Updated 9/16/2025

Part 1: Theory
L03: Basics of entropy (1/7)
[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 ... ?

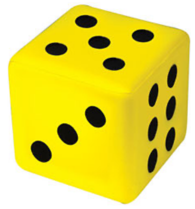
Let's try to quantify "information"

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 ?

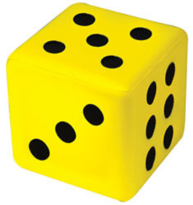
Let's try to quantify "information"

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**}.

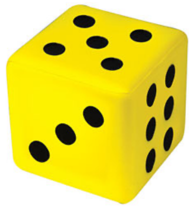
00 01 10 11

- "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**}.

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.

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

$$\left.\begin{array}{l}\\ \\ \end{array}\right\} \quad \begin{array}{l} \log_4(4) = 1 \\ \log_4(64) = 3 \end{array}$$

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)

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

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

?

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

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?

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

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

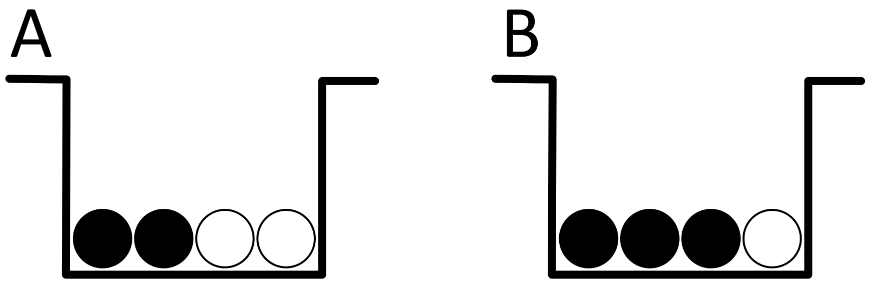
$$\log_2(8{,}174{,}891{,}806) \approx 32.93$$

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.



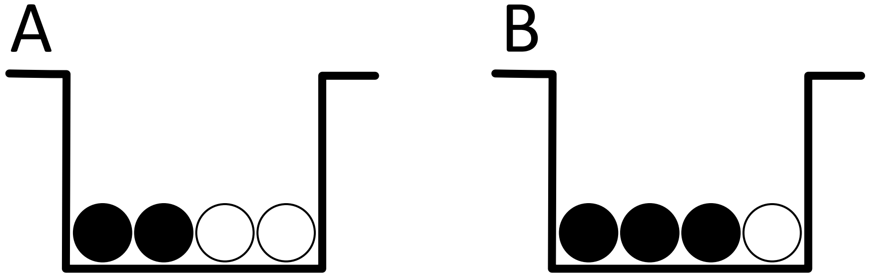
We randomly draw a ball from both hats. Let U_A, U_B be the color of the ball.

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

?

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.



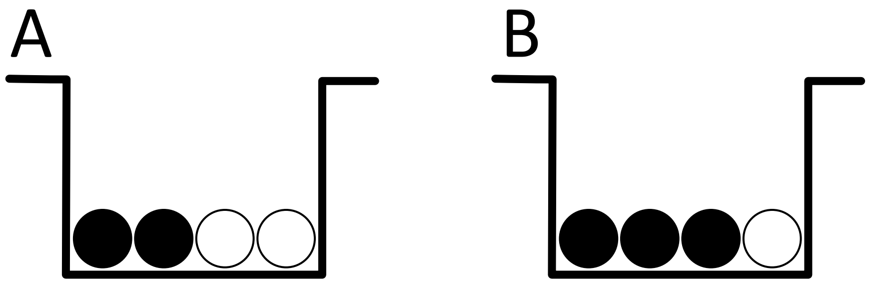
We randomly draw a ball from both hats. Let U_A, U_B be the color of the ball.

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

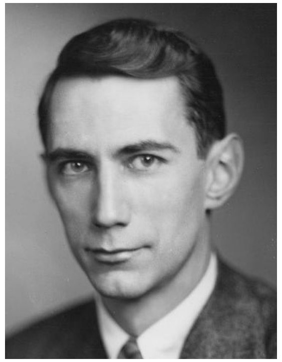


We randomly draw a ball from both hats. Let U_A, U_B be the color of the ball.

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

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

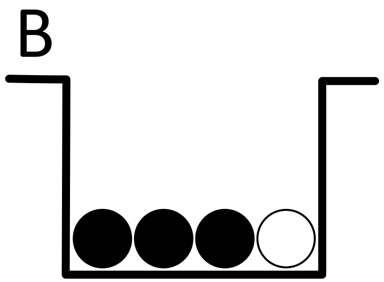
⇒ 3) A proper measure of information should take into account the (possibly different) probabilities of the various outcomes.



This was the key insight of Claude Shannon [1948]

"Fixing" Hartley's information measure

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



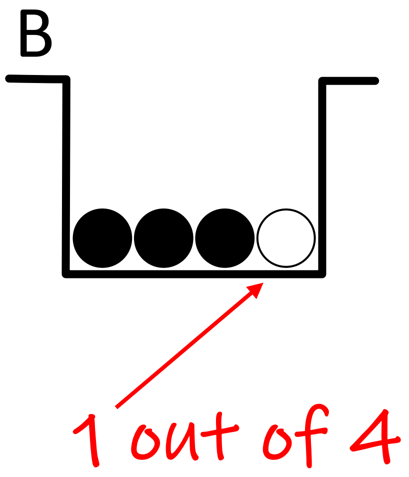
$\underline{U_B = \text{white}}$:

What does Hartley tell us about the information we get after learning $\boldsymbol{U}_{\mathrm{B}}$=white ?

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 for U_{B}:



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

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

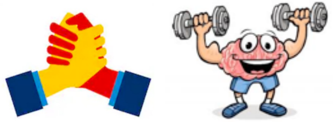
That's the result of 1 out of $n = 4$ possible outcomes.

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

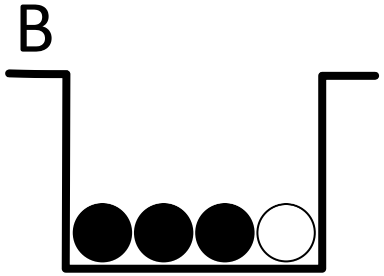
?

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 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.

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

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

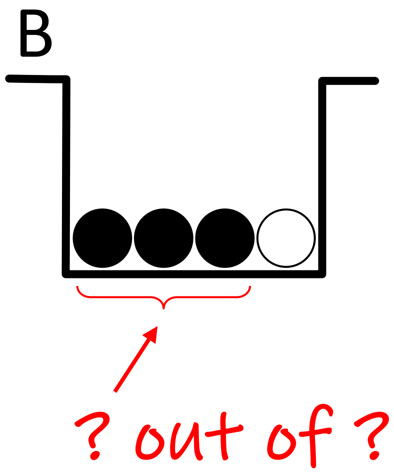
$\underline{U_B = \text{black}}$:

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.

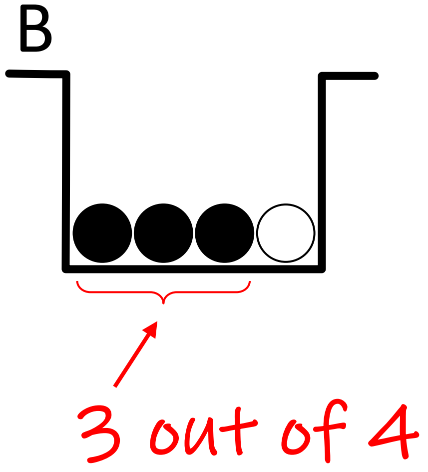
$$H_0(U_{\mathrm{B}}) = \lg(4) = 2 \text{ bits}$$

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

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

"Fixing" Hartley's information measure

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



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

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

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

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

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

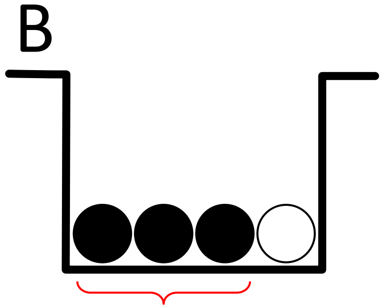
$\underline{U_B = \text{black}}$:

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 $p = \frac{1}{4}$ 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 $p = \frac{3}{4}$ chance to draw a black ball.

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

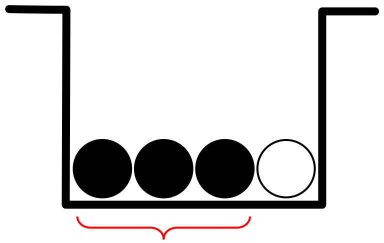
$$H_0(U_{\mathrm{B}}) = \quad ?$$

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:

B



3 out of 4
= 1 out of 4/3

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

#total balls / #black balls

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

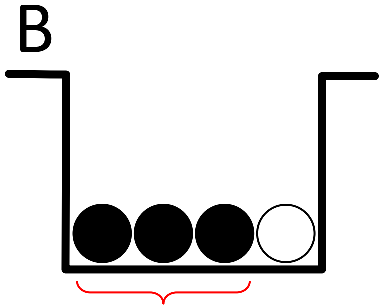
That's the result of 1 out of $n = 4/3$ possible outcomes.

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

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

U_B = white:

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

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

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

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

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

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

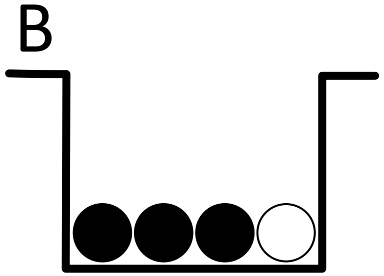
$$H_0(U_{\mathrm{B}}) = \lg\left(\frac{4}{3}\right) = 0.415 \text{ bits}$$

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.

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

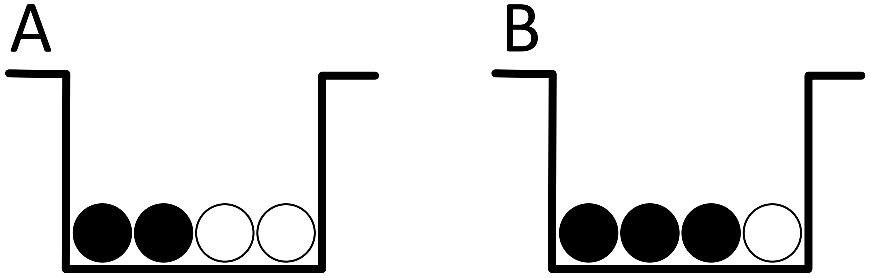
That's our expected amount of information we learn.

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

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

"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.

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

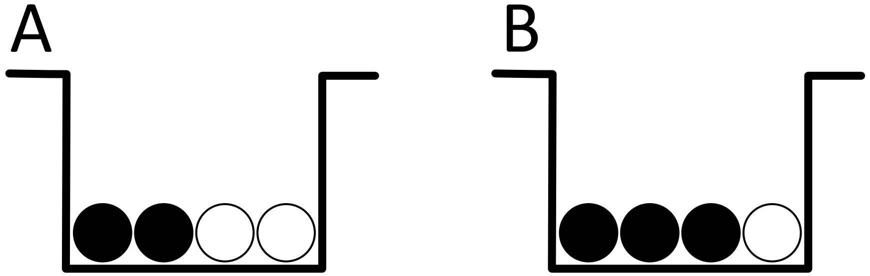
What would we get for hat A instead of hat B ?

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

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

"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.

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

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

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

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

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

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>

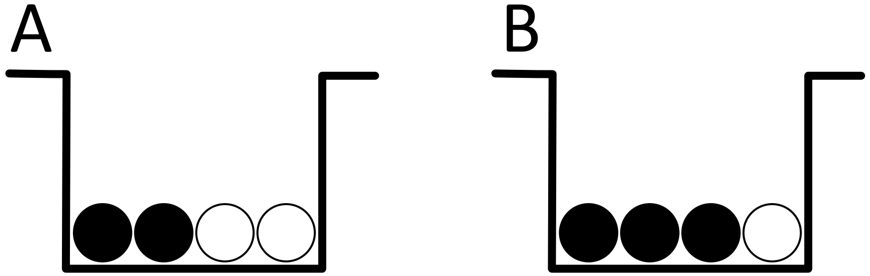
1 bit for hat A

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

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

"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.

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

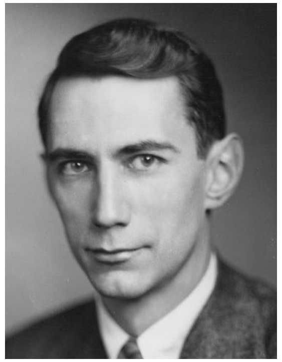
$\underline{U_B = \text{black}}$: $\lg\left(\frac{1}{p}\right)$

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

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

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

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

Shannon's entropy

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

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

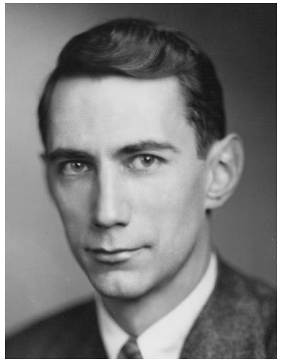
p_i = probability of the i-th possible outcome

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

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

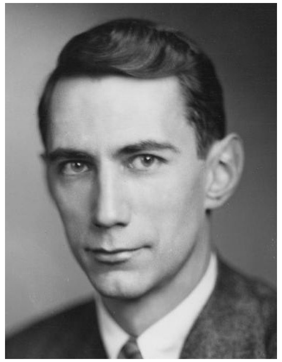
Uncertainty: Normalized number of outcomes, for option i to be "1 out of … outcomes"

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)

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

p_i = probability of the i-th possible outcome

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

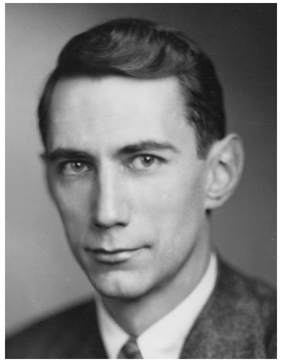
1) The number of possible outcomes should be linked to "information"

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/

Updated 9/19/2025

Part 1: Theory
L04: Basics of entropy (2/7)
[measures of information, intuition behind entropy]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

9/18/2025

Pre-class conversations

- Last class recapitulation (incl. grouping rule)
- To be posted: Online Python notebook (feedback *very* welcome, also possibly useful for your own scribes)
- Any feedback on organization on course website (Canvas, Piazza)?

- Today:
  - Keep pen & paper ready for hands-on calculus, logarithm
    - also see Schneider's "Information Theory Primer, With an Appendix on Logarithms"
  - 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

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?

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

Shannon entropy for unbiased outcomes

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

$$H(\boldsymbol{p}) = \boxed{\sum_{i=1}^{r} p_i \cdot \lg\left(\frac{1}{p_i}\right)} = \underbrace{\left(\sum_{i=1}^{r} p_i\right)}_{1} \cdot \lg\left(\frac{1}{p_i}\right) = \lg\left(\frac{1}{p_i}\right) \; ?$$

Shannon entropy for unbiased outcomes = Hartley measure

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

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

$$= \ ?$$

number of outcomes

Entropy is exactly the Hartley information measure for unbiased outcomes ☺

Shannon entropy for unbiased outcomes = Hartley measure

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

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

$$= \lg(8) = 3$$

number of outcomes

Entropy is exactly the Hartley information measure for unbiased outcomes ☺

Characterization of the Hartley information measure

Shannon entropy for uniform sampling from n choices.

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

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.

Screenshot source: https://en.wikipedia.org/wiki/Hartley_function . For a derivation, see Sect 2.2.2 of "Klir. Uncertainty and Information: Foundations of Generalized Information Theory", 2006. https://onlinelibrary.wiley.com/doi/book/10.1002/0471755575

Learning partial information

EXAMPLE 2: We roll an unbiased 8-sided die with sides {1,2, ...,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**, ...,**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**, ...,**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**}
- After the message: There are 4 choices: {**2**,**4**,**6**,**8**}

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**, ...,**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: { 1, **2**, 3, **4**, 5, **6**, 7, **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**, ...,**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

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

Before: { **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8** }

$\downarrow$ 000 001 010 011 100 101 110 111

After: { 1, **2**, 3, **4**, 5, **6**, 7, **8** }

000 001 010 011 100 101 110 111

"Grouping rule": Dividing the outcomes into two (last bit), 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**, ...,**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.

$\lg(2) = 1$

+1 bit

$\lg(4) = 2$

+1 bit

$\lg(8) = 3$

Learning partial information

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

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

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

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**, ...,**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"!

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

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

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides {**1**,**2**, ...,**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?

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

$H(U_2|U_1) =$?

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

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

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides {**1**,**2**, ...,**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?

?

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

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

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

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

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides {**1**,**2**, ...,**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?

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

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

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

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

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

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**, ...,**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?

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

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

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

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

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

$$\boxed{\begin{aligned} &H(\{U_1, U_2, U_3, U_4\}) \\ &\quad = H(U_1) + H(U_2|U_1) + H\left(U_3|U_{1,2}\right) + H(U_4|U_{1-3}) \end{aligned}}$$

This is called the "chain rule"

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides {**1**,**2**, ...,**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?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$
$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$
$$H\left(U_3|U_{1,2}\right) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$
$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$
$$H(\{U_1, U_2, U_3, U_4\}) = 1 \text{ bit}$$

$$H(\{U_1, U_2, U_3, U_4\})$$
$$= H(U_1) + H(U_2|U_1) + H\left(U_3|U_{1,2}\right) + H(U_4|U_{1-3})$$
$$= \lg\left(\frac{8}{7}\right) + \lg\left(\frac{7}{6}\right) + \lg\left(\frac{6}{5}\right) + \lg\left(\frac{5}{4}\right)$$
$$= \text{?}$$

Learning partial information

EXAMPLE 3: We roll an unbiased 8-sided die with sides {**1**,**2**, ...,**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?

$$H(U_1) = \lg\left(\frac{8}{7}\right) = 0.193 \text{ bits}$$
$$H(U_2|U_1) = \lg\left(\frac{7}{6}\right) = 0.222 \text{ bits}$$
$$H\left(U_3|U_{1,2}\right) = \lg\left(\frac{6}{5}\right) = 0.263 \text{ bits}$$
$$H(U_4|U_{1-3}) = \lg\left(\frac{5}{4}\right) = 0.322 \text{ bits}$$
$$H(\{U_1, U_2, U_3, U_4\}) = 1 \text{ bit}$$

$$H(\{U_1, U_2, U_3, U_4\})$$
$$= H(U_1) + H(U_2|U_1) + H\left(U_3|U_{1,2}\right) + H(U_4|U_{1-3})$$
$$= \lg\left(\frac{8}{7}\right) + \lg\left(\frac{7}{6}\right) + \lg\left(\frac{6}{5}\right) + \lg\left(\frac{5}{4}\right)$$
$$= \lg\left(\frac{8}{\cancel{7}} \cdot \frac{\cancel{7}}{\cancel{6}} \cdot \frac{\cancel{6}}{\cancel{5}} \cdot \frac{\cancel{5}}{4}\right) = \lg\left(\frac{8}{4}\right) = 1 \text{ bit}$$

Again, the logarithm ☺

Maximum Entropy distributions (with a bit of Calculus)

Binary Entropy Function

Biased coin flip: 

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

$H_{\mathrm{B}}(p) =$?

Binary Entropy Function

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

$$H_{\mathrm{B}}(p) = -p \cdot \lg(p) - (1-p) \cdot \lg(1-p)$$

How to choose p in order to maximize entropy ?

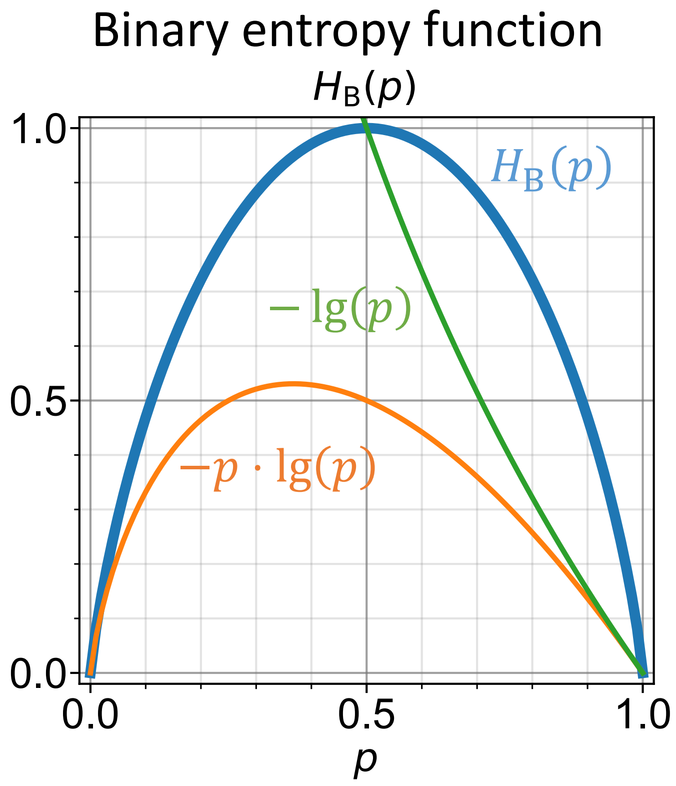


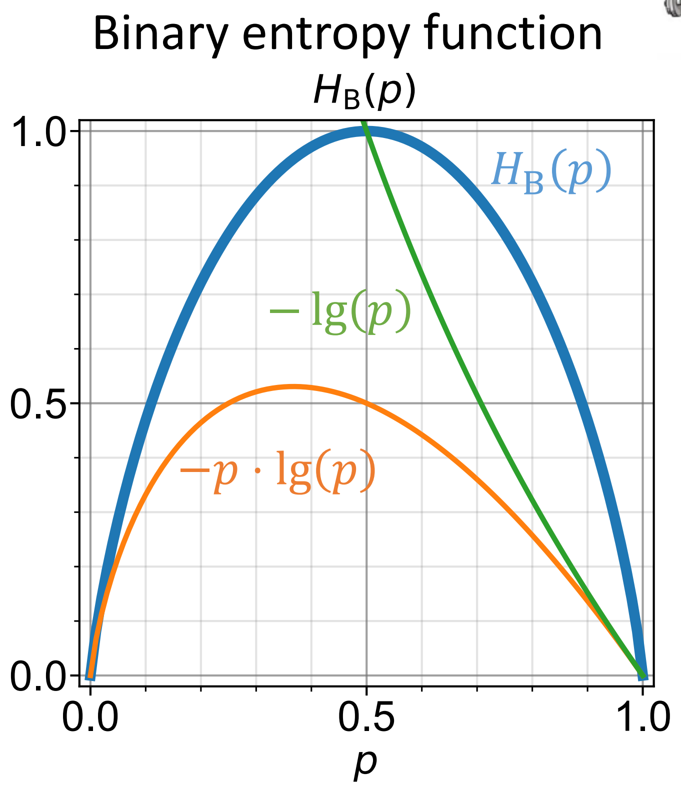
Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Binary Entropy Function

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

$$H_\mathrm{B}(p) = -p \cdot \lg(p) - (1-p) \cdot \lg(1-p)$$

How to choose p in order to maximize entropy?



$$\frac{dH}{dp} = ?$$

Understanding "change of basis"

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

How do you derive that?

Calculus cheat sheet

$\ln(x)' =$

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

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

$\lg(1-x)' =$

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

?

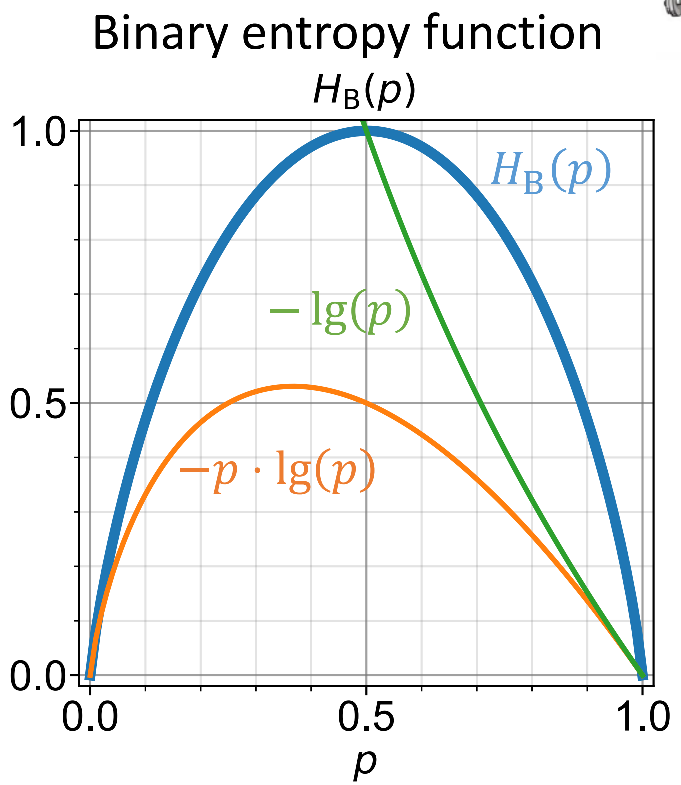
Binary Entropy Function

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

$$H_\mathrm{B}(p) = -p \cdot \lg(p) - (1-p) \cdot \lg(1-p)$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = ?$$



Understanding "change of basis"

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

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

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

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

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

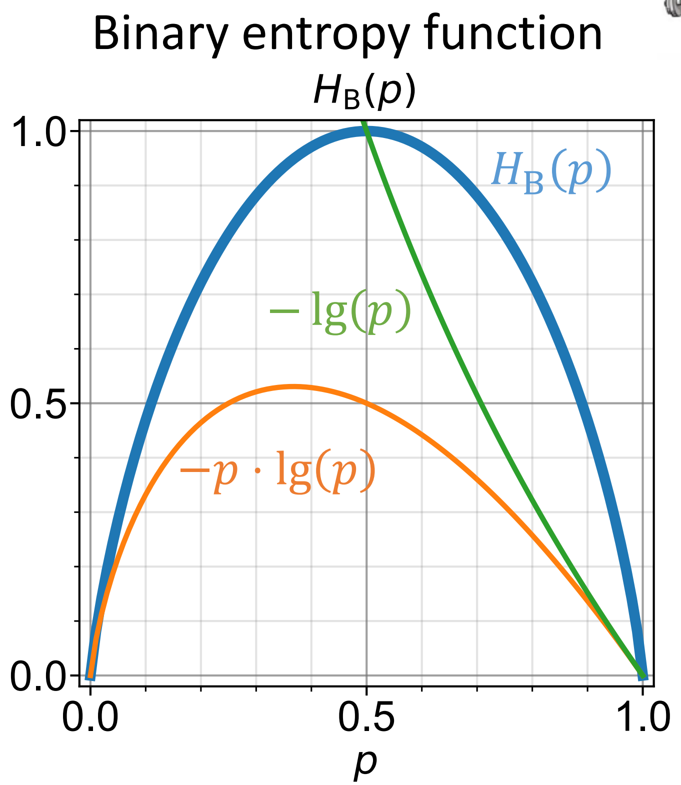
X is a Bernoulli RV with $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$

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

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = ?$$

$$\frac{1}{x \cdot \ln(2)}$$



Calculus cheat sheet

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

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

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

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

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

Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Binary Entropy Function

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

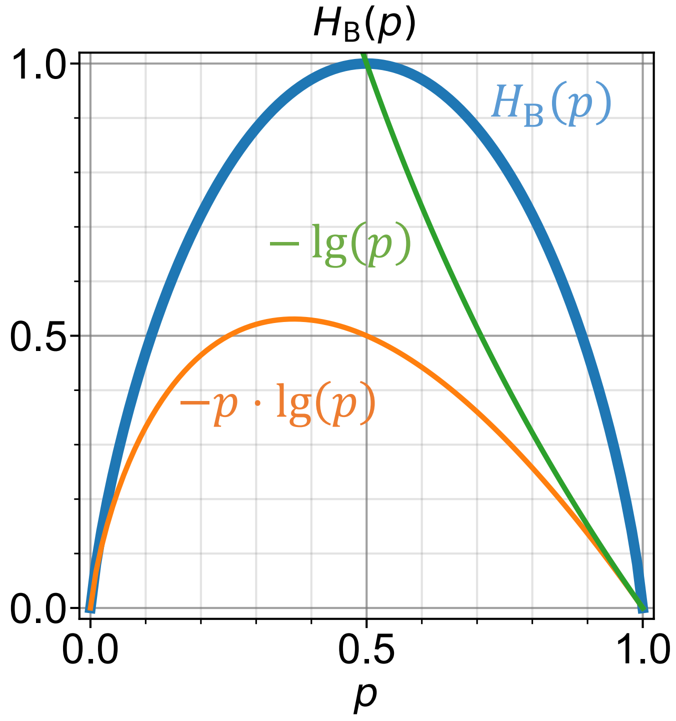
$$H_\mathrm{B}(p) = \underbrace{-p \cdot \lg(p)}\,\underbrace{-\,(1-p)\cdot \lg(1-p)}$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = \overbrace{\cancel{-\frac{1}{\ln(2)}} - \lg(p)}\,\overbrace{+\cancel{\frac{1}{\ln(2)}} + \lg(1-p)}$$

$\frac{1}{x\cdot\ln(2)}$

Binary entropy function



Calculus cheat sheet

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

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

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

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

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

Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Binary Entropy Function

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

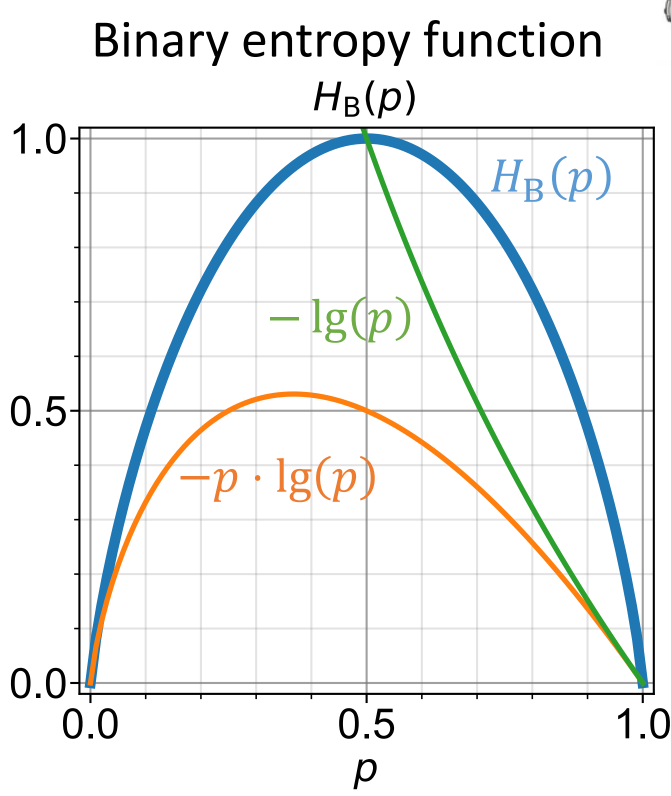
$$H_{\mathrm{B}}(p) = \underbrace{-p \cdot \lg(p)} \underbrace{-\,(1-p) \cdot \lg(1-p)}$$

How to choose p in order to maximize entropy?

$$\frac{dH}{dp} = \overbrace{-\cancel{\frac{1}{\ln(2)}} - \lg(p)} \overbrace{+\cancel{\frac{1}{\ln(2)}} + \lg(1-p)} = 0$$

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

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



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 $p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \end{cases}$

$$H_{\mathrm{B}}(p) = \underbrace{-p \cdot \lg(p)}\ \underbrace{-\ (1-p) \cdot \lg(1-p)}$$

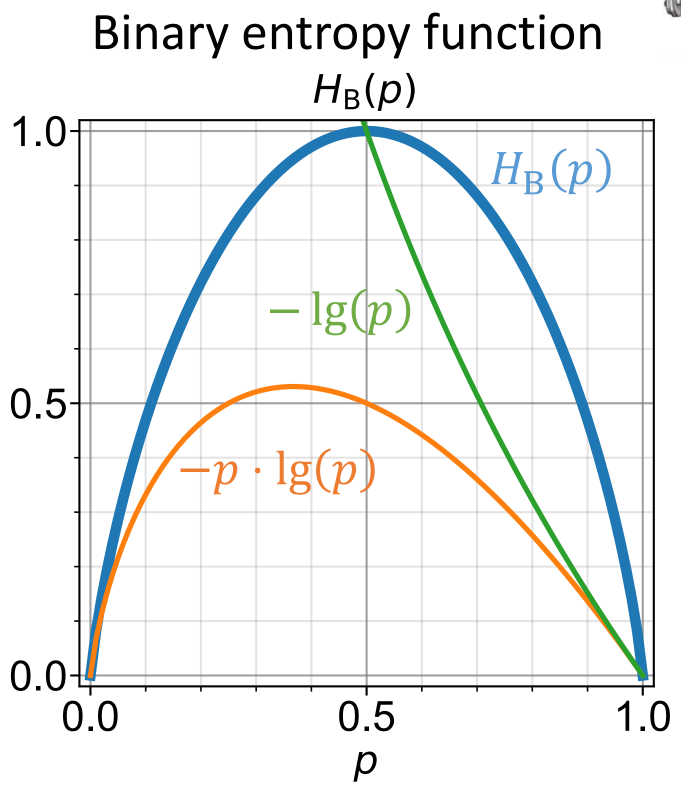
How to choose p in order to maximize entropy?

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

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

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

$$\frac{1}{x \cdot \ln(2)}$$



Calculus cheat sheet

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

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

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

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

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

Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

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:

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

...

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!

Interaction with https://chatgpt.com/ (ChatGPT 4o mini, 9/2024)

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 $\mathbf{p} = (p_1, \ldots, p_n)$ 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 $\mathbf{p} = (p_1, \ldots, p_n)$ maximizes the entropy?

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

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: $\max_{\mathbf{p}} [H(\mathbf{p})]$ 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 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 $\mathbf{p} = (p_1, \ldots, p_n)$ maximizes the entropy?

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

Form the Lagrangian:

$$J(\mathbf{p}, \lambda) = -\sum_{i=1}^{n} p_i \cdot \lg(p_i) + \lambda \underbrace{\left(\sum_{i=1}^{n} p_i - 1\right)}_{=0}$$

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

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

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

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: $\max\limits_{\mathbf{p}} [H(\mathbf{p})]$ s.t. $\sum_{i=1}^{n} p_i = 1$

Form the Lagrangian: $\frac{\partial\,(p_1 \cdot \lg p_1 + \cancel{p_2 \cdot \lg p_2 + \dots})}{\partial p_1}$

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

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

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

we are done ☺, all p_i are identical!

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

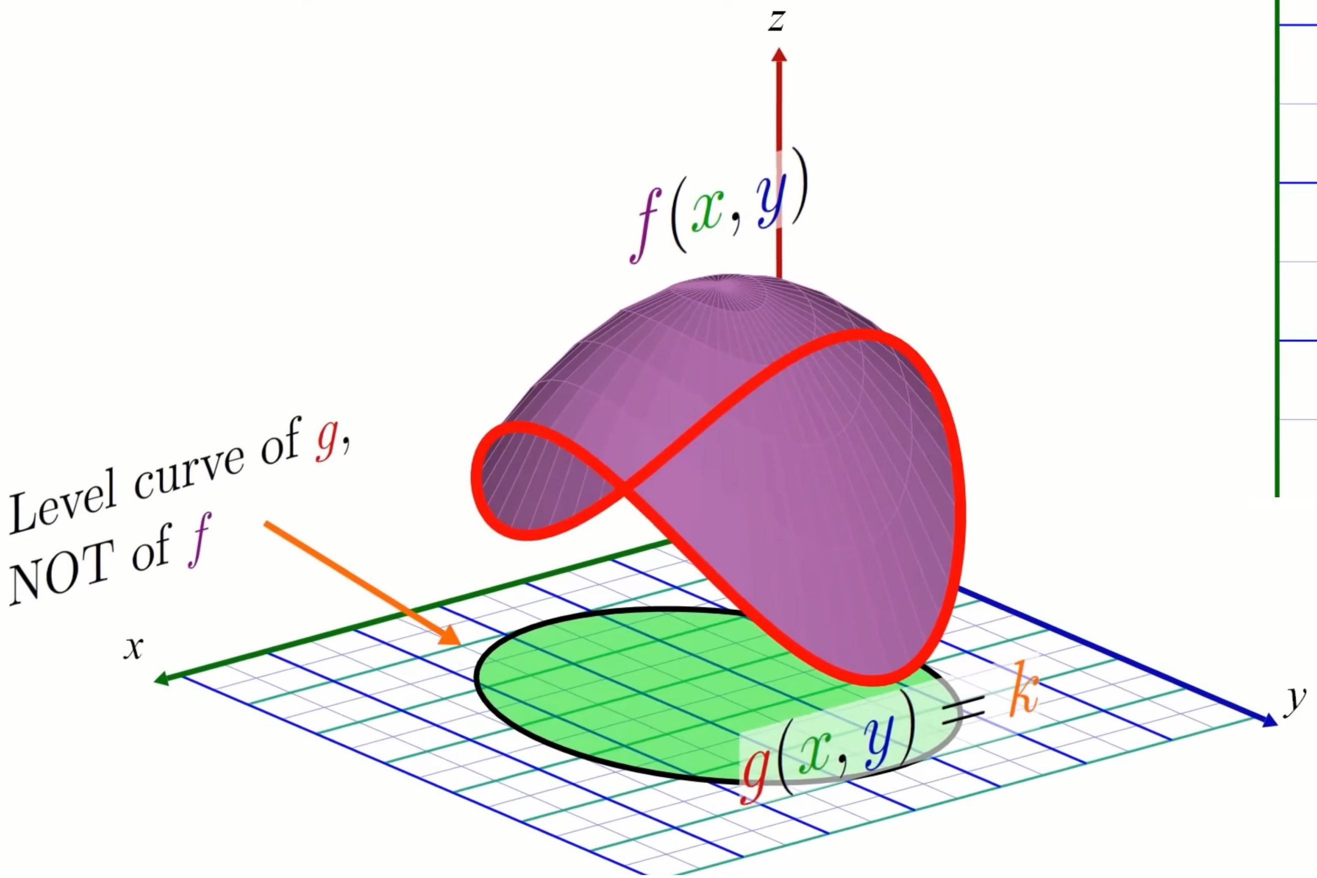
Calculus exercise

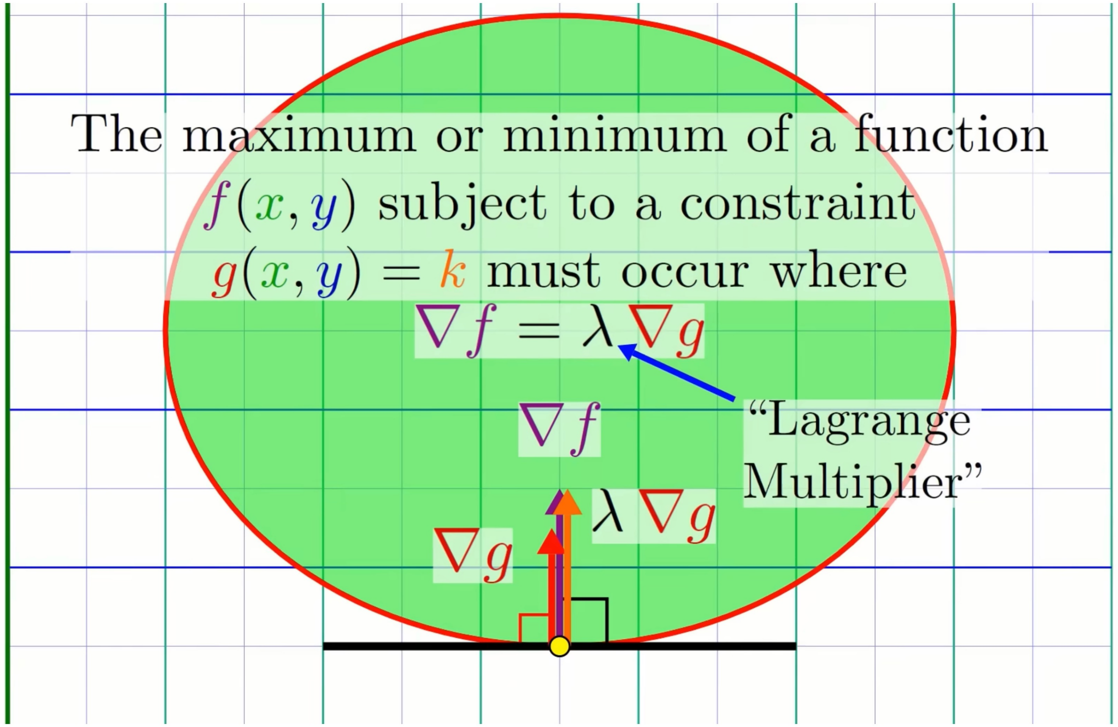
$$\ln(x)' = \frac{1}{x}$$

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

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

Nice video on Optimization w/ Lagrangian Multipliers





Serpentine Integral: "Understanding Lagrange Multipliers Visually", https://www.youtube.com/watch?v=5A39Ht9Wcu0
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

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**, ...,**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**, ...,**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?

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

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

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

$$H(U_2|U_1) \overset{?}{=} H(U_2) = 1$$

Learning partial information

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

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

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

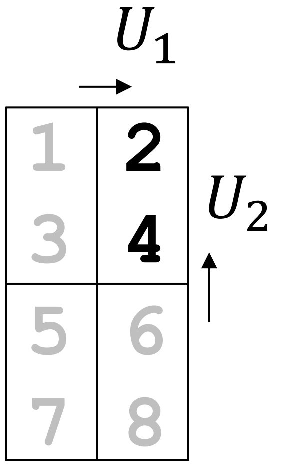
$$H(U_2|U_1) = \lg\left(\frac{4}{2}\right) = 1 \text{ bit}$$

messages are independent

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

the events are independent

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



Learning partial information

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

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 →	
1	**2**	U_2 ↑
3	**4**	
5	6	
7	8	

$p(x, y) = p(x) \cdot p(y)$

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

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

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:

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

Alternative notation: $p(X) = p_X(x)$. Also: $\mathbb{E}_p[\ldots]$ or $\mathbb{E}_X[\ldots]$ or $\mathbb{E}_{X\sim p}[\ldots]$ 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.

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

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

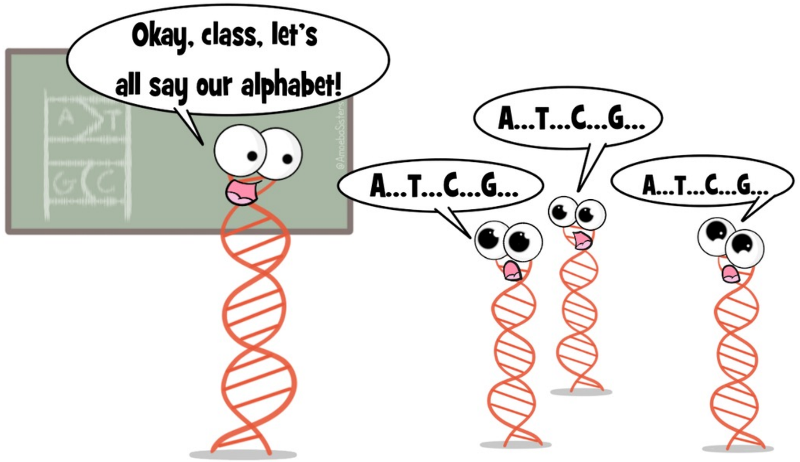


Figure source: https://www.amoebasisters.com/parameciumparlorcomics/dna-alphabet

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:

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

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

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:

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

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:

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

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

Updated 9/23/2025

Part 1: Theory
L05: Basics of entropy (3/7)
[joint entropy, conditional entropy, mutual information, cross entropy]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

9/22/2025

Pre-class conversations

- Last class recapitulation
- "Serendipity"
- Why even old slide decks still get updated
- Any questions on organization or projects? Thoughts on interactivity?

- Today:
  - Keep pen & paper ready for hands-on calculus, logarithm
  - Intuition behind entropy with examples continued
  - Together with the general principles of entropy

Joint entropy, Conditional entropy, Mutual information (continued)

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:

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

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

$$\boxed{H(Y|X) = \quad ?}$$

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:

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

$H(X,Y) =$?

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

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

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:

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

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

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

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

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

$$\boxed{H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y)}$$

... 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:

$$\mathbb{P}(A,B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A)$$
$$= \mathbb{P}(A|B) \cdot \mathbb{P}(B)$$

Learning partial information

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

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**, ...,**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

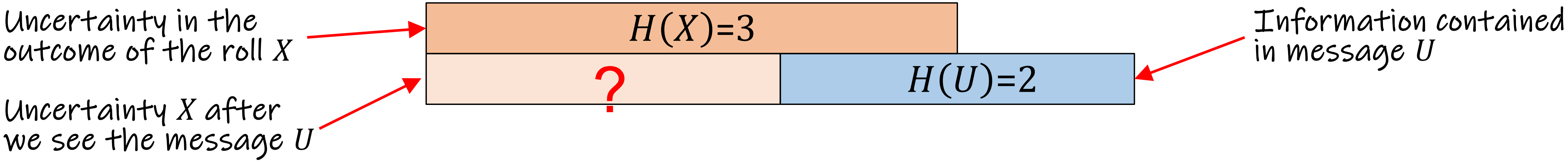
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}$={**1**,**2**, ...,**8**} .

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

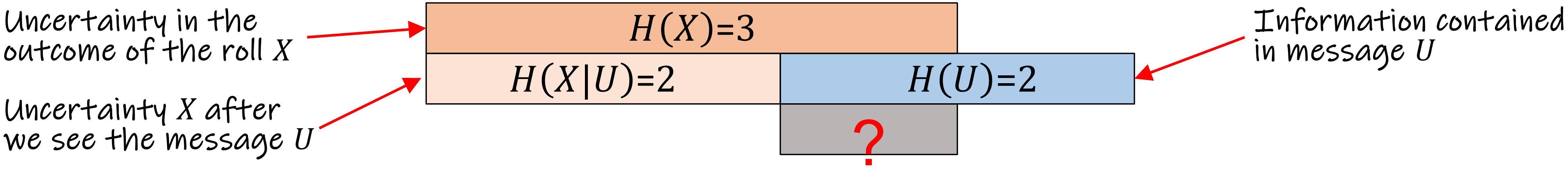
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}$={**1**,**2**, ...,**8**} .

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

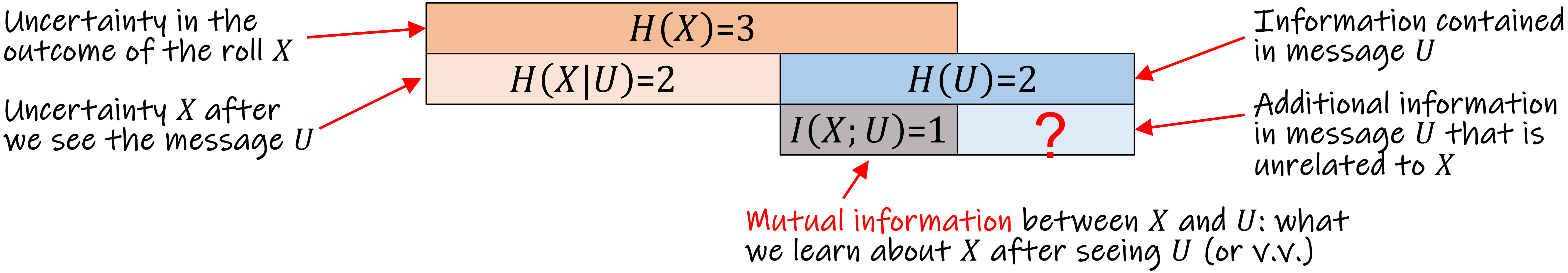
EXAMPLE 5: We again roll the unbiased 8-sided die with sides $\mathcal{X}$={**1**,**2**, ...,**8**} .

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?



Bar diagrams inspired by Fig 8.1 in "MacKay. Information Theory, Inference, and learning Algorithms. Cambridge University Press, 2002." https://www.inference.org.uk/itprnn/book.pdf

Learning partial information

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

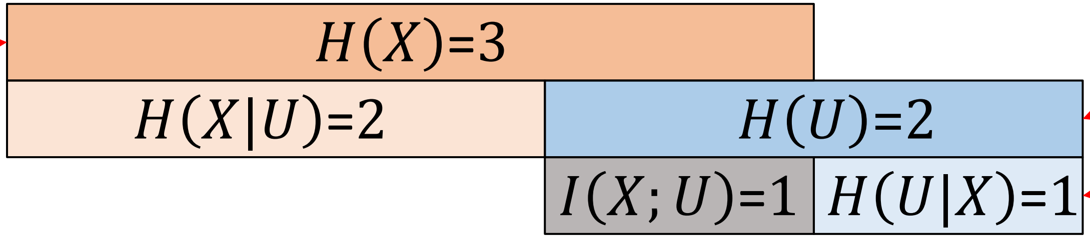
- We still learn 3-2=1 bit about the roll of the die X.
- We also learn 1 bit about the election outcome.

We learned 2 bits (U contains 2 bits)

How do these numbers add up?

Uncertainty in the outcome of the roll X

Uncertainty X after we see the message U



Information contained in message U

Additional information in message U that is unrelated to X

Mutual information between X and U: what we learn about X after seeing U (or v.v.)

Bar diagrams inspired by Fig 8.1 in "MacKay. Information Theory, Inference, and learning Algorithms. Cambridge University Press, 2002." https://www.inference.org.uk/itprnn/book.pdf

Mutual information

Given two RVs X and Y, mutual information is the amount of information that Y provides about X (thus when Y is observed, but X is not).

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

Is this function symmetric in X and Y ?

Mutual information

Given two RVs X and Y, mutual information is the amount of information that Y provides about X (thus when Y is observed, but X is not).

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

Conditional entropy: the amount of information needed to describe the outcome of RV Y given that we know the value of another RV X.

$$= H(X) - \big(H(X,Y) - H(Y)\big)$$

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

?

Mutual information

Given two RVs X and Y, mutual information is the amount of information that Y provides about X (thus when Y is observed, but X is not).

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

Conditional entropy: the amount of information needed to describe the outcome of RV Y given that we know the value of another RV X.

$$= H(X) - \big(H(X,Y) - H(Y)\big)$$

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

symmetric in X and Y!

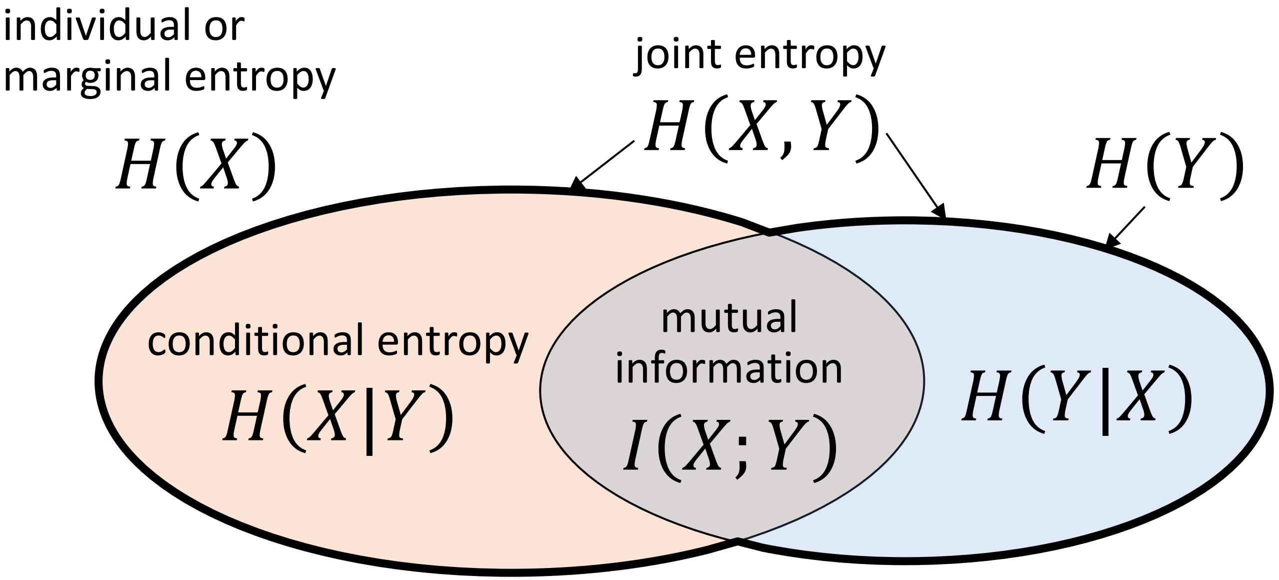
Thus, $I(X;Y) = I(Y;X)$

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

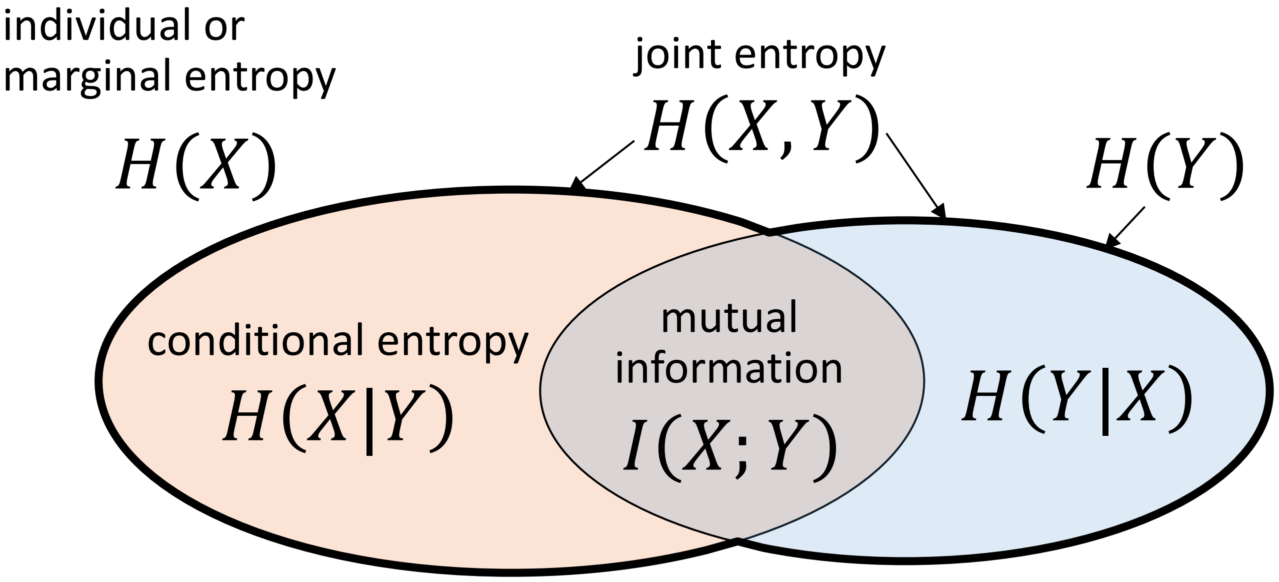
That's why it is called "mutual" information (it does not "prefer" X or Y).

Reduction of the uncertainty of one RV once we observe the other.

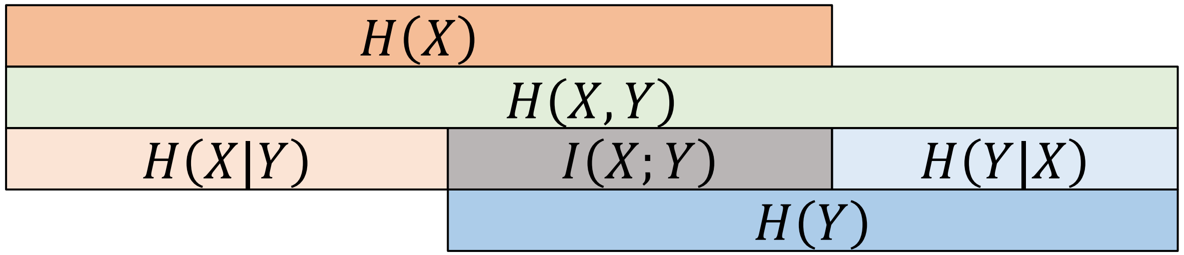
Entropy, conditional entropy, mutual information



Entropy, conditional entropy, mutual information



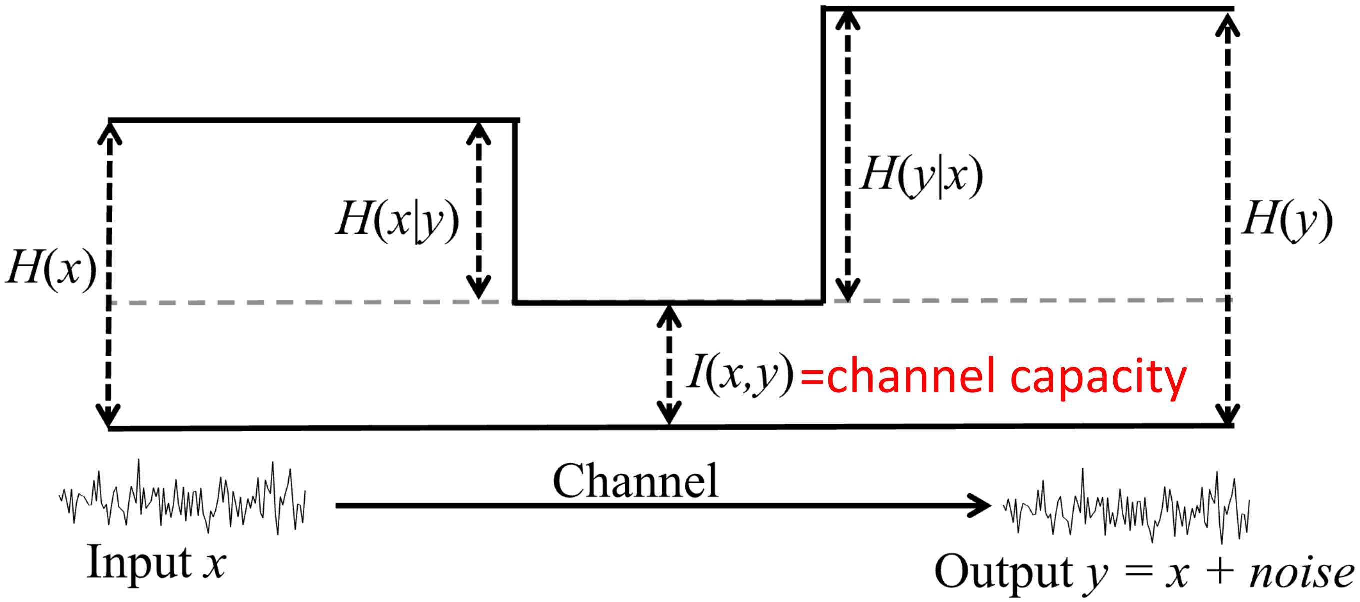
Basically the difference from
knowing things separately vs. jointly



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

The bar diagrams are inspired by Fig 8.1 in "MacKay. Information Theory, Inference, and learning Algorithms. Cambridge University Press, 2002." https://www.inference.org.uk/itprnn/book.pdf.
In particular, see the Interesting discussion and explanation in the solution to exercise 8.8 for why VENN diagrams (with more than 2 variables) can be misleading.

Entropy, conditional entropy, mutual information



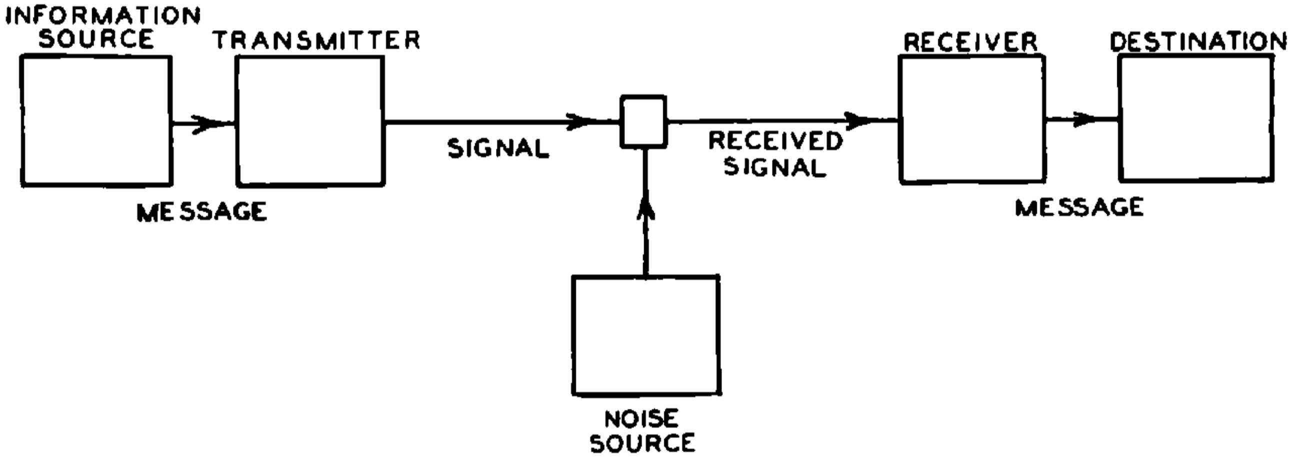


Fig. 1—Schematic diagram of a general communication system.

Figure sources: https://en.wikipedia.org/wiki/Mutual_information / Shannon. A Mathematical Theory of Communication. 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/

Self-information

What is $I(X; X)$?

How much does X tell us about itself?

Self-information

What is $I(X;X)$?

How much does X tell us about itself?

$$I(X;X) = H(X) - \underbrace{H(X|X)}_{=0}$$

$= 0$ no uncertainty (entropy) left)

$$I(X;X) = H(X)$$

We learn from X everything about X

Entropy is "self-information".

Relative entropy
= KL divergence
(≠ Cross-Entropy)

Relative Entropy = KL divergence (≠ Cross-Entropy)

The relative entropy (or KL divergence) of a distribution p with respect to a distribution q defined on the alphabet $\mathcal{X}$ of RV X is:

$$\boxed{D_{\mathrm{KL}}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{p(x)}{q(x)}\right)}$$

$\mathbb{E}_p[\ldots]$ also written as $\mathbb{E}_{X\sim p}[\ldots]$ for the expected value operator w.r. to the distribution p

It measures the inefficiency ("divergence", think "difference") for assuming a distribution q instead of a true distribution p for RV.

If we use q to construct a binary code, the expected message length is called cross-entropy:

$$\boxed{H(p||q) = \quad ?}$$

Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Relative Entropy = KL divergence (≠ Cross-Entropy)

The relative entropy (or KL divergence) of a distribution p with respect to a distribution q defined on the alphabet $\mathcal{X}$ of RV X is:

$$\boxed{D_{\mathrm{KL}}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{p(x)}{q(x)}\right)}$$

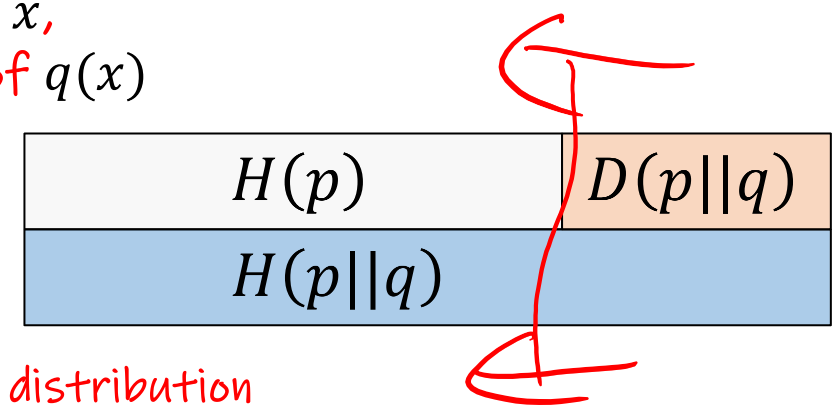
$\mathbb{E}_p[\ldots]$ also written as $\mathbb{E}_{X\sim p}[\ldots]$ for the expected value operator w.r. to the distribution p

It measures the inefficiency ("divergence", think "difference") for assuming a distribution q instead of a true distribution p for RV.

If we use q to construct a binary code, the expected message length is called cross-entropy:

$$\boxed{H(p||q) = D_{\mathrm{KL}}(p||q) + H(p)}$$

$$= \mathbb{E}_p\left[\lg\left(\frac{1}{q(X)}\right)\right] = \underbrace{\sum_{x\in\mathcal{X}} p(x)\cdot\overbrace{\lg\left(\frac{1}{q(x)}\right)}^{\text{my surprise for seeing } x\text{, given my assumption of } q(x)}}_{\text{my expected surprise given } p \text{ as the true distribution}}$$



Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Relative Entropy = KL divergence (≠ Cross-Entropy)

Relative entropy (or KL divergence) :

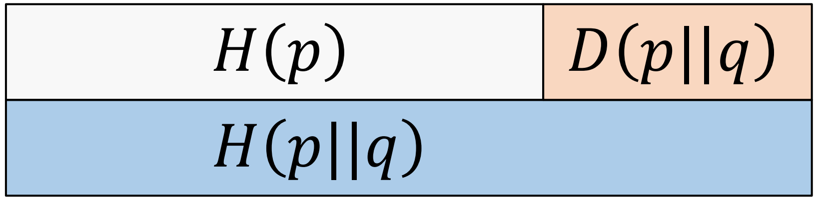
$$D_{\mathrm{KL}}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \boxed{\sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{p(x)}{q(x)}\right)}$$

Which of those should / do we commonly use in ML? And why?

?

Cross-entropy:

$$H(p||q) = \mathbb{E}_p\left[\lg\left(\frac{1}{q(X)}\right)\right] = \boxed{\sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{1}{q(x)}\right)}$$



Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Relative Entropy = KL divergence (≠ Cross-Entropy)

Relative entropy (or KL divergence) :

$$D_{\mathrm{KL}}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \boxed{\sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{p(x)}{q(x)}\right)}$$

We don't know $p(x)$. $D=$

But we have samples $\{x^{(1)}, \dots x^{(N)}\}$

Sample distribution as approximation of true distribution p

$$\approx -\frac{1}{N}\cdot\sum_i \lg\left(q\left(x^{(i)}\right)\right)$$

Have you seen this before (not in this class) ?

Cross-entropy:

$$H(p||q) = \mathbb{E}_p\left[\lg\left(\frac{1}{q(X)}\right)\right] = \boxed{\sum_{x\in\mathcal{X}} p(x)\cdot\lg\left(\frac{1}{q(x)}\right)}$$

$H(p)$	$D(p\|\|q)$
$H(p\|\|q)$	

Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Relative Entropy = KL divergence (≠ Cross-Entropy)

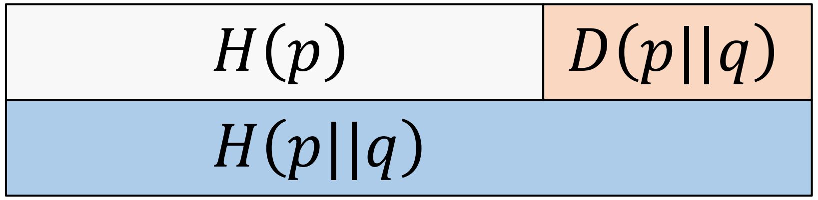
Sample distribution as approximation
of true distribution p

$$\approx -\frac{1}{N} \cdot \sum_i \lg\left(q_\theta\left(x^{(i)}\right)\right)$$

Parameterized model
distribution q_θ

Cross-entropy:

$$H(p||q) = \mathbb{E}_p\left[\lg\left(\frac{1}{q(X)}\right)\right] = \boxed{\sum_{x \in \mathcal{X}} p(x) \cdot \lg\left(\frac{1}{q(x)}\right)}$$



Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Relative Entropy = KL divergence (≠ Cross-Entropy)

$$L(\theta|\mathcal{D}) = \prod_i q_\theta(x^{(i)})$$

Likelihood of seeing the dataset $\mathcal{D} = \{x^{(1)}, \ldots x^{(N)}\}$ under model q_θ

$\lg \prod_i$

$$\ell(\theta|\mathcal{D}) = \sum_i \lg\left(q_\theta(x^{(i)})\right)$$

Log-likelihood

$$NLL(\theta|\mathcal{D}) = -\sum_i \lg\left(q_\theta(x^{(i)})\right)$$

Negative Log-likelihood

$$\approx -\frac{1}{N} \cdot \sum_i \lg\left(q_\theta(x^{(i)})\right)$$

Cross-entropy = average per-sample loss

Cross-entropy:

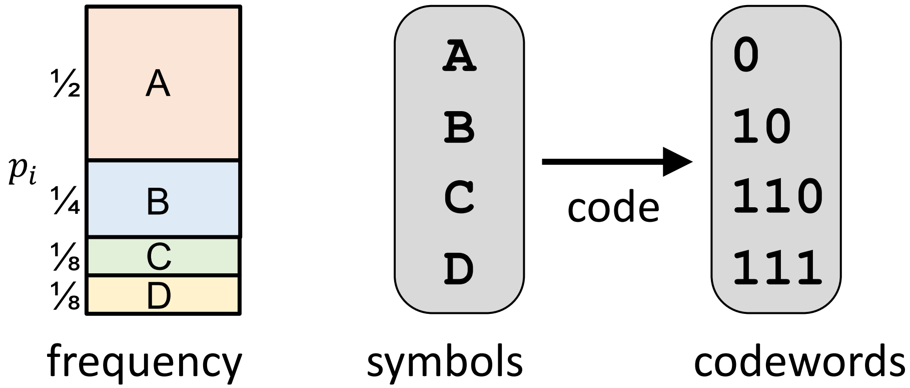
$$H(p||q) = \mathbb{E}_p\left[\lg\left(\frac{1}{q(X)}\right)\right] = \boxed{\sum_{x \in \mathcal{X}} p(x) \cdot \lg\left(\frac{1}{q(x)}\right)}$$

$H(p)$	$D(p\|\|q)$
$H(p\|\|q)$	

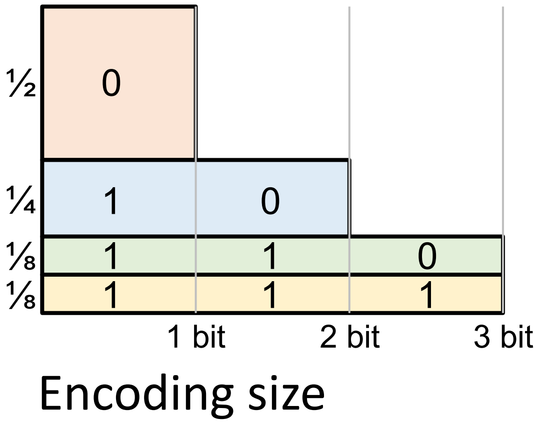
Cross-entropy is usually written as $H(p,q)$, but that notation hides its asymmetry and looks too similar to joint entropy. I prefer the notation $H(p||q)$ which captures the asymmetry with a similar notation as $D_{\mathrm{KL}}(p||q)$. Another non-standard notation is $H_p(q)$ which shows that p is the true distribution, whereas q determines the assumed surprise.

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :

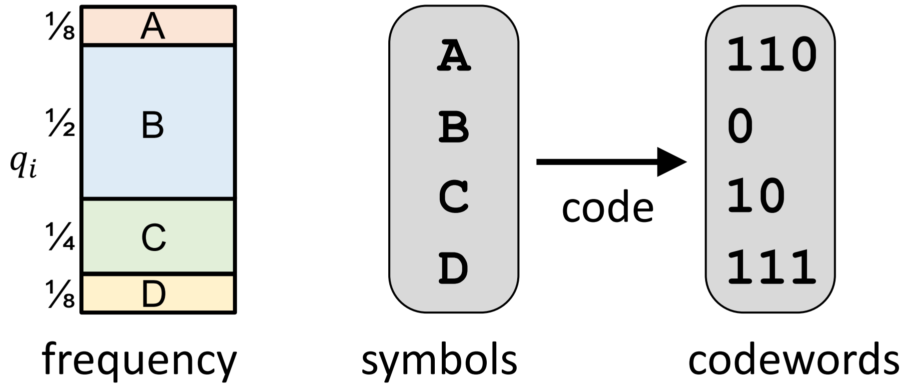


½ · 1
¼ · 2 = 1.75 bits!
⅛ · 3
⅛ · 3

$\lg\left(\frac{1}{2}\right) = -1$
$\lg\left(\frac{1}{4}\right) = -2$
$\lg\left(\frac{1}{8}\right) = -3$

Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

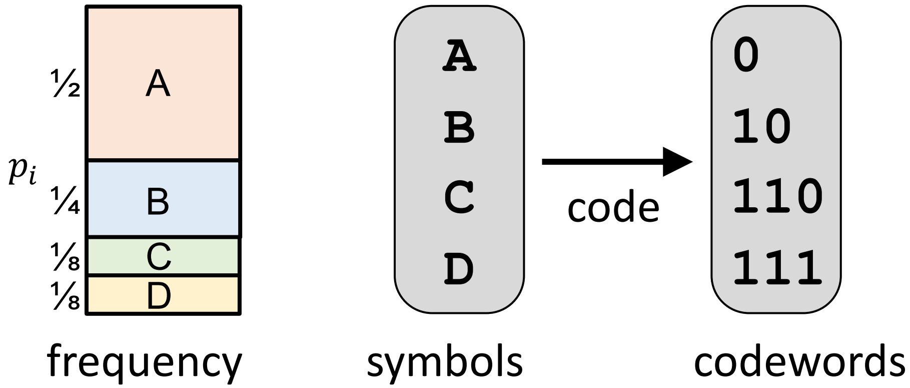
- What if we assume following distribution:



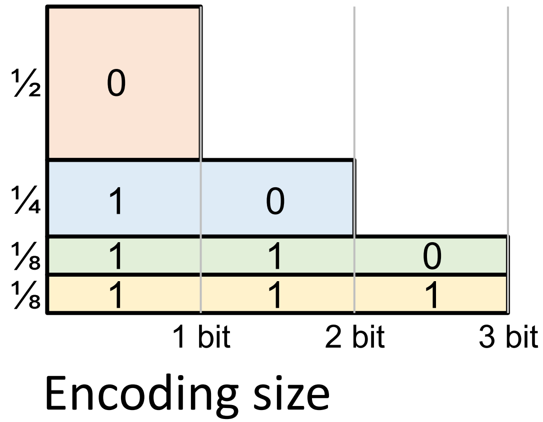
What is our expected message length per symbol if we use that code, but p is the actual distribution ?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

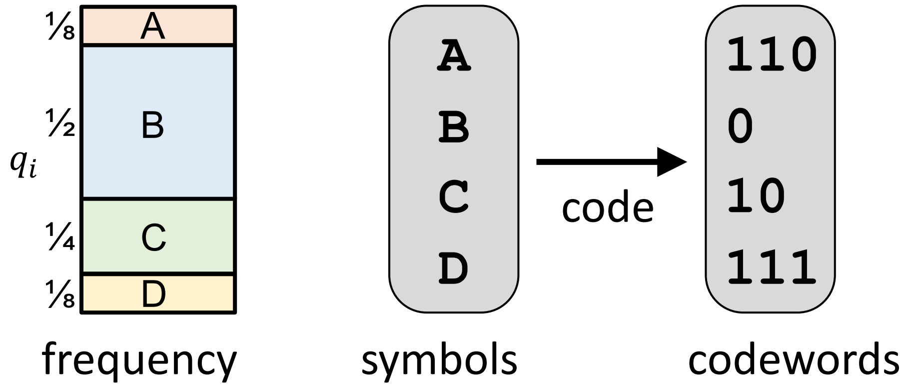
$\lg\left(\frac{1}{2}\right) = -1$

$\lg\left(\frac{1}{4}\right) = -2$

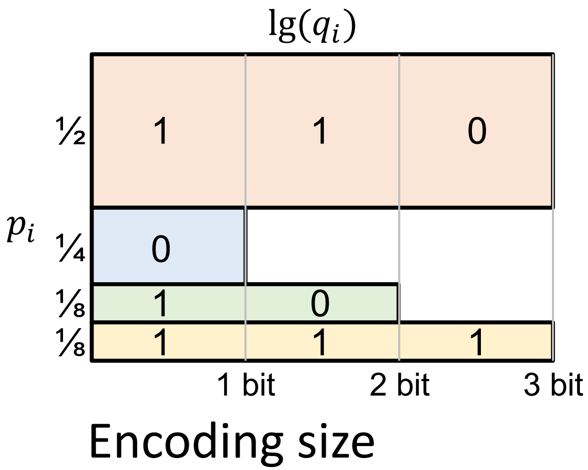
$\lg\left(\frac{1}{8}\right) = -3$

Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

- What if we assume following distribution:



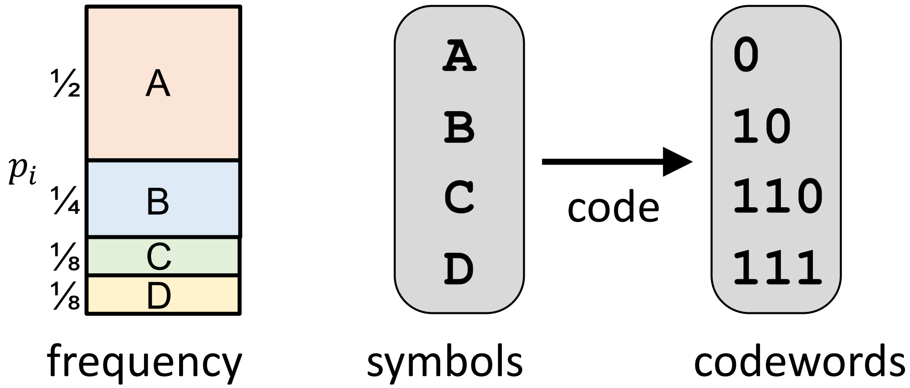
Our new expected message length per symbol:



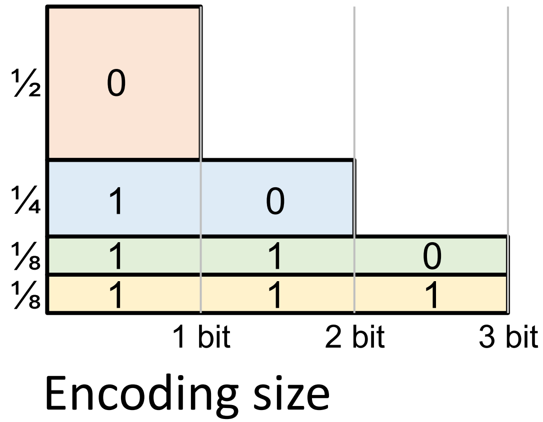
What is the formula we need to evaluate ?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :

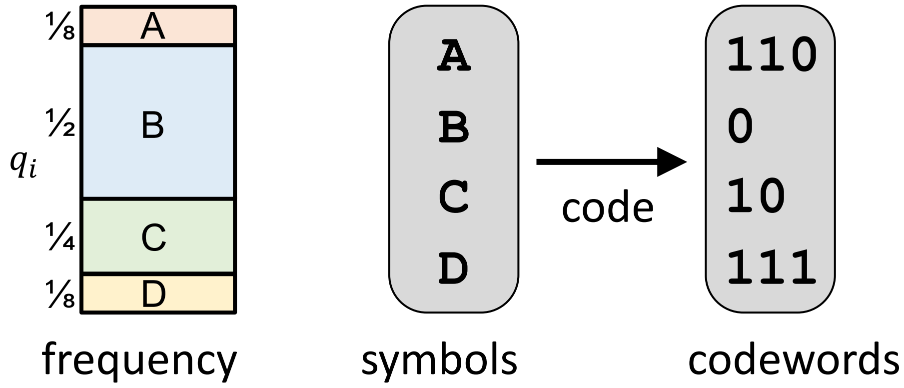


½ · 1
¼ · 2 = 1.75 bits!
⅛ · 3
⅛ · 3

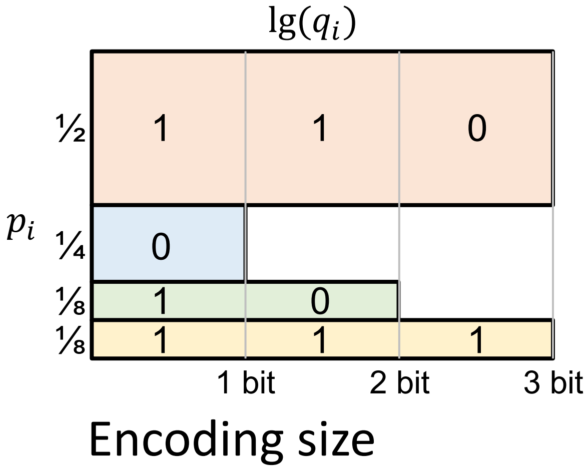
$\lg\left(\frac{1}{2}\right) = -1$
$\lg\left(\frac{1}{4}\right) = -2$
$\lg\left(\frac{1}{8}\right) = -3$

Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

- What if we assume following distribution:



Our new expected message length per symbol:



= 2.375 bits!

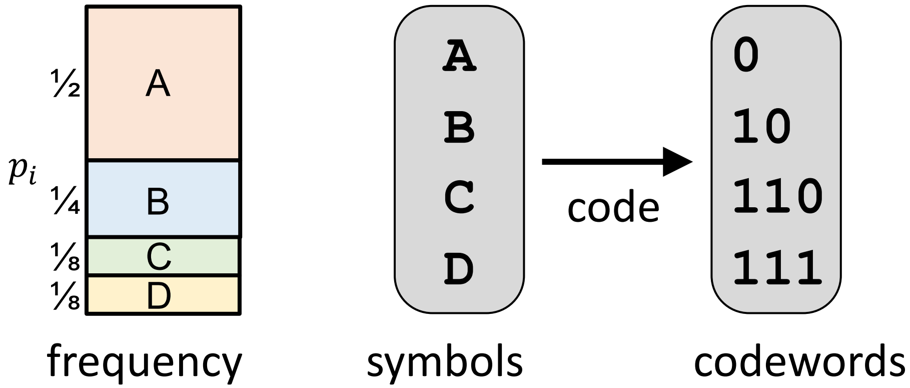
$-\sum_i p_i \cdot \lg(q_i)$

What is this formula called

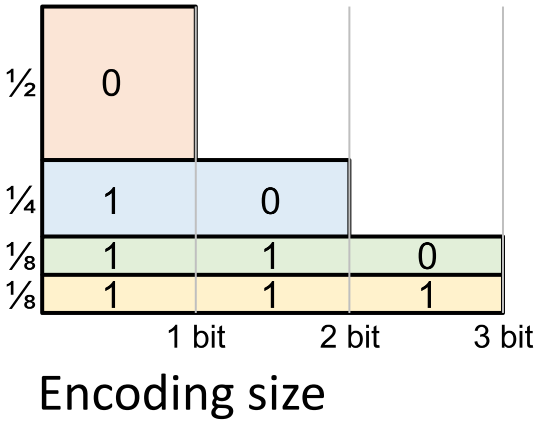
?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$\frac{1}{2} \cdot 1$

$\frac{1}{4} \cdot 2$ = 1.75 bits!

$\frac{1}{8} \cdot 3$

$\frac{1}{8} \cdot 3$

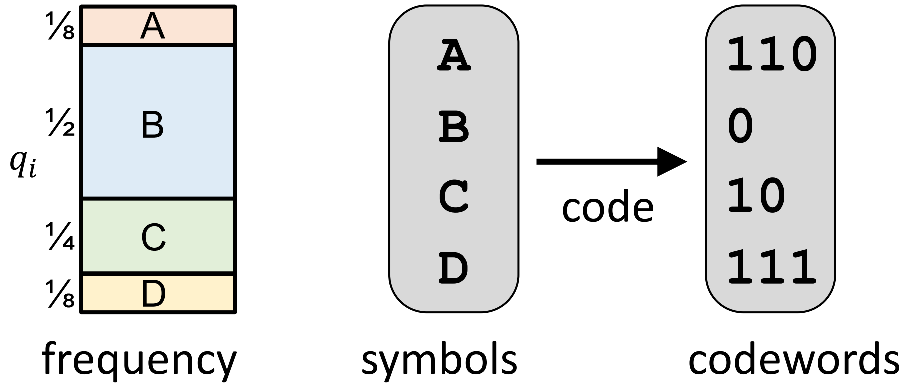
$\lg\left(\frac{1}{2}\right) = -1$

$\lg\left(\frac{1}{4}\right) = -2$

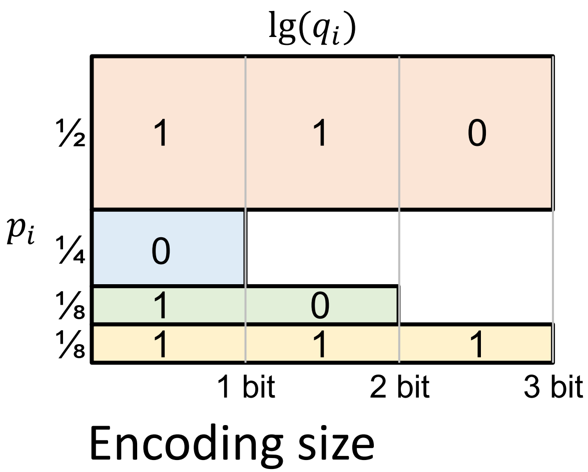
$\lg\left(\frac{1}{8}\right) = -3$

Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

- What if we assume following distribution:



Our new expected message length per symbol:



= 2.375 bits!

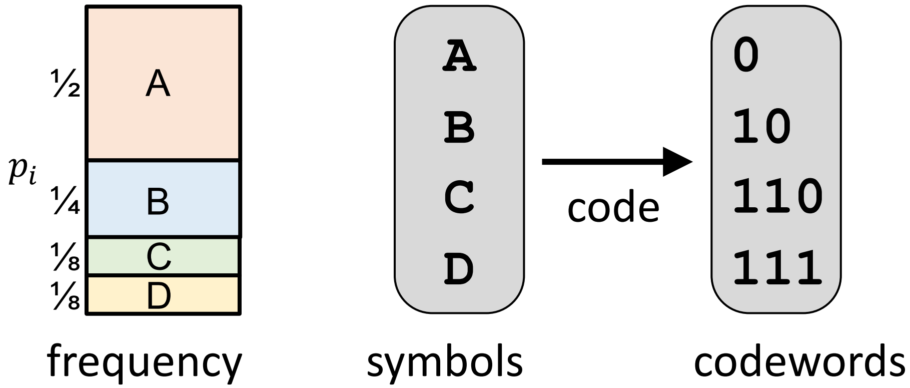
$-\sum_i p_i \cdot \lg(q_i)$

Cross entropy $H(p||q)$ ☺

Which distribution q minimizes $H(p||q)$?

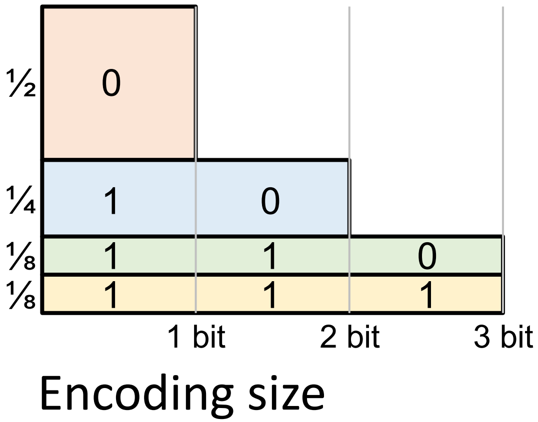
Compressing messages via variable length codes

- Assume we have the following symbol frequency:



Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

New expected length :



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

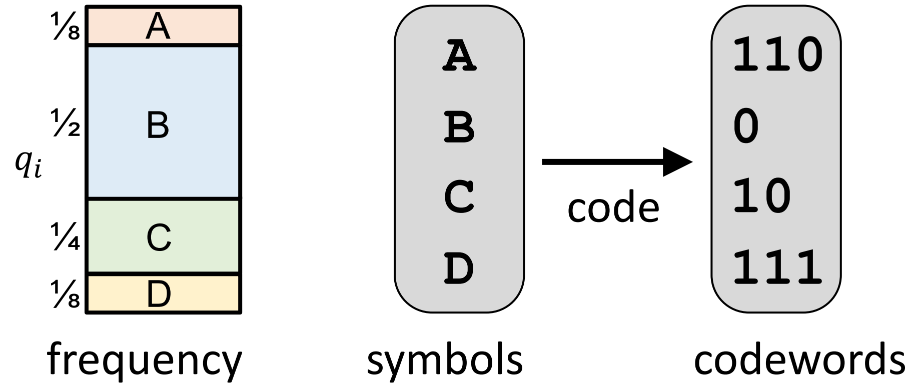
⅛ · 3

$\lg\left(\frac{1}{2}\right) = -1$

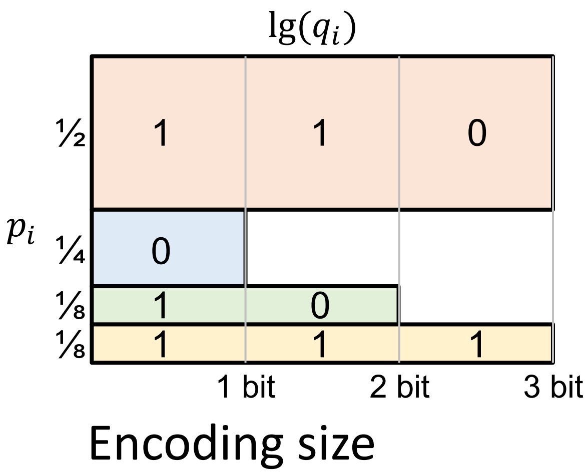
$\lg\left(\frac{1}{4}\right) = -2$

$\lg\left(\frac{1}{8}\right) = -3$

- What if we assume following distribution:



Our new expected message length per symbol:



= 2.375 bits!

$-\sum_i p_i \cdot \lg(q_i)$

Cross entropy $H(p||q)$ ☺

$q = p$ minimizes $H(p||q)$

Properties of Relative Entropy = KL divergence

1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

$$\mathbb{E}_{p}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

EXAMPLE : $\mathbf{u} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\quad \mathbf{p} = \begin{pmatrix} p \\ \bar{p} \end{pmatrix}$ $\quad \bar{p} = 1 - p$

	$D_{\mathrm{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\mathrm{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	?	?
$p = 0$	?	?
$p = 0.01$	?	?

Properties of Relative Entropy = KL divergence


1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\text{KL}}(p||q) \neq D_{\text{KL}}(q||p)$$

$$\mathbb{E}_{p}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

EXAMPLE : $\mathbf{u} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\quad \mathbf{p} = \begin{pmatrix} p \\ \bar{p} \end{pmatrix}$ $\quad \bar{p} = 1 - p$

	$D_{\text{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\text{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	0	0
$p = 0$	?	?
$p = 0.01$	?	?

Properties of Relative Entropy = KL divergence

1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

$$\mathbb{E}_{p}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

EXAMPLE : $\mathbf{u} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\quad \mathbf{p} = \begin{pmatrix} p \\ \bar{p} \end{pmatrix}$ $\quad \bar{p} = 1 - p$

	$D_{\mathrm{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\mathrm{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	0	0
$p = 0$	1	∞
$p = 0.01$	?	?

Properties of Relative Entropy = KL divergence

1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

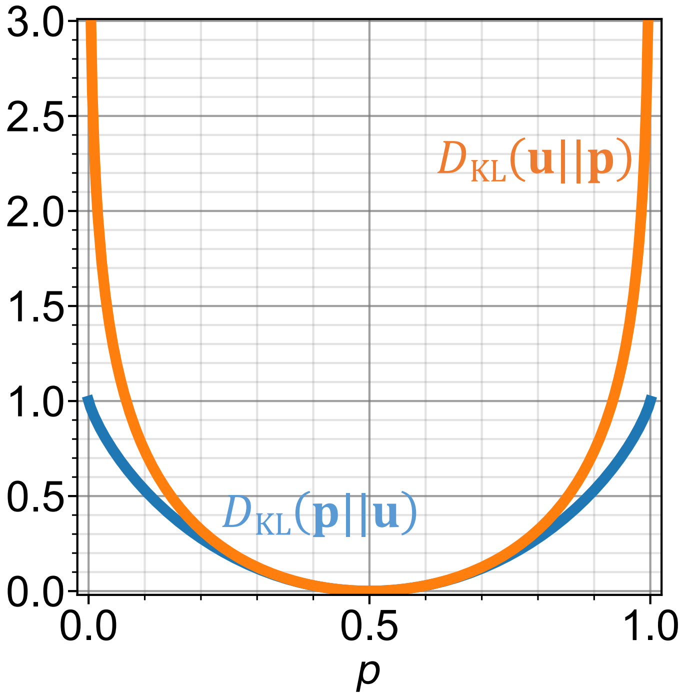
$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

$\mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$

What about cross entropies $H(\mathbf{p}||\mathbf{u})$ and $H(\mathbf{p}||\mathbf{u})$?

EXAMPLE : $\mathbf{u} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\quad \mathbf{p} = \begin{pmatrix} p \\ \bar{p} \end{pmatrix}$ $\quad \bar{p} = 1 - p$

	$D_{\mathrm{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\mathrm{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	0	0
$p = 0$	1	∞
$p = 0.01$	0.92	2.33
	$\underbrace{.01\lg\left(\frac{.01}{.5}\right)}_{-0.06} + \underbrace{.99\lg\left(\frac{.99}{.5}\right)}_{0.96}$	$\underbrace{.5\lg\left(\frac{.5}{.01}\right)}_{2.82} + \underbrace{.5\lg\left(\frac{.5}{.99}\right)}_{-0.49}$



Properties of Relative Entropy = KL divergence

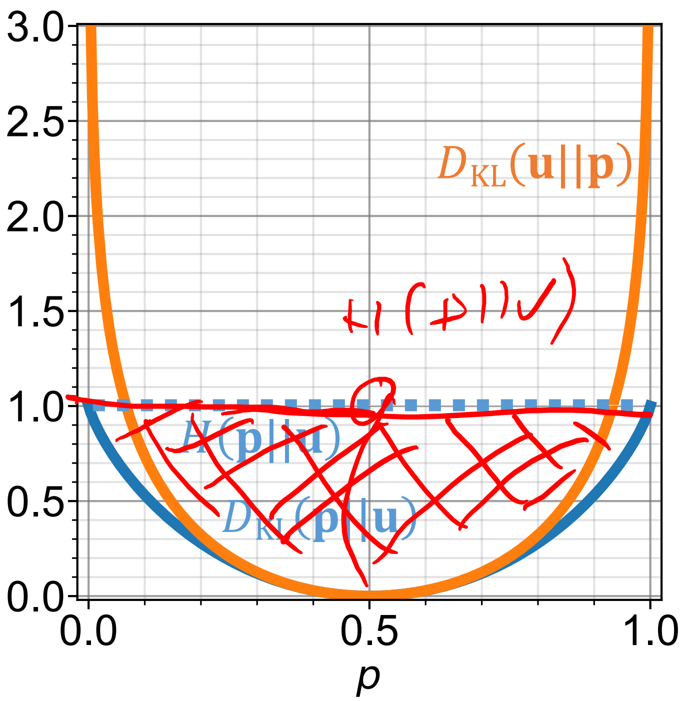
1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\text{KL}}(p||q) \neq D_{\text{KL}}(q||p)$$

$$\mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

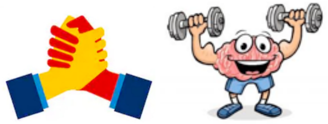
EXAMPLE : $\mathbf{u} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\quad \mathbf{p} = \begin{pmatrix} p \\ \bar{p} \end{pmatrix}$ $\quad \bar{p} = 1 - p$

	$D_{\text{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\text{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	0	0
$p = 0$	1	∞
$p = 0.01$	0.92	2.33
	$\underbrace{.01\lg\left(\frac{.01}{.5}\right)}_{-0.06} + \underbrace{.99\lg\left(\frac{.99}{.5}\right)}_{0.96}$	$\underbrace{.5\lg\left(\frac{.5}{.01}\right)}_{2.82} + \underbrace{.5\lg\left(\frac{.5}{.99}\right)}_{-0.49}$



$H(\mathbf{p}||\mathbf{u}) = D(\mathbf{p}||\mathbf{u}) + H(\mathbf{p})$

Properties of Relative Entropy = KL divergence



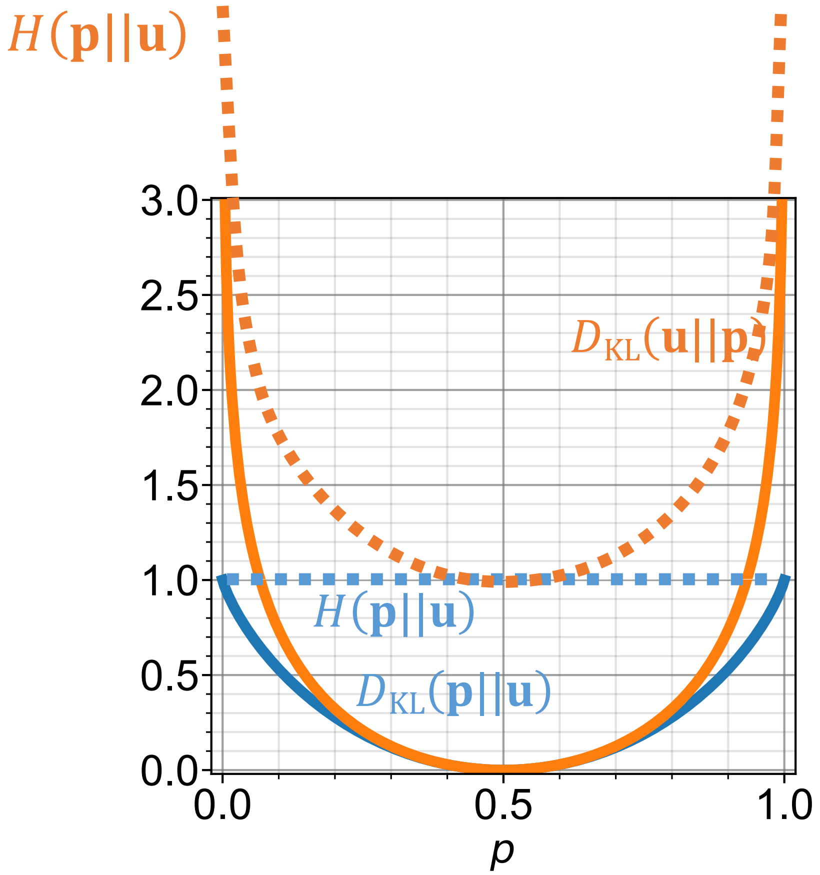
1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

$$\mathbb{E}_{p}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

EXAMPLE : $\mathbf{u} = \binom{0.5}{0.5}$ $\quad \mathbf{p} = \binom{p}{\bar{p}}$ $\quad \bar{p} = 1 - p$

	$D_{\mathrm{KL}}(\mathbf{p}\|\|\mathbf{u})$	$D_{\mathrm{KL}}(\mathbf{u}\|\|\mathbf{p})$
$p = 0.5$	0	0
$p = 0$	1	∞
$p = 0.01$	0.92	2.33
	$\underbrace{.01\lg\left(\frac{.01}{.5}\right)}_{-0.06} + \underbrace{.99\lg\left(\frac{.99}{.5}\right)}_{0.96}$	$\underbrace{.5\lg\left(\frac{.5}{.01}\right)}_{2.82} + \underbrace{.5\lg\left(\frac{.5}{.99}\right)}_{-0.49}$



$$H(\mathbf{p}||\mathbf{u}) = D(\mathbf{p}||\mathbf{u}) + H(\mathbf{p})$$

$$H(\mathbf{u}||\mathbf{p}) = D(\mathbf{u}||\mathbf{p}) + 1$$

Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Properties of Relative Entropy = KL divergence

1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

2. $D_{\mathrm{KL}}(p||p) =$?

Properties of Relative Entropy = KL divergence

1. Relative entropy is asymmetric (does not satisfy triangle inequality, thus not a metric):

$$D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$$

2. $D_{\mathrm{KL}}(p||p) = 0$

3. $D_{KL}(p||q) \geq 0$ for all distributions p, q (equality only holds for $p = q$)

We will prove that next (with Jensen's inequality)

Commuting functions: an apparent digression

- Do functions commute with taking the expectation?

$$\mathbb{E}[f(X)] = f(\mathbb{E}[X]) \quad ?$$

Commuting functions: an apparent digression

- Do functions commute with taking the expectation?

$$\cancel{\mathbb{E}[f(X)] = f(\mathbb{E}[X])}$$

- No! This only holds for linear functions:

$$f(x) = ax + b$$
$$\mathbb{E}[ax + b] = a\mathbb{E}[x] + b$$

- Jensen's inequality for convex f:

?

Commuting functions: an apparent digression

- Do functions commute with taking the expectation?

$$\cancel{\mathbb{E}[f(X)] = f(\mathbb{E}[X])}$$

- No! This only holds for linear functions:

$$f(x) = ax + b$$
$$\mathbb{E}[ax + b] = a\mathbb{E}[x] + b$$

- Jensen's inequality for convex f:

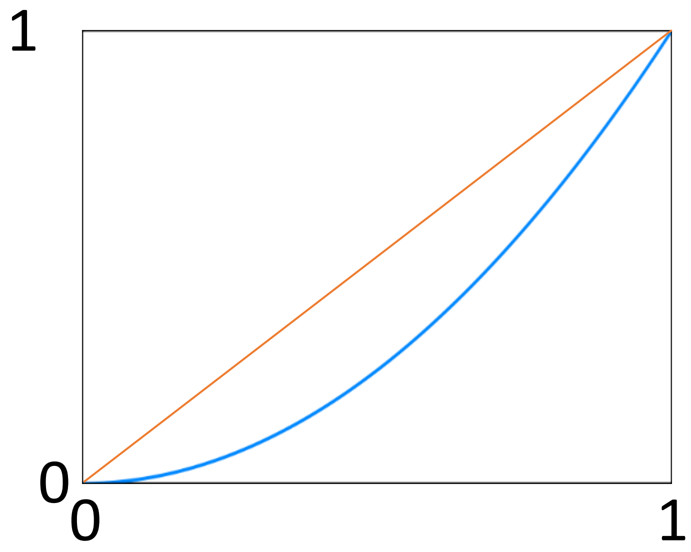
$$\boxed{\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])}$$

- Example $f(x) = x^2$:

  Consider the interval $0 \leq x \leq 1$:

  $f(\mathbb{E}[X]) =$?

  $\mathbb{E}[f(X)] =$?



Commuting functions: an apparent digression

- Do functions commute with taking the expectation?

$$\cancel{\mathbb{E}[f(X)] = f(\mathbb{E}[X])}$$

- No! This only holds for linear functions:

$$f(x) = ax + b$$
$$\mathbb{E}[ax + b] = a\mathbb{E}[x] + b$$

- Jensen's inequality for convex f:

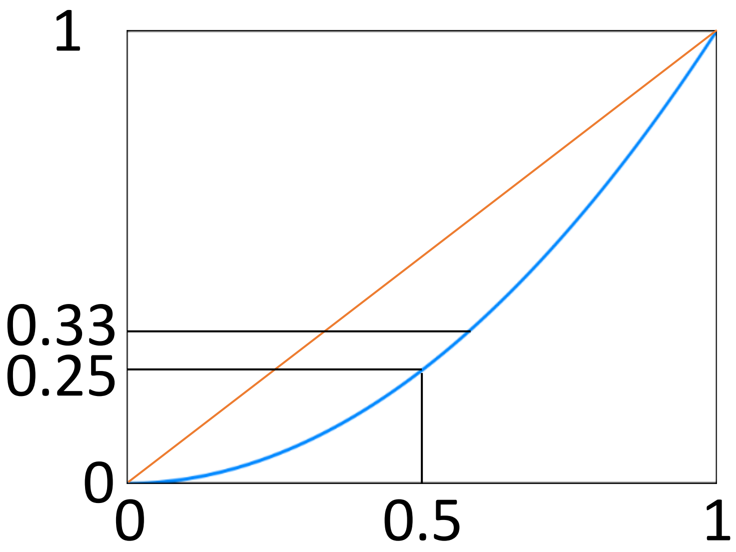
$$\boxed{\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])}$$

- Example $f(x) = x^2$:

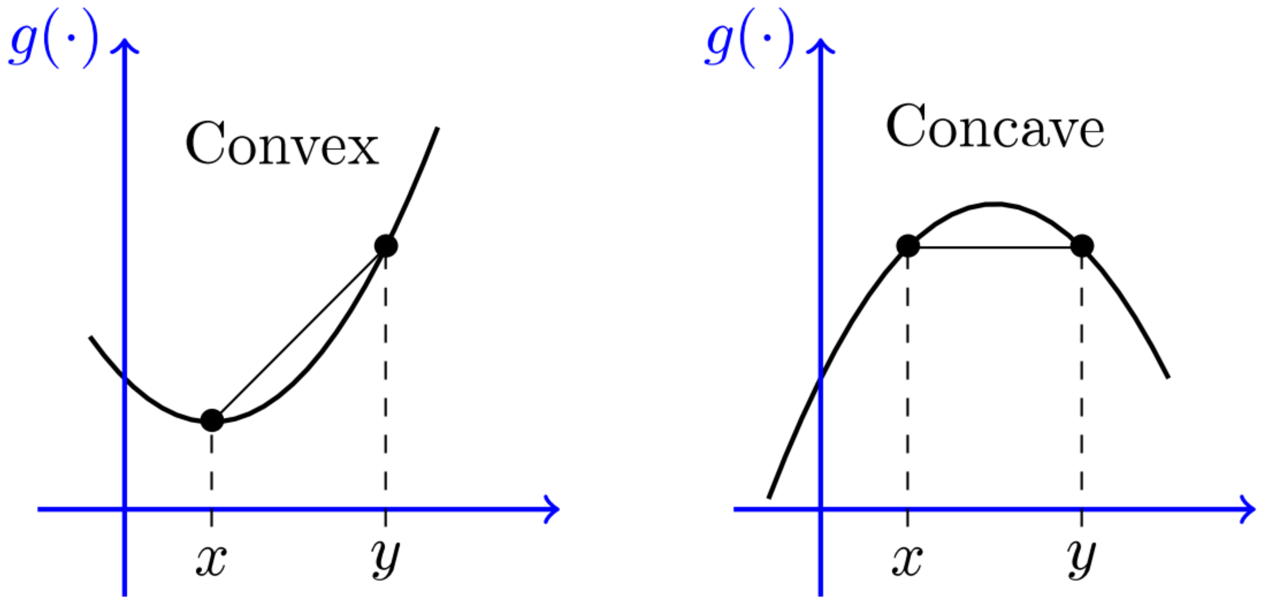
Consider the interval $0 \leq x \leq 1$:

$$f(\mathbb{E}[X]) = \ f(\mathbb{E}[X]) = f(0.5) = 0.25$$

$$\mathbb{E}[f(X)] = \frac{\int_0^1 f(x)}{1-0} = \frac{x^3}{3}\Big|_0^1 = 0.33$$



Background: Convex / Concave function



Definition 6.3

Consider a function $g : I \to \mathbb{R}$, where I is an interval in $\mathbb{R}$. We say that g is a **convex** function if, for any two points x and y in I and any $\alpha \in [0, 1]$, we have

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

We say that g is **concave** if

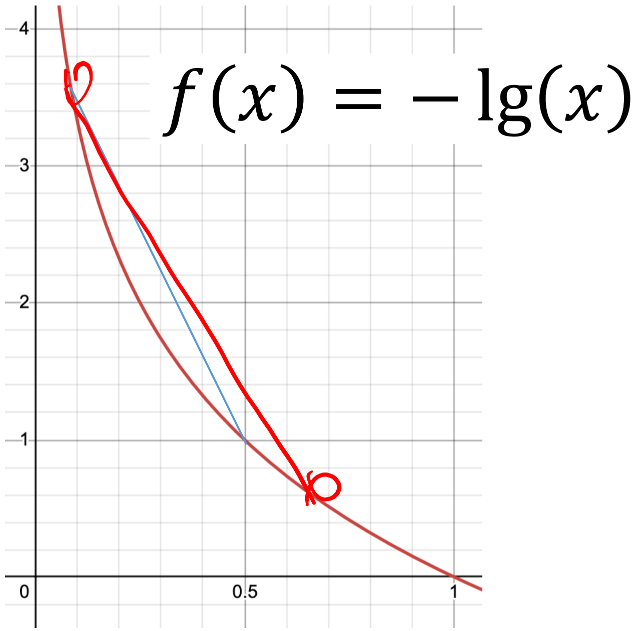
$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y).$$

Figure source: https://www.probabilitycourse.com/chapter6/6_2_5_jensen%27s_inequality.php

Information inequality $D_{KL}(p||q) \geq 0$

Ingredients:

1. $-\lg(x)$ is convex
2. Jensen's inequality $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$



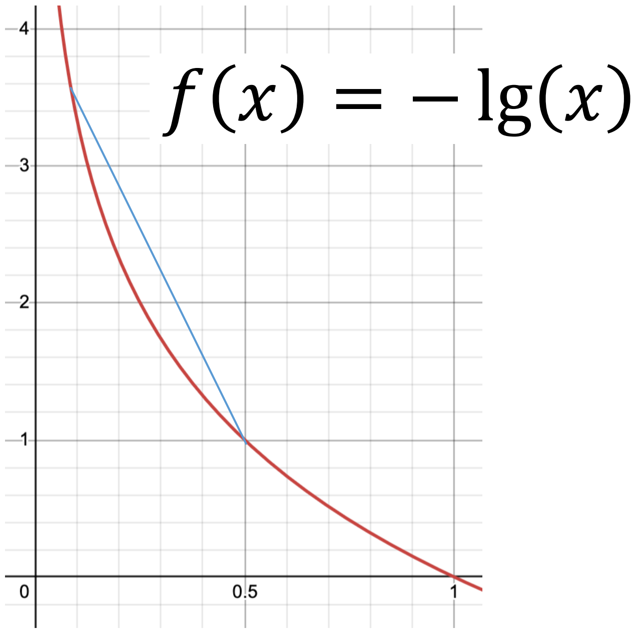
$$D_{KL}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$

$$= \quad ?$$

Information inequality $D_{KL}(p||q) \geq 0$

Ingredients:

1. $-\lg(x)$ is convex
2. Jensen's inequality $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$



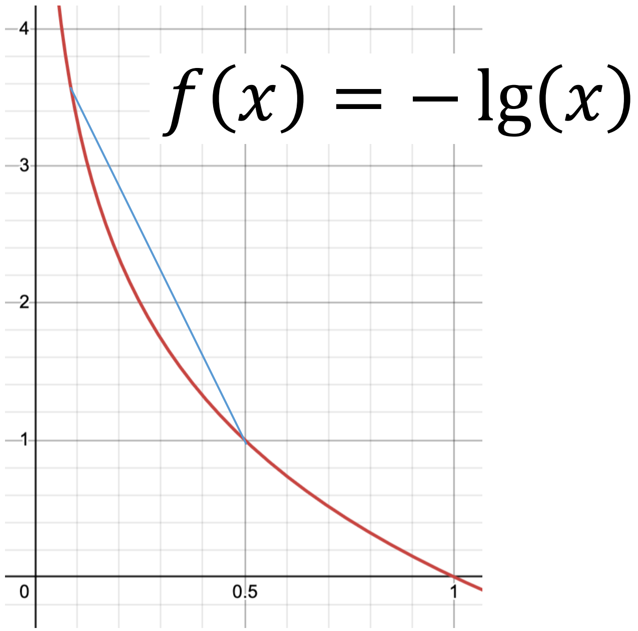
$$D_{KL}(p||q) = \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right]$$
$$= \mathbb{E}_p\left[-\lg\left(\frac{q(X)}{p(X)}\right)\right]$$
$$\geq -\lg\left(\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right]\right) = -\lg\underbrace{\left(\sum_x \cancel{p(x)} \cdot \frac{q(x)}{\cancel{p(x)}}\right)}_{=1} = 0$$

$D_{KL}(p||q) = 0$ iff ?

Information inequality $D_{KL}(p||q) \geq 0$

Ingredients:

1. $-\lg(x)$ is convex
2. Jensen's inequality $\quad \mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$



$$
\begin{aligned}
D_{KL}(p||q) &= \mathbb{E}_p\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] \\
&= \mathbb{E}_p\left[-\lg\left(\frac{q(X)}{p(X)}\right)\right] \\
&\geq -\lg\left(\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right]\right) = -\lg\underbrace{\left(\sum_x \cancel{p(x)} \cdot \frac{q(x)}{\cancel{p(x)}}\right)}_{=1} = 0
\end{aligned}
$$

$D_{KL}(p||q) = 0$ iff $q(x) = p(x)$ for all x.

Updated 10/1/2025

Part 1: Theory
L06: Basics of entropy (4/7)
[mutual information, multivariate entropies]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

9/25/2025

Pre-class conversations

- Last class recapitulation
- Slide decks: please continue checking for errors / inconsistencies / unclear details
- Scribes? Your own Python scripts could be part of your next scribes!
- Please share pointers to work using information theory in your area of expertise that you find interesting / we may add that as topic.

- Today:
  - Mutual information, multivariate entropies, Markov Chains,
  - Next time: Data Processing inequality, sufficient statistics

Mutual information as relative entropy and thus ≥ 0

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

$$\boxed{I(X;Y) := H(X) - H(X|Y)} \qquad \geq 0 \ ?$$

notation $x \in \mathcal{X}, y \in \mathcal{Y}$

Mutual information as relative entropy and thus ≥ 0

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

notation $x \in \mathcal{X}, y \in \mathcal{Y}$

$$= \sum_x p(x) \cdot \lg\left(\frac{1}{p(x)}\right) \quad - \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x|y)}\right)$$

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

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

$$= \sum_{x,y} p(x,y) \cdot \lg\left(\boxed{\frac{p(x,y)}{p(x) \cdot p(y)}}\right) \quad = \quad ?$$

ratio between joint distribution and product of marginals

Mutual information as relative entropy and thus ≥ 0

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

notation $x \in \mathcal{X}, y \in \mathcal{Y}$

$$= \sum_{x} p(x) \cdot \lg\left(\frac{1}{p(x)}\right) \quad - \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x|y)}\right)$$

$$= \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x)}\right) \quad - \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x|y)}\right)$$

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

$$= \sum_{x,y} p(x,y) \cdot \lg\left(\boxed{\frac{p(x,y)}{p(x) \cdot p(y)}}\right) \quad = D_{\mathrm{KL}}(p(x,y) || p(x) \cdot p(y)) \geq 0$$

When equality?

Mutual information is the relative entropy (KL divergence) between joint distribution and product of their marginal distributions!

Mutual information as relative entropy and thus ≥ 0

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

$$\boxed{I(X;Y) := H(X) - H(X|Y)}$$

notation $x \in \mathcal{X}, y \in \mathcal{Y}$

$$= \sum_{x} p(x) \cdot \lg\left(\frac{1}{p(x)}\right) \quad - \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x|y)}\right)$$

$$= \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x)}\right) \quad - \sum_{x,y} p(x,y) \cdot \lg\left(\frac{1}{p(x|y)}\right)$$

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

$$= \sum_{x,y} p(x,y) \cdot \lg\left(\boxed{\frac{p(x,y)}{p(x) \cdot p(y)}}\right) \quad = D_{\mathrm{KL}}(p(x,y) || p(x) \cdot p(y)) \geq 0$$

equality when X and Y are independent!

alternative notation: $D_{\mathrm{KL}}(p_{X,Y} || p_X \cdot p_Y)$

Mutual information is the relative entropy (KL divergence) between joint distribution and product of their marginal distributions!

Mutual information as relative entropy

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

$$I(X;Y) = \sum_{x,y} p(x,y) \cdot \overbrace{\lg\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right)}^{\text{What is that ?}} = D_{\mathrm{KL}}(p(x,y) || p(x) \cdot p(y)) \geq 0$$

Mutual information is the relative entropy (KL divergence) between joint distribution and product of their marginal distributions!

Mutual information as relative entropy

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

PMI (pointwise mutual information) is a measure of association

>0: values (x, y) co-occur more often together

=0: independent at that point (x, y)

<0: values (x, y) co-occur less often together

$$I(X;Y) = \sum_{x,y} p(x,y) \cdot \overbrace{\lg\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right)} = D_{\mathrm{KL}}(p(x,y) || p(x) \cdot p(y)) \geq 0$$

Mutual information is the relative entropy (KL divergence) between joint distribution and product of their marginal distributions!

Mutual information as relative entropy

Given two RVs X and Y, mutual information is the amount of information that Y provides about X when Y is observed, but X is not.

PMI (pointwise mutual information) is a measure of association

MI is then the average over all joint events (MI), i.e. the expected PMI under the distribution

While PMI may be negative, MI is ≥ 0

$$I(X;Y) = \sum_{x,y} p(x,y) \cdot \lg\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right) = D_{\mathrm{KL}}(p(x,y) || p(x) \cdot p(y)) \geq 0$$

Mutual information is the relative entropy (KL divergence) between joint distribution and product of their marginal distributions!

Conditioning reduces entropy, in expectation

$H(X|Y) \leq H(X)$ (follows from $I(X;Y) = H(X) - H(X|Y) \geq 0$)

The nonnegativity of mutual information implies that on average the entropy of X conditioned on the observation $Y = y$ is $\leq$ than the entropy of X (which intuitively makes sense: getting more information only reduces uncertainty, in expectation).

But importantly, the inequality is applied to averaged quantities.

It is still possible that there is new rare evidence y for which:

$$H(X) < H(X|Y = y)$$

Can you think of an exapmle?

Example taken from Cover, Thomas. Elements of Information Theory (book, 2nd ed). Theorem 2.6.5. https://doi.org/10.1002/047174882X
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Conditioning reduces entropy, in expectation

$H(X|Y) \leq H(X)$ (follows from $I(X;Y) = H(X) - H(X|Y) \geq 0$)

The nonnegativity of mutual information implies that on average the entropy of X conditioned on the observation $Y = y$ is $\leq$ than the entropy of X (which intuitively makes sense: getting more information only reduces uncertainty, in expectation).

But importantly, the inequality is applied to averaged quantities.

It is still possible that there is new rare evidence y for which:

$$H(X) < H(X|Y = y)$$

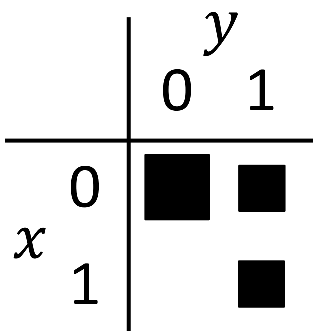
EXAMPLE: in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.

Example taken from Cover, Thomas. Elements of Information Theory (book, 2nd ed). Theorem 2.6.5. https://doi.org/10.1002/047174882X

New concrete evidence may increase entropy

EXAMPLE 6: Consider the joint ensemble (X, Y) with Boolean domains $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and following joint distribution.

$p(x,y)$	$y=0$	$y=1$
$x=0$	½	¼
$x=1$	0	¼



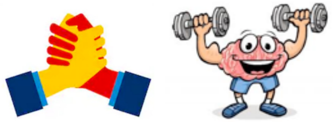
$H(X) =$?

$H(X|y=0) =$?

$H(X|y=1) =$?

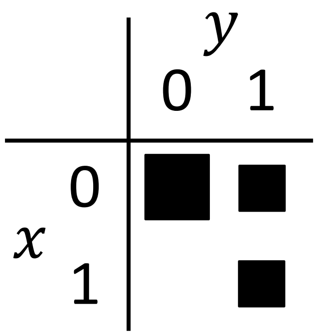
$H(X|Y) =$?

New concrete evidence may increase entropy



EXAMPLE 6: Consider the joint ensemble (X, Y) with Boolean domains $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and following joint distribution.

$p(x,y)$	$y=0$	$y=1$	Σ
$x=0$	½	¼	3/4
$x=1$	0	¼	1/4
Σ	1/2	1/2	



$$H(X) = \frac{3}{4}\lg\left(\frac{4}{3}\right) + \frac{1}{4}\lg(4) = 0.811$$

$$H(X|y=0) = 0$$

$$H(X|y=1) = 1$$

$$H(X|Y) = \frac{1}{2}\underbrace{H(X|y=0)}_{0} + \frac{1}{2}\underbrace{H(X|y=1)}_{1} = 0.5$$

$$\underbrace{H(X|y=0)}_{0} < \underbrace{H(X|Y)}_{0.5} \leq \underbrace{H(X)}_{0.811} < \underbrace{H(X|y=1)}_{1}$$

Three-term (multivariate) entropies, conditional mutual information, interaction information

Conditioning & chain rules

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

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

$$H(X,Y) \quad = \qquad ?$$

Two-variable chain rule

Conditioning & chain rules

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

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

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

Two-variable chain rule

$$H(X,Y|Z) = \quad ?$$

Conditional chain rule.

Conditioning & chain rules

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

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

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

Two-variable chain rule

$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z=z)]} = \quad ?$$

Notice the implied precedence rule

Conditional chain rule.

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditioning & chain rules

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

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

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

Two-variable chain rule

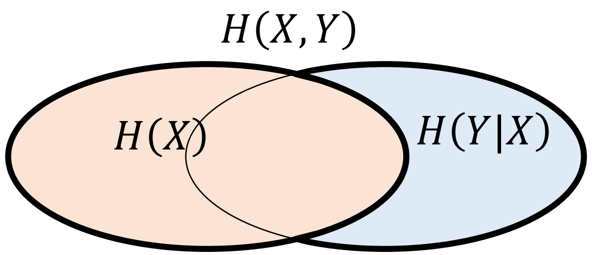
$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z = z)]} = H(X|Z) + H(Y|X,Z)$$

Notice the implied precedence rule

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditional chain rule.

Conditioning on an event creates a new probability space where the same probability concepts apply.



Conditioning & chain rules

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

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

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

Two-variable chain rule

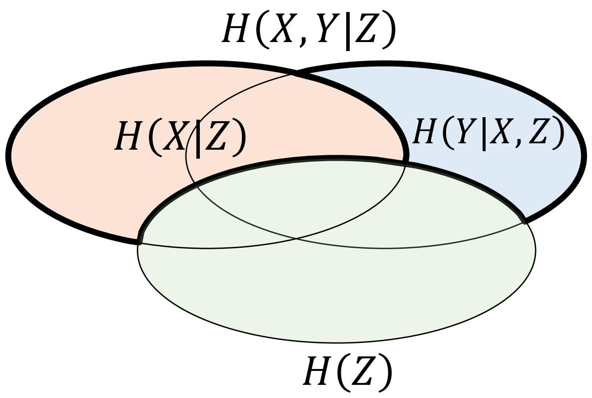
$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z=z)]} = H(X|Z) + H(Y|X,Z)$$

Notice the implied precedence rule

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditional chain rule.

Conditioning on an event creates a new probability space where the same probability concepts apply.



Conditioning & chain rules

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

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

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

Two-variable chain rule

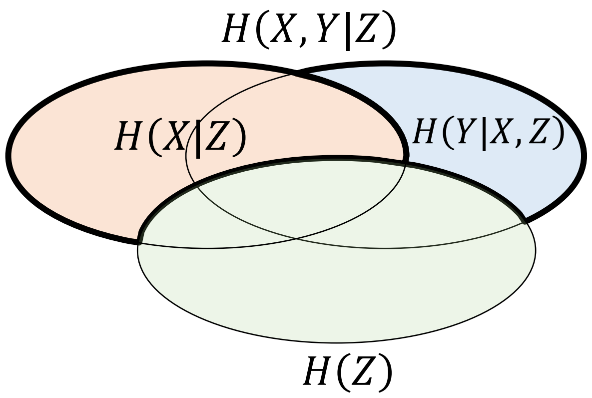
$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z=z)]} = H(X|Z) + H(Y|X,Z)$$

Notice the implied precedence rule

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditional chain rule.

Conditioning on an event creates a new probability space where the same probability concepts apply.



$$H(X,Y|Z) \; ? \; H(X|Z) + H(Y|Z)$$

$\leq$ or $\geq$

Conditioning & chain rules

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

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

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

Two-variable chain rule

$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z=z)]} = H(X|Z) + H(Y|X,Z)$$

Notice the implied precedence rule

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditional chain rule.

Conditioning on an event creates a new probability space where the same probability concepts apply.

$$H(X,Y|Z) \leq H(X|Z) + H(Y|Z)$$

Equality holds if X and Y are conditionally independent, given Z (Proof similar to unconditional case).

$$H(X,Y,Z) = \text{?}$$

Three-variable chain rule

Conditioning & chain rules

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

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

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

Two-variable chain rule

$$\underbrace{H(X,Y|Z)}_{\mathbb{E}_Z[H(X,Y|Z=z)]} = H(X|Z) + H(Y|X,Z)$$

Notice the implied precedence rule

Conditional joint entropy $H(X,Y|Z)$: expected joint entropy of X and Y together, given that Z is known

Conditional chain rule.

Conditioning on an event creates a new probability space where the same probability concepts apply.

$$H(X,Y|Z) \leq H(X|Z) + H(Y|Z)$$

Equality holds if X and Y are conditionally independent, given Z (Proof similar to unconditional case).

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

Three-variable chain rule

Conditional mutual information & interaction information

$$\underbrace{I(X;Y|Z)}_{\mathbb{E}_Z[I(X;Y|Z=z)]} = \quad ?$$

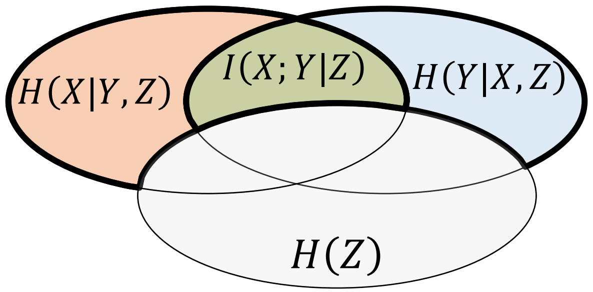
Conditional mutual information $I(X;Y|Z)$:
expected mutual information of X and Y,
given Z is known

Conditional mutual information & interaction information

$$\underbrace{I(X;Y|Z)}_{\mathbb{E}_Z[I(X;Y|Z=z)]} = H(X|Z) + H(Y|Z) - \underbrace{H(X,Y|Z)}_{H(Y|Z) + \underbrace{H(X|Y,Z)}_{H(X|(Y,Z))}}$$

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

Conditional mutual information $I(X;Y|Z)$: expected mutual information of X and Y, given Z is known



$$J(X;Y;Z) = \quad ?$$

Interaction information (often called "mutual information"*): measures the negated influence of a variable Z on the amount of information shared between X and Y.

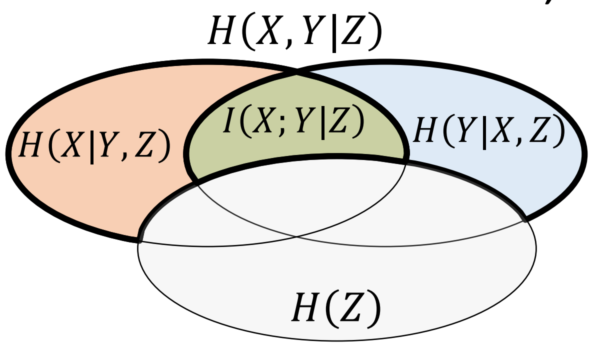
* Alternative notations include $\mathcal{J}(X;Y;Z)$ and $R(X;Y;Z)$. We don't recommend calling it "mutual information" and thus also replace the more common notation $I(X;Y;Z)$ with $J(X;Y;Z)$. Some sources prefer not to even define that measure at all (we will discuss the reasons in a bit) https://en.wikipedia.org/wiki/Interaction_information .

Conditional mutual information & interaction information

$$\underbrace{I(X;Y|Z)}_{\mathbb{E}_Z[I(X;Y|Z=z)]} = H(X|Z) + H(Y|Z) - \underbrace{H(X,Y|Z)}_{H(Y|Z)+H(X|Y,Z)}$$

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

Conditional mutual information $I(X;Y|Z)$: expected mutual information of X and Y, given Z is known



$$J(X;Y;Z) = I(X;Y) - I(X;Y|Z)$$

Interaction information (often called "mutual information"*): measures the negated influence of a variable Z on the amount of information shared between X and Y.

Is it symmetric in all the variables ?

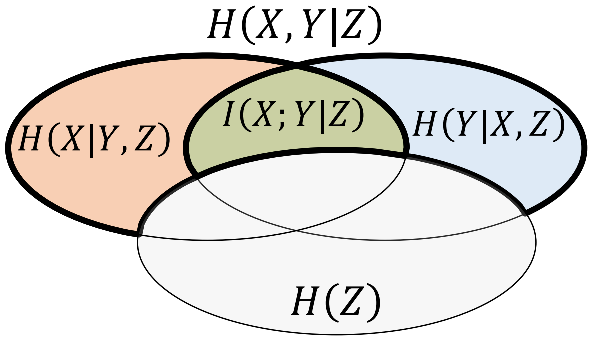
* Alternative notations include $\mathcal{J}(X;Y;Z)$ and $R(X;Y;Z)$. We don't recommend calling it "mutual information" and thus also replace the more common notation $I(X;Y;Z)$ with $J(X;Y;Z)$. Some sources prefer not to even define that measure at all (we will discuss the reasons in a bit) https://en.wikipedia.org/wiki/Interaction_information .

Conditional mutual information & interaction information

$$\underbrace{I(X;Y|Z)}_{\mathbb{E}_Z[I(X;Y|Z=z)]} = H(X|Z) + H(Y|Z) - \underbrace{H(X,Y|Z)}_{H(Y|Z)+H(X|Y,Z)}$$

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

Conditional mutual information $I(X;Y|Z)$: expected mutual information of X and Y, given Z is known



$$J(X;Y;Z) = I(X;Y) - I(X;Y|Z)$$

Interaction information (often called "mutual information"*): measures the negated influence of a variable Z on the amount of information shared between X and Y.

$$= \overbrace{H(X) - H(X|Y)} - \overbrace{(H(X|Z) - H(X|Y,Z))}$$

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

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

(...) thus symmetric in all 3 variables!

* Alternative notations include $\mathcal{J}(X;Y;Z)$ and $R(X;Y;Z)$. We don't recommend calling it "mutual information" and thus also replace the more common notation $I(X;Y;Z)$ with $J(X;Y;Z)$. Some sources prefer not to even define that measure at all (we will discuss the reasons in a bit) https://en.wikipedia.org/wiki/Interaction_information .

Interaction information example


PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z
0	0	0
0	1	1
1	0	1
1	1	0

Interaction information example

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼
0	0	1	0
…	…	…	0

Thus any 2 variables functionally determine the 3rd, e.g. $(x, z) \rightarrow y$!

Interaction information example

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼
0	0	1	0
…	…	…	0

Thus any 2 variables functionally determine the 3rd, e.g. $(x, z) \rightarrow y$!

$H(X) =$?

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

$H(X|Y,Z) =$
$I(X;Y|Z) =$?

$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) =$?

Interaction information example

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

$H(X) = 1$

Similarly, $H(Y) = 1$ and $H(Z) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

Similarly, all variables are pairwise independent

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$

Thus, if Z is observed, then X and Y become dependent:
(knowing $X = x$ and $Z = z$, tells you what Y is: $y = z - x \bmod 2$)

Thus the conditional mutual information is bigger than the unconditional mutual information: $I(X;Y|Z) > I(X;Y)$

$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \text{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

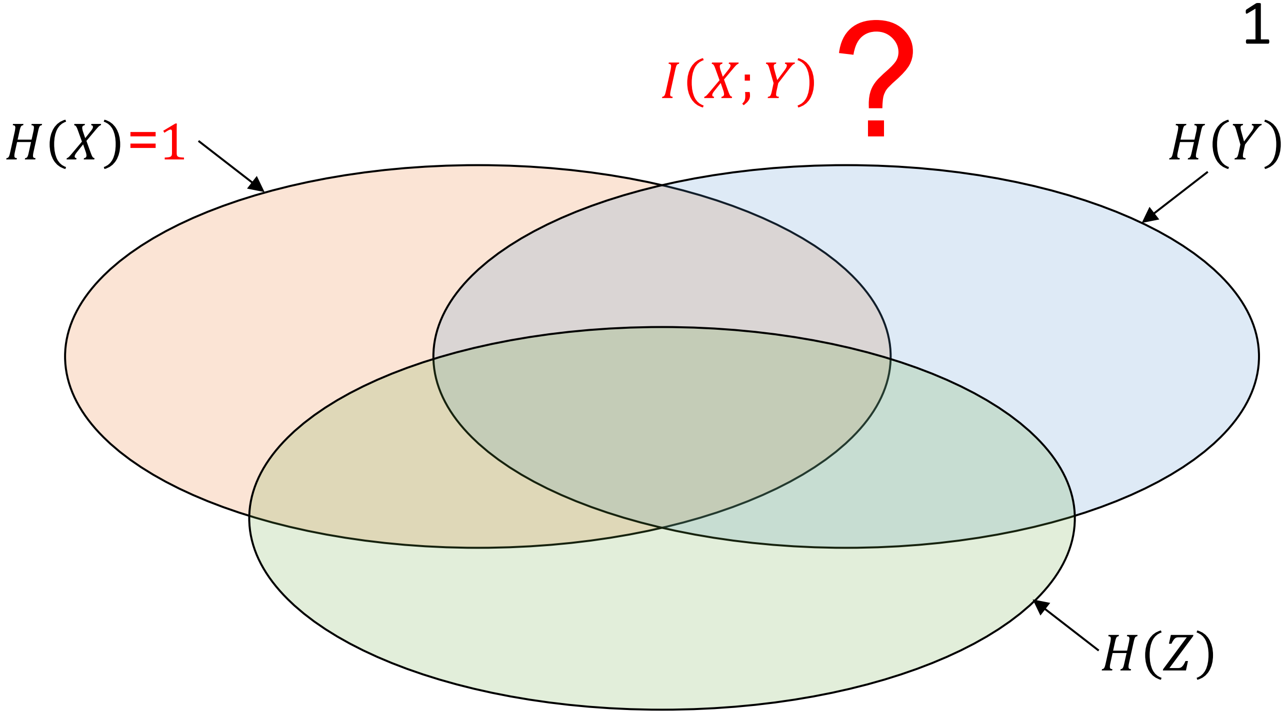
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

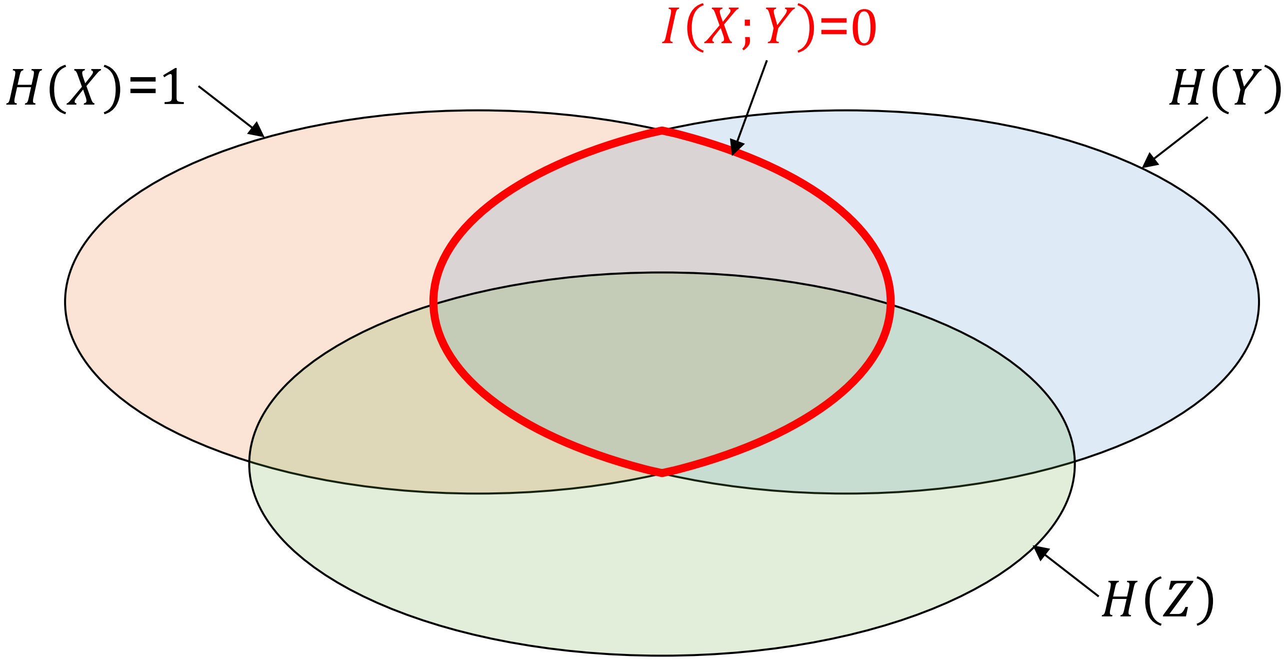
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

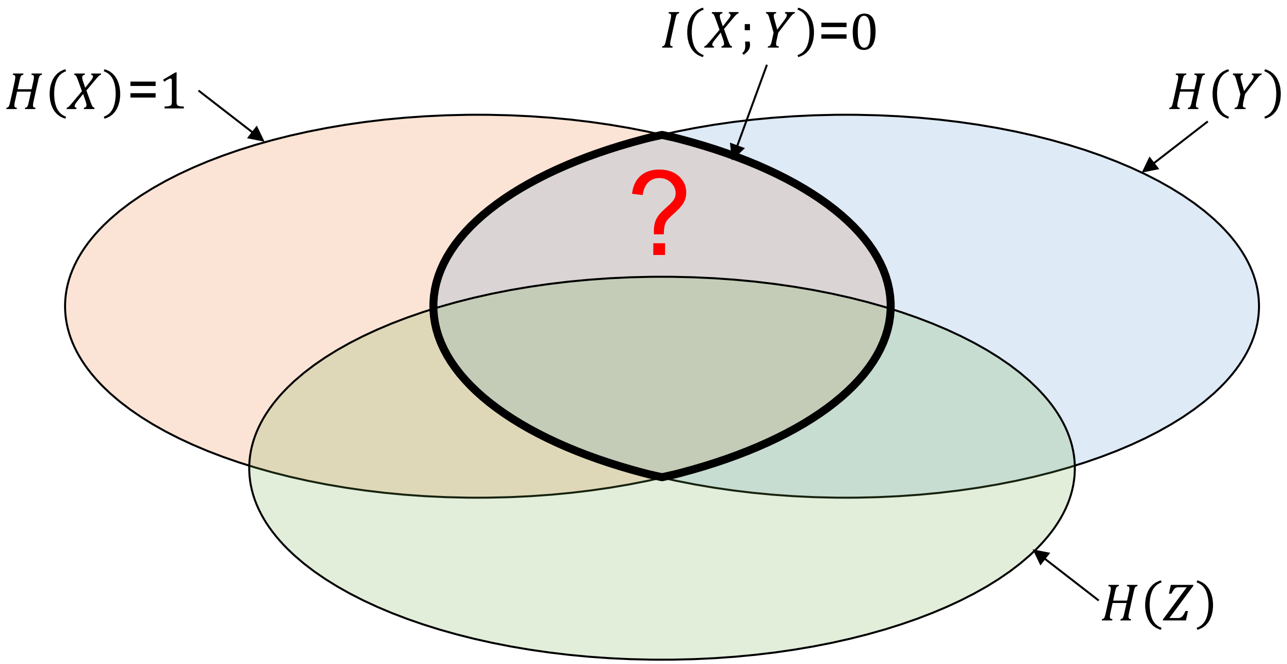
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

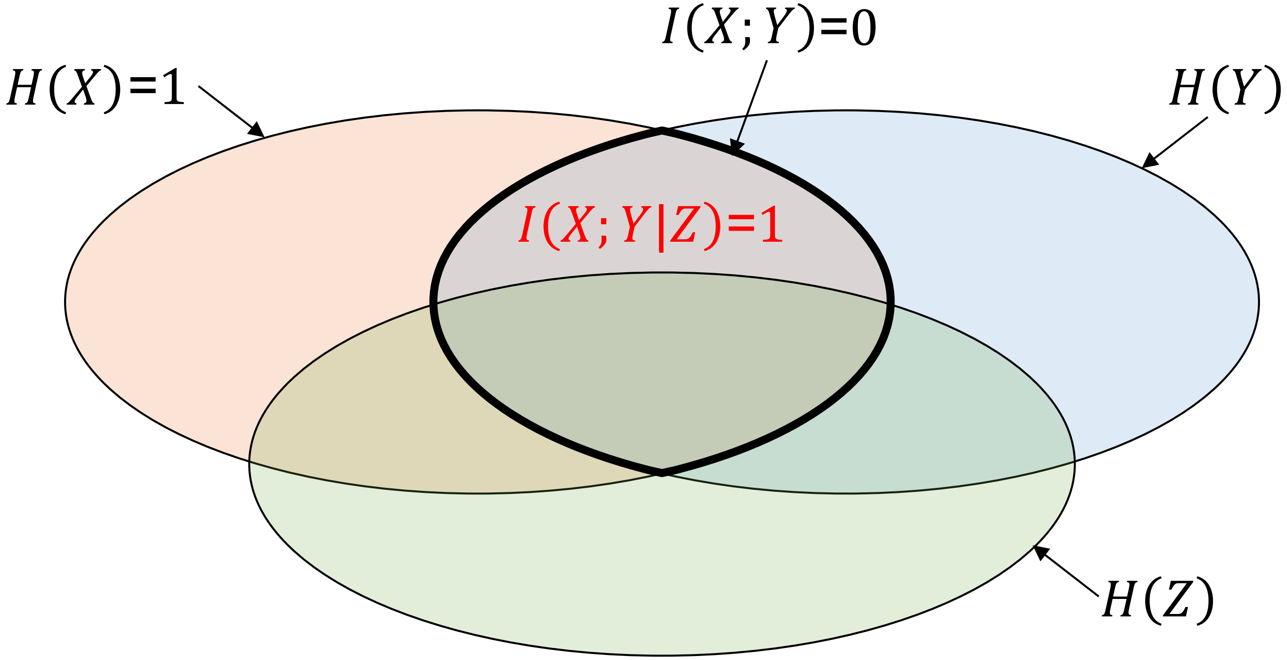
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

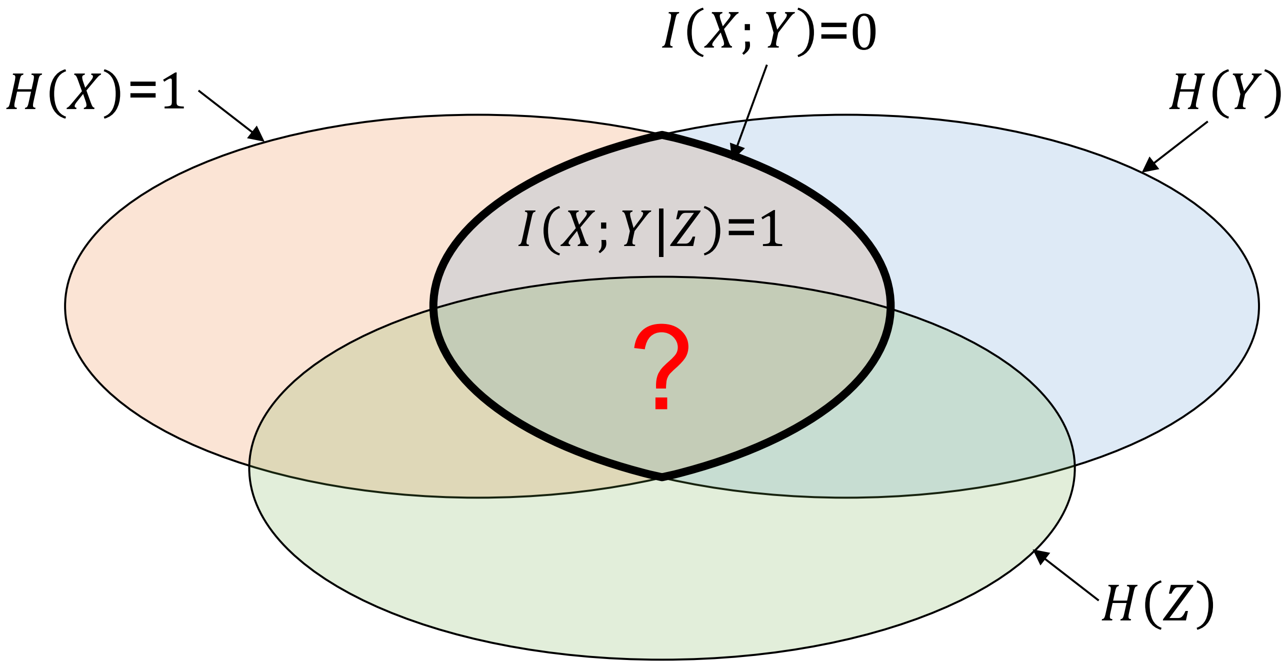
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

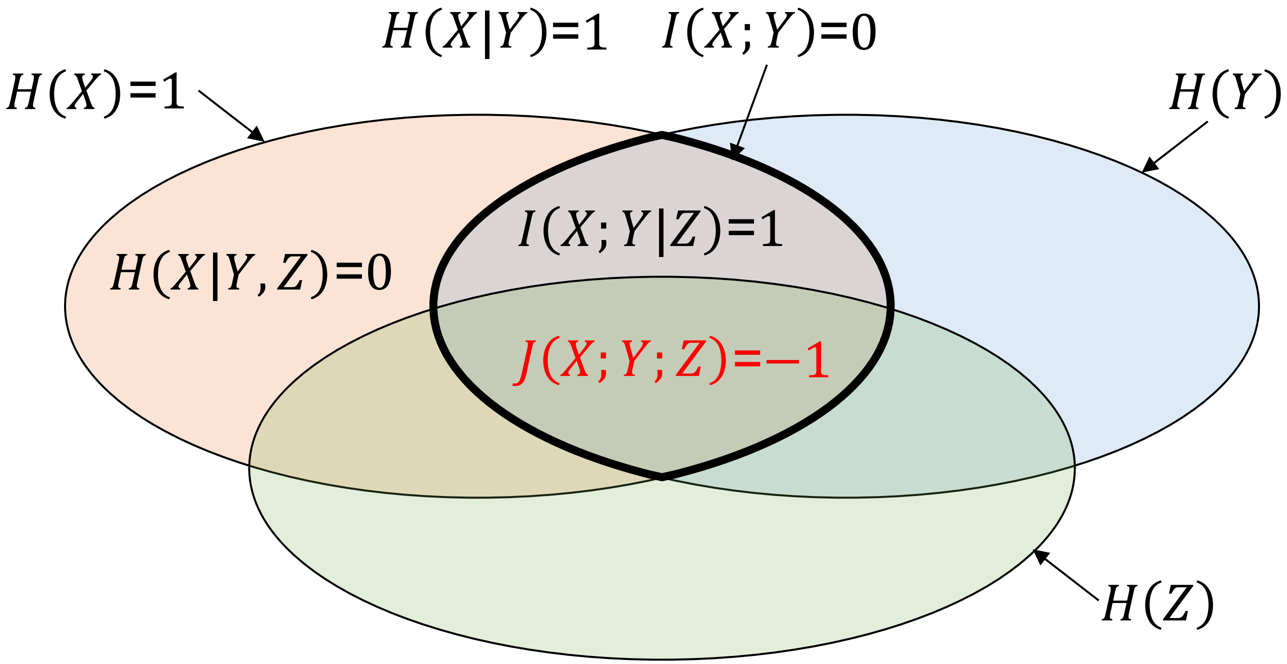
$H(X) = 1$

$H(X|Y) = H(X) = 1$

$I(X;Y) = 0$

$H(X|Y,Z) = 0$

$I(X;Y|Z) = 1$



$J(X;Y;Z) = I(X;Y) - I(X;Y|Z) = -1$

⇒ VENN diagrams applied to joint entropies with ≥ 2 variables can mislead

[MacKay'02] discourages VENN diagrams and does not name $J(X;Y;Z)$

Figure 8.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy.

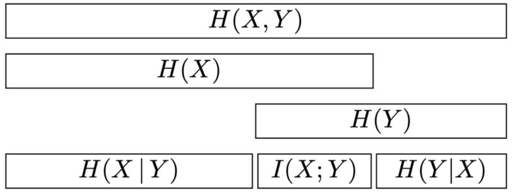
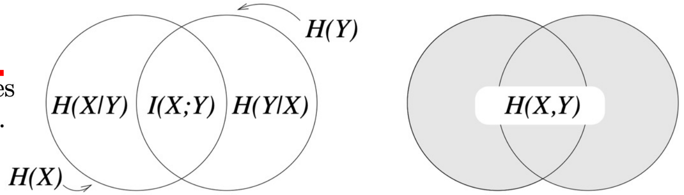


Figure 8.2. A misleading representation of entropies (contrast with figure 8.1).



Exercise 8.8.[3, p.143] Many texts draw figure 8.1 in the form of a Venn diagram (figure 8.2). Discuss why this diagram is a misleading representation of entropies. Hint: consider the three-variable ensemble XYZ in which $x \in \{0,1\}$ and $y \in \{0,1\}$ are independent binary variables and $z \in \{0,1\}$ is defined to be $z = x + y \bmod 2$. ...

Source: Section 8 from [MacKay'02]. Information Theory, Inference, and learning Algorithms. Cambridge University Press, 2002." https://www.inference.org.uk/itprnn/book.pdf

[MacKay'02] discourages VENN diagrams and does not name $I(X;Y;Z)$

Figure 8.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy.

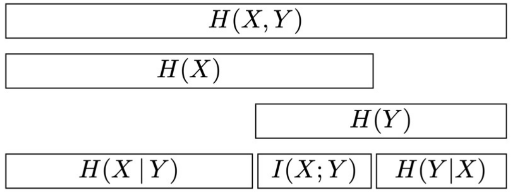
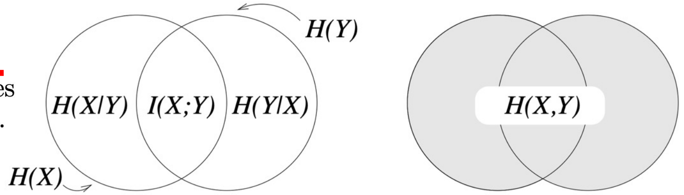
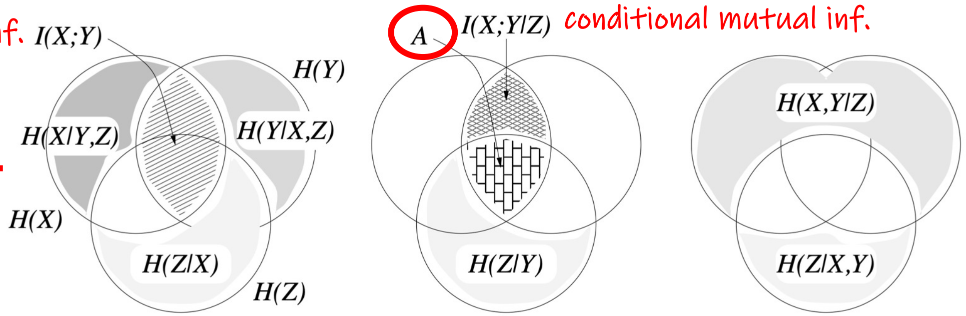


Figure 8.2. A misleading representation of entropies (contrast with figure 8.1).



unconditional mutual inf.

Figure 8.3. A misleading representation of entropies, continued.



Exercise 8.8.[3, p.143] Many texts draw figure 8.1 in the form of a Venn diagram (figure 8.2). Discuss why this diagram is a misleading representation of entropies. Hint: consider the three-variable ensemble XYZ in which $x \in \{0,1\}$ and $y \in \{0,1\}$ are independent binary variables and $z \in \{0,1\}$ is defined to be $z = x + y \bmod 2$. ...

Solution to exercise 8.8 (p.141). The depiction of entropies in terms of Venn diagrams is misleading for at least two reasons.

First, one is used to thinking of Venn diagrams as depicting sets; but what are the 'sets' $H(X)$ and $H(Y)$ depicted in figure 8.2, and what are the objects that are members of those sets? I think this diagram encourages the novice student to make inappropriate analogies. For example, some students imagine that the random outcome (x, y) might correspond to a point in the diagram, and thus confuse entropies with probabilities.

The conditional mutual information between X and Y given Z is the average over z of the above conditional mutual information.

$$I(X;Y\,|\,Z) = H(X\,|\,Z) - H(X\,|\,Y,Z). \tag{8.10}$$

No other 'three-term entropies' will be defined. For example, expressions such as $I(X;Y;Z)$ and $I(X\,|\,Y;Z)$ are illegal. But you may put conjunctions of arbitrary numbers of variables in each of the three spots in the expression $I(X;Y\,|\,Z)$ – for example, $I(A,B;C,D\,|\,E,F)$ is fine: it measures how much information on average c and d convey about a and b, assuming e and f are known.

Secondly, the depiction in terms of Venn diagrams encourages one to believe that all the areas correspond to positive quantities. In the special case of two random variables it is indeed true that $H(X\,|\,Y)$, $I(X;Y)$ and $H(Y\,|\,X)$ are positive quantities. But as soon as we progress to three-variable ensembles, we obtain a diagram with positive-looking areas that may actually correspond to negative quantities. Figure 8.3 correctly shows relationships such as

$$H(X) + H(Z\,|\,X) + H(Y\,|\,X,Z) = H(X,Y,Z). \tag{8.31}$$

But it gives the misleading impression that the conditional mutual information $I(X;Y\,|\,Z)$ is *less than* the mutual information $I(X;Y)$. In fact the area labelled A can correspond to a *negative* quantity. Consider the joint ensemble (X,Y,Z) in which $x \in \{0,1\}$ and $y \in \{0,1\}$ are independent binary variables and $z \in \{0,1\}$ is defined to be $z = x + y \bmod 2$. Then clearly $H(X) = H(Y) = 1$ bit. Also $H(Z) = 1$ bit. And $H(Y\,|\,X) = H(Y) = 1$ since the two variables are independent. So the mutual information between X and Y is zero. $I(X;Y) = 0$. However, if z is observed, X and Y become dependent — knowing x, given z, tells you what y is: $y = z - x \bmod 2$. So $I(X;Y\,|\,Z) = 1$ bit. Thus the area labelled A must correspond to -1 bits for the figure to give the correct answers.

The above example is not at all a capricious or exceptional illustration. The binary symmetric channel with input X, noise Y, and output Z is a situation in which $I(X;Y) = 0$ (input and noise are independent) but $I(X;Y\,|\,Z) > 0$ (once you see the output, the unknown input and the unknown noise are intimately related!).

The Venn diagram representation is therefore valid only if one is aware that positive areas may represent negative quantities. With this proviso kept in mind, the interpretation of entropies in terms of sets can be helpful (Yeung, 1991).

[Cover,Thomas'06] on three-term entropies

2.25 *Venn diagrams*. There isn't really a notion of mutual information common to three random variables. Here is one attempt at a definition: Using Venn diagrams, we can see that the mutual information common to three random variables X, Y, and Z can be defined by

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z).$$

This quantity is symmetric in X, Y, and Z, despite the preceding asymmetric definition. Unfortunately, $I(X;Y;Z)$ is not necessarily nonnegative. Find X, Y, and Z such that $I(X;Y;Z) < 0$, and prove the following two identities:

(a) $I(X;Y;Z) = H(X,Y,Z) - H(X) - H(Y) - H(Z) + I(X;Y) + I(Y;Z) + I(Z;X)$.

(b) $I(X;Y;Z) = H(X,Y,Z) - H(X,Y) - H(Y,Z) - H(Z,X) + H(X) + H(Y) + H(Z)$.

The first identity can be understood using the Venn diagram analogy for entropy and mutual information. The second identity follows easily from the first.

Source: [Cover, Thomas'06]. Elements of Information Theory (book, 2nd ed). Theorem 2.6.5. https://doi.org/10.1002/047174882X

[Yeung'08] disagrees and heavily uses "information diagrams"

3.5 Information Diagrams

We have established in Section 3.3 a one-to-one correspondence between Shannon's information measures and set theory. Therefore, it is valid to use an *information diagram*, which is a variation of a Venn diagram, to represent the relationship between Shannon's information measures.

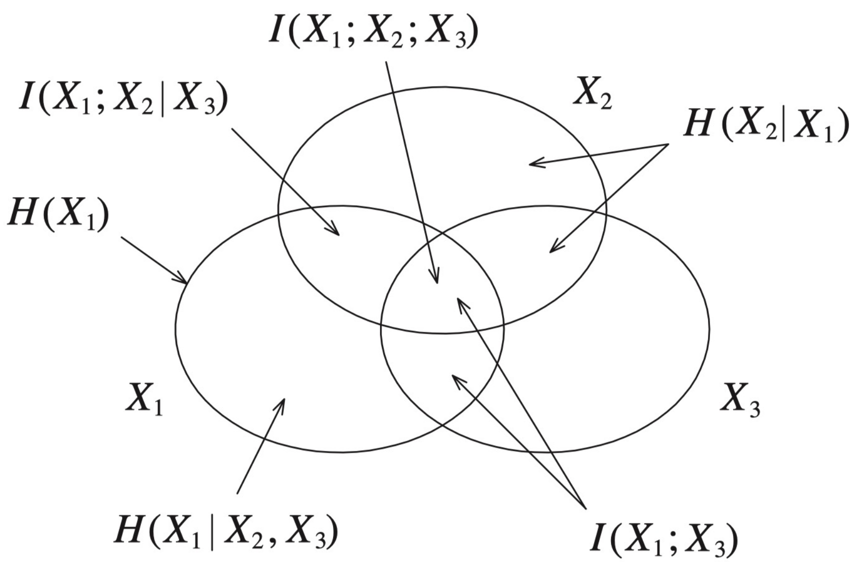


Fig. 3.4. The generic information diagram for X_1, X_2, and X_3.

Sources: Yeung, Information Theory and Network Coding, 2008. Section 3.5 Information Diagrams. http://iest2.ie.cuhk.edu.hk/~whyeung/tempo/main2.pdf
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

[Abramson 1963]

Although Figure 5-19 is an important aid in remembering relationships among the quantities we have defined, it can also be somewhat deceptive. The mutual information $I(A; B)$ was shown to be nonnegative; the mutual information $I(A; B; C)$, however, can be negative. This means that the intersection of the three circles of Figure 5-19*a* can be *negative!* To show this we present an example.

parity example

Example 5-12. Consider the three binary alphabets A, B, C. Let a_i and b_j be selected as 0 or 1, each with probability $\frac{1}{2}$ and each independently of the other. Finally, we assume that c_k is selected as 0 if a_i equals b_j and as 1 if a_i does not equal b_j. Some of the probabilities of these three random variables are given in Table 5-3.

Using this table, we calculate

$$
\begin{aligned}
I(A; B) &= 0 \text{ bits} \\
I(A; B/C) &= 1 \text{ bit} \\
I(A; B; C) &= I(A; B) - I(A; B/C) = -1 \text{ bit}
\end{aligned}
$$

It is clear why we get such an answer. Since A and B are statistically independent, $I(A; B) = 0$, and B provides no information about A. If we already know C, however, learning B tells us which A was chosen, and therefore provides us with one bit of information.

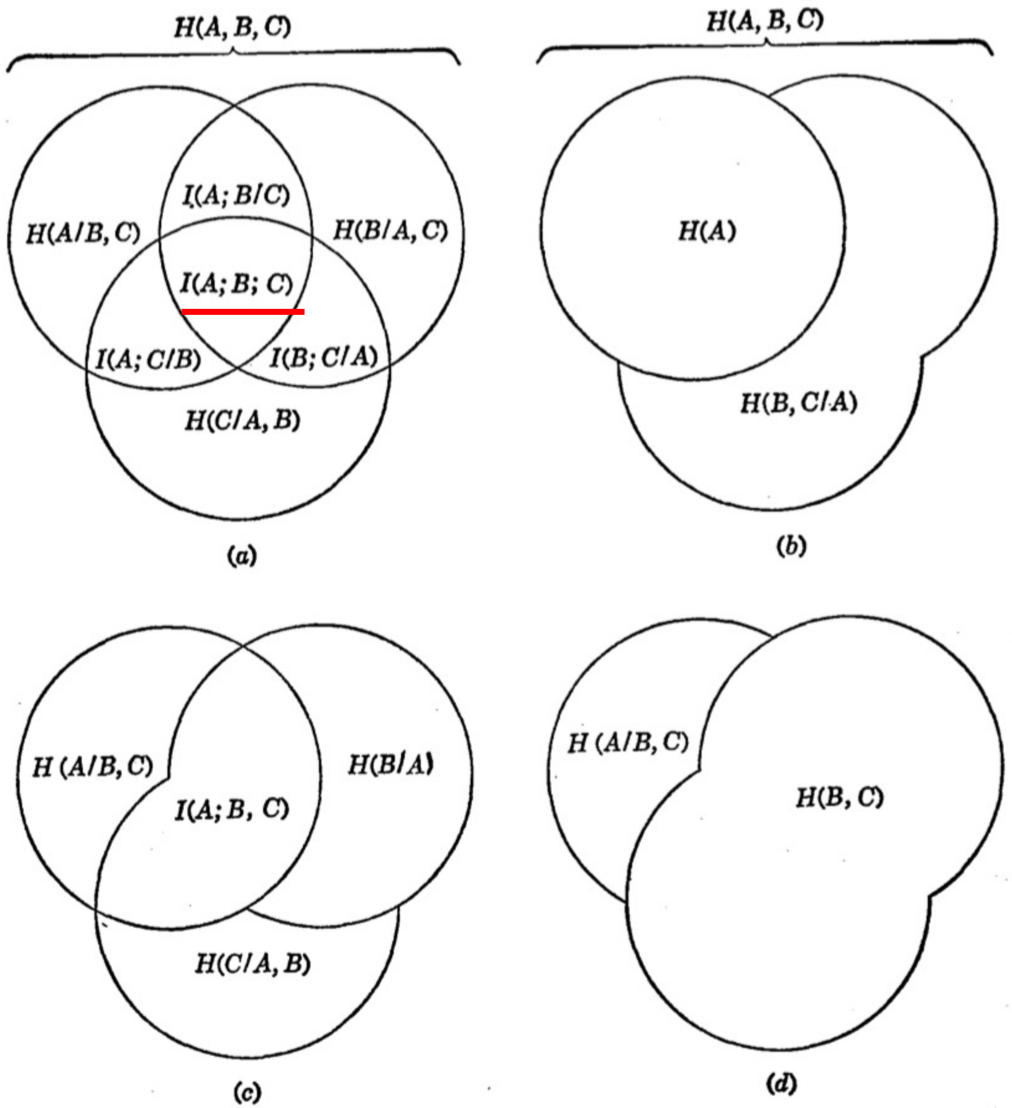


FIGURE 5-19. Some information relationships.

Sources: Abramson. Information Theory and Coding, 1963. https://archive.org/details/informationtheor0000abra
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

[Ellerman'21] does not like negative values and changes the definitions completely

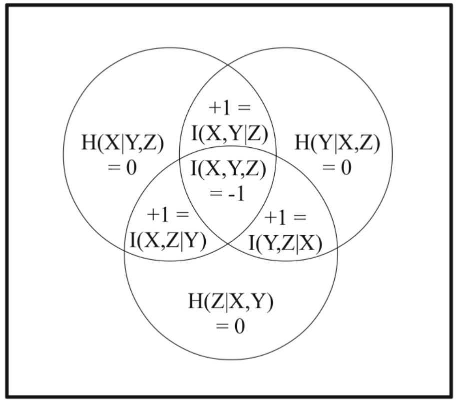
2.2 Logical Entropy, Not Shannon Entropy, Is a (Non-negative) Measure

4.2 An Example of Negative Mutual Information for Shannon Entropy

Norman Abramson gives an example [1, pp. 130–131] where the Shannon mutual information of three variables is negative.[3] William Feller gives a similar concrete example that we will use [11, Exercise 26, p. 143]. Any probability theory textbook example to show that pair-wise independence does not imply mutual independence for three or more random variables would do as well.

One fair die is thrown first and the result is recorded odd as 1 (the number of the face up mod 2) or even as 0. Then the same is done with a second fair die so the outcome space if $U = \{(0,0), (0,1), (1,0), (1,1)\} = \{0,1\} \times \{0,1\}$ (first die on the left and second die on the right). Let X be the random variable for the outcome (0 or 1) of the first throw, Y for the second throw, and Z for the sum $X + Y$ mod 2. Since Z is a function of X and Y, the outcome space is $U \times U = (X \times Y)^2$. So many Venn diagrams are illustrated rather symbolically, e.g., with circles for $h(X)$, $h(Y)$, and $h(Z)$, that it will be useful to give the actual Venn/box diagrams for this example.

Fig. 4.6 Negative 'area' $I(X, Y, Z)$ in Venn diagram



It is unclear how negative mutual information for three variables can be interpreted. Indeed, as Imre Csiszar and Janos Körner remark:

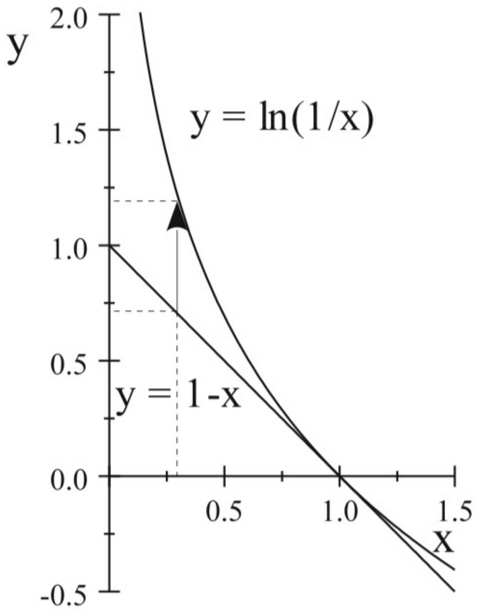
> The set-function analogy might suggest to introduce further information quantities corresponding to arbitrary Boolean expressions of sets. E.g., the "information quantity" corresponding to $\mu(A \cap B \cap C) = \mu(A \cap B) - \mu((A \cap B) - C)$ would be $I(X, Y) - I(X, Y|Z)$; this quantity has, however, no natural intuitive meaning. [7, pp. 53–4]

Sources: Ellerman. New Foundations for Information Theory: Logical Entropy and Shannon Entropy, Springer, 2021. https://doi.org/10.1007/978-3-030-86552-8
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

[Ellerman'21] does not like negative values and changes the definitions completely

2.2 Logical Entropy, Not Shannon Entropy, Is a (Non-negative) Measure

Fig. 2.2 The dit-bit transform $1 - p \rightsquigarrow \ln\left(\frac{1}{p}\right)$ (natural logs)



Abstract This book presents a new foundation for information theory where the notion of information is defined in terms of distinctions, differences, distinguishability, and diversity. The direct measure is logical entropy which is the quantitative measure of the distinctions made by a partition. Shannon entropy is a transform or re-quantification of logical entropy for Claude Shannon's "mathematical theory of communications." The interpretation of the logical entropy of a partition is the two-draw probability of getting a distinction of the partition (a pair of elements distinguished by the partition) so it realizes a dictum of Gian-Carlo Rota: $\frac{Probability}{Subsets} \approx \frac{Information}{Partitions}$. Andrei Kolmogorov suggested that information should be defined independently of probability, so logical entropy is first defined in terms of the set of distinctions of a partition and then a probability measure on the set defines the quantitative version of logical entropy. We give a history of the logical entropy formula that goes back to Corrado Gini's 1912 "index of mutability" and has been rediscovered many times.

Keywords Information-as-distinctions · Logical entropy · History of the formula

Sources: Ellerman. New Foundations for Information Theory: Logical Entropy and Shannon Entropy, Springer, 2021. https://doi.org/10.1007/978-3-030-86552-8
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

[Ellerman'21] does not like negative values and changes the definitions completely

1.5 Brief History of the Logical Entropy Formula

The logical entropy formula $\boxed{h(p) = \sum_i p_i (1 - p_i) = 1 - \sum_i p_i^2}$ is the probability of getting distinct values $u_i \neq u_j$ in two independent samplings of the random variable u. The complementary measure $1 - h(p) = \sum_i p_i^2$ is the probability that the two drawings yield the same value from U. Thus $1 - \sum_i p_i^2$ is a measure of heterogeneity or diversity in keeping with our theme of information as distinctions, while the complementary measure $\sum_i p_i^2$ is a measure of homogeneity or concentration. Historically, the formula can be found in either form depending on the particular context. The p_i's might be relative shares such as the relative share of organisms of the ith species in some population of organisms, and then the interpretation of p_i as a probability arises by considering the random choice of an organism from the population.

According to I. J. Good, the formula has a certain naturalness: "If $p_1, \ldots, p_t$ are the probabilities of t mutually exclusive and exhaustive events, any statistician of this century who wanted a measure of homogeneity would have take about two seconds to suggest $\sum p_i^2$ which I shall call ρ." [13, p. 561] As noted by Bhargava and Uppuluri [4], the formula $1 - \sum p_i^2$ was used by Gini in 1912 [10] as a measure of "mutability" or diversity. But another development of the formula

"index of mutability" and has been rediscovered many times. In addition to being defined as a (probability) measure in the sense of measure theory (unlike Shannon entropy), logical entropy is always non-negative (unlike three-way Shannon mutual information) and finitely-valued for countable distributions. One perhaps surprising result is that in spite of decades of MaxEntropy efforts based maximizing Shannon entropy, the maximization of logical entropy gives a solution that is closer to the uniform distribution in terms of the usual (Euclidean) notion of distance. When

In sum, the main argument of the book is that information is about distinctions, differences, distinguishability, and diversity, and that logical entropy is its direct measure—while Shannon entropy is a requantification of information-as-distinctions that is fundamental for the theory of coding and communications.

Sources: Ellerman. New Foundations for Information Theory: Logical Entropy and Shannon Entropy, Springer, 2021. https://doi.org/10.1007/978-3-030-86552-8
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

EXAMPLE 3 (CONTINUED):

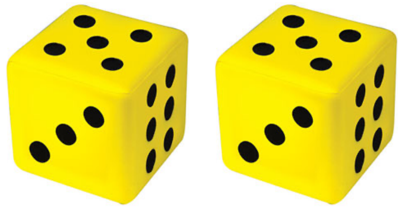
roll two fair dice with 6 sides

$$\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$$

A = "1st roll is odd"

B = "2nd roll is odd"

C = "sum of rolls is odd"



second die

first die	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,2)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

EXAMPLE 3 (CONTINUED):

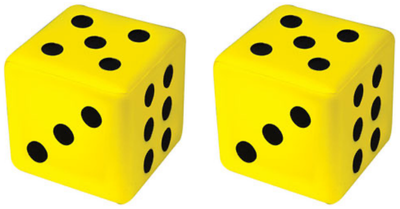
roll two fair dice with 6 sides

$$\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$$

A = "1st roll is odd"

B = "2nd roll is odd"

C = "sum of rolls is odd"



second die

first die	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,2)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

When VENN diagrams confuse more than help (?)

PARITY EXAMPLE: Consider the joint ensemble (X, Y, Z) with Boolean domains $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$. X and Y are independent uniform binary variables. And let Z be the XOR of X and Y: $z = \mathrm{XOR}(x, y)$, or equally, $z = x + y \bmod 2$.

x	y	z	p
0	0	0	¼
0	1	1	¼
1	0	1	¼
1	1	0	¼

EXAMPLE 3 (CONTINUED):

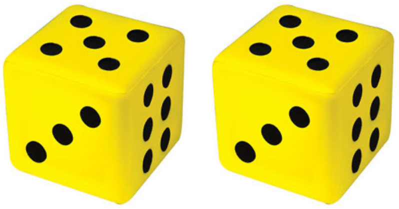
roll two fair dice with 6 sides

$$\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$$

A = "1st roll is odd"

B = "2nd roll is odd"

C = "sum of rolls is odd"



second die

first die	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

> 26. *Pairwise but not totally independent events.* Two dice are thrown and three events are defined as follows: A means "odd face with first die"; B means "odd face with second die"; finally, C means "odd sum" (one face even, the other odd). If each of the 36 sample points has probability $\frac{1}{36}$, then any two of the events are independent. The probability of each is $\frac{1}{2}$. Nevertheless, the three events cannot occur simultaneously.

Fig source: Feller. An Introduction to Probability Theory and its Applications, Volume I, 1968. Chapter V. https://archive.org/details/introductiontopr0001fell/page/n7/mode/2up

Interaction information

$J(X;Y;Z)$

- measures the negated influence of a variable Z on the mutual information between X and Y.* (And it is symmetric across the variables)
- It is positive when Z decreases/inhibits (i.e., accounts for or explains some of) the correlation between X and Y (e.g., that happens in Markov Chains).
- It is negative when Z increases/facilitates the correlation (e.g., when X and Y are independent but not conditionally independent given Z, that's our last parity example).

$$J(X;Y;Z) = I(X;Y) - I(X;Y|Z)$$

can be used to define it recursively
for more than 3 variables

$$= H(X) + H(Y) + H(Z) - \big(H(X,Y) + H(X,Z) + H(Y,Z)\big) + H(X,Y,Z)$$

* Alternative notations include $\mathcal{J}(X;Y;Z)$ and $R(X;Y;Z)$. We don't recommend calling it "mutual information" and thus also replace the more common notation $I(X;Y;Z)$ with $J(X;Y;Z)$.
For more details, see https://en.wikipedia.org/wiki/Interaction_information .

Updated 10/1/2025

Part 1: Theory
L07: Basics of entropy (5/7)
[interaction information, Markov chains, data processing inequality]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

9/29/2025

Pre-class conversations

- Last class recapitulation
- Slide decks / Piazza posts: please continue checking for errors / inconsistencies / unclear details etc. (e.g. earlier incorrect links)
- I am working through scribes (to be renamed to mini projects). For your next mini projects, Python scripts could be part!

- Today:
  - Markov Chains, Data Processing inequality
  - Next time: sufficient statistics, information inequalities

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

$I(X;Y) = \sum_{x,y} p_{xy} \cdot \lg\left(\frac{p_{xy}}{p_{x*} \cdot p_{*y}}\right)$

$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$

moment-based dependencies
that capture linearity

$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: 2 RVs $X, Y \in \{0,1\}$

$$I(X;Y) = \sum_{x,y} p_{xy} \cdot \lg\left(\frac{p_{xy}}{p_{x*} \cdot p_{*y}}\right)$$

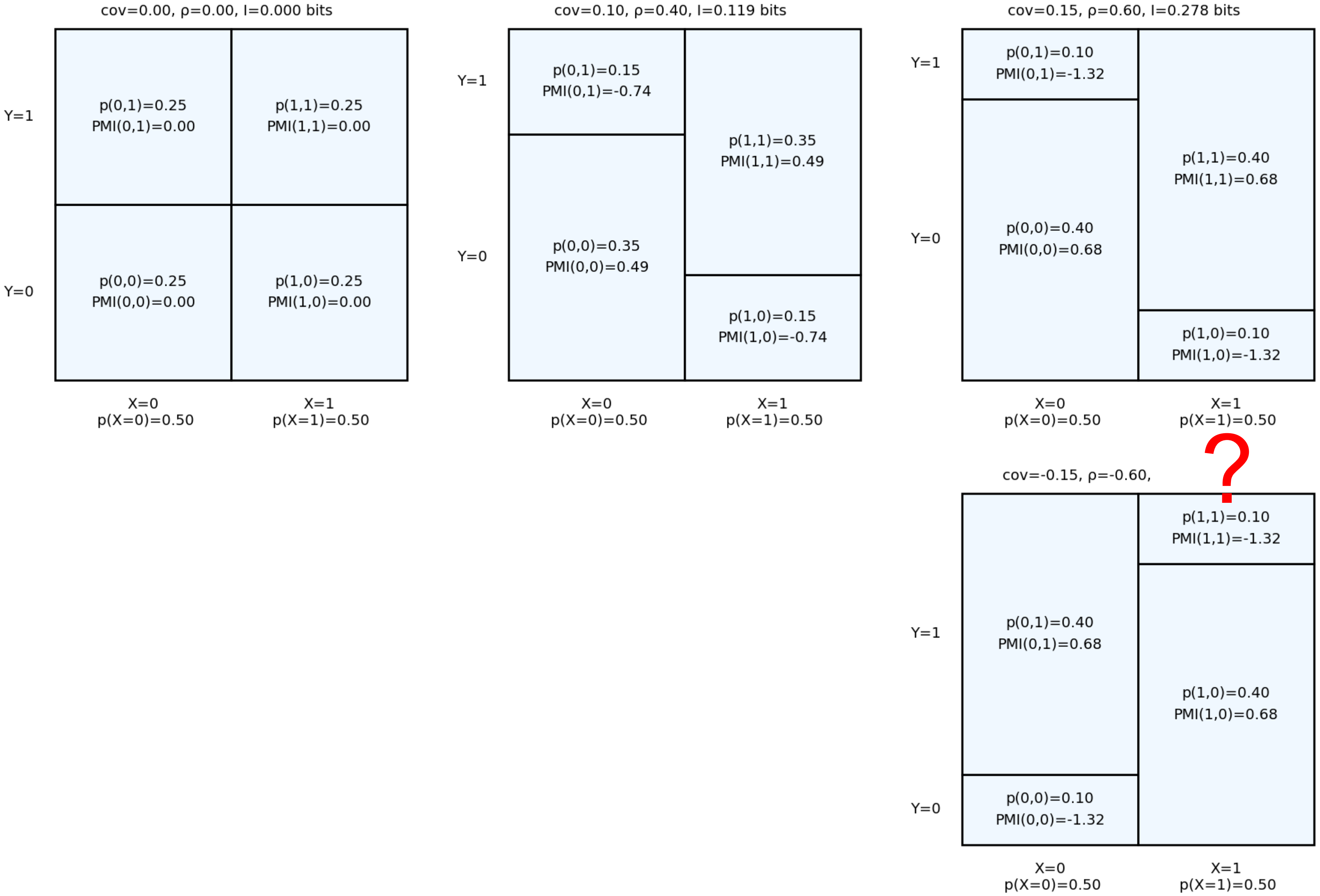
$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$= p_{11} - p_{1*} \cdot p_{*y}$$

moment-based dependencies
that capture linearity

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\sqrt{p_{1*} \cdot \overline{p_{1*}} \cdot p_{*1} \cdot \overline{p_{*1}}}}$$



Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: 2 RVs $X, Y \in \{0,1\}$

$$I(X;Y) = \sum_{x,y} p_{xy} \cdot \lg\left(\frac{p_{xy}}{p_{x*} \cdot p_{*y}}\right)$$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$= p_{11} - p_{1*} \cdot p_{*y}$$

moment-based dependencies that capture linearity

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\sqrt{p_{1*} \cdot \overline{p_{1*}} \cdot p_{*1} \cdot \overline{p_{*1}}}}$$

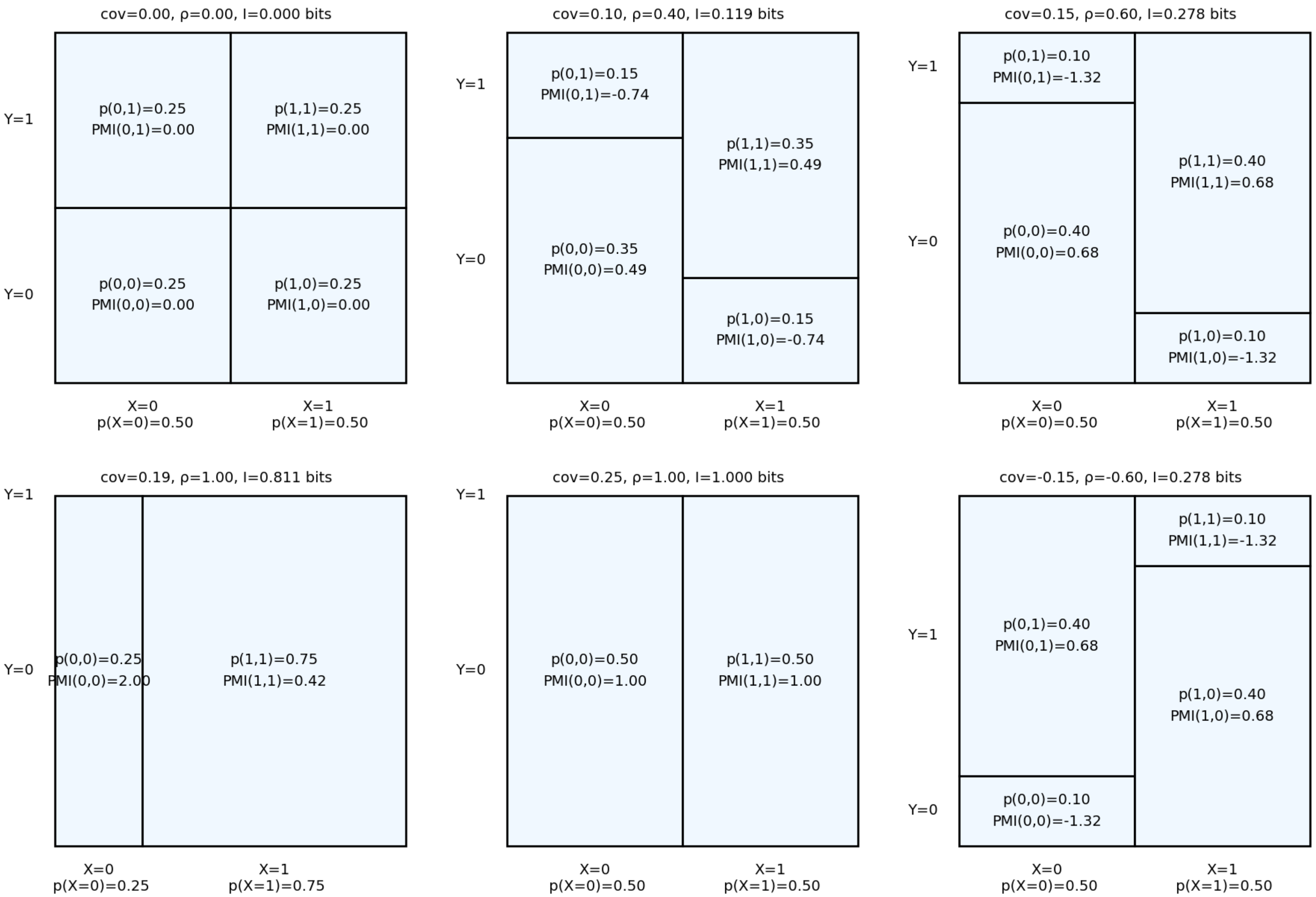


Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: 2 RVs $X, Y \in \{0,1\}$

$$I(X;Y) = \sum_{x,y} p_{xy} \cdot \lg\left(\frac{p_{xy}}{p_{x*} \cdot p_{*y}}\right)$$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$= p_{11} - p_{1*} \cdot p_{*y}$$

moment-based dependencies that capture linearity

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\sqrt{p_{1*} \cdot \overline{p_{1*}} \cdot p_{*1} \cdot \overline{p_{*1}}}}$$

Is it possible to have a non-linear dependence, and thus $\mathrm{Cov}(X,Y) = 0$, but $I(X;Y) > 0$?

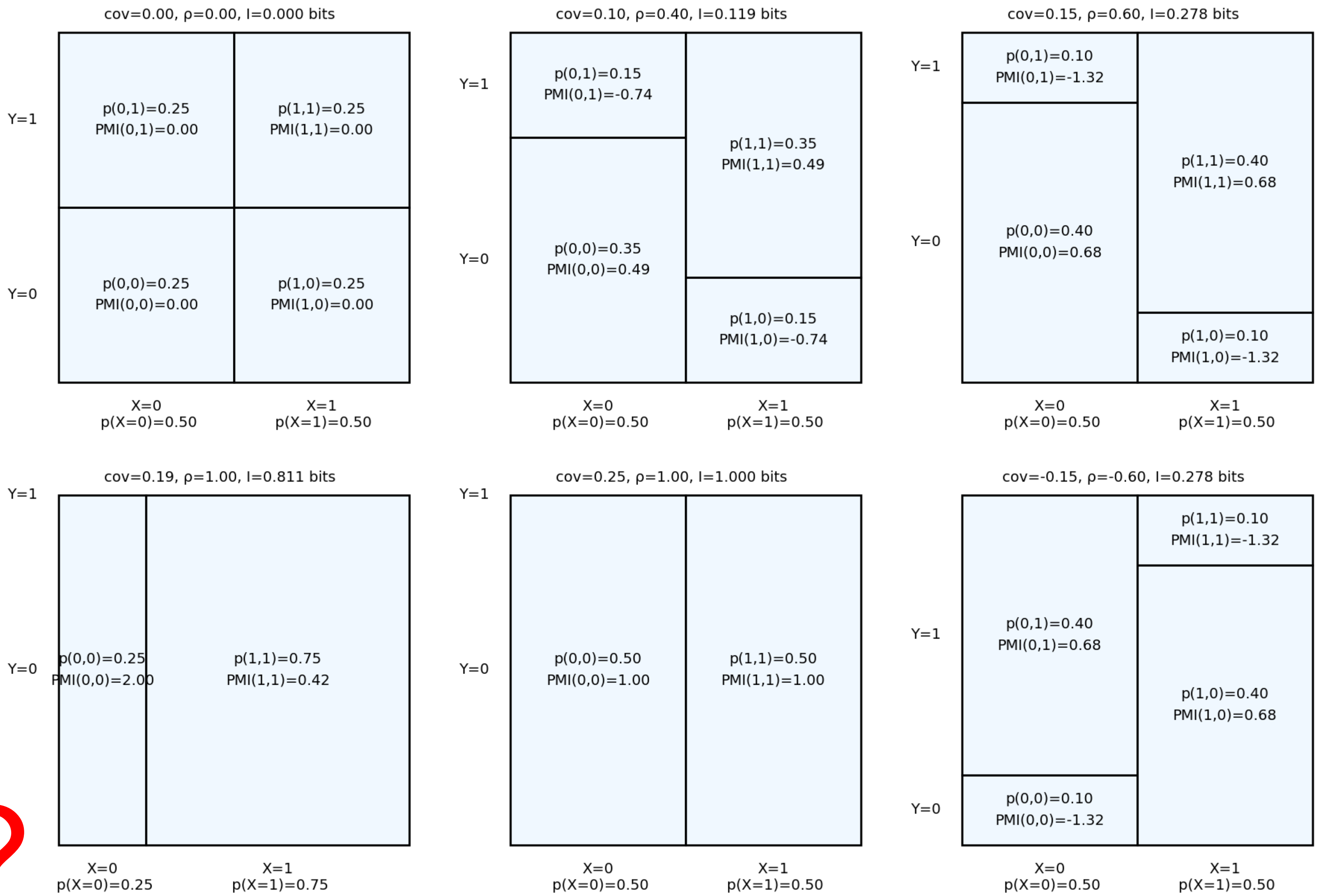


Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: 2 RVs $X, Y \in \{0,1\}$

$$I(X;Y) = \sum_{x,y} p_{xy} \cdot \lg\left(\frac{p_{xy}}{p_{x*} \cdot p_{*y}}\right)$$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$= p_{11} - p_{1*} \cdot p_{*y}$$

moment-based dependencies
that capture linearity

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\sqrt{p_{1*} \cdot \overline{p_{1*}} \cdot p_{*1} \cdot \overline{p_{*1}}}}$$

Is it possible to have a non-linear dependence, and thus
$\mathrm{Cov}(X,Y) = 0$, but $I(X;Y) > 0$?

Yes, it is possible to have a non-linear dependence, and thus
$\mathrm{Cov}(X,Y) = 0$, but $I(X;Y) > 0$

But it is <u>not possible for two binary variables!</u>

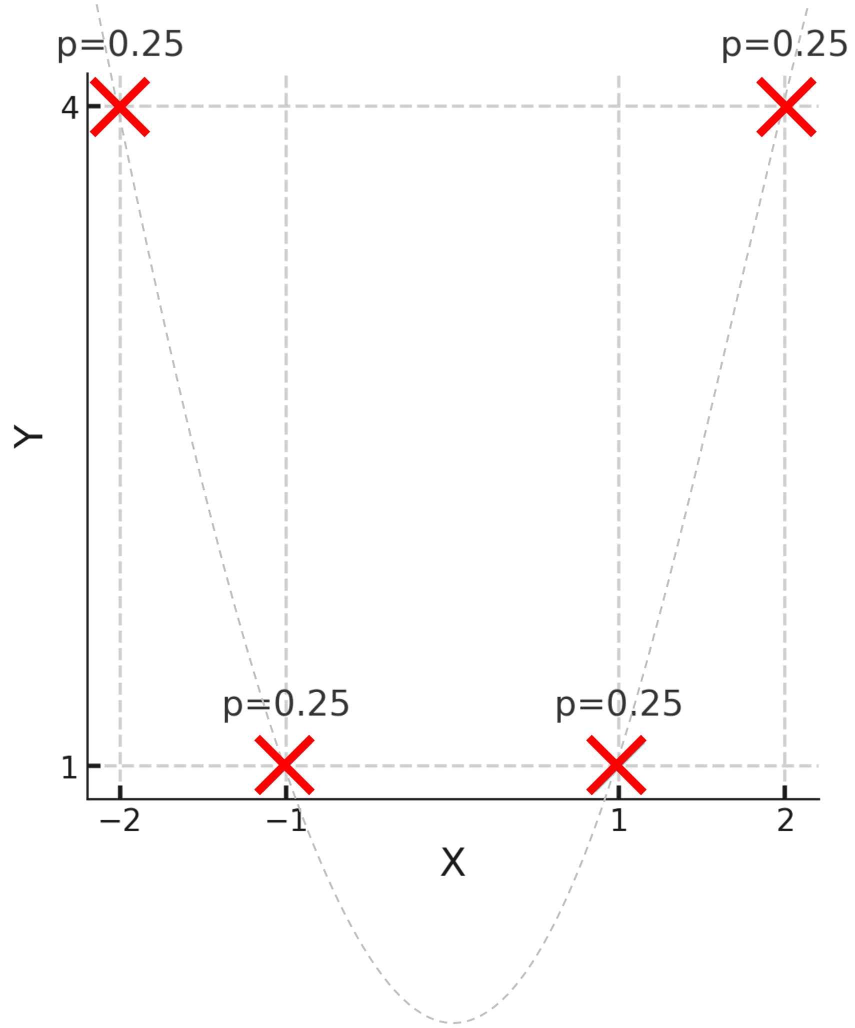
Two binary variables are determined by a 2×2 joint probability table.

If they are dependent, then the joint distribution deviates from the product of marginals. This deviation always shows up as a nonzero covariance.

Figure source: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: Let X take values from $\{-2,-1,1,2\}$ uniformly, and let $Y = X^2$



Covariance

$\mathbb{E}[X] =$?

$\mathbb{E}[Y] =$?

$\mathbb{E}[XY] =$?

$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$

$=$?

Mutual information

$H(X) =$?

$H(Y) =$?

$H(Y|X) =$?

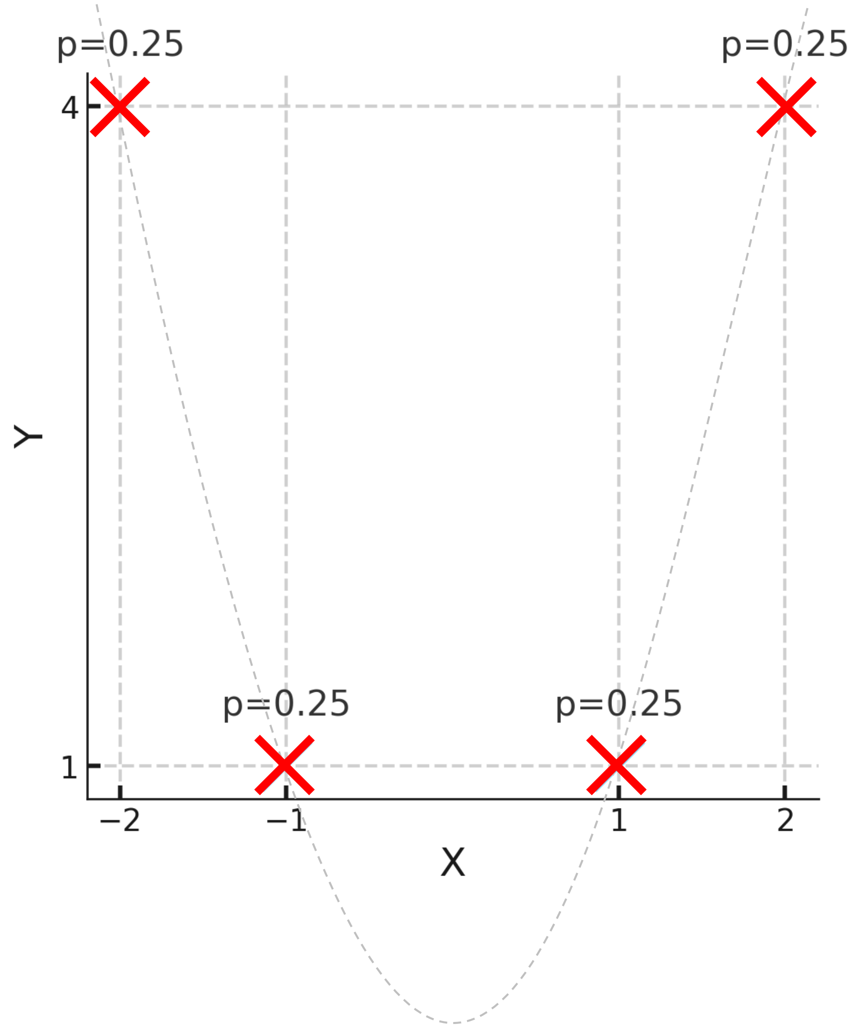
$I(X;Y) = H(Y) - H(Y|X)$

$=$?

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: Let X take values from $\{-2, -1, 1, 2\}$ uniformly, and let $Y = X^2$

Notice: X has 4 possible outcomes, Y has 2.



Covariance

$$\mathbb{E}[X] = \quad 0$$

$$\mathbb{E}[Y] = \quad 2.5$$

$$\mathbb{E}[XY] = \frac{-8-1+1+8}{4} = 0$$

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$= 0$$

Mutual information

$$H(X) = 2$$

$$H(Y) = 1$$

$H(Y|X) = 0$ (since X determines Y)

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

$$= 1$$

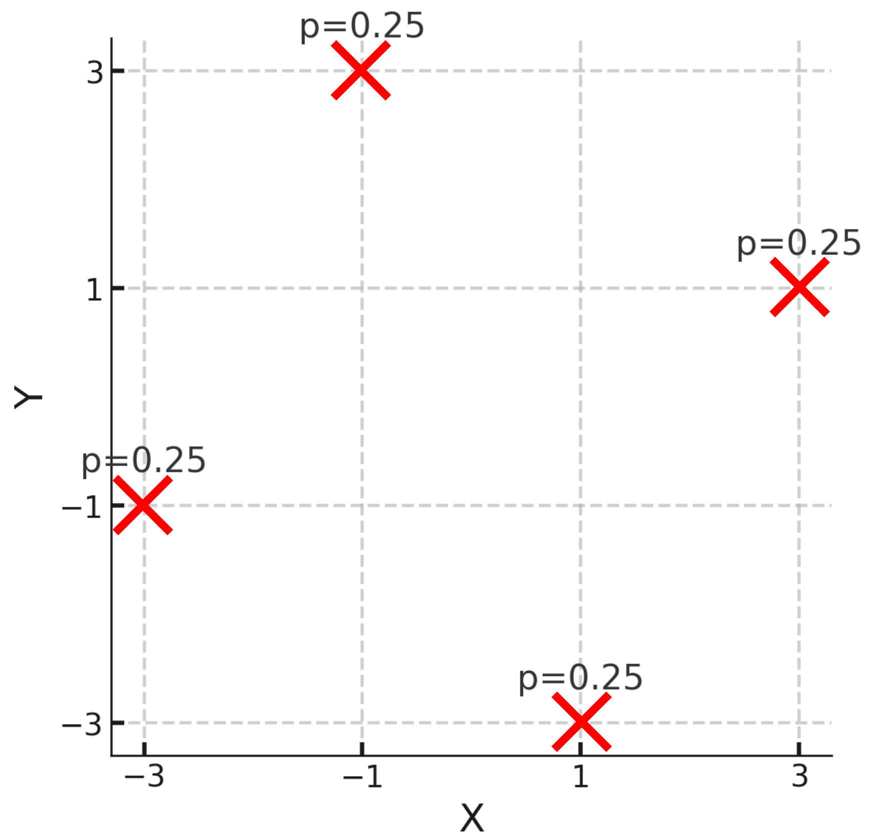
Can we rearrange the points as to increase mutual information to 2 bits?

?

Mutual Information I vs. Covariance $\mathbf{Cov}$ (Correlation ρ)

EXAMPLE: Let (X, Y) take values from $\{(-3,-1),(-1,3),(1,-3),(3,1)\}$ uniformly.

Notice: X has 4 possible outcomes, now Y also has 4.



Covariance

$$\mathbb{E}[X] = 0$$

$$\mathbb{E}[Y] = 0$$

$$\mathbb{E}[XY] = \frac{3-3-3+3}{4} = 0$$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = 0$$

Mutual information

$$H(X) = 2$$

$$H(Y) = 2$$

$$H(Y|X) = 0$$

$$I(X;Y) = H(Y) - H(Y|X) = 2$$

$X \mapsto Y$ is a bijection (it is deterministic and invertible)

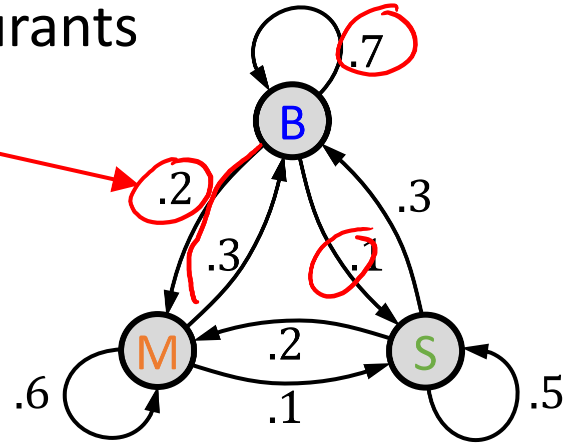
Markov chains

Markov Chain

$0.2 + 0.1 + 0.7 = 1$

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c|c} & \begin{array}{ccc} B & M & S \end{array} \\ \begin{array}{c} B \\ M \\ S \end{array} & \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \end{array}$$

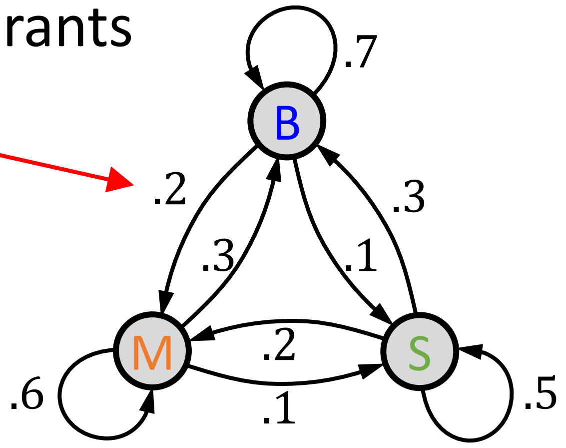
Is this matrix row-stochastic or column-stochastic ?

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

j

$$i \quad \mathbf{P} = \begin{array}{c} \\ B \\ M \\ S \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} B & M & S \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array} \begin{array}{c} \Sigma \\ 1 \\ 1 \\ 1 \\ \\ \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

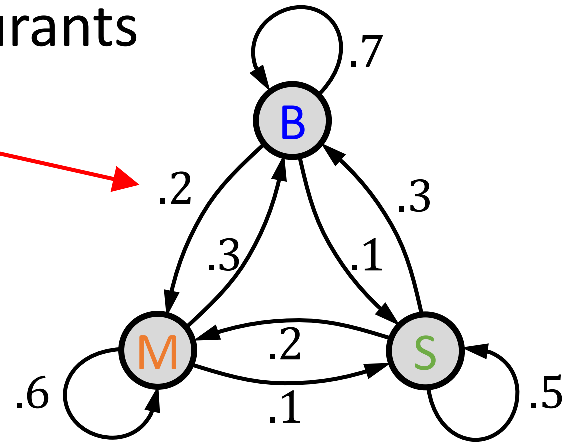
How to find the stationary distribution $\boldsymbol{\mu}$?

?

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c c} & \begin{array}{cccc} B & M & S & \Sigma \end{array} \\ \begin{array}{c} B \\ M \\ S \end{array} & \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \\ \Sigma & \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:

probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?

By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system algebraically:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \sum_i \mu_i P_{i,j} \text{ for all } j$$

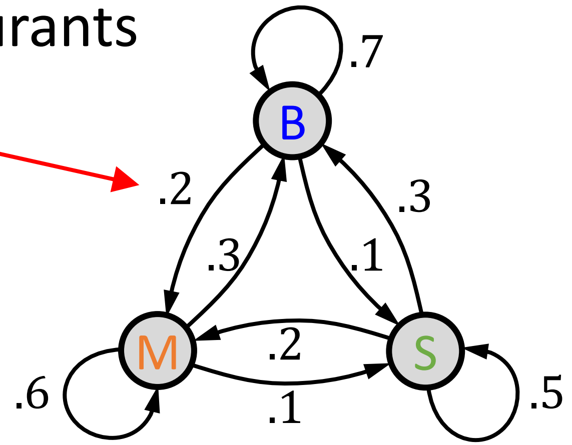
$\boldsymbol{\mu}$ is a left eigenvector for $\mathbf{P}$ (right eigenvector for $\mathbf{P}^{\mathrm{T}}$) with eigenvalue 1. And 1 is the largest left eigenvalue for a row-stochastic matrix (Perron-Frobenius theorem)

If graph is disconnected, then we would have multiple linearly independent eigenvectors with eigenvalue 1.

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[\mathrm{M}|\mathrm{B}] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{B} \\ \mathrm{M} \\ \mathrm{S} \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{B} & \mathrm{M} & \mathrm{S} \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array} \begin{array}{c} \Sigma \\ 1 \\ 1 \\ 1 \\ \\ \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \, | X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?

By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system algebraically:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \textstyle\sum_i \mu_i P_{i,j} \text{ for all } j$$

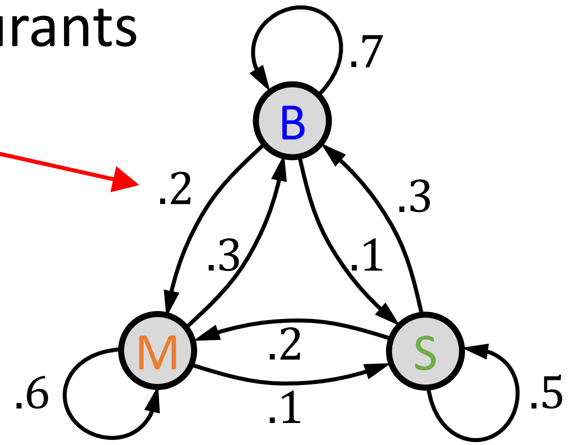
$$\begin{aligned} \mu_{\mathrm{B}} &= 0.7 \cdot \mu_{\mathrm{B}} + 0.3 \cdot \mu_{\mathrm{M}} + 0.3 \cdot \mu_{\mathrm{S}} \quad &(1) \\ \mu_{\mathrm{M}} &= 0.2 \cdot \mu_{\mathrm{B}} + 0.6 \cdot \mu_{\mathrm{M}} + 0.2 \cdot \mu_{\mathrm{S}} \quad &(2) \\ \mu_{\mathrm{S}} &= 0.1 \cdot \mu_{\mathrm{B}} + 0.1 \cdot \mu_{\mathrm{M}} + 0.5 \cdot \mu_{\mathrm{S}} \quad &(3) \end{aligned}$$

3 equations and 3 unknowns. So can we solve it ?

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[\text{M}|\text{B}] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ \text{B} \\ \text{M} \\ \text{S} \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} \text{B} & \text{M} & \text{S} \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array} \begin{array}{c} \Sigma \\ 1 \\ 1 \\ 1 \\ \\ \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?
By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system algebraically:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \textstyle\sum_i \mu_i P_{i,j} \text{ for all } j$$

$$\begin{aligned}
\mu_B &= 0.7 \cdot \mu_B + 0.3 \cdot \mu_M + 0.3 \cdot \mu_S && (1)\\
\mu_M &= 0.2 \cdot \mu_B + 0.6 \cdot \mu_M + 0.2 \cdot \mu_S && (2)\\
&\cancel{\mu_S = 0.1 \cdot \mu_B + 0.1 \cdot \mu_M + 0.5 \cdot \mu_S} && (3)\\
\mu_B &+ \mu_M + \mu_S = 1 && (4)
\end{aligned}$$

$$\begin{aligned}
0 &= -3\mu_B + 3\mu_M + 3\mu_S && (1)\\
0 &= 2\mu_B - 4\mu_M + 2\mu_S && (2)\\
1 &= \mu_B + \mu_M + \mu_S && (4)
\end{aligned}$$

$$\begin{aligned}
2 \cdot (4) - (2):&\quad 2 = 6\mu_M && \Rightarrow \mu_M = 1/3 && (5)\\
3 \cdot (4) - (1):&\quad 3 = 6\mu_B && \Rightarrow \mu_B = 1/2 && (6)\\
(5),(6) \rightarrow (4):&\quad 1 = 1/3 + 1/2 + \mu_S && \Rightarrow \mu_S = 1/6
\end{aligned}$$

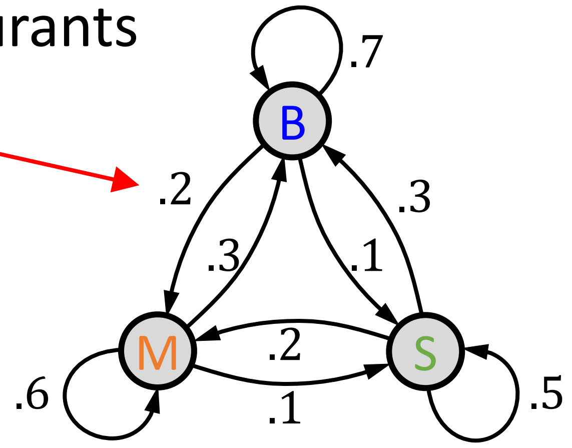
In general, solving n equations in n unknowns takes $\mathcal{O}(n^3)$. So is there a more efficient (practical) way ?

Entropy rates of Markov Chains

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c|ccc|c} i \backslash j & B & M & S & \Sigma \\ \hline B & .7 & .2 & .1 & 1 \\ M & .3 & .6 & .1 & 1 \\ S & .3 & .2 & .5 & 1 \\ \hline \Sigma & 1.3 & 1.0 & .7 & \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

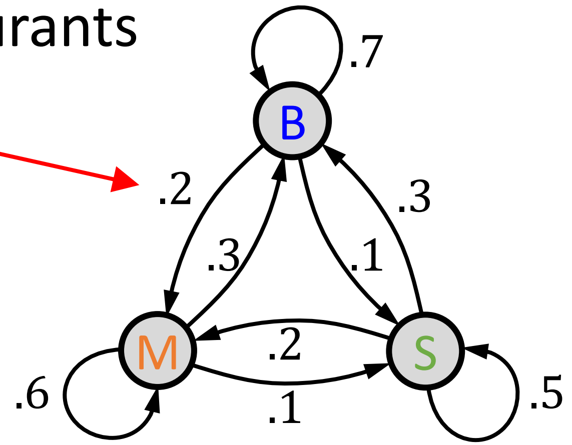
How to find the stationary distribution $\boldsymbol{\mu}$?

?

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[\text{M}|\text{B}] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ \text{B} \\ \text{M} \\ \text{S} \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} \text{B} & \text{M} & \text{S} \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array} \begin{array}{c} \Sigma \\ 1 \\ 1 \\ 1 \\ \\ \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?
By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \textstyle\sum_i \mu_i P_{i,j} \text{ for all } j$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/6 \end{pmatrix} \qquad H(\boldsymbol{\mu}) = 1.460$$

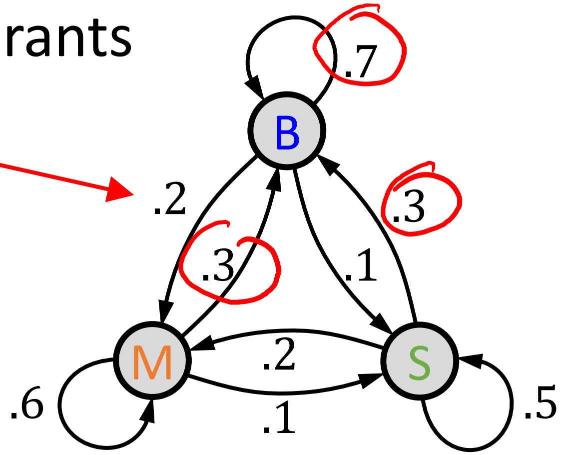
What would be the state transition matrix $\mathbf{P}'$ with same stationary distribution $\boldsymbol{\mu}$ if there was no memory: $\mathbb{P}[X_{n+1} = j \,|X_n = i] = \mathbb{P}[X_{n+1} = j\,]$

?

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[\text{M}|\text{B}] = 0.2$



State transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{ccc} \text{B} & \text{M} & \text{S} \end{array} & \Sigma \\ \begin{array}{c} \text{B} \\ \text{M} \\ \text{S} \end{array} & \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} & \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \\ \Sigma & \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} & \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:

probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?

By finding the largest eigenvector of $\mathbf{P}$,

i.e. solving an equation system:

transpose

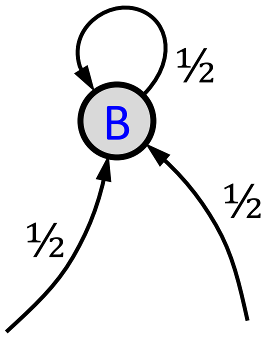
$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \sum_i \mu_i P_{i,j} \text{ for all } j$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/6 \end{pmatrix} \qquad H(\boldsymbol{\mu}) = 1.460$$

What would be the state transition matrix $\mathbf{P}'$ with same stationary distribution $\boldsymbol{\mu}$ if there was no memory: $\mathbb{P}[X_{n+1} = j \,|X_n = i] = \mathbb{P}[X_{n+1} = j\,]$

$$\mathbf{P}' = \begin{array}{cc} & \begin{array}{ccc} \text{B} & \text{M} & \text{S} \end{array} & \Sigma \\ \begin{array}{c} \text{B} \\ \text{M} \\ \text{S} \end{array} & \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 1/2 & 1/3 & 1/6 \\ 1/2 & 1/3 & 1/6 \end{pmatrix} & \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \\ \Sigma & \begin{array}{ccc} 1.5 & 1.0 & .5 \end{array} & \end{array}$$

$P_{i,j}{}' = \mu_j$

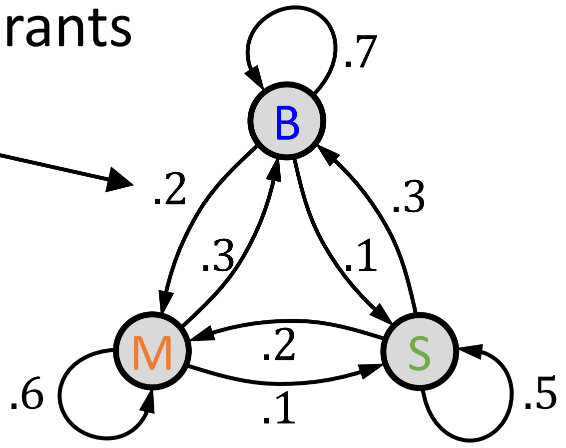


Which process ($\mathbf{P}$ or $\mathbf{P}'$) has a higher "entropy rate" ?

Markov Chain

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$
\begin{array}{cc}
 & \begin{array}{cccc} B & M & S & \Sigma \end{array} \\
\mathbf{P} = \begin{array}{c} B \\ M \\ S \end{array} & \left(\begin{array}{ccc} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{array}\right) \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \\
\Sigma & \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array}
\end{array}
$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?

By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \textstyle\sum_i \mu_i P_{i,j}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/6 \end{pmatrix} \qquad H(\boldsymbol{\mu}) = 1.460$$

$\mathbf{P}$

$\mathbf{P}'$

```
A random sample:
BSSSSSSBBBBBBBBMMMMMMMBMMMMMMBMBMBBBBBBBBBBSSMMB
 BBBBBBBBBBBBBBBMBBBBBBBBBBBBSBBBSSBMMBBBBBBSMMM
 BBBBBBBBBBSSSBBSBBSSSSBBMMBSBBSSSBBSSSSSBBMMMM
 MBBBBSSSMBBSBBBMMMMMMMMBBSSMMBSBMMMMBBBBMBSSBM
 MBBBBBMMMMMBBBSSBBBBSBBBBMMMMMMBBBBSBBMSMMMBBB
 BBMSBBBBBSBBMBBBMBBMBBMMMBBMBBBBMBBBBBBMSSMMBB
 SSSBBBMMMSBBBBBMMBBSBMMMMMSBBBBBSSSMSBBBBMBBSB
 BBBBBBBBBMMMMBBBBBMMSSSSMSSBMBBBBMBBMBSSBSSSSS
 SSSSSBBBBSMMMMMMMBMBMMMMBBBBBBSSSMMMBSBMBBBMM
 MMSSBMBSMBBSSSSBBMMBMMMMMMMMBMMMBBBBBBBMMBBBMM
 MMMMBMSMMBBBBBBBBBBBBBBBMMMMMSBBBBMBBBBBBMBBBBB
 BMMMSSSBBBBBBMMMMBSMBBMMMBSSBMBMBBBSSBBBSBMMMMS
 MMBMMBMMMMBMMBSSBBBMMBMMMBBBBSBBBBBBBBBBBSSMMMM
 MMMMSBBBSBBBBBBMMBBBBBBBBBBBBSSMBMMMBMSBBBSSBB
 BBBBBBBBBBBBBBMMMMBMBBBBBMMMBBMMMBBBMMBBBBBSMBS
 MMMMBSSSBBMMBBBBBBSSMMBBBBSMBBBBBBBBBBBSBBMMSSBS
 SSBBBBBBBBBBBBBMMMMMMMMMMMMBBMBBBBBBBSBSSBBBBBBBS
 SSBBMMBBMBMMMMMMMBBBSSMMMMMBBBBBBBBBBSSSSMSBBB
 BMMBBBMMMMSMBMSSBBBBMMMMBBSSBMBMBBBBBBBBBBBBBBM
 MMMBBBBBBBBBBMMBBBMMMMBMMMSSSMMMBBBSSMMBBBBSMBM
 BBBBMMMMBBBBBSBBBBMMBMMBMBBMBBBBBBBBMMMSSBBBMMMS
```

```
A random sample:
SBSMBBBBBMBBBBSBBBBBMBMBBMMMMBBBMBBBBMBBMBSMBBM
 BBMBBBSMMSSBSSMMSBBSMBBMMBMSBBMMMMBMMBBMSSMBBM
 MMBSSSBSBMBBBMMBBBBSMBBBMMBBBBBBBMSBBMMBBBBMBB
 BSBMBBMBMBMBSSBBMSMBSMBMBBBMBMSMBMSSMBBBMSMMBB
 SBBSBMBBBBBBSMBBMBMMBBSMBMMMBBMSBMMSBBMBBSBMBBM
 MSSBMMMBBBBMBBBMBBBBBMBMMSBSBBBBBSMBSMBMMBBMBB
 BMSSMMBMBBMMMMMSBSSSBMMBBSBSMBMBSSBMBBBSMMBBMM
 BBBMBMBMSMSMSMMBMBBSMMBBMMMBMSBBBBMBMSMMSBBBMM
 BBBMSMBMBSMMSSBBBMMBMMMMBSSBBMMMBBMBBBSBBMSSSM
 MMSMBSBMSBBMSMBBBMMMBBBBBBBBBBSBMBSMMMBBSBMBBBB
 MMSMBMBBMBBMMBBBMSMMBSBMMMBBSSMMBBBBMMMMBMBBBM
 SBMMBBMBBBMBBMBMBSSMBMSMSBBBMMMBMBBBBBMBBBBBSB
 MBMMSMBBBMBMBMMBBMBBBMSBMMBMBMBBBBSSMBBMBBBBBS
 BMMMMSBMMMBMMSSMBSBMBMMBSMSBBMMMBBBMBMBMBBMBMM
 MBBMBMSMBBSMSBBBMBMBBMMMBBSMMMBBSSBMBBSMBBBBMB
 BBBMMBSBMBSBMMBSBBSBBMMBMBSSMMBMMMMBMBBMSBMMMB
 BBSMBBBBBMBMMBBBBBBBBBMMBBBMSMMBMBMMSBMBMSBBBB
 BMSBMSMMBMSSBSBBBMSSBMBSSBSBMBSBBBMMSBBBMBSMBB
 SMMSMSSMBMMBSBMBBMSBBMBBMMBSSSBBBBBMMMMMMBSSBB
 MSMSBMMMMSMBBBBMBMMMBBBBMMBBSBMBSBMBSMSBBSMBMB
 BMMMBBMBBBBBMMBBSSSBMMBMBBBBBMBBMBBBBSMBBSBBBBM
```

Which process ($\mathbf{P}$ or $\mathbf{P}'$) has a higher "entropy rate" ?

Markov Chains and information measures

$X \to Y \to Z$ is a Markov chain if ?

Markov Chains and information measures

$X \rightarrow Y \rightarrow Z$ is a Markov chain if $\boxed{X \perp Z | Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

$$p(x, y, z) = \quad ?$$

Markov Chains and information measures

$X \rightarrow Y \rightarrow Z$ is a Markov chain if $\boxed{X \perp Z|Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

$$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|y)$$

In general, $p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|x,y)$

$$p(x,z|y) = \quad ?$$

Markov Chains and information measures

$X \to Y \to Z$ is a Markov chain if $\boxed{X \perp Z | Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

$p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$

In general, $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|x, y)$

$p(x, z|y) = p(x|y) \cdot p(z|y)$

In general, $p(x, z|y) = p(x|y) \cdot p(z|x, y)$

$I(X; Z|Y) =$?

MI beween X and Z
if we know Y

Markov Chains and information measures

$X \to Y \to Z$ is a Markov chain if $\boxed{X \perp Z | Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

$p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$

In general, $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|x, y)$

$p(x, z|y) = p(x|y) \cdot p(z|y)$

In general, $p(x, z|y) = p(x|y) \cdot p(z|x, y)$

$I(X; Z|Y) = 0$

What does this mean for the interaction information $J(X; Y; Z)$?

?

Markov Chains and information measures

$X \to Y \to Z$ is a Markov chain if $\boxed{X \perp Z|Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|y)$ In general, $p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|x,y)$

$p(x,z|y) = p(x|y) \cdot p(z|y)$ In general, $p(x,z|y) = p(x|y) \cdot p(z|x,y)$

$I(X;Z|Y) = 0$

What does this mean for the interaction information $J(X;Y;Z)$?

$$J(X;Z;Y) = I(X;Z) - \underbrace{I(X;Z|Y)}_{=0} = I(X;Z) \geq 0$$

- Recall: $J(X;Z;Y)$ measures the negated influence of a variable Y on the amount of information shared between X and Z (their mutual information).
- It is positive when Y decreases/inhibits (i.e., accounts for or explains some of) the correlation between X and Z (that happens here in Markov chains).
- It is negative when Y increases/facilitates the correlation (e.g., when X and Y are independent yet not conditionally independent given Z, see earlier parity example).

Markov Chains and information measures

$X \rightarrow Y \rightarrow Z$ is a Markov chain if $\boxed{X \perp Z|Y}$, and thus:

Intuitively: The future depends only on the current state (not the previous ones)

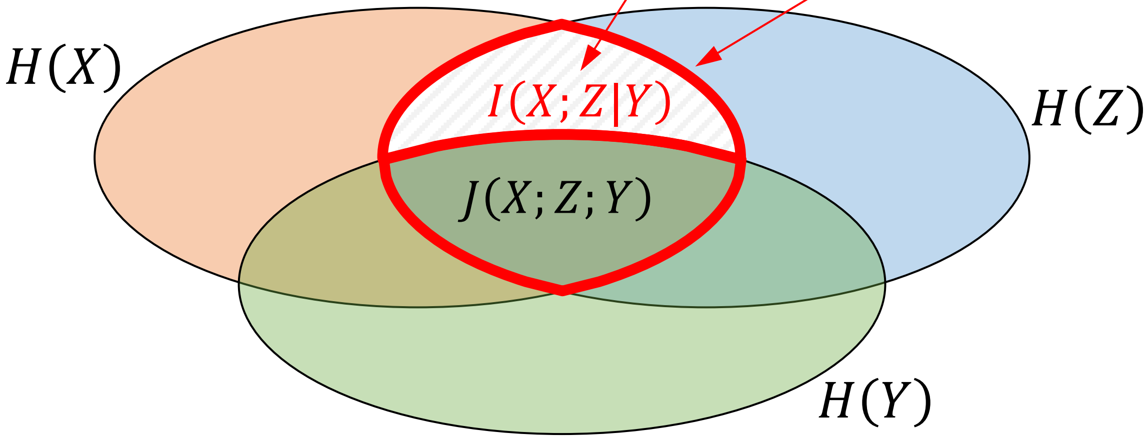
$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|y)$ In general, $p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|x,y)$

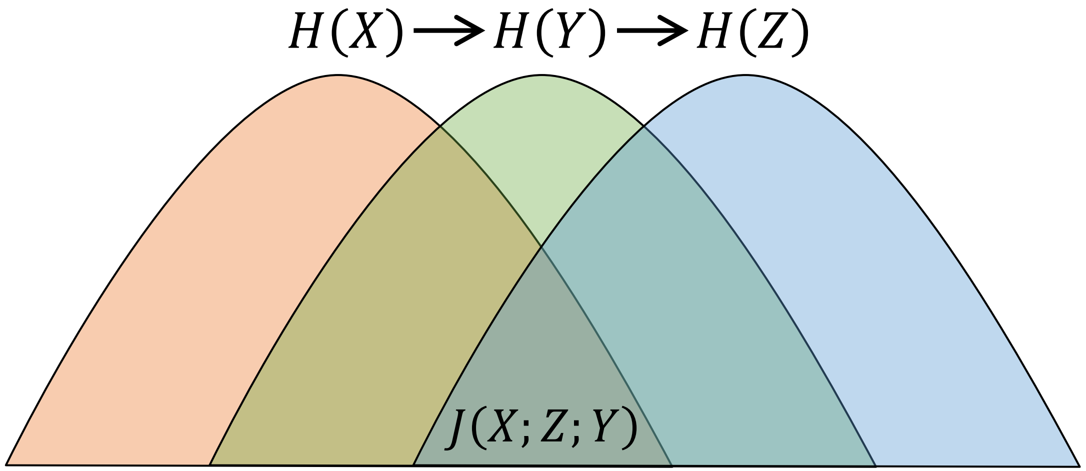
$p(x,z|y) = p(x|y) \cdot p(z|y)$ In general, $p(x,z|y) = p(x|y) \cdot p(z|x,y)$

$I(X;Z|Y) = 0$

What does this mean for the interaction information $J(X;Y;Z)$?

$$J(X;Z;Y) = I(X;Z) - \underbrace{I(X;Z|Y)}_{=0} = I(X;Z) \geq 0$$





Markov Chains and stationary stochastic processes

$X \rightarrow Y \rightarrow Z$ is a Markov chain if $\boxed{X \perp Z|Y}$, and thus:

$p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$ In general, $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|x, y)$

$p(x, z|y) = p(x|y) \cdot p(z|y)$ In general, $p(x, z|y) = p(x|y) \cdot p(z|x, y)$

$I(X; Z|Y) = 0$

A discrete stochastic process $(X_1, X_2, \ldots)$ is a Markov chain if

?

Markov Chains and stationary stochastic processes

$X \to Y \to Z$ is a Markov chain if $\boxed{X \perp Z | Y}$, and thus:

$p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$ In general, $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|x, y)$

$p(x, z|y) = p(x|y) \cdot p(z|y)$ In general, $p(x, z|y) = p(x|y) \cdot p(z|x, y)$

$I(X; Z|Y) = 0$

A discrete stochastic process $(X_1, X_2, \ldots)$ is a Markov chain if each RV depends only on the one preceding it and is conditionally independent of all the other preceding RVs

$$\mathbb{P}[x_{n+1}|x_n, x_{n-1}, \ldots, x_1] = \mathbb{P}[x_{n+1}|x_n]$$

A stochastic process $\{X_i\} = (X_1, X_2, \ldots)$ is stationary if ...

?

Markov Chains and stationary stochastic processes

$X \to Y \to Z$ is a Markov chain if $\boxed{X \perp Z | Y}$, and thus:

$$p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$$

In general, $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|x, y)$

$$p(x, z|y) = p(x|y) \cdot p(z|y)$$

In general, $p(x, z|y) = p(x|y) \cdot p(z|x, y)$

$$I(X; Z|Y) = 0$$

A discrete stochastic process $(X_1, X_2, \dots)$ is a Markov chain if each RV depends only on the one preceding it and is conditionally independent of all the other preceding RVs

$$\mathbb{P}[x_{n+1} | x_n, x_{n-1}, \dots, x_1] = \mathbb{P}[x_{n+1} | x_n]$$

A stochastic process $\{X_i\} = (X_1, X_2, \dots)$ is stationary if the joint distribution of any subsequence is invariant w.r.t. shifts in the time index (say by ℓ steps)

$$\mathbb{P}[(x_1, x_2, \dots, x_k)] = \mathbb{P}[(x_{1+\ell}, x_{2+\ell}, \dots, x_{k+\ell})]$$

Entropy rate for stationary Markov Chain

The entropy rate of a stochastic process $\{X_i\} =: \mathcal{X}$ is the average entropy per symbol:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \cdot H(X_1, X_2, \ldots, X_n) \qquad H(X_1, X_2, \ldots, X_n) \to n \cdot H(\mathcal{X})$$

Notice a slight inconsistency in notation inherited from [Cover, Thomas]: This symbol $\mathcal{X}$ represents the entire stochastic process as a whole. Thus literally, $H(\mathcal{X})$ should be $H(X_1, X_2, \ldots, X_n)$, and not $\frac{1}{n}$ of that. To clarify, some textbooks write $H'(\mathcal{X})$ for the entropy rate.

For a stationary stochastic process, this is equal to the rate of information innovation

$$H(\mathcal{X}) = \quad ?$$

Entropy rate for stationary Markov Chain

The entropy rate of a stochastic process $\{X_i\} =: \mathcal{X}$ is the average entropy per symbol:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \cdot H(X_1, X_2, \dots, X_n) \qquad H(X_1, X_2, \dots, X_n) \to n \cdot H(\mathcal{X})$$

Notice a slight inconsistency in notation inherited from [Cover, Thomas]: This symbol $\mathcal{X}$ represents the entire stochastic process as a whole. Thus literally, $H(\mathcal{X})$ should be $H(X_1, X_2, \dots, X_n)$, and not $\frac{1}{n}$ of that. To clarify, some textbooks write $H'(\mathcal{X})$ for the entropy rate.

For a stationary stochastic process, this is equal to the rate of information innovation

$$H(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, \dots, X_1)$$

For a stationary Markov Chain, the entropy rate is

$$H(\mathcal{X}) = \quad ?$$

Entropy rate for stationary Markov Chain

The entropy rate of a stochastic process $\{X_i\} =: \mathcal{X}$ is the average entropy per symbol:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \cdot H(X_1, X_2, \ldots, X_n) \qquad H(X_1, X_2, \ldots, X_n) \to n \cdot H(\mathcal{X})$$

Notice a slight inconsistency in notation inherited from [Cover, Thomas]: This symbol $\mathcal{X}$ represents the entire stochastic process as a whole. Thus literally, $H(\mathcal{X})$ should be $H(X_1, X_2, \ldots, X_n)$, and not $\frac{1}{n}$ of that. To clarify, some textbooks write $H'(\mathcal{X})$ for the entropy rate.

For a stationary stochastic process, this is equal to the rate of information innovation

$$H(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1)$$

For a stationary Markov Chain, the entropy rate is

$H(\mathcal{X}) = H(X_2 | X_1)$ where the conditional entropy is calculated using the <u>stationary distribution</u> (!) **?** $H(X_2)$

$\leq$ or $\geq$

Entropy rate for stationary Markov Chain

The entropy rate of a stochastic process $\{X_i\} =: \mathcal{X}$ is the average entropy per symbol:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \cdot H(X_1, X_2, \dots, X_n) \qquad H(X_1, X_2, \dots, X_n) \to n \cdot H(\mathcal{X})$$

Notice a slight inconsistency in notation inherited from [Cover, Thomas]: This symbol $\mathcal{X}$ represents the entire stochastic process as a whole. Thus literally, $H(\mathcal{X})$ should be $H(X_1, X_2, \dots, X_n)$, and not $\frac{1}{n}$ of that. To clarify, some textbooks write $H'(\mathcal{X})$ for the entropy rate.

For a stationary stochastic process, this is equal to the rate of information innovation

$$H(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, \dots, X_1)$$

For a stationary Markov Chain, the entropy rate is

(conditioning cannot increase the entropy)

$H(\mathcal{X}) = H(X_2|X_1)$ where the conditional entropy is calculated using the <u>stationary distribution</u> (!) $\leq H(X_2)$

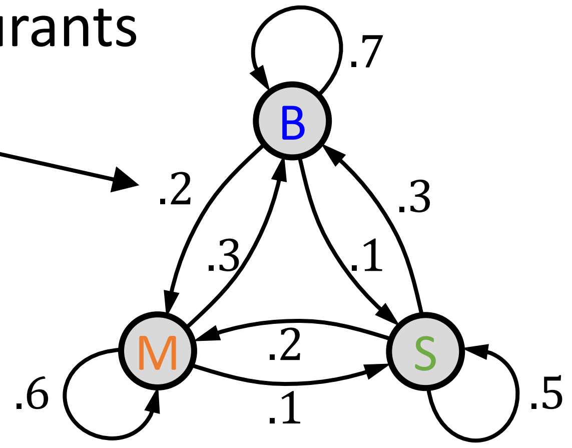
$$= \sum_i \mu_i \cdot H(X_2 | X_1 = i) = \sum_i \mu_i \cdot H(\mathbf{P}_{i:}) = \mathbb{E}_{i\sim\mu}[H(\mathbf{P}_{i:})]$$

$$= -\sum_i \mu_i \cdot P_{ij} \cdot \lg(P_{ij})$$

Markov Chain (cont.)

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$\begin{array}{c} \\ \mathbf{P} = \begin{array}{c} B \\ M \\ S \end{array} \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} B & M & S \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \\ \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array} \begin{array}{c} \Sigma \\ 1 \\ 1 \\ 1 \\ \\ \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?
By finding the largest eigenvector of $\mathbf{P}$,
i.e. solving an equation system:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \sum_i \mu_i P_{i,j}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/6 \end{pmatrix} \qquad H(\boldsymbol{\mu}) = 1.460$$

```
A random sample:
BSSSSSSBBBBBBBBMMMMMMMBMMMMMMBMBMBBBBBBBBBBSSMMB
 BBBBBBBBBBBBBBBMBBBBBBBBBBBBSBBBSSBMMBBBBBBBSMMM
 BBBBBBBBBBSSSBBSBBSSSSBBMMBSBBSSSBBSSSSSBBMMMM
 MBBBBSSSMBBSBBBMMMMMMMMBBSSMMBSBMMMMBBBBMBSSBM
 MBBBBBMMMMMBBBSSBBBBSBBBBMMMMMMBBBBSBBMSMMMBBB
 BBMSBBBBBSBBMBBBMBBMBBMMMBBMBBBBMBBBBBBMSSMMBB
 SSSBBBMMMSBBBBBMMBBSBMMMMMSBBBBBSSSMSBBBBMBBSB
 BBBBBBBBBMMMMBBBBBMMSSSSMSSBMBBBBMBBMBSSBSSSSS
 SSSSSBBBBSMMMMMMMBMBMMMMBBBBBBBSSSMMMBSBMBBBMM
 MMSSBMBSMBBSSSSBBMMBMMMMMMMMBMMMBBBBBBBMMBBBMM
 MMMMBMSMMBBBBBBBBBBBBBBMMMMMSBBBBMBBBBBBMBBBBB
 BMMMSSSBBBBBMMMMBSMBBMMMBSSBMBMBBBSSBBBSBMMMMS
```

```
A random sample:
SBSMBBBBBMBBBBSBBBBBMBMBBMMMMBBBMBBBBMBBMBSMBBM
 BBMBBBSMMSSBSSMMSBBSMBBMMBMSBBMMMMBMMBBMSSMBBM
 MMBSSSBSBMBBBMMBBBBSMBBBMMBBBBBBBMSBBMMBBBBMBB
 BSBMBBMBMBMBSSBBMSMBSMBMBBBMBMSMBMSSMBBBMSMMBB
 SBBSBMBBBBBSMBBMBMMBBSMBMMMBBMSBMMSBBMBBSBMBBM
 MSSBMMMBBBBMBBBMBBBBBMBMMSBSBBBBBSMBSMBMMBBMBB
 BMSSMMBMBBMMMMMSBSSSBMMBBSBSMBMBSSBMBBBSMMBBMM
 BBBMBMBMSMSMSMMBMBBSMMBBMMMBMSBBBBMBMSMMSBBBMM
 BBBMSMBMBSMMSSBBBMMBMMMMBSSBBMMMBBMBBBSBBMSSSM
 MMSMBSBMSBBMSMBBBMMMBBBBBBBBBSBMBSMMMBBSBMBBBB
 MMSMBMBBMBBMMBBBMSMMBSBMMMBBSSMMBBBBMMMMBMBBBM
 SBMMBBMBBBMBBMBMBSSMBMSMSBBBMMMBMBBBBBMBBBBBSB
```

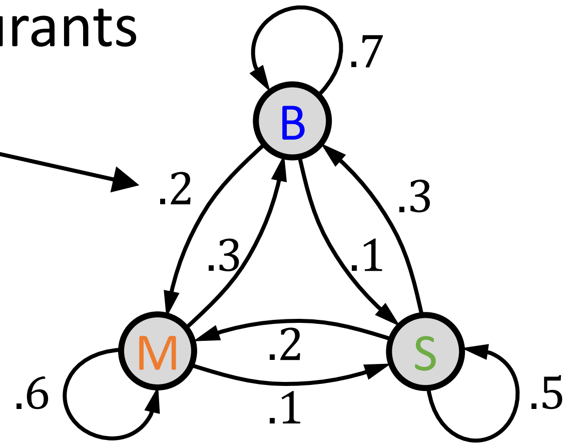
Entropy rate of $\mathbf{P}$:

$$H(\mathbf{P}) = \quad ?$$

Markov Chain (cont.)

EXAMPLE: restaurants

$\mathbb{P}[M|B] = 0.2$



State transition matrix $\mathbf{P}$:

$$\begin{array}{cc} & \begin{array}{cccc} B & M & S & \Sigma \end{array} \\ \mathbf{P} = \begin{array}{c} B \\ M \\ S \end{array} & \left(\begin{array}{ccc} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{array}\right) \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \\ \Sigma & \begin{array}{ccc} 1.3 & 1.0 & .7 \end{array} \end{array}$$

row-stochastic

$P_{i,j} = \mathbb{P}[X_{n+1} = j \,|X_n = i]$:
probability of choosing j after i

$\mathbf{P}_{i,:}$ row vector (probability distribution)

How to find the stationary distribution $\boldsymbol{\mu}$?
By finding the largest eigenvector of $\mathbf{P}$, i.e. solving an equation system:

transpose

$$\boldsymbol{\mu} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\mu} \qquad \mu_j = \sum_i \mu_i P_{i,j}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/6 \end{pmatrix} \qquad H(\boldsymbol{\mu}) = 1.460$$

```
A random sample:
BSSSSSSBBBBBBBBMMMMMMBMMMMMMBMBMBBBBBBBBBBSSMMB
 BBBBBBBBBBBBBBBMBBBBBBBBBBBBSBBBSSBMMBBBBBBBSMMM
 BBBBBBBBBBSSSBBSBBSSSSBBMMBSBBSSSBBSSSSSSBBMMMM
 MBBBBSSSMBBSBBBMMMMMMMBBSSMMBSBMMMMBBBBMBSSBM
 MBBBBBMMMMMBBBSSBBBBSBBBBMMMMMMBBBBSBBMSMMMBBB
 BBMSBBBBBSBBMBBBMBBMBBMMMBBMBBBBMBBBBBBMSSMMBB
 SSSBBBMMMSBBBBBBMMBBSBMMMMMSBBBBBSSSMSBBBBMBBSB
 BBBBBBBBBMMMMBBBBBBMMSSSSMSSBMBBBBMBBMBSSBSSSSS
 SSSSSBBBBSMMMMMMMBMBMMMMBBBBBBBSSSMMMBSBMBBBMM
 MMSSBMBSMBBSSSSBBMMBMMMMMMMMBMMMBBBBBBBMMBBBMM
 MMMMBMSMMBBBBBBBBBBBBBBBMMMMMSBBBBMBBBBBBMBBBBB
 BMMMSSSBBBBBMMMMBSMBBMMMBSSBMBMBBBSSBBBSBMMMMS
```

```
A random sample:
SBSMBBBBBMBBBBSBBBBBMBMBBMMMMBBBMBBBBMBBMBSMBBM
 BBMBBBSMMSSBSSMMSBBSMBBMMBMSBBMMMMBMMBBMSSMBBM
 MMBSSSBSBMBBBMMBBBBSMBBBMMBBBBBBBMSBBMMBBBBMBB
 BSBMBBMBMBMBSSBBMSMBSMBMBBBMBMSMBMSSMBBBMSMMBB
 SBBSBMBBBBBBSMBBMBMMBBSMBMMMBBMSBMMSBBMBBSBMBBM
 MSSBMMMBBBBMBBBMBBBBBMBMMSBSBBBBBSMBSMBMMBBMBB
 BMSSMMBMBBMMMMMSBSSSBMMBBSBSMBMBSSBMBBBSMMBBMM
 BBBMBMBMSMSMSMMBMBBSMMBBMMMBMSBBBBMBMSMMSBBBMM
 BBBMSMBMBSMMSSBBBMMBMMMMBSSBBMMMBBMBBBSBBMSSSM
 MMSMBSBMSBBMSMBBBMMMBBBBBBBBSBMBSMMMBBSBMBBBB
 MMSMBMBBMBBMMBBBMSMMBSBMMMBBSSMMBBBBMMMMBMBBBM
 SBMMBBMBBBMBBMBMBSSMBMSMSBBBMMMBMBBBBBMBBBBBSB
```

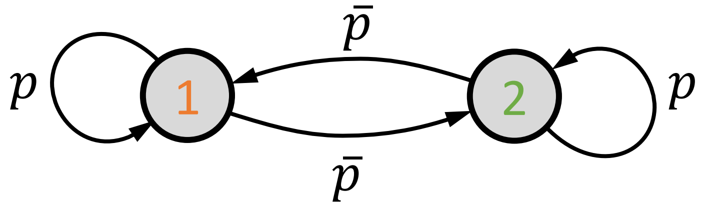
Entropy rate of $\mathbf{P}$:

$$H(\mathbf{P}) = \mathbb{E}_{i\sim\mu}[H(\mathbf{P}_{i:})]$$

$$= \sum_i \mu_i \cdot H(\mathbf{P}_{i:}) = 1.258$$

Markov Chain

Example: A simple two-state Markov Chain

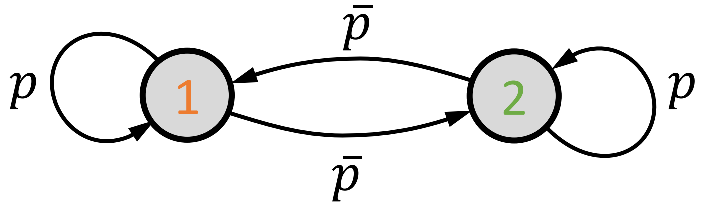


$\bar{p} = 1 - p$

$\mathbf{P} =$?

Markov Chain

EXAMPLE: A simple two-state Markov Chain



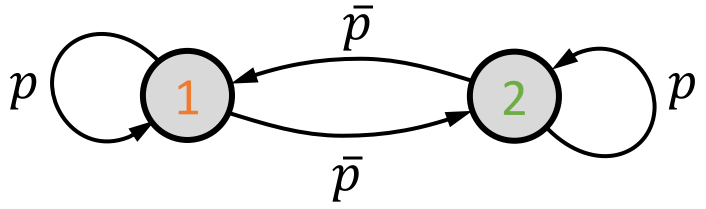
$$\bar{p} = 1 - p$$

$$\mathbf{P} = \begin{pmatrix} p & \bar{p} \\ \bar{p} & p \end{pmatrix}$$

$$\boldsymbol{\mu} = \ ?$$

Markov Chain

EXAMPLE: A simple two-state Markov Chain



$$\bar{p} = 1 - p$$

$$\mathbf{P} = \begin{pmatrix} p & \bar{p} \\ \bar{p} & p \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

$$H(\boldsymbol{\mu}) = \ ?$$

$p = 0.95$:

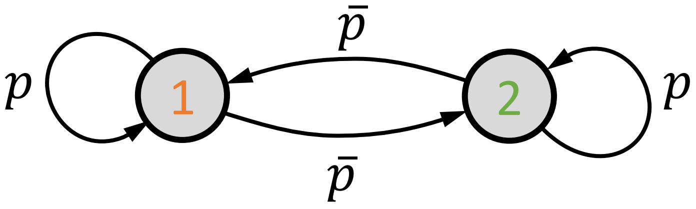
```
A random sample:
AAAAAAABBBBBBBAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAA
 AAAAAAAAAAAAABBBBBBBBBAAAAAAAAAAAAABBBBBBABBBBBBBAAAABAAAABBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBAAAAAAAAAAAAAABBBBA
 AABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAABB
 BBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAABBBBBABBBBBBBBBBBBBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAA
 AAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBB
 BBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAABBBBBBBBBBBBBBBBBBAAAABBAAAAAAAAAABBBBBBBBBBBBBBBBB
 BBBAAAAAAAAABBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAABBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

$p = 0.05$:

```
A random sample:
ABABABABABABABABABABABABABABABABABABABAABABAABABABABABABABABABAB
 AAABABABABABABABABABABABABABABABABABABABABAABABABABABABABABABAB
 ABABABABABABABABABABABABABABAAABABABABABABABABBABABABABAABABABA
 BABABABABABABABABABAABABABABABABABABABABABABABBABABABABABABABAB
 ABABBABABABABABAABAABABABABABABABABABABABABBAABABABABABABABABAB
 ABABABABABBABABABABABABABABABABABABABABABABABABBABABABABABABABA
 BABABABABABABABABABBABABAABABABABAABABBABABABABABABABABABABABAB
 ABABABABABABABABABABABABABABABABABBABABABABABABABABABABABABABAB
 AABABABABABBABABABABABABABABABABABABABABABABAABABABBABABABABABA
 BBABABABABABABABABABABABABABABABABABABABABABABABABABABABABABABB
 ABABABAABABABABABABABABABABABABABABABABABABBABABABABABABABABBAB
 ABAABABABBAABABABABABABABABABABABABABABABABABABABABABABABABBABA
 BABABABABABABABABABABAABABABABABABABABAABABABABABABABABAABBABAB
 ABABABABABABABABAABABABABABABABABABABABABABABABABABAAABABABABA
 BABABABBABABABABABABABABABABABABABABABBABABABABABABABABABABABAB
 ABABABABABAABABBABABABABABABABAABABABABABABABABABABABAB
```

Markov Chain

EXAMPLE: A simple two-state Markov Chain



$\bar{p} = 1 - p$

$$\mathbf{P} = \begin{pmatrix} p & \bar{p} \\ \bar{p} & p \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

$$H(\boldsymbol{\mu}) = 1$$

$$H(\mathbf{P}) = \text{?}$$

$p = 0.95$:

```
A random sample:
AAAAAAABBBBBBBAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAA
 AAAAAAAAAAAAABBBBBBBBBAAAAAAAAAAAABBBBBBABBBBBBBBAAABAAAABBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBAAAAAAAAAAAAAAABBBBA
 AABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAABB
 BBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAABBBBBABBBBBBBBBBBBBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAA
 AAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBB
 BBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAABBBBBBBBBBBBBBBBBAAAABBAAAAAAAAAABBBBBBBBBBBBBBBBB
 BBBAAAAAAAAABBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAABBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBB
```

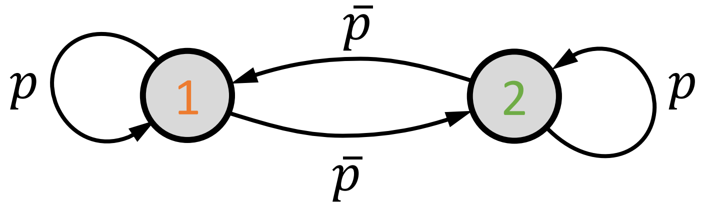
$p = 0.05$:

```
A random sample:
ABABABABABABABABABABABABABABABABABABAABABAABABABABABABABABABAB
 AAABABABABABABABABABABABABABABABABABABABAABABABABABABABABABABAB
 ABABABABABABABABABABABABABAAABABABABABABABABABBABABABAABABABA
 BABABABABABABABABABAABABABABABABABABABABABABABABBABABABABABABAB
 ABABBABABABABAABAABABABABABABABABABABABABABBAABABABABABABABABAB
 ABABABABABBABABABABABABABABABABABABABABABABABABBABABABABABABA
 BABABABABABABABABABABBABABAABABABAABABBABABABABABABABABABABAB
 ABABABABABABABABABABABABABABABABBABABABABABABABABABABABABABAB
 AABABABABABBABABABABABABABABABABABABABABAABABABBABABABABABABA
 BBABABABABABABABABABABABABABABABABABABABABABABABABABABABABABB
 ABABABAABABABABABABABABABABABABABABABABABABBABABABABABABABBAB
 ABAABABABBAABABABABABABABABABABABABABABABABABABABABABABABBABA
 BABABABABABABABABABABABAABABABABABABAABABABABABABABABAABBABAB
 ABABABABABABABABABAABABABABABABABABABABABABABABABABAABABABABA
 BABABABBABABABABABABABABABABABABABBABABABABABABABABABABABAB
 ABABABABABAABABBABABABABABABAABABABABABABABABABABABAB
```

Code: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/IT_illustration.ipynb

Markov Chain

EXAMPLE: A simple two-state Markov Chain



$\bar{p} = 1 - p$

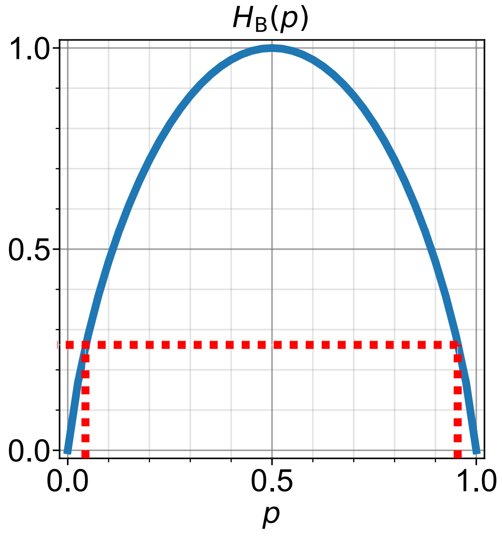
$$\mathbf{P} = \begin{pmatrix} p & \bar{p} \\ \bar{p} & p \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

$$H(\boldsymbol{\mu}) = 1$$

$$H(\mathbf{P}) = \mathbb{E}_{i \sim \mu}[H(\mathbf{P}_{i:})]$$

$$= H_{\mathrm{B}}(p)$$



$H_{\mathrm{B}}(p) = 0.286$

$p = 0.95$:

```
A random sample:
AAAAAAABBBBBBBAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAA
 AAAAAAAAAAAAABBBBBBBBBBAAAAAAAAAAAABBBBBABBBBBBBBAAABAAAAABBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBAAAAAAAAAAAAAAABBBBA
 AABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAABB
 BBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAABBBBABBBBBBBBBBBBBBBBB
 BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAA
 AAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBB
 BBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAABBBBBBBBBBBBBBBBBAAAAABAAAAAAAAAAABBBBBBBBBBBBBBBBB
 BBBAAAAAAAAABBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAABBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

$p = 0.05$:

```
A random sample:
ABABABABABABABABABABABABABABABABABABABAABABAABABABABABABABABABAB
 AAABABABABABABABABABABABABABABABABABABABABAABABABABABABABABABAB
 ABABABABABABABABABABABABABABAAABABABABABABABBABABABABAABABABA
 BABABABABABABABABAABABABABABABABABABABABABABABBABABABABABABABAB
 ABABBABABABABAABAABABABABABABABABABABABABBAABABABABABABABABAB
 ABABABABABBABABABABABABABABABABABABABABABABABABBABABABABABABABA
 BABABABABABABABABABABABBABABAABABABAABABBABABABABABABABABABABAB
 ABABABABABABABABABABABABABABABABABBABABABABABABABABABABABABAB
 AABABABABBABABABABABABABABABABABABABABAABABABBABABABABABABA
 BBABABABABABABABABABABABABABABABABABABABABABABABABABABABABABABB
 ABABABAABABABABABABABABABABABABABABABABABABBABABABABABABABBAB
 ABAABABABBAABABABABABABABABABABABABABABABABABABABABABABABBABA
 BABABABABABABABABABABAABABABABABABABABAABABABABABABABABAABBAB
 ABABABABABABABABAABABABABABABABABABABABABABABABABABABABAABABABA
 BABABABBABABABABABABABABABABABABABABBABABABABABABABABABABABAB
 ABABABABABAABABBABABABABABABABAABABABABABABABABABABAB
```

Updated 10/2/2025

Part 1: Theory
L08: Basics of entropy (6/7)
[Data processing inequality]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

10/2/2025

Pre-class conversations

35%: **Mini projects (former name: "flipped class scribes")**: Students "illustrate" 7 lectures of their own choice with mini projects that independently explore some side issues motivated by the topics covered in class and their own questions. Graduate theory classes often ask students to scribe the lecture content. However, we change the rule of the game Rather than scribing (= repeating and summarizing) the content of the class, I ask you to *"illustrate" some interesting aspect in the covered topics with imaginative and ideally tricky illustrating examples*. ~~Only to be consistent with that standard notation, we refer to these illustrations as "scribes".~~ Those illustrations are great if they in turn can help other students practice and solidify their understanding of the topics discussed. Please start either from our PPTX template or use your own template (as long as you include slide numbers), and submit it as PDF to Canvas, naming it "cs7840-fa25-[YOUR NAME]-scribe[NUMBER]-[SOME DESCRIPTIVE TITLE].PDF".

- Last class recapitulation
- Python scripts ...
- Renaming scribes to mini projects).
- Project ideas: Talk to me often. I can't meet right after class for next 4 times, but before or in my office via email coordination.
  - Feel free to explore how to use information theory to your current work
  - And feel free to explore "crazy" ideas; that's what these projects are for

- Today:
  - (Markov Chains) -> Data Processing inequality, Sufficient statistics
  - Today or next time: information inequalities

Data Processing Inequality

Data Processing Inequality for $X \rightarrow Y \rightarrow Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \rightarrow Y \rightarrow Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \ ? \ I(X;Z)$$

$\leq$ or $\geq$

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \;\geq\; I(X;Z)$$

COROLLARY: If $Z = f(Y)$, then $I(X;Y) \geq I\big(X;f(Y)\big)$. Thus functions of Y cannot increase the information about X. In other words, no processing of Y, deterministic or random, can increase the information that Y contains about X (unless you add additional outside information).

This follows from $X \to Y \to f(Y)$ forming a Markov chain.

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \geq I(X;Z)$$

PROOF:

$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = \ ?$$

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \geq I(X;Z)$$

PROOF:

$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - H(X|Y,Z) = \underbrace{H(X) + (-H(X|Z)}_{?} + \underbrace{H(X|Z)) - H(X|Y,Z)}_{?}$$

$$= \quad ? \quad + \quad ?$$

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z | Y$), then

$$I(X;Y) \geq I(X;Z)$$

PROOF:

$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - H(X|Y,Z) = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{?}$$

(similarly, from symmetry)

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \geq I(X;Z)$$

PROOF:

$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - H(X|Y,Z) = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$ (similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{?}$$

Data Processing Inequality for $X \to Y \to Z$

Intuitively, the data processing inequality states that no clever transformation of a received representation Y can increase the information about the original information X.

THEOREM: Suppose we have a Markov chain $X \to Y \to Z$ (and thus $X \perp Z|Y$), then

$$I(X;Y) \;\geq\; I(X;Z)$$

PROOF:

$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - H(X|Y,Z) = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

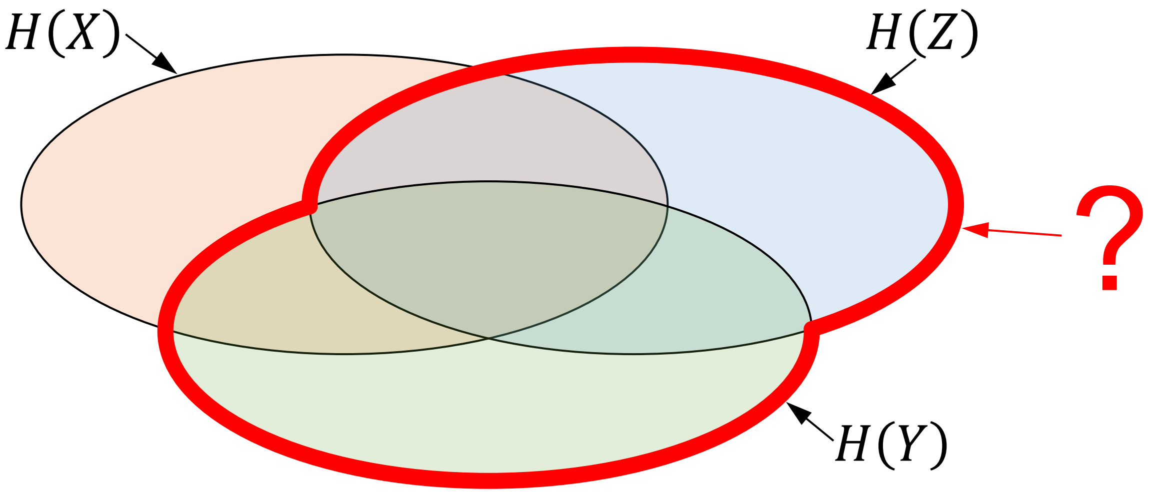
$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$ (similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \qquad I(X;Z) \qquad + \qquad I(X;Y|Z)$$

$$= \qquad I(X;Y) \qquad + \qquad \underbrace{I(X;Z|Y)}_{=0}$$

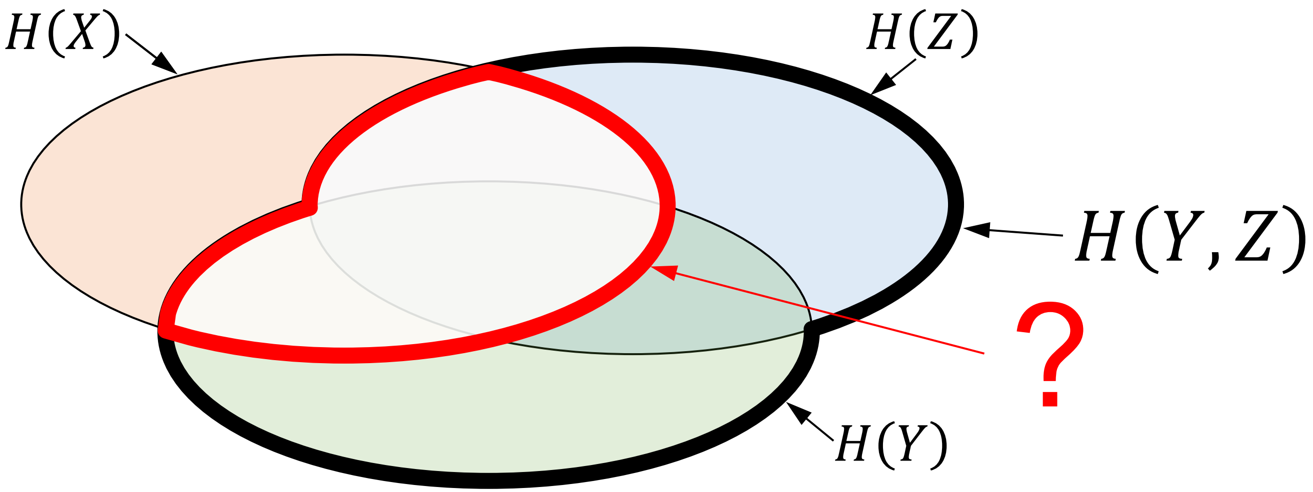
(similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$

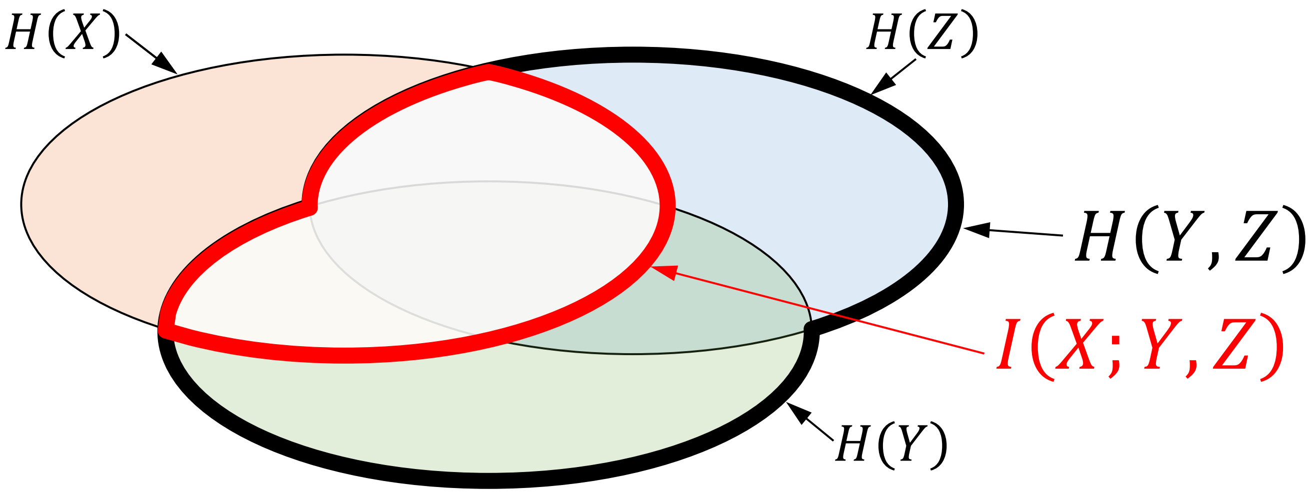
(similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

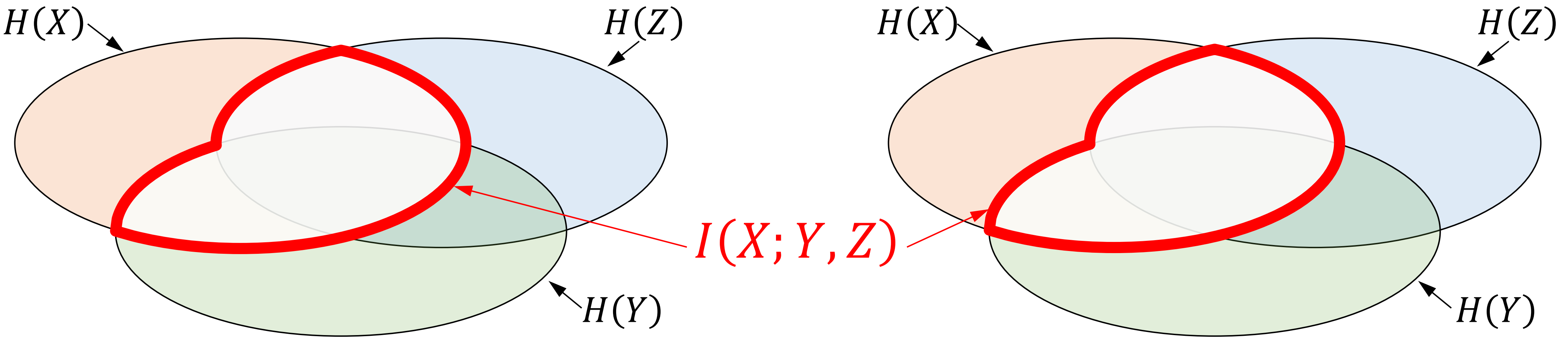
$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$ (similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$ since mutual information is always non-negative

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

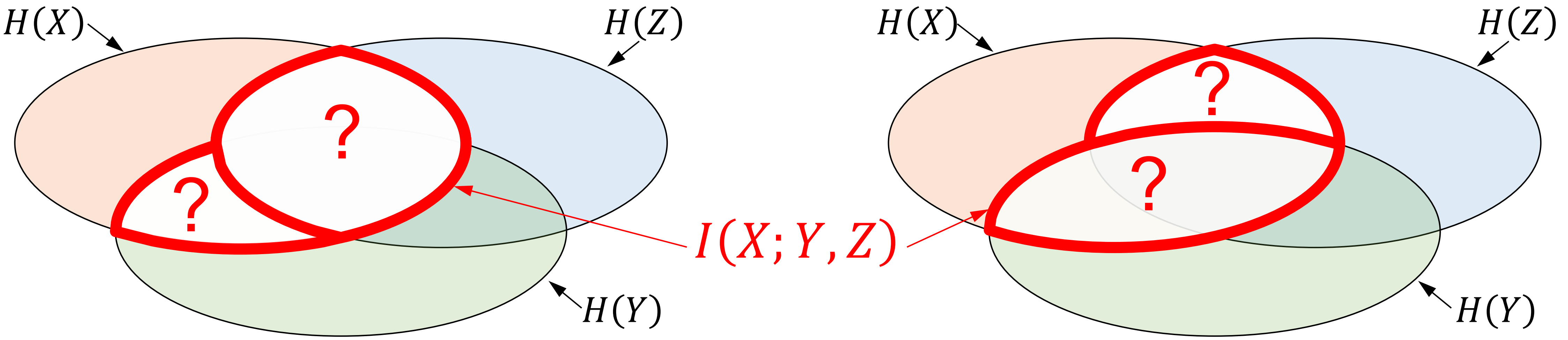
$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$ (similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$ since mutual information is always non-negative

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$

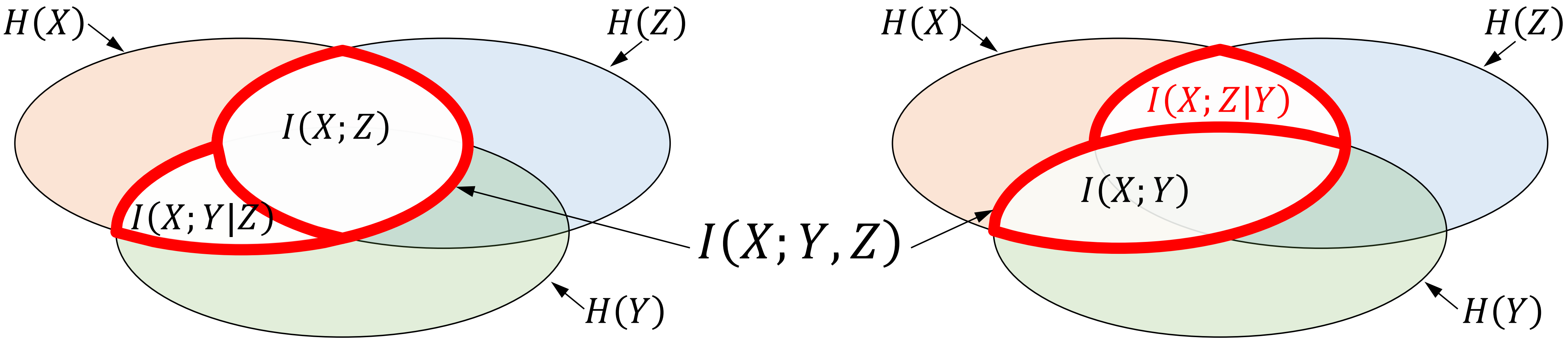
(similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \qquad I(X;Z) \qquad + \qquad I(X;Y|Z)$$

$$= \qquad I(X;Y) \qquad + \qquad \underbrace{I(X;Z|Y)}_{=0}$$

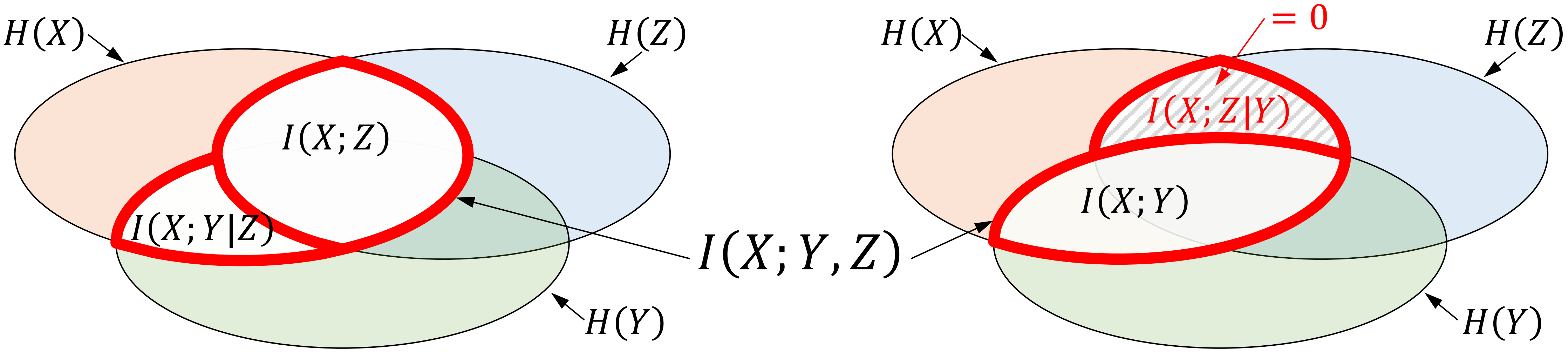
(similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$



$$\underbrace{I(X;Y,Z)}_{I(X;(Y,Z))} = H(X) - \underbrace{H(X|Y,Z)}_{I(X|(Y,Z))} = \underbrace{H(X) + \big(-H(X|Z)}_{} + \underbrace{H(X|Z)\big) - H(X|Y,Z)}_{}$$

$$= \quad I(X;Z) \quad + \quad I(X;Y|Z)$$

$$= \quad I(X;Y) \quad + \quad \underbrace{I(X;Z|Y)}_{=0}$$

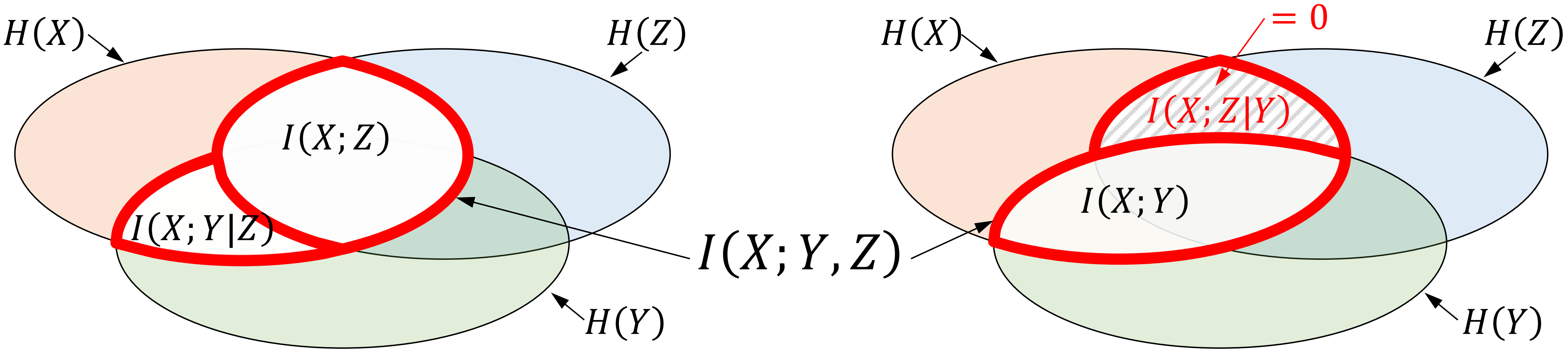
(similarly, from symmetry)

$$\Rightarrow I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

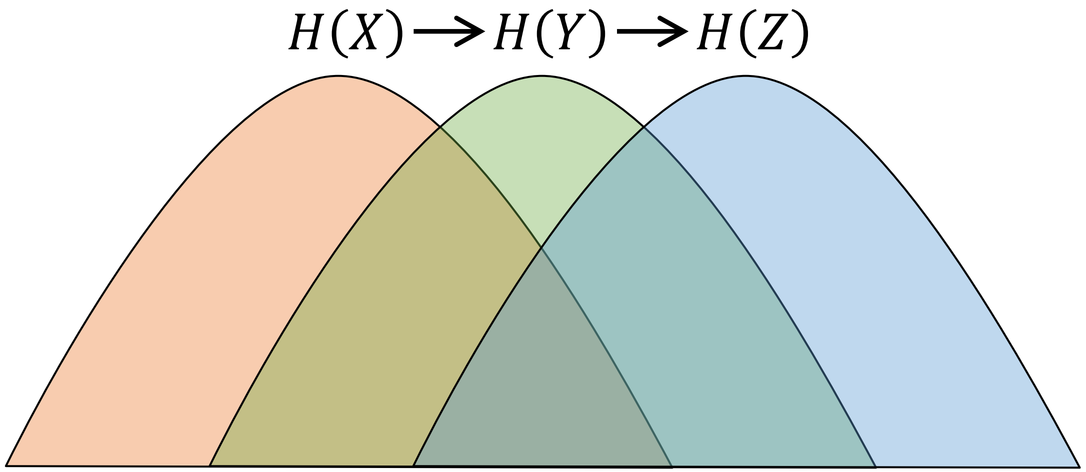
$$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

since mutual information is always non-negative

Data Processing Inequality for $X \to Y \to Z$

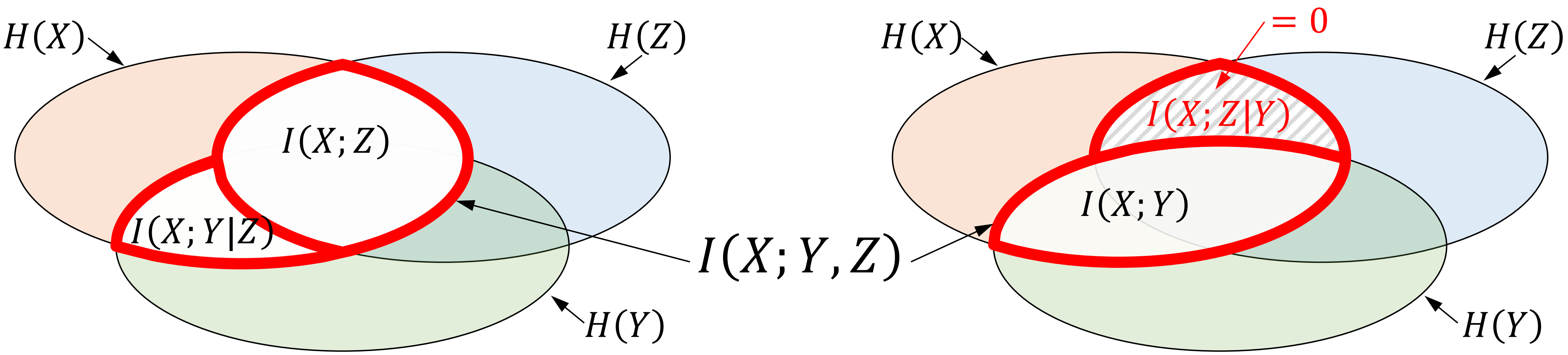


$J(X;Z;Y) =$

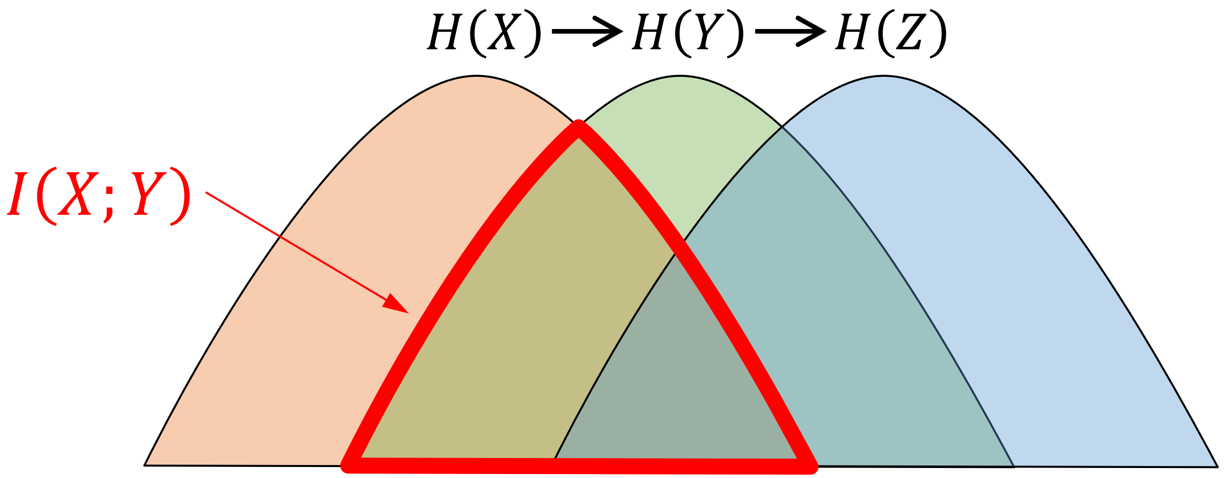


$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$

Data Processing Inequality for $X \rightarrow Y \rightarrow Z$

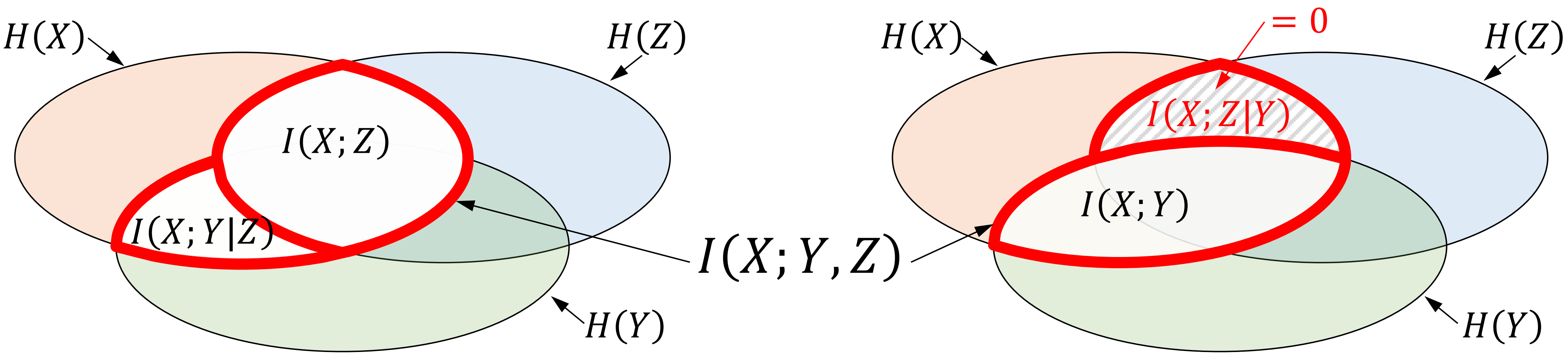


$J(X;Z;Y) =$

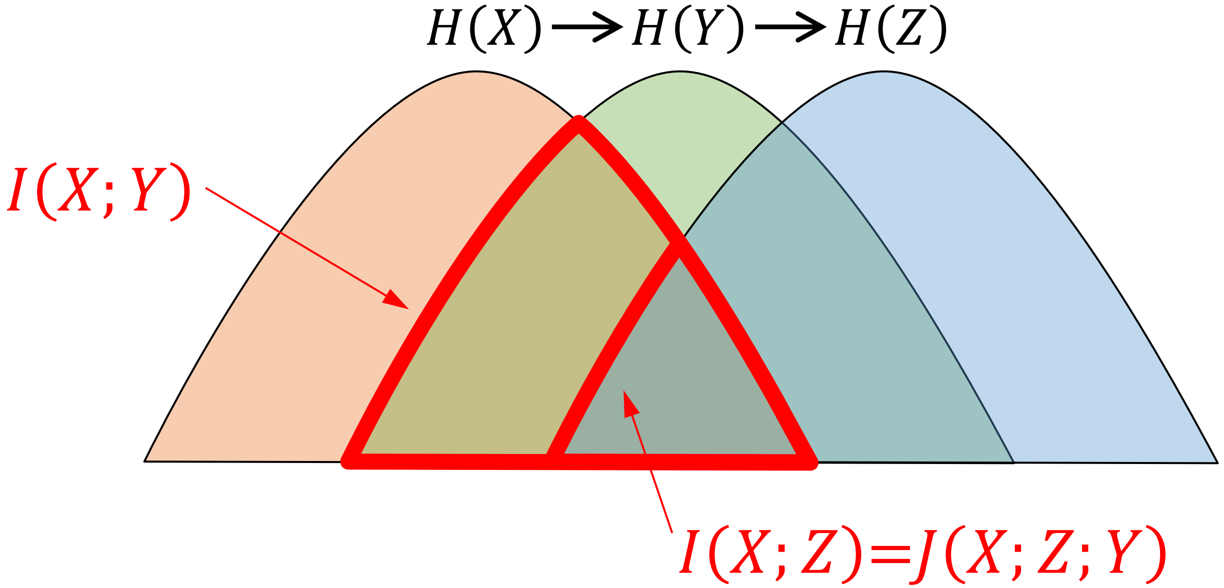


$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$

Data Processing Inequality for $X \rightarrow Y \rightarrow Z$



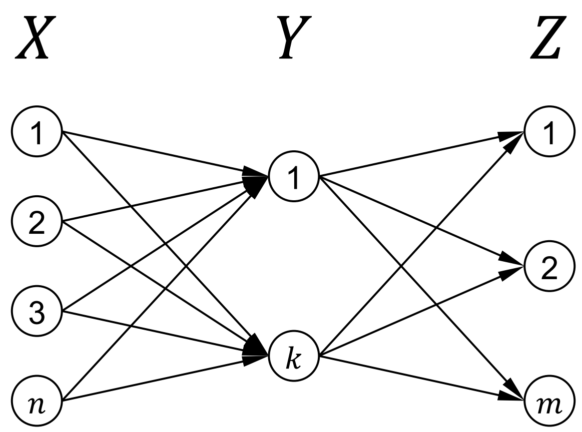
$J(X;Z;Y) =$



$\Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$

Bottleneck $X \to Y \to Z$

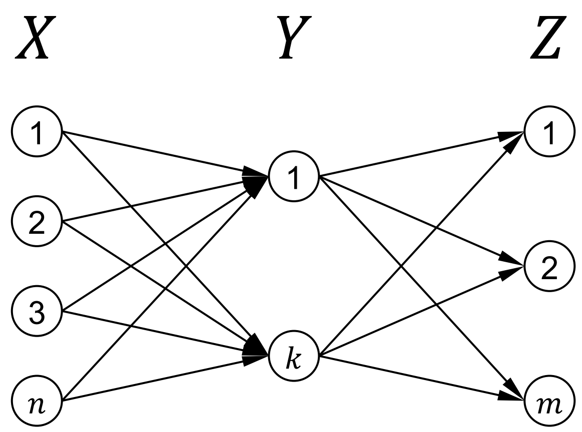
EXAMPLE: suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. In other words, $X \to Y \to Z$ with $p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|y)$, and $x \in [n], y \in [k], z \in [m]$.



How can we upper bound $I(X;Z)$?

Bottleneck $X \to Y \to Z$

EXAMPLE: suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. In other words, $X \to Y \to Z$ with $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$, and $x \in [n], y \in [k], z \in [m]$.



How can we upper bound $I(X;Z)$?

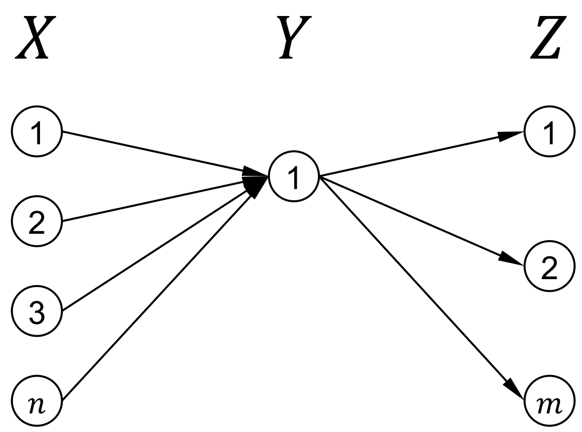
$$
\begin{aligned}
I(X;Z) \le I(X;Y) &= H(Y) - H(Y|X) \\
&\le H(Y) \\
&\le \lg(k)
\end{aligned}
$$

$\Rightarrow$ The dependence between X and Z is limited by the size k of the bottleneck.

What if $k = 1$?

Bottleneck $X \to Y \to Z$

EXAMPLE: suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. In other words, $X \to Y \to Z$ with $p(x, y, z) = p(x) \cdot p(y|x) \cdot p(z|y)$, and $x \in [n], y \in [k], z \in [m]$.



How can we upper bound $I(X; Z)$?

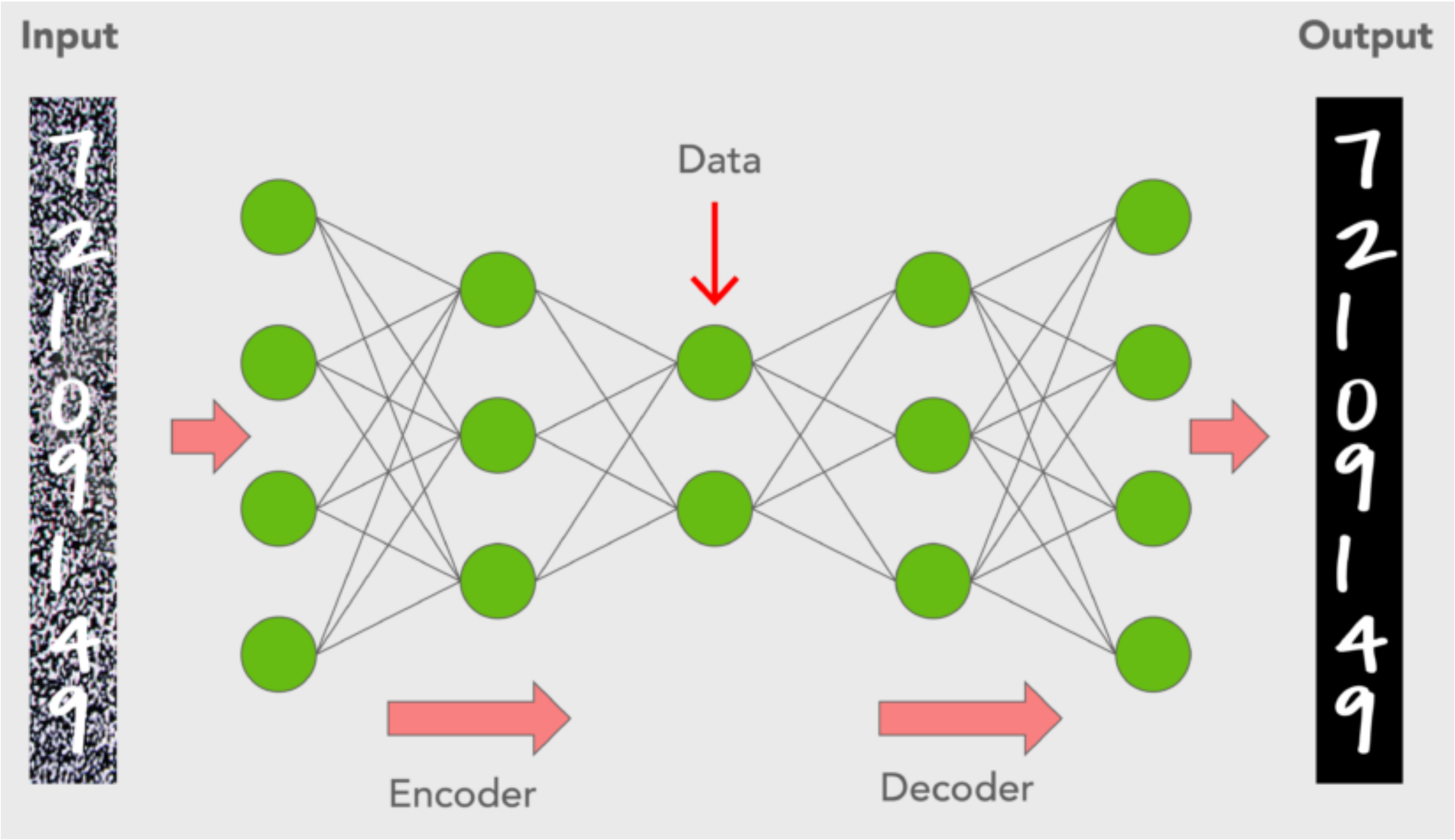
$$I(X;Z) \leq I(X;Y) = H(Y) - H(Y|X)$$
$$\leq H(Y)$$
$$\leq \lg(k)$$

⇒ The dependence between X and Z is limited by the size k of the bottleneck.

What if $k = 1$? ⇒ $I(X;Z) \leq \lg 1 = 0.$ ⇒ X and Z are independent.

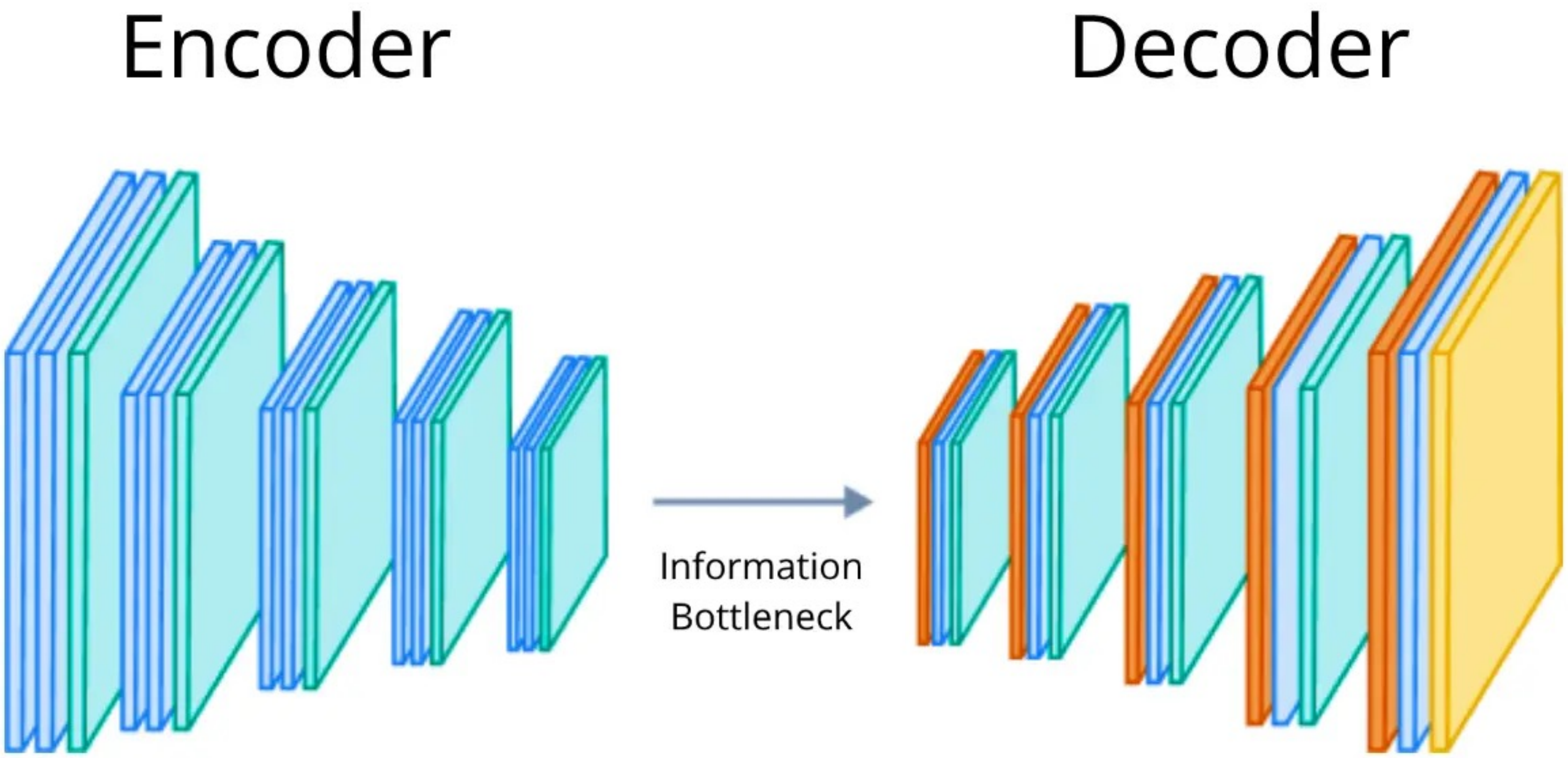
Example 2.16 from: [Cover, Thomas'06]. Elements of Information Theory (book, 2nd ed). https://doi.org/10.1002/047174882X
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Autoencoders

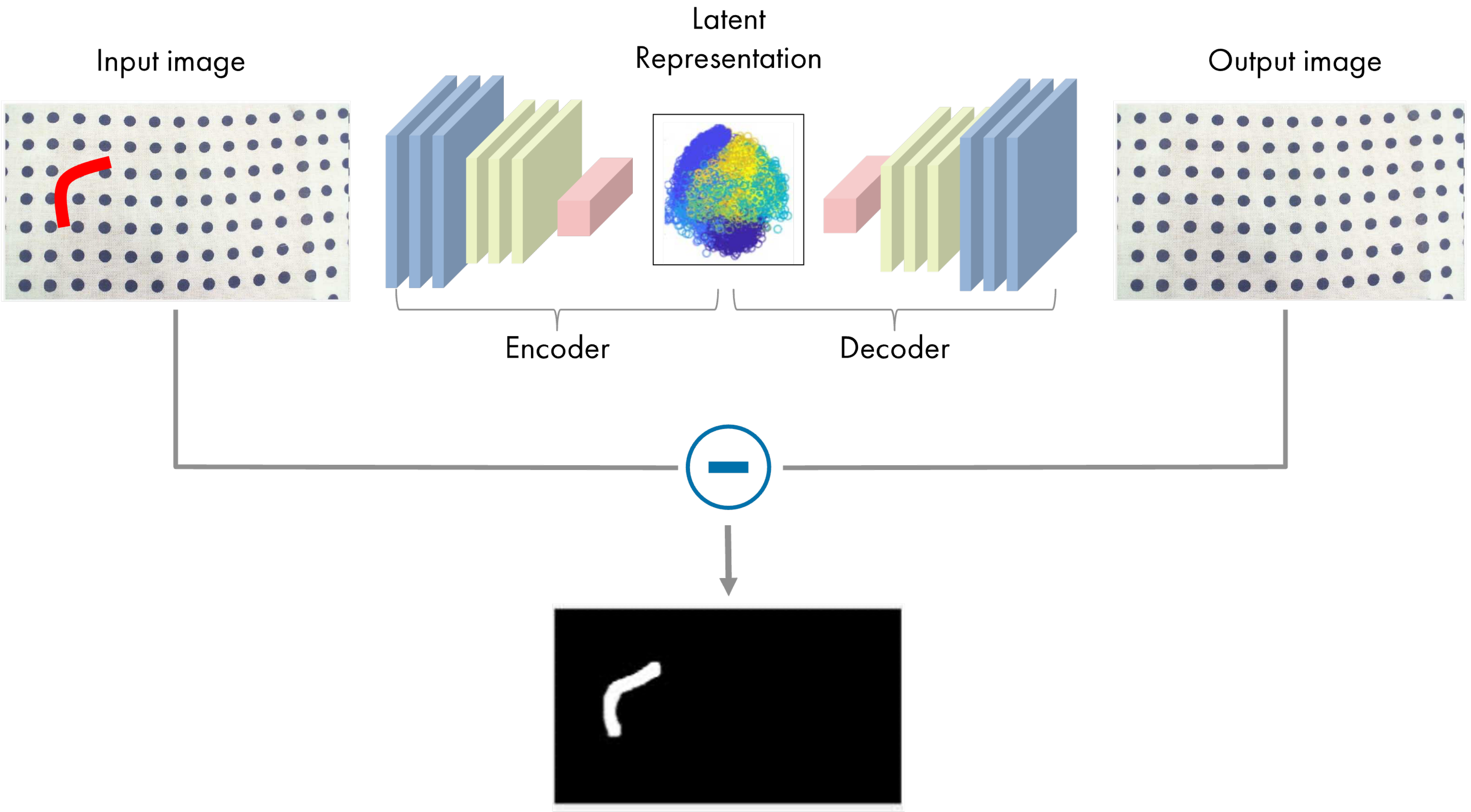


Picture source: https://medium.com/@khotpavankumar27/autoencoder-s-the-deep-generative-models-9645f6e94942

Autoencoders

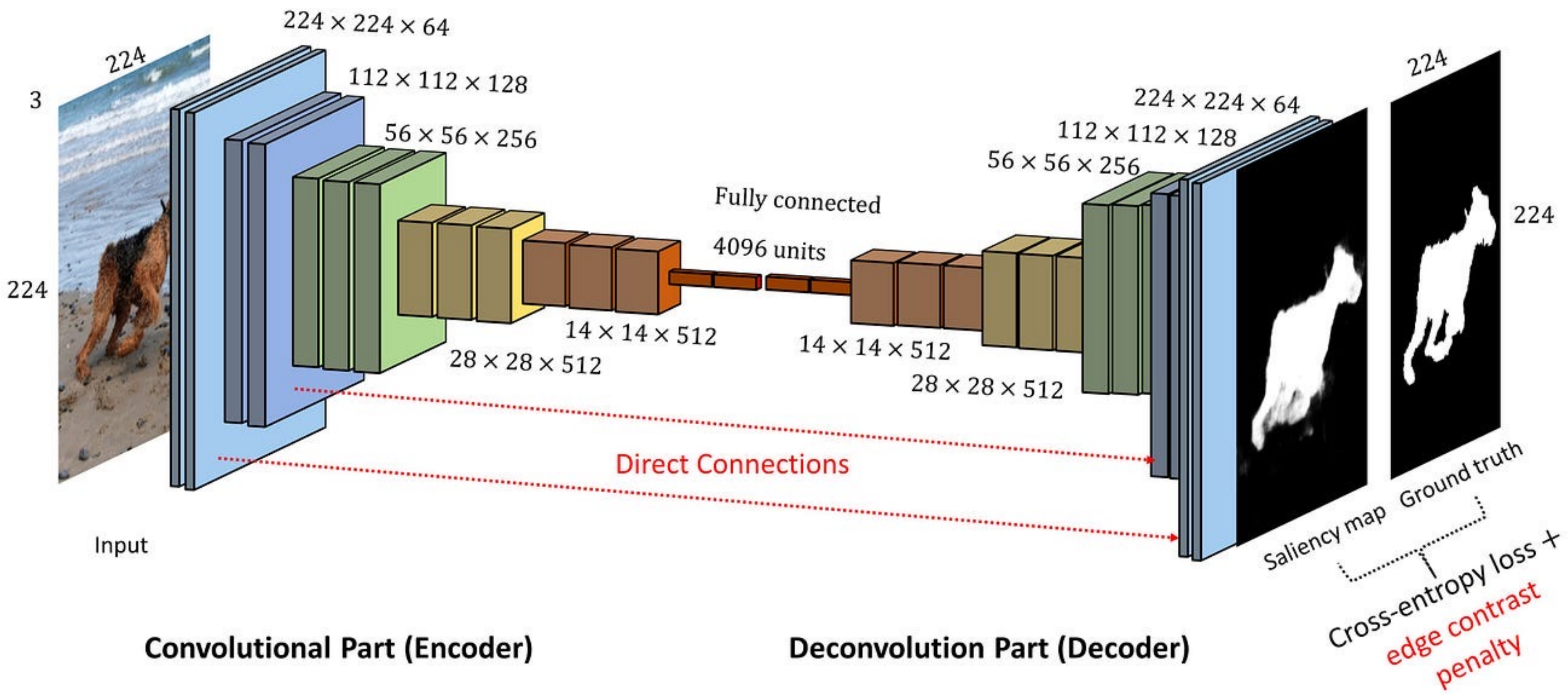


Autoencoders



Picture source: https://www.mathworks.com/discovery/autoencoder.html

Autoencoders



Knowledge Distillation vs. Data processing inequality?

Concept of distillation [edit]

Knowledge transfer from a large model to a small one somehow needs to teach the latter without loss of validity. If both models are trained on the same data, the smaller model may have insufficient capacity to learn a concise knowledge representation compared to the large model. However, some information about a concise knowledge representation is encoded in the pseudolikelihoods assigned to its output: when a model correctly predicts a class, it assigns a large value to the output variable corresponding to such class, and smaller values to the other output variables. The distribution of values among the outputs for a record provides information on how the large model represents knowledge. Therefore, the goal of economical deployment of a valid model can be achieved by training only the large model on the data, exploiting its better ability to learn concise knowledge representations, and then distilling such knowledge into the smaller model, by training it to learn the soft output of the large model.[1]

What is Knowledge Distillation?

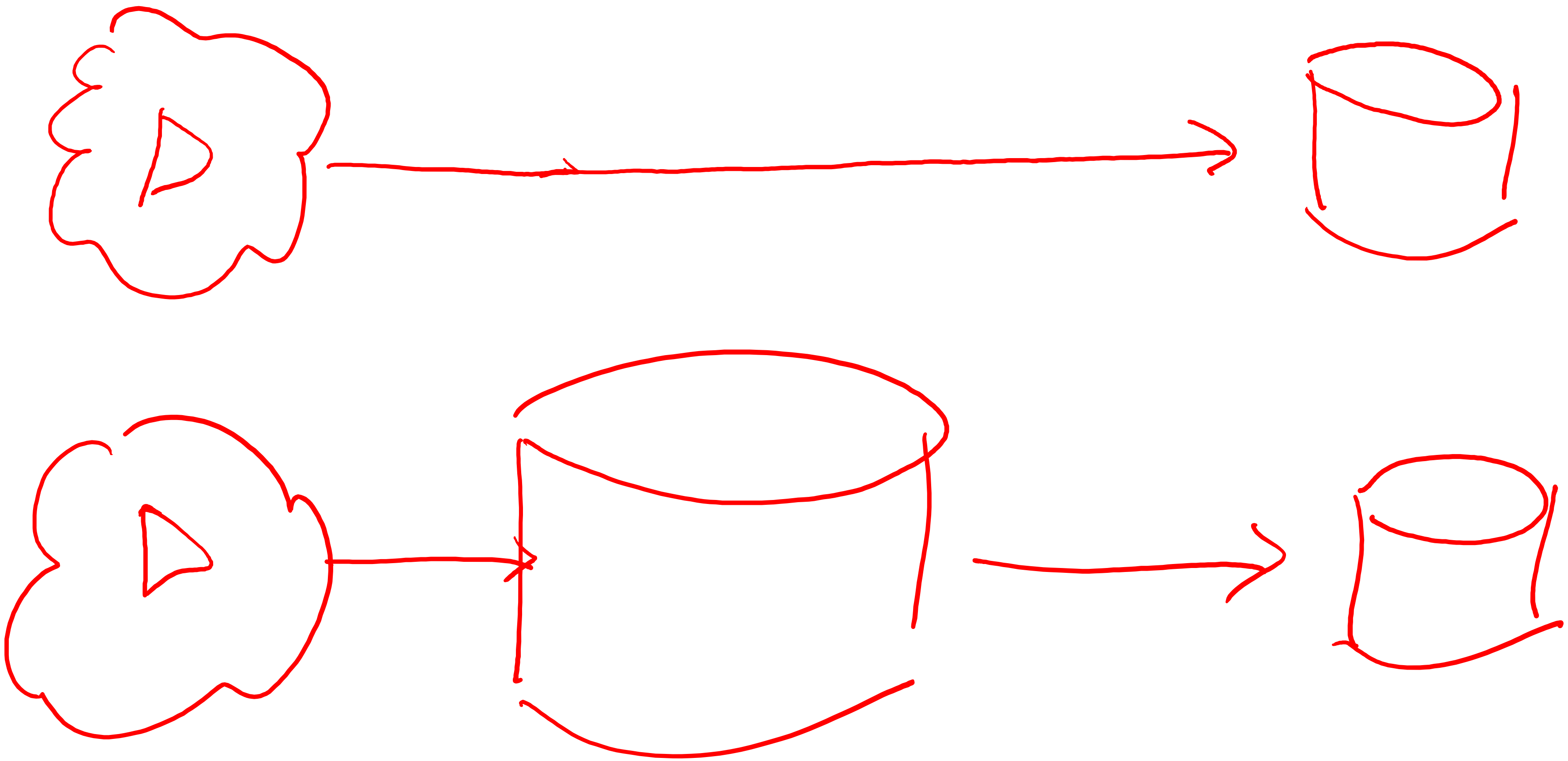
Knowledge distillation is a powerful technique in machine learning that allows us to transfer knowledge from a large, complex model to a smaller, simpler one. By doing so, we can reduce the memory footprint and computational requirements of the model without significant performance loss.

The fundamental idea behind knowledge distillation is to leverage the soft probabilities or logits of a larger "teacher network" along with the available class labels to train a smaller "student network". These soft probabilities provide more information than just the class labels, enabling the student network to learn more effectively.

1. **Offline Distillation**: Imagine an aspiring author learning from an already published, successful book. The published book (the teacher model) is complete and fixed. The new writer (the student model) learns from this book, attempting to write their own based on the insights gained. *In the context of neural networks*, this is like using a fully trained, sophisticated neural network to train a simpler, more efficient network. The student network learns from the established knowledge of the teacher without modifying it.

Screenshots from https://en.wikipedia.org/wiki/Knowledge_distillation , https://www.scaler.com/topics/nlp/distilbert/, https://medium.com/@aadityaura_26777/quantization-vs-distillation-in-neural-networks-a-comparison-8ef522e4fbec ,
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Knowledge Distillation vs. Data processing inequality?



Sufficient statistics

Following part builds on text, notation and examples from several sources, in particular:

[Casella,Berger'24] Statistical inference (2nd ed), 2024: Ch 6 Principles of Data Reduction. https://doi.org/10.1201/9781003456285

[Fithian'24] Statistics 210a: Theoretical Statistics, Berkeley, 2014: Lecture 4 sufficiency. https://stat210a.berkeley.edu/fall-2024/reader/sufficiency.html

[Scott'11] EECS 564: Estimation, Filtering, and Detection, University of Michigan, 2011: Lecture 5 Sufficient statistics. https://web.eecs.umich.edu/~cscott/past_courses/eecs564w11/index.html

[Cover,Thomas'06] Elements of Information Theory (2nd ed), 2006: Ch 2.9 Sufficient Statistics. https://www.doi.org/10.1002/047174882X

Parameter estimation

Suppose the probability distribution of a random variable X is determined by a parameter θ:

$X \sim f_\theta(x)$ Think of this as a conditional distribution: $f_\theta(x) = p(x|\theta)$

EXAMPLE: If X is a discrete Bernoulli RV, then its pmf (probability mass function) is parameterized by p:

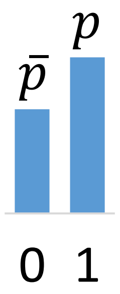
$f_p(x) =$?

Parameter estimation

Suppose the probability distribution of a random variable X is determined by a parameter θ:

$X \sim f_\theta(x)$ Think of this as a conditional distribution: $f_\theta(x) = p(x|\theta)$

EXAMPLE: If X is a discrete Bernoulli RV, then its pmf (probability mass function) is parameterized by p:



$$f_p(x) = \begin{cases} p & \text{if } x = 1 \\ \bar{p} & \text{if } x = 0 \end{cases} \qquad \bar{p} \coloneqq 1 - p$$

EXAMPLE: If X is a continuous Normal RV, then its pdf (probability density function) is parameterized by (μ, σ^2):

$$f_{(\mu,\sigma^2)}(x) = \text{?}$$

The parameter can also be a vector

Parameter estimation

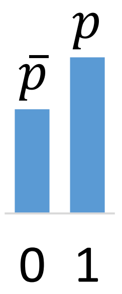
Suppose the probability distribution of a random variable X is determined by a parameter θ:

$$X \sim f_\theta(x)$$

Think of this as a conditional distribution: $f_\theta(x) = p(x|\theta)$

EXAMPLE: If X is a discrete Bernoulli RV, then its pmf (probability mass function) is parameterized by p:

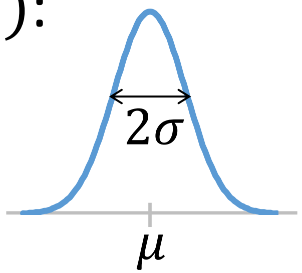
$$f_p(x) = \begin{cases} p & \text{if } x = 1 \\ \bar{p} & \text{if } x = 0 \end{cases}$$



EXAMPLE: If X is a continuous Normal RV, then its pdf (probability density function) is parameterized by (μ, σ^2):

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter can also be a vector



In statistical inference, we assume the functional form of f is known, but θ is hidden. We then observe a realization (a sample) $\mathbf{x}$ of iid RV's $\mathbf{X}$ and want to guess θ ("estimate θ").

Independent and Identically Distributed

Parameter estimation

Statistical Science
2001, Vol. 16, No. 3, 199–231
https://doi.org/10.1214/ss/1009213726

Statistical Modeling: The Two Cultures

Leo Breiman

Statistics

ML

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

In essence, when you "do statistics", you want to infer the process by which data you have was generated.

When you "do machine learning", you want to predict what future data will look like w.r.t. some variable.

In statistical inference, we assume the functional form of f is known, but θ is hidden. We then observe a realization (a sample) $\mathbf{x}$ of iid RV's $\mathbf{X}$ and want to guess θ ("estimate θ").

Independent and Identically Distributed

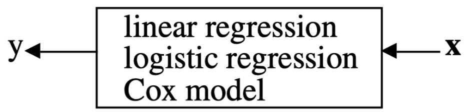
Parameter estimation

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

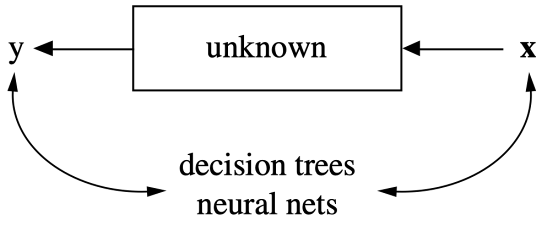
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;
- Kept statisticians from using more suitable algorithmic models;
- Prevented statisticians from working on exciting new problems;

basically, why ML took over ...

In statistical inference, we assume the functional form of f is known, but θ is hidden. We then observe a realization (a sample) $\mathbf{x}$ of iid RV's $\mathbf{X}$ and want to guess θ ("estimate θ").

Independent and Identically Distributed

Parameter estimation

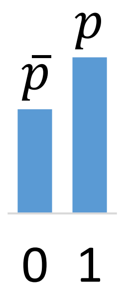
Suppose the probability distribution of a random variable X is determined by a parameter θ:

$$X \sim f_\theta(x)$$

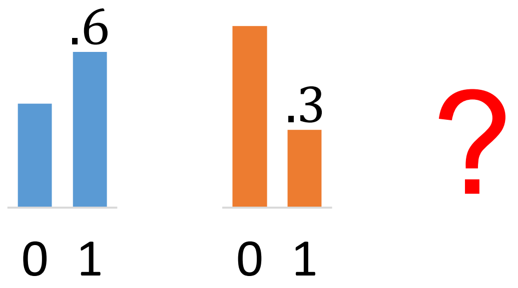
Think of this as a conditional distribution: $f_\theta(x) = p(x|\theta)$

EXAMPLE: If X is a discrete Bernoulli RV, then its pmf (probability mass function) is parameterized by p:

$$f_p(x) = \begin{cases} p & \text{if } x = 1 \\ \bar{p} & \text{if } x = 0 \end{cases}$$



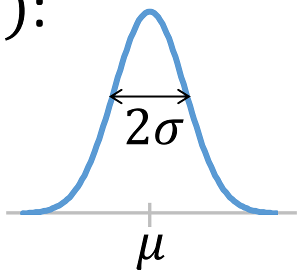
$\mathbf{x} = (1,1,0,1,1,1,0,0,1,1)$



EXAMPLE: If X is a continuous Normal RV, then its pdf (probability density function) is parameterized by (μ, σ^2):

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter can also be a vector



In statistical inference, we assume the functional form of f is known, but θ is hidden. We then observe a realization (a sample) $\mathbf{x}$ of iid RV's $\mathbf{X}$ and want to guess θ ("estimate θ").

Independent and Identically Distributed

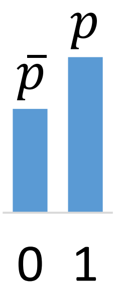
Parameter estimation

Suppose the probability distribution of a random variable X is determined by a parameter θ:

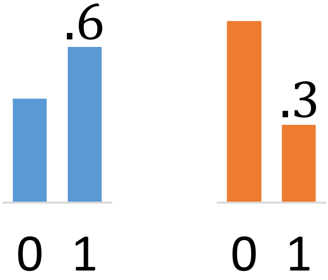
$X \sim f_\theta(x)$ Think of this as a conditional distribution: $f_\theta(x) = p(x|\theta)$

EXAMPLE: If X is a discrete Bernoulli RV, then its pmf (probability mass function) is parameterized by p:

$$f_p(x) = \begin{cases} p & \text{if } x = 1 \\ \bar{p} & \text{if } x = 0 \end{cases}$$



$\mathbf{x} = (1,1,0,1,1,1,0,0,1,1)$

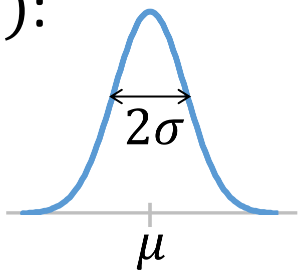


?

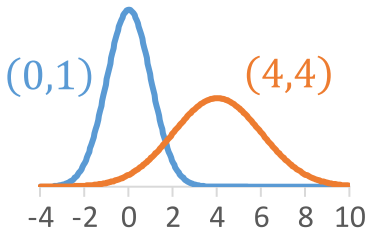
EXAMPLE: If X is a continuous Normal RV, then its pdf (probability density function) is parameterized by (μ, σ^2):

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter can also be a vector



$\mathbf{x} = (5.2, 2.5, 0.3, 4.2)$



?

In statistical inference, we assume the functional form of f is known, but θ is hidden. We then observe a realization (a sample) $\mathbf{x}$ of iid RV's $\mathbf{X}$ and want to guess θ ("estimate θ").

Independent and Identically Distributed

Sufficient statistics

Given a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ with unknown parameter θ, we would like to compress the measurements $\boldsymbol{x}$ into a low-dimensional statistic without affecting the quality of the possible inference about θ (i.e. we do not want to loose relevant information about θ).

In other words, we are interested in whether there exists a sufficient statistic $T(\mathbf{X})$ where the dimension of $\mathbf{t} = T(\mathbf{x})$ is $m < n$, s.t. $\mathbf{t}$ carries all the useful information from $\mathbf{x}$ about θ.

If such a sufficient statistic exists, then for the purpose of studying θ, we could discard the raw measurement $\mathbf{x}$ and retain only the compressed statistic $\mathbf{t}$.

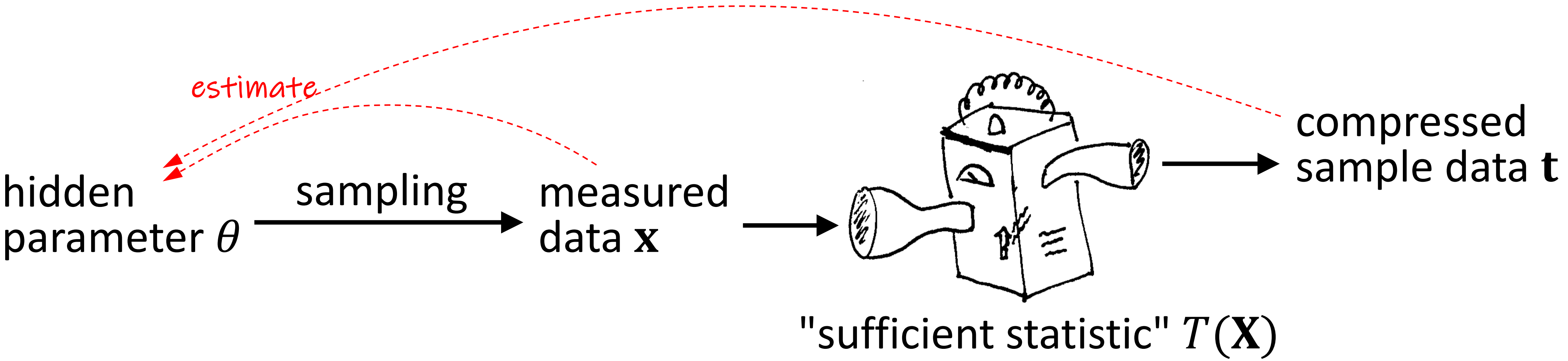
$$\begin{tikzcd} \text{hidden parameter } \theta \arrow[r, "\text{sampling}"] & \text{measured data } \mathbf{x} \arrow[l, bend right, dashed, "\text{estimate}"'] \end{tikzcd}$$

Sufficient statistics

Given a sample $\boldsymbol{x} = (x_1, \dots, x_n)$ with unknown parameter θ, we would like to compress the measurements $\boldsymbol{x}$ into a low-dimensional statistic without affecting the quality of the possible inference about θ (i.e. we do not want to loose relevant information about θ).

In other words, we are interested in whether there exists a sufficient statistic $T(\mathbf{X})$ where the dimension of $\mathbf{t} = T(\mathbf{x})$ is $m < n$, s.t. $\mathbf{t}$ carries all the useful information from $\mathbf{x}$ about θ.

If such a sufficient statistic exists, then for the purpose of studying θ, we could discard the raw measurement $\mathbf{x}$ and retain only the compressed statistic $\mathbf{t}$.



Sufficient statistics in the eyes of information theory

can also be a vector

Given a family of distributions $\{f_\theta(x)\}$ indexed by a parameter θ. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from f_θ, and $T(\mathbf{X})$ be a statistic (a quantity computed from the values in the sample).

Then $\theta \to \mathbf{X} \to T(\mathbf{X})$ forms a Markov chain $\theta \perp T(\mathbf{X}) \mid \mathbf{X}$

From the data processing inequality, we thus know

?

Sufficient statistics in the eyes of information theory

can also be a vector

Given a family of distributions $\{f_\theta(x)\}$ indexed by a parameter θ. Let $\mathbf{X} = (X_1, \dots, X_n)$ be an iid sample from f_θ, and $T(\mathbf{X})$ be a statistic (a quantity computed from the values in the sample).

Then $\theta \to \mathbf{X} \to T(\mathbf{X})$ forms a Markov chain $\theta \perp T(\mathbf{X}) \mid \mathbf{X}$

From the data processing inequality, we thus know

$$I(\theta; T(\mathbf{X})) \leq I(\theta; \mathbf{X})$$

A statistic is sufficient for θ if it preserves all the information in $\mathbf{X}$ about θ:

?

Sufficient statistics in the eyes of information theory

can also be a vector

Given a family of distributions $\{f_\theta(x)\}$ indexed by a parameter θ. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from f_θ, and $T(\mathbf{X})$ be a statistic (a quantity computed from the values in the sample).

Then $\theta \to \mathbf{X} \to T(\mathbf{X})$ forms a Markov chain $\theta \perp T(\mathbf{X}) \mid \mathbf{X}$

From the data processing inequality, we thus know

$$I(\theta; T(\mathbf{X})) \leq I(\theta; \mathbf{X})$$

A statistic is sufficient for θ if it preserves all the information in $\mathbf{X}$ about θ:

$$I\big(\theta; T(\mathbf{X})\big) = I(\theta; \mathbf{X})$$

PRACTICAL DEFINITION: A function $T(\mathbf{X})$ is said to be a sufficient statistic relative to the family $\{f_\theta(x)\}$ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ:

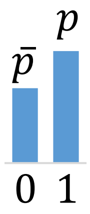
$\boxed{\theta \perp \mathbf{X} \mid T(\mathbf{X})}$ In other words, $\theta \to T(\mathbf{X}) \to \mathbf{X}$ also forms a Markov chain

This follows from our earlier definition of Markov chains (we just need to avoid to the temptation to equate the arrows with direction of time!)

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then, given a fixed n, what could be a sufficient statistic $T(\mathbf{X})$ for p



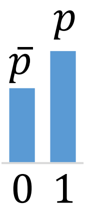
?

$$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$$
$$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$$

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



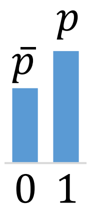
$$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$$
$$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$$

PROOF:

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \dots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$

$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$

PROOF: We know that $p \rightarrow \mathbf{X} \rightarrow k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \rightarrow k \rightarrow \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] =$?

$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] =$?

joint probability

$\mathbb{P}_p[T(\mathbf{X}) = k] =$?

$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] =$?

Updated 10/8/2025

Part 1: Theory
L09: Basics of entropy (7/7)
[Data processing inequality]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

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

10/6/2025

Pre-class conversations

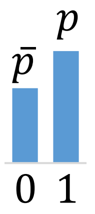
- Last class recapitulation
- Project ideas: Talk to me often. I can't meet right after class today, but before or in my office via email coordination.
  - I will look at the Piazza project posts before THU

- Today:
  - Sufficient statistics, information inequalities
  - Today or next time: start of compression

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



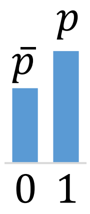
$$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$$
$$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$$

PROOF:

Example Sufficient statistics


This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$. Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$

$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$

PROOF: We know that $p \rightarrow \mathbf{X} \rightarrow k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \rightarrow k \rightarrow \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] =$?

$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] =$?

joint probability

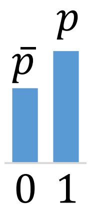
$\mathbb{P}_p[T(\mathbf{X}) = k] =$?

$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] =$?

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$
$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$

PROOF: We know that $p \rightarrow \mathbf{X} \rightarrow k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \rightarrow k \rightarrow \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \quad ?$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] = \quad ?$$

joint probability

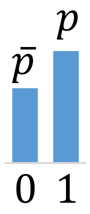
$$\mathbb{P}_p[T(\mathbf{X}) = k] = \quad ?$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = \quad ?$$

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$
$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$

PROOF: We know that $p \to \mathbf{X} \to k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \to k \to \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \boxed{\prod_i^n p^{x_i}(1-p)^{\bar{x}_i} = p^k(1-p)^{n-k}}$$

Very important later: Notice that the density $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ depends on $\mathbf{x}$ only through $k = T(\mathbf{X})$. Thus, $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ could be written as some function $g(T(\mathbf{x}), \theta)$, which is key to what happens next.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] =$$

joint probability

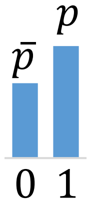
$$\mathbb{P}_p[T(\mathbf{X}) = k] =$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = \text{?}$$

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$$
$$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$$

PROOF: We know that $p \to \mathbf{X} \to k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \to k \to \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \boxed{\prod_i^n p^{x_i}(1-p)^{\bar{x}_i} = p^k(1-p)^{n-k}}$$

Very important later: Notice that the density $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ depends on $\mathbf{x}$ only through $k = T(\mathbf{X})$. Thus, $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ could be written as some function $g(T(\mathbf{x}), \theta)$, which is key to what happens next.

$$\underbrace{\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k]}_{\text{joint probability}} = \begin{cases} \mathbb{P}_p[\mathbf{X} = \mathbf{x}] & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

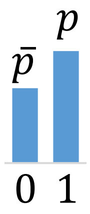
$$\mathbb{P}_p[T(\mathbf{X}) = k] = \binom{n}{k} \cdot p^k(1-p)^{n-k} \quad \text{binomial distribution}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = ?$$

Example Sufficient statistics

This is the parameter θ

EXAMPLE: Given a sample $\mathbf{x}$ of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$.
Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).



$$\mathbf{x}_1 = (1,1,0,1,1,1,0,0,1,1)$$
$$\mathbf{x}_2 = (1,0,1,1,1,1,0,0,1,1)$$

PROOF: We know that $p \to \mathbf{X} \to k$ forms a Markov chain from the fact that k is calculated from $\mathbf{X}$. To prove that k is a sufficient statistic for p, it is enough to show that $p \to k \to \mathbf{X}$ also forms a Markov chain.

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \boxed{\prod_i^n p^{x_i}(1-p)^{\bar{x}_i} = p^k(1-p)^{n-k}}$$

Very important later: Notice that the density $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ depends on $\mathbf{x}$ only through $k = T(\mathbf{X})$. Thus, $\mathbb{P}_p[\mathbf{X} = \mathbf{x}]$ could be written as some function $g(T(\mathbf{x}), \theta)$, which is key to what happens next.

$$\underbrace{\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k]}_{\text{joint probability}} = \begin{cases} \mathbb{P}_p[\mathbf{X} = \mathbf{x}] & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}_p[T(\mathbf{X}) = k] = \binom{n}{k} \cdot p^k(1-p)^{n-k} \quad \text{binomial distribution}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = \frac{\cancel{p^k(1-p)^{n-k}}}{\binom{n}{k} \cdot \cancel{p^k(1-p)^{n-k}}} = \begin{cases} \binom{n}{k}^{-1} & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

Thus, we have shown that $\mathbb{P}_p[\mathrm{X}|k] = \mathbb{P}[\mathrm{X}|k]$ is independent of p.
Concretely, all sequences x with k 1's (and n−k 0's) are equally likely.

Factorization Theorem

In the previous example, we had to guess the sufficient statistic and work out the conditional pmf $\mathbb{P}[\mathbf{X} \mid T(\mathbf{X}) = T(\mathbf{x})]$ by hand. This can become quite difficult in general.

As we will see next, we didn't really need to go to the trouble of calculating the conditional distribution. Once we noticed that the density $\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}]$ (also $f_\theta(\mathbf{x})$) depends on $\mathbf{x}$ only through $T(\mathbf{x})$, we could have concluded that the statistics $T(\mathbf{X})$ was sufficient.

The easiest way to identify and verify sufficient statistics is to show that the density $f_\theta(\mathbf{x})$ factorizes into a part that involves only the parameter θ and $T(\mathbf{x})$, and a part that involves only $\mathbf{x}$. This can be used as a working definition of sufficiency.

THEOREM: Let $f_\theta(\mathbf{x})$ (or $f(\mathbf{x}|\theta)$) denote the joint distribution of a data set $\mathbf{X}$, given parameter θ. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(T(\mathbf{x}), \theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ, $f_\theta(\mathbf{x})$ factorizes into:

$$f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta) \cdot h(\mathbf{x})$$

Notice that the unknown parameter θ interacts with the data $\mathbf{x}$ only via the statistic $T(\mathbf{x})$, and $h(\mathbf{x})$ is independent of θ.

This was $\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = p^k(1-p)^{n-k}$ in the previous example.

Example Sufficient statistics via factorization

EXAMPLE: Given a sample of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$. Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).

Can you find the factorization $f_p(\mathbf{x}) = g(T(\mathbf{x}), p) \cdot h(\mathbf{x})$ in our earlier proof ?

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \prod_i^n p^{x_i}(1-p)^{\overline{x_i}} = p^k(1-p)^{n-k}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] = \begin{cases} \mathbb{P}_p[\mathbf{X} = \mathbf{x}] & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}_p[T(\mathbf{X}) = k] = \binom{n}{k} \cdot p^k(1-p)^{n-k}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = \frac{\cancel{p^k(1-p)^{n-k}}}{\binom{n}{k} \cdot \cancel{p^k(1-p)^{n-k}}} = \begin{cases} \binom{n}{k}^{-1} & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

Thus, we have shown that $\mathbb{P}_p[\mathrm{X}|k] = \mathbb{P}[\mathrm{X}|k]$ is independent of p.

Example Sufficient statistics via factorization

EXAMPLE: Given a sample of n iid Bernoulli RVs $X_1, \ldots, X_n$ with unknown $\mathbb{P}[X_i = 1] = p$. Then $k = T(\mathbf{X}) = \sum_i X_i$ is a sufficient statistic for θ (assuming n is fixed).

Can you find the factorization $f_p(\mathbf{x}) = g(T(\mathbf{x}), p) \cdot h(\mathbf{x})$ in our earlier proof ?

We prove that by showing that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = k$ is independent of θ.

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \prod_i^n p^{x_i}(1-p)^{\overline{x_i}} = \underbrace{p^k(1-p)^{n-k}}_{g(k,p)} \cdot \underbrace{1}_{h(\mathbf{x})}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \wedge T(\mathbf{X}) = k] = \cdots$$

$$\mathbb{P}_p[T(\mathbf{X}) = k] = \binom{n}{k} \cdot p^k(1-p)^{n-k}$$

$$\mathbb{P}_p[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = k] = \frac{\mathbb{P}_p[\mathbf{X}=\mathbf{x}, T(\mathbf{X})=k]}{\mathbb{P}_p[k]} = \frac{\cancel{p^k(1-p)^{n-k}}}{\binom{n}{k} \cdot \cancel{p^k(1-p)^{n-k}}} = \begin{cases} \binom{n}{k}^{-1} & \text{if } \sum_i^n x_i = k \\ 0 & \text{otherwise} \end{cases}$$

Thus, we have shown that $\mathbb{P}_p[\mathrm{X}|k] = \mathbb{P}[\mathrm{X}|k]$ is independent of p.

Proof Factorization Theorem (1/2)

PROOF (DISCRETE CASE): sufficient statistics $\Leftrightarrow$ factorization $f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta) \cdot h(\mathbf{x})$

FIRST DIRECTION sufficient statistics $\Rightarrow$ factorization:
Assume $T(\mathbf{X})$ to be a sufficient statistics, i.e. $\theta \perp \mathbf{X}|T(\mathbf{X})$.
Let $f_\theta(\mathbf{x}, T(\mathbf{x}) = t)$ be the joint pdf of $\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t]$.

$f_\theta(\mathbf{x}) =$?

since T is a function of $\mathbf{X}$, and as long as $t = T(\mathbf{X})$

chain rule

by the definition of sufficient statistics $\theta \perp \mathbf{x}|t$

because t is a function of $\mathbf{x}$: $t = T(\mathbf{x})$

Proof Factorization Theorem (1/2)

PROOF (DISCRETE CASE): sufficient statistics $\Leftrightarrow$ factorization $f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta) \cdot h(\mathbf{x})$

FIRST DIRECTION sufficient statistics $\Rightarrow$ factorization:

Assume $T(\mathbf{X})$ to be a sufficient statistics, i.e. $\theta \perp \mathbf{X} | T(\mathbf{X})$.

Let $f_\theta(\mathbf{x}, T(\mathbf{x}) = t)$ be the joint pdf of $\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t]$.

$$
\begin{aligned}
f_\theta(\mathbf{x}) &= f_\theta(\mathbf{x}, t) && \text{since } T \text{ is a function of } \mathbf{X}\text{, and as long as } t = T(\mathbf{X}) \\
&= g_\theta(t) \cdot h_\theta(\mathbf{x}|t) && \text{chain rule} \\
&= \underbrace{g_\theta(t)}_{g(T(\mathbf{x}),\theta)} \cdot \underbrace{h(\mathbf{x}|t)}_{h(\mathbf{x})} && \text{by the definition of sufficient statistics } \theta \perp \mathbf{x}|t
\end{aligned}
$$

$h(\mathbf{x})$ because t is a function of $\mathbf{x}$: $t = T(\mathbf{x})$

$g(T(\mathbf{x}), \theta)$

This was $\mathbb{P}_p[\mathbf{X} = \mathbf{x}] = \prod_i^n p^{x_i}(1-p)^{\bar{x}_i} = p^k(1-p)^{n-k}$ in the previous example.

Proof Factorization Theorem (2/2)

OTHER DIRECTION: factorization $\Rightarrow$ sufficient statistics:

Assume $f_\theta(\mathbf{x}) = g(t, \theta) \cdot h(\mathbf{x})$.

We need to show that the conditional probability distribution $f_\theta(\mathbf{x}|t)$ of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ, i.e. $f_\theta(\mathbf{x}|t) = f(\mathbf{x}|t)$.

$$f_\theta(t) = \quad ?$$

$$f_\theta(\mathbf{x}|t) = \quad ?$$

Proof Factorization Theorem (2/2)


OTHER DIRECTION: factorization $\Rightarrow$ sufficient statistics:

Assume $f_\theta(\mathbf{x}) = g(t, \theta) \cdot h(\mathbf{x})$.

We need to show that the conditional probability distribution $f_\theta(\mathbf{x}|t)$ of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ, i.e. $f_\theta(\mathbf{x}|t) = f(\mathbf{x}|t)$.

$f_\theta(t) =$? — definition of marginal probability distribution

since t is a function of $\mathbf{x}$

using our assumption

factoring out a common factor

$$f_\theta(\mathbf{x}|t) = \frac{f_\theta(\mathbf{x}, t)}{f_\theta(t)} = \frac{f_\theta(\mathbf{x})}{f_\theta(t)}$$

definition of conditional probability distribution

? — does not depend on θ, hence T is a sufficient statistic

Proof Factorization Theorem (2/2)

OTHER DIRECTION: factorization ⇒ sufficient statistics:

Assume $f_\theta(\mathbf{x}) = g(t, \theta) \cdot h(\mathbf{x})$.

We need to show that the conditional probability distribution $f_\theta(\mathbf{x}|t)$ of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ, i.e. $f_\theta(\mathbf{x}|t) = f(\mathbf{x}|t)$.

$f_\theta(t) = \sum_{\mathbf{x}:T(\mathbf{x})=t} f_\theta(\mathbf{x}, t)$ — definition of marginal probability distribution

$= \sum_{\mathbf{x}:T(\mathbf{x})=t} f_\theta(\mathbf{x})$ — since t is a function of $\mathbf{x}$

$= \sum_{\mathbf{x}:T(\mathbf{x})=t} g(t, \theta) \cdot h(\mathbf{x})$ — using our assumption

$= g(t, \theta) \cdot \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})$ — factoring out a common factor

$$f_\theta(\mathbf{x}|t) = \frac{f_\theta(\mathbf{x}, t)}{f_\theta(t)} = \frac{f_\theta(\mathbf{x})}{f_\theta(t)}$$ — definition of conditional probability distribution

?

does not depend on θ, hence T is a sufficient statistic

Proof Factorization Theorem (2/2)

OTHER DIRECTION: factorization ⇒ sufficient statistics:

Assume $f_\theta(\mathbf{x}) = g(t,\theta) \cdot h(\mathbf{x})$.

We need to show that the conditional probability distribution $f_\theta(\mathbf{x}|t)$ of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ, i.e. $f_\theta(\mathbf{x}|t) = f(\mathbf{x}|t)$.

$$f_\theta(t) = \textstyle\sum_{\mathbf{x}:T(\mathbf{x})=t} f_\theta(\mathbf{x},t)$$ definition of marginal probability distribution

$$= \textstyle\sum_{\mathbf{x}:T(\mathbf{x})=t} f_\theta(\mathbf{x})$$ since t is a function of $\mathbf{x}$

$$= \textstyle\sum_{\mathbf{x}:T(\mathbf{x})=t} g(t,\theta) \cdot h(\mathbf{x})$$ using our assumption

$$= g(t,\theta) \cdot \textstyle\sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})$$ factoring out a common factor

$$f_\theta(\mathbf{x}|t) = \frac{f_\theta(\mathbf{x},t)}{f_\theta(t)} = \frac{f_\theta(\mathbf{x})}{f_\theta(t)}$$ definition of conditional probability distribution

$$= \frac{\cancel{g(t,\theta)} \cdot h(\mathbf{x})}{\cancel{g(t,\theta)} \cdot \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})}$$ does not depend on θ, hence T is a sufficient statistic

Sufficient Statistics & Factorization Theorem

The concept of sufficient statistics is due to Sir Ronald Fisher around 1920, thus before the advent of information theory.

The factorization theorem is also varyingly called:

- Fisher's factorization theorem
- Fisher-Neyman factorization theorem
- Neyman-Fisher factorization theorem
- Halmos-Savage factorization theorem

Sir Ronald Fisher (1890–1962)

Fisher, "On the mathematical foundations of theoretical statistics", Philosophical Transactions of the Royal Society A, 1922. https://doi.org/10.1098/rsta.1922.0009 . See also https://en.wikipedia.org/wiki/Sufficient_statistic, and references in https://encyclopediaofmath.org/wiki/Factorization_theorem

Normal (Gaussian) distribution: (μ, σ^2) are sufficient statistics

Example 6.2.9 (Normal sufficient statistic, both parameters unknown) Again assume that $X_1, \dots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2)$ but, unlike Example 6.2.4, assume that both μ and σ^2 are unknown so the parameter vector is $\theta = (\mu, \sigma^2)$. Now when using the Factorization Theorem, any part of the joint pdf that depends on either μ or σ^2 must be included in the g function. From (6.2.1) it is clear that the pdf depends on the sample $\mathbf{x}$ only through the two values $T_1(\mathbf{x}) = \bar{x}$ and $T_2(\mathbf{x}) = s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$. Thus, we can define $h(\mathbf{x}) = 1$ and

$$\begin{aligned} g(\mathbf{t}|\theta) &= g(t_1, t_2|\mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\left(n(t_1-\mu)^2 + (n-1)t_2\right)/(2\sigma^2)\right). \end{aligned}$$

Then it can be seen that

$$f(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x}). \tag{6.2.5}$$

Thus, by the Factorization Theorem, $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\bar{X}, S^2)$ is a sufficient statistic for (μ, σ^2) in this normal model. ‖

Example 6.2.9 demonstrates that, for the normal model, the common practice of summarizing a data set by reporting only the sample mean and variance is justified. The sufficient statistic $(\bar{X}, S^2)$ contains all the information about (μ, σ^2) that is available in the sample.

Exponential Family

The definition in terms of one *real-number* parameter can be extended to one *real-vector* parameter

$$\boldsymbol{\theta} \equiv [\,\theta_1,\, \theta_2,\, \ldots,\, \theta_s\,]^{\mathsf{T}} \,.$$

A family of distributions is said to belong to a vector exponential family if the probability density function (or probability mass function, for discrete distributions) can be written as

$$f_X(x \mid \boldsymbol{\theta}) = h(x)\, g(\boldsymbol{\theta})\, \exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(x)\Big)$$

- $T(x)$ is a *sufficient statistic* of the distribution. For exponential families, the sufficient statistic is a function of the data that holds all information the data x provides with regard to the unknown parameter values. This means that, for any data sets x and y, the likelihood ratio is the same, that is $\frac{f(x;\theta_1)}{f(x;\theta_2)} = \frac{f(y;\theta_1)}{f(y;\theta_2)}$ if $T(x) = T(y)$. This is true even if x and y are not equal to each other. The dimension of $T(x)$ equals the number of parameters of θ and encompasses all of the information regarding the data related to the parameter θ. The sufficient statistic of a set of independent identically distributed data observations is simply the sum of individual sufficient statistics, and encapsulates all the information needed to describe the posterior distribution of the parameters, given the data (and hence to derive any desired estimate of the parameters). (This important property is discussed further below.)

Exponential families have a large number of properties that make them extremely useful for statistical analysis. In many cases, it can be shown that *only* exponential families have these properties. Examples:

- Exponential families are the only families with sufficient statistics that can summarize arbitrary amounts of independent identically distributed data using a fixed number of values. (Pitman–Koopman–Darmois theorem)

IID

Source: https://en.wikipedia.org/wiki/Exponential_family

Aggregates in Databases

Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals*

JIM GRAY Gray@Microsoft.com
SURAJIT CHAUDHURI SurajitC@Microsoft.com
ADAM BOSWORTH AdamB@Microsoft.com
ANDREW LAYMAN AndrewL@Microsoft.com
DON REICHART DonRei@Microsoft.com
MURALI VENKATRAO MuraliV@Microsoft.com
Microsoft Research, Advanced Technology Division, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

FRANK PELLOW Pellow@vnet.IBM.com
HAMID PIRAHESH Pirahesh@Almaden.IBM.com
IBM Research, 500 Harry Road, San Jose, CA 95120

ICDE Influential Paper Awards

ICDE 2006

Jim Gray, Adam Bosworth, Andrew Layman, Hamid Pirahesh

Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total, ICDE 1996

Citation: This seminal paper defined a simple SQL construct that enables one to efficiently compute aggregations over all combinations of group-by columns in a single query, where previous approaches required multiple queries. This feature has had significant impact on industry and is now incorporated in all major database systems.

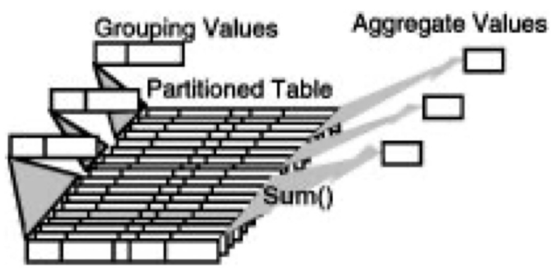


Figure 2. The GROUP BY relational operator partitions a table into groups. Each group is then aggregated by a function. The aggregation function summarizes some column of groups returning a value for each group.

Consider aggregating a two dimensional set of values $\{X_{ij} \mid i = 1, \ldots, I;\ j = 1, \ldots, J\}$. Aggregate functions can be classified into three categories:

Distributive: Aggregate function $F()$ is distributive if there is a function $G()$ such that $F(\{X_{i,j}\}) = G(\{F(\{X_{i,j} \mid i = 1, \ldots, I\}) \mid j = 1, \ldots J\})$. COUNT(), MIN(), MAX(), SUM() are all distributive. In fact, $F = G$ for all but COUNT(). $G =$ SUM() for the COUNT() function. Once order is imposed, the cumulative aggregate functions also fit in the distributive class.

Algebraic: Aggregate function $F()$ is algebraic if there is an M-tuple valued function $G()$ and a function $H()$ such that $F(\{X_{i,j}\}) = H(\{G(\{X_{i,j} \mid i = 1, \ldots, I\}) \mid j = 1, \ldots, J\})$. Average(), standard deviation, MaxN(), MinN(), center_of_mass() are all algebraic. For Average, the function $G()$ records the sum and count of the subset. The $H()$ function adds these two components and then divides to produce the global average. Similar techniques apply to finding the N largest values, the center of mass of group of objects, and other algebraic functions. The key to algebraic functions is that a fixed size result (an M-tuple) can summarize the sub-aggregation.

Holistic: Aggregate function $F()$ is holistic if there is no constant bound on the size of the storage needed to describe a sub-aggregate. That is, there is no constant M, such that an M-tuple characterizes the computation $F(\{X_{i,j} \mid i = 1, \ldots, I\})$. Median(), MostFrequent() (also called the Mode()), and Rank() are common examples of holistic functions.

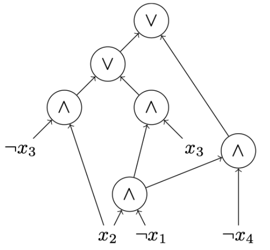
We know of no more efficient way of computing super-aggregates of holistic functions than the 2^N-algorithm using the standard GROUP BY techniques. We will not say more about cubes of holistic functions.

Circuits: another form of the factorization theorem?

Tractable Circuits in Database Theory

Antoine Amarilli
LTCI, Télécom Paris
Institut Polytechnique de Paris
antoine.amarilli@telecom-paris.fr

Florent Capelli
Université d'Artois, CNRS, UMR 8188
Centre de Recherche en Informatique de Lens
capelli@cril.fr



x_1	x_2	x_3	x_4
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	1	0	0
1	1	0	1

Figure 1: A DNNF and its satisfying valuations

Table 1: Tractable tasks for circuit classes. The input circuit C is of size s, depth d, and n variables; k is the number of solutions to output (for Enum and Sampling).

Task	Circuit class	Complexity	Ref.
WMC	d-DNNF	$O(ns)$	[94]
WMC in ring	d-D	$O(ns)$	[107]
ApproxWMC	DNF	FPRAS	[86, 127]
ApproxWMC	SDNNF	FPRAS	[23]
Enum	d-SDNNF	$O(s+nk)$	[8]
Sampling	d-DNNF	$O(s+dnk)$	[123]

ABSTRACT

This work reviews how database theory uses tractable circuit classes from knowledge compilation. We present relevant query evaluation tasks, and notions of tractable circuits. We then show how these tractable circuits can be used to address database tasks. We first focus on Boolean provenance and its applications for aggregation tasks, in particular probabilistic query evaluation. We study these for Monadic Second Order (MSO) queries on trees, and for safe Conjunctive Queries (CQs) and Union of Conjunctive Queries (UCQs). We also study circuit representations of query answers, and their applications to enumeration tasks: both in the Boolean setting (for MSO) and the multivalued setting (for CQs and UCQs).

Table 2: From Boolean circuits to relational circuits.

Boolean Circuits	Relational Circuits
NNF	$\{\bar{\cup}, \bowtie\}$
DNNF	$\{\bar{\cup}, \times\}$ or $\{\cup, \times\}$ if smooth
d-DNNF	$\{\bar{\uplus}, \times\}$ or $\{\uplus, \times\}$ if smooth
dec-DNNF	$\{\mathrm{dec}, \times\}$

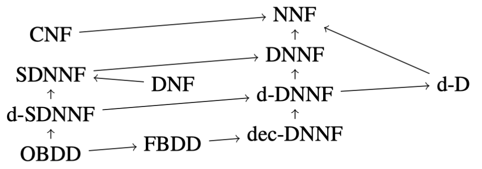


Figure 2: Inclusions between circuit classes. Arrows denote inclusion (i.e., linear-time transformations); they are all known to be strict in the sense that reverse arrows do not exist, except inclusions involving d-Ds which are not separated from d-DNNF [106] and only conditionally separated from NNF. Most results are in [6, 14].

TRACTABLE BOOLEAN AND ARITHMETIC CIRCUITS*

ADNAN DARWICHE+

Darwiche. Tractable Boolean and Arithmetic Circuits, 2022. https://arxiv.org/pdf/2202.02942,
Amarilli, Capelli, Tractable Circuits in Database Theory, 2024. https://sigmodrecord.org/?smd_process_download=1&download_id=13771
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Information Inequalities

Best reference:
[Yeung'08] Yeung, Information Theory and Network Coding, 2008. Ch 2.6, 2.7, 13, 14, 15
http://iest2.ie.cuhk.edu.hk/~whyeung/tempo/main2.pdf

Basic inequalities

Shannon's information measures refer to entropy, conditional entropy, mutual information, and conditional mutual information (but not interaction information!).

They can be expressed as linear combinations of entropies:

$$H(X|Y) = \quad ?$$

$$I(X;Y) = \quad ?$$

$$I(X;Y|Z) = \quad ?$$

Basic inequalities

Shannon's information measures refer to entropy, conditional entropy, mutual information, and conditional mutual information (but not interaction information!).

They can be expressed as linear combinations of entropies:

$$H(X|Y) = H(X,Y) - H(Y)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$$

by repeated expansion of conditional entropies; also holds if we replace variables with sets of variables (grouping rule)

They are also special cases of conditional mutual information.

$$H(X) = ?$$
$$H(X|Z) = ?$$
$$I(X;Y) = ?$$

Assume φ to be degenerate RV that takes on a constant value with probability 1

Basic inequalities

Shannon's information measures refer to entropy, conditional entropy, mutual information, and conditional mutual information (but not interaction information!).

They can be expressed as linear combinations of entropies:

$$H(X|Y) = H(X,Y) - H(Y)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$$

by repeated expansion of conditional entropies; also holds if we replace variables with sets of variables (grouping rule)

They are also special cases of conditional mutual information.

$$H(X) = I(X;X|\varphi)$$
$$H(X|Z) = I(X;X|Z)$$
$$I(X;Y) = I(X;Y|\varphi)$$

Assume φ to be degenerate RV that takes on a constant value with probability 1

With the basic inequalities we refer to the fact that all Shannon's information measures are non-negative (because conditional mutual information is ≥ 0).

$$I(U;V|W) \geq 0$$

U, V, W can be arbitrary joint entropies

Shannon-type inequalities Γ_n (and constraints)

Shannon-type inequalities are inequalities on information measures implied by the basic inequalities and possibly additional constraints on the joint distribution of the RVs involved.

EXAMPLE: data-processing inequality for $X \to Y \to Z$:

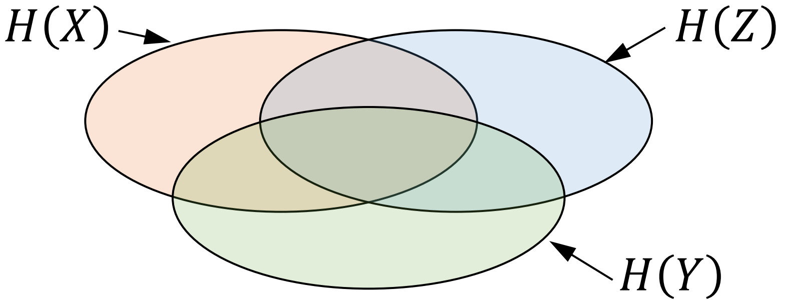
From $I(X;Z|Y) = 0$ and basic inequalities, we derived $I(X;Y) \geq I(X;Z)$

not a basic inequality

Information inequalities are the inequalities that govern the impossibilities in information theory. They imply that certain things cannot happen. For this reason, they are sometimes referred to as the laws of information theory.

EXAMPLE : $n = 3$ variables with given $k = 2^3 - 1 = 7$ joint entropies:

$$H(X) = 2 \quad H(X,Y) = 4 \quad H(X,Y,Z) = 5$$
$$H(Y) = 3 \quad H(X,Z) = 4$$
$$H(Z) = 4 \quad H(Y,Z) = 4$$



Find 3 RVs that fulfill those constraints ?

Shannon-type inequalities Γ_n (and constraints)

Shannon-type inequalities are inequalities on information measures implied by the basic inequalities and possibly additional constraints on the joint distribution of the RVs involved.

EXAMPLE: data-processing inequality for $X \to Y \to Z$:

not a basic inequality

From $I(X;Z|Y) = 0$ and basic inequalities, we derived $I(X;Y) \geq I(X;Z)$

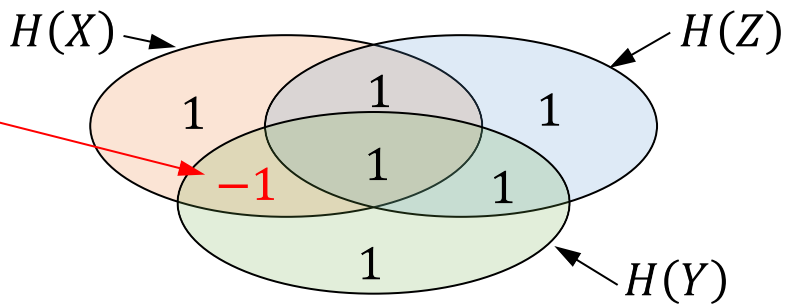
Information inequalities are the inequalities that govern the impossibilities in information theory. They imply that certain things cannot happen. For this reason, they are sometimes referred to as the laws of information theory.

EXAMPLE : $n = 3$ variables with given $k = 2^3 - 1 =$ 7 joint entropies:

$$
\begin{array}{lll}
H(X) = 2 & H(X,Y) = 4 & H(X,Y,Z) = 5 \\
H(Y) = 3 & H(X,Z) = 4 & \\
H(Z) = 4 & H(Y,Z) = 4 &
\end{array}
$$

$I(X;Y|Z) \ngeq 0$

not possible ☹



Almost all the information inequalities known to date are Shannon-type inequalities and thus implied by the basic inequalities.

Applications in Databases

Worst-case Optimal Join Algorithms

HUNG Q. NGO, University at Buffalo, SUNY
ELY PORAT, Bar-Ilan University
CHRISTOPHER RÉ, Stanford University
ATRI RUDRA, University at Buffalo, SUNY

PODS'12, JACM'18

Efficient join processing is one of the most fundamental and well-studied tasks in database research. In this work, we examine algorithms for natural join queries over many relations and describe a new algorithm to process these queries optimally in terms of worst-case data complexity. Our result builds on recent work by Atserias, Grohe, and Marx, who gave bounds on the size of a natural join query in terms of the sizes of the individual relations in the body of the query. These bounds, however, are not constructive: they rely on Shearer's entropy inequality, which is information-theoretic. Thus, the previous results leave open the question of whether there exist algorithms whose runtimes achieve these optimal bounds. An answer to this question may be interesting to database practice, as we show in this article that any project-join style plans, such as ones typically employed in a relational database management system, are asymptotically slower than the optimal for some queries. We present an algorithm whose runtime is worst-case optimal for all natural join queries. Our result may be of independent interest, as our algorithm also yields a constructive proof of the general fractional cover bound by Atserias, Grohe, and Marx without using Shearer's inequality. This bound implies two famous inequalities in geometry: the Loomis-Whitney inequality and its generalization, the Bollobás-Thomason inequality. Hence, our results algorithmically prove these inequalities as well. Finally, we discuss how our algorithm can be used to evaluate full conjunctive queries optimally, to compute a relaxed notion of joins and to optimally (in the worst-case) enumerate all induced copies of a fixed subgraph inside of a given large graph.

Decision Problems in Information Theory

Mahmoud Abo Khamis
relationalAI, Berkeley, CA, USA
Phokion G. Kolaitis
UC Santa Cruz and IBM Research - Almaden, Santa Cruz, CA, USA
Hung Q. Ngo
relationalAI, Berkeley, CA, USA
Dan Suciu
University of Washington, Seattle, WA, USA

ICALP'20

B.2 Application to Relational Query Evaluation

The problem of bounding the number of copies of a graph inside of another graph has a long and interesting history [17, 4, 14, 35]. The subgraph homomorphism problem is a special case of the relational query evaluation problem, in which case we want to find an upper bound on the output size of a full conjunctive query. Using the entropy argument from [14], *Shearer's lemma* in particular, Atserias, Grohe, and Marx [5] established a tight upper bound on the answer to a full conjunctive query over a database. Note that Shearer's lemma is a Shannon-type inequality. Their result was extended to include functional dependencies and more generally degree constraints in a series of recent work in database theory [19, 2, 3]. All these results can be cast as applications of Shannon-type inequalities. For a simple example, let $R(X,Y), S(Y,Z), T(Z,U)$ be three binary relations (tables), each with N tuples, then their join $R(X,Y) \bowtie S(Y,Z) \bowtie T(Z,U)$ can be as large as N^2 tuples. However, if we further know that the functional dependencies $XZ \to U$ and $YU \to X$ hold in the output, then one can prove that the output size is $\le N^{3/2}$, by using the following Shannon-type information inequality:

$$h(XY) + h(YZ) + h(ZU) + h(X|YU) + h(U|XZ) \ge 2h(XYZU) \tag{24}$$

Ngo, Porat, Re, Rudra. Worst-case Optimal Join Algorithms, JAC 2018 (PODS 2012). https://doi.org/10.1145/3180143 ,
Khamis, Kolaitis, Ngo, Suciu, "Decision Problems in Information Theory", ICALP 2020. https://doi.org/10.4230/LIPIcs.ICALP.2020.106
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Applications in Databases

1.1 The problems

In the history of query evaluation in database, logic and constraint satisfaction areas, there are three research threads which have yielded spectacular results recently.

Thread 1: size bound for conjunctive queries. From the seminal work of Grohe and Marx [33], Atserias, Grohe, and Marx [8], and Gottlob, Lee, Valiant and Valiant [30], we now know of a deep connection between the output size bound of a conjunctive query with (or without) functional dependencies (FD) and information theory. In particular, we can derive tight output size bounds by solving a convex optimization problem whose variables are marginal entropies. Briefly, the bound works as follows. Consider a conjunctive query Q represented by a multi-hypergraph $\mathcal{H} = (V, \mathcal{E})$, where $V = [n]$ is identified with the set of variables $A_1, \ldots, A_n$. To each hyperedge $F \in \mathcal{E}$ there is an input relation R_F whose attributes are $(A_i)_{i \in F}$. A function $h : 2^V \to \mathbb{R}_+$ is called *entropic* if there exists a joint distribution on n variables such that $h(F)$ is the marginal entropy of the distribution on the variables in F, for every non-empty set $F \subseteq V$; by convention, $h(\emptyset) = 0$. Let Γ_n^* denote the set of all n-variable entropic functions.[1] Let CC denote the set of "cardinality constraints" of the form $N_F = |R_F|$, obtained from the input database instance. Let FD denote the set of "FD constraints" of the form $X \to Y$, where $\emptyset \subseteq X \subset Y \subseteq V$.[2] From the cardinality- and FD-constraints, we define two classes of set functions:

What do Shannon-type Inequalities, Submodular Width, and Disjunctive Datalog have to do with one another?

Mahmoud Abo Khamis
LogicBlox Inc.

Hung Q. Ngo
LogicBlox Inc.

Dan Suciu
LogicBlox Inc. and
University of Washington

PODS'17

ABSTRACT

Recent works on bounding the output size of a conjunctive query with functional dependencies and degree bounds have shown a deep connection between fundamental questions in information theory and database theory. We prove analogous output bounds for *disjunctive datalog rules*, and answer several open questions regarding the tightness and looseness of these bounds along the way. The bounds are intimately related to Shannon-type information inequalities. We devise the notion of a "proof sequence" of a specific class of Shannon-type information inequalities called "Shannon flow inequalities". We then show how a proof sequence can be used as symbolic instructions to guide an algorithm called PANDA, which answers disjunctive datalog rules within the size bound predicted. We show that PANDA can be used as a black-box to devise algorithms matching precisely the fractional hypertree width and the submodular width runtimes for aggregate and conjunctive queries *with* functional dependencies and degree bounds.

Our results improve upon known results in three ways. First, our bounds and algorithms are for the much more general class of disjunctive datalog rules, of which conjunctive queries are a special case. Second, the runtime of PANDA matches precisely the submodular width bound, while the previous algorithm by Marx has a runtime that is polynomial in this bound. Third, our bounds and algorithms work for queries with input cardinality bounds, functional dependencies, *and* degree bounds.

Overall, our results showed a deep connection between three seemingly unrelated lines of research; and, our results on proof sequences for Shannon flow inequalities might be of independent interest.

Abo Khamis, Ngo, Suciu. What Do Shannon-type Inequalities, Submodular Width, and Disjunctive Datalog Have to Do with One Another? PODS 2017. https://doi.org/10.1145/3034786.3056105

Applications in Databases

Applications of Information Inequalities
to Database Theory Problems *

Dan Suciu
University of Washington
Seattle, USA
suciu@cs.washington.edu

June 6, 2024

LICS'23 keynote

Abstract

The paper describes several applications of information inequalities to problems in database theory. The problems discussed include: upper bounds of a query's output, worst-case optimal join algorithms, the query domination problem, and the implication problem for approximate integrity constraints. The paper is self-contained: all required concepts and results from information inequalities are introduced here, gradually, and motivated by database problems.

Suciu. Applications of Information Inequalities to Database Theory Problems, LICS keynote 2023. https://arxiv.org/pdf/2304.11996 ,
slides: https://homes.cs.washington.edu/~suciu/talk-lics-2023.pdf ,
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R?

- $h(X) \leq h(XY) \leq h(XYZ)$
  Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
  Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.
- $h(XYZ|X) \geq h(XYZ|XY)$
  Max frequency(X) is $\geq$ max frequency(XY).
- Careful! $h(XZ) + h(YZ) \geq h(XYZ) + h(Z)$,
  but $\underbrace{|\Pi_{XZ}(R)|}_{3} \cdot \underbrace{|\Pi_{YZ}(R)|}_{3} \not\geq \underbrace{|R|}_{5} \cdot \underbrace{|\Pi_Z(R)|}_{2}$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Source: Dan Suciu. CS294-248: Special topics in databases, Berkeley fall 2023. See also videos: https://berkeley-cs294-248.github.io/
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Information inequalities Γ_n^*

Information inequalities are the inequalities that govern the impossibilities in information theory. They imply that certain things cannot happen. For this reason, they are sometimes referred to as the laws of information theory.

An information inequality or identity involves (linear combinations of) Shannon's information measures only (and possibly with constant terms) and is said to always hold if it holds for any joint distribution for the random variables involved.

There exist laws in information theory that are not implied by the basic inequalities (called non-Shannon-type inequalities). This celebrated result was published by [Zhang,Yeung'98]

PROPOSITION: The following information inequality always holds on any list of five random variables X, Y, Z, U, V, but is not implied by the basic inequalities:

$$H(X) + H(Y) + I(U;V|X) + I(U;V|Y) + 2I(U;V|Z) + I(U,V;Z) \geq H(X,Z) + 2I(U;V)$$

Key proof insight: $I(XY;Z|UV) = 0$ can be assumed for a different argument

Quoted from: [Yeung'08] Information Theory and Network Coding, 2008. http://iest2.ie.cuhk.edu.hk/~whyeung/tempo/main2.pdf / [Zhang,Yeung'98]. On characterization of entropy function via information inequalities, 1998. https://doi.org/10.1109/18.681320 / For an accessible proof, see [Cramer, Fehr'15] The Mathematical Theory of Information, and Applications, lecture notes. https://staff.science.uva.nl/c.schaffner/courses/infcom/2014/notes/CramerFehr.pdf

A formalization of entropic vectors

{1, 2, ... n}

- Given a set of n RVs $\Theta = \{X_1, \dots, X_n\}$, written as $\Theta = \{X_i\}, i \in [n]$.
- Associated with Θ are $2^n - 1$ joint entropies $H(X_1)$,..., $H(X_1, \dots, X_n)$, written as $H_\Theta(\alpha) = H(X_\alpha)$ for any subset of $[n]$. Call the function $H_\Theta(\alpha), \alpha \in 2^{[n]}$ the entropy function of Θ.
- Example: $H(X_1, X_2, X_4)$ is $H_\Theta(\alpha)$ for $\alpha = \{1,2,4\}$.
- Together, the joint entropies form a point in the $2^n - 1$ dimensional entropy space $\mathbb{R}^{2^n-1}$.
- In turn, a point in that space is called entropic if the point corresponds to the entropy function H_Θ of some set Θ of n RVs. Let $\Gamma_n^* \subset \mathbb{R}^{2^n-1}$ be the set of all entropic points.
- How does that space $\Gamma_n^* \subset \mathbb{R}^{2^n-1}$ look like?

Our earlier EXAMPLE: $n = 3$, thus $k = 2^3 - 1 = 7$ joint entropies, representing a point in $\mathbb{R}^7$

$H(X) = 2$	$H(X,Y) = 4$	$H(X,Y,Z) = 5$
$H(Y) = 3$	$H(X,Z) = 4$	
$H(Z) = 4$	$H(Y,Z) = 4$	

?

Entropic vectors

- Given a set of n RVs $\Theta = \{X_1, \ldots, X_n\}$, written as $\Theta = \{X_i\}, i \in [n]$.
- Associated with Θ are $2^n - 1$ joint entropies $H(X_1)$,..., $H(X_1, \ldots, X_n)$, written as $H_\Theta(\alpha) = H(X_\alpha)$ for any subset of $[n]$. Call the function $H_\Theta(\alpha), \alpha \in 2^{[n]}$ the entropy function of Θ.
- Example: $H(X_1, X_2, X_4)$ is $H_\Theta(\alpha)$ for $\alpha = \{1,2,4\}$.
- Together, the joint entropies form a point in the $2^n - 1$ dimensional entropy space $\mathbb{R}^{2^n-1}$.
- In turn, a point in that space is called entropic if the point <u>corresponds to the entropy function H_Θ of some set Θ of n RVs</u>. Let $\Gamma_n^* \subset \mathbb{R}^{2^n-1}$ be the set of all entropic points.
- How does that space $\Gamma_n^* \subset \mathbb{R}^{2^n-1}$ look like?

Our earlier EXAMPLE: $n = 3$, thus $k = 2^3 - 1 = 7$ joint entropies, representing a point in $\mathbb{R}^7$

$$H(X) = 2 \qquad H(X,Y) = 4 \qquad H(X,Y,Z) = 5$$
$$H(Y) = 3 \qquad H(X,Z) = 4$$
$$H(Z) = 4 \qquad H(Y,Z) = 4$$

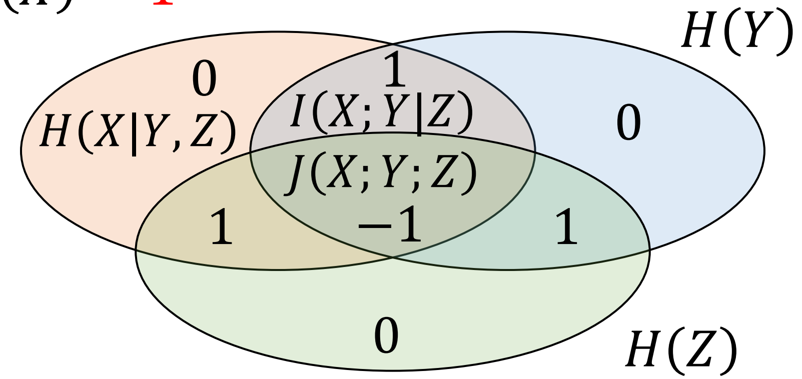
$I(X;Y|Z) \ngeq 0$

Thus this point $(2,3,4,4,4,5) \notin \Gamma_3^*$?

A subtlety: entropic vectors Γ_n^* vs. almost entropic vectors $\bar{\Gamma}_n^*$

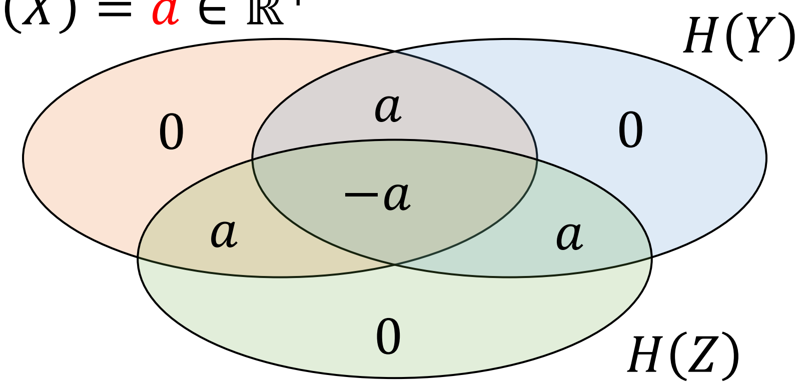
Our earlier "parity example":

$H(X) = 1$



More generally (from basic inequalities):

$H(X) = a \in \mathbb{R}^+$



However, a more careful analysis shows that all variables X, Y, Z need to be uniform for this example to work, which implies only discrete particular entropies as possible.

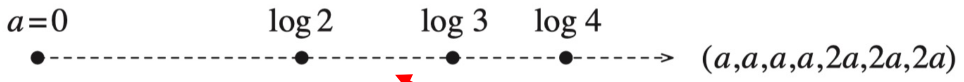


Fig. 15.2. The values of a for which $(a, a, a, 2a, 2a, 2a, 2a)$ is in Γ_3.

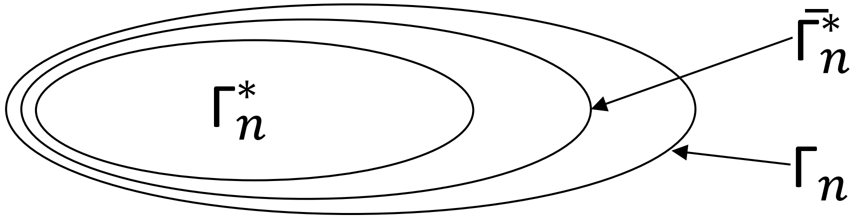
Γ_n^* set of all entropic vectors

$\bar{\Gamma}_n^*$ set of all almost entropic vectors: defined as topological closure of Γ_n^*

Γ_n subset of vectors that fulfill the Shannon inequalities

The closure of a subset S of points in a topological space consists of all points in S together with all limit points of S.

Intuitively, it is possible to create a mixture model that models any rational number. The "closure" extends that to the real numbers.



For details see: [Yeung'08] Yeung, Information Theory and Network Coding, 2008. Ch 15.1 http://iest2.ie.cuhk.edu.hk/~whyeung/tempo/main2.pdf