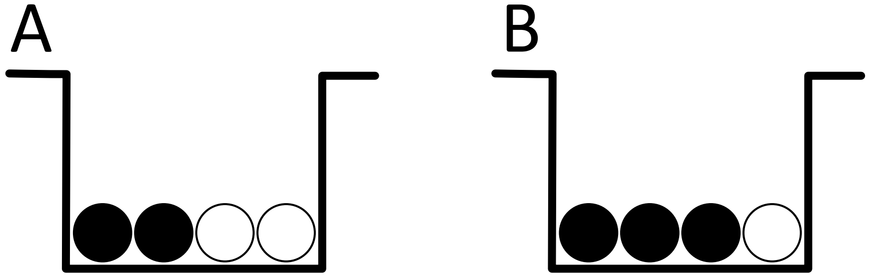
<u>Let's do "in expectation":</u>

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

Numeric example from: Moser. Information Theory (lecture Notes, 6th ed). https://moser-isi.ethz.ch/cgi-bin/request_script.cgi?script=it
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

"Fixing" Hartley's information measure

Let's analyze the possible outcomes:



$\underline{U_B = \text{white}}$:

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

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

$$H_0(U_\mathrm{B}) = \lg(4) = 2 \text{ bits}$$

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

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.

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

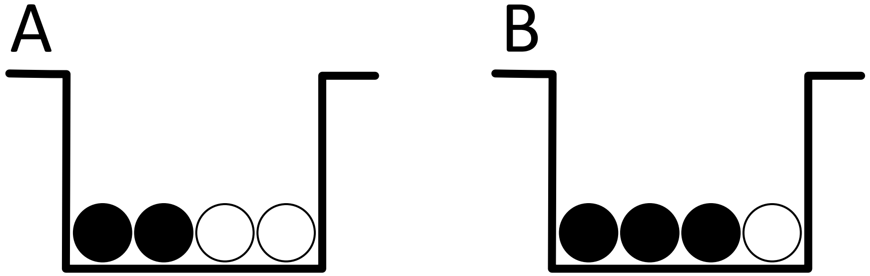
What would we get for hat A instead of hat B ?

<u>Let's do "in expectation":</u>

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

"Fixing" Hartley's information measure

Let's analyze the possible outcomes:



$\underline{U_B = \text{white}}$:

There is a $p = ¼$ chance to draw a white ball.

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

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

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

There is a $p = ¾$ chance to draw a black ball.

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

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

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

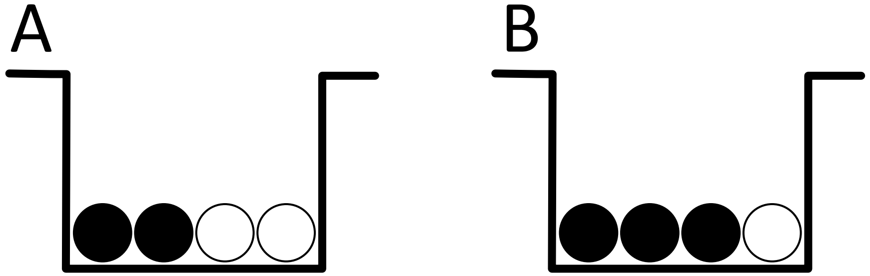
<u>Let's do "in expectation":</u>

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

1 bit for hat A

0.811 bits hat B

"Fixing" Hartley's information measure



Let's analyze the possible outcomes:

<u>U_B = white</u>:

There is a $p = ¼$ chance to draw a white ball.

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

$$H_0(U_\mathrm{B}) = \lg(4) = 2 \text{ bits}$$

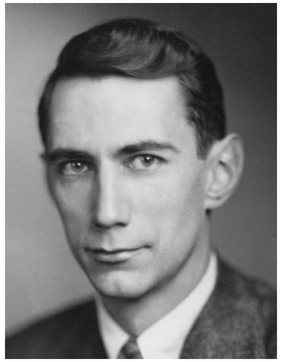
<u>U_B = black</u>: $\lg\left(\frac{1}{p}\right)$

There is a $p = ¾$ chance to draw a black ball.

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

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

This is Claude Shannon's measure of information



<u>Let's do "in expectation"</u>:

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

1 bit for hat A

= 0.811 bits hat B

Numeric example from: Moser. Information Theory (lecture Notes, 6th ed). https://moser-isi.ethz.ch/cgi-bin/request_script.cgi?script=it

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

Shannon's entropy

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

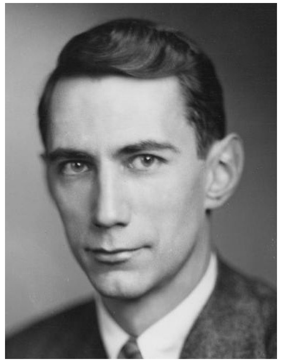
$$H(\boldsymbol{p}) = \boxed{\sum_{i=1}^{r} p_i \cdot \underbrace{\lg\left(\frac{1}{p_i}\right)}_{H_0(U)}} = -\sum_{i=1}^{r} p_i \cdot \lg(p_i) = \mathbb{E}\left[\lg\left(\frac{1}{p_i}\right)\right]$$

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)

$$H(\boldsymbol{p}) = \underbrace{\boxed{\sum_{i=1}^{r} p_i \cdot \lg\left(\frac{1}{p_i}\right)}}_{H_0(U)} = -\sum_{i=1}^{r} p_i \cdot \lg(p_i) = \mathbb{E}\left[\lg\left(\frac{1}{p_i}\right)\right]$$

p_i = probability of the i-th possible outcome

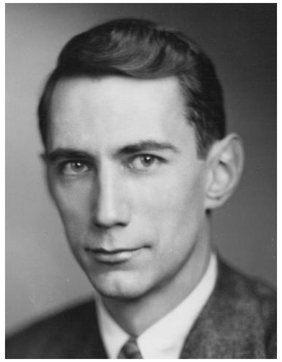
Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

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/

Shannon's entropy

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

$$H(\boldsymbol{p}) = \underbrace{\boxed{\sum_{i=1}^{r} p_i \cdot \lg\left(\frac{1}{p_i}\right)}}_{H_0(U)} = -\sum_{i=1}^{r} p_i \cdot \lg(p_i) = \mathbb{E}\left[\lg\left(\frac{1}{p_i}\right)\right]$$

p_i = probability of the i-th possible outcome

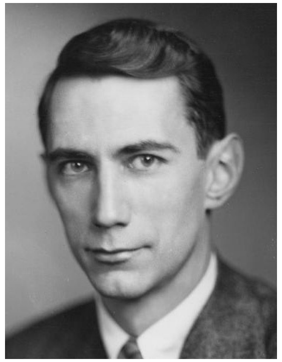
Uncertainty: Normalized number of outcomes, for option i to be "1 out of ... outcomes"

1) The number of possible outcomes should be linked to "information"
2) Information is additive in some sense

$\Big\}\ H_0$

3) A proper measure of information should take into account the different probabilities of the outcomes.

$\Big\}\ H$ (items 1–3)



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/