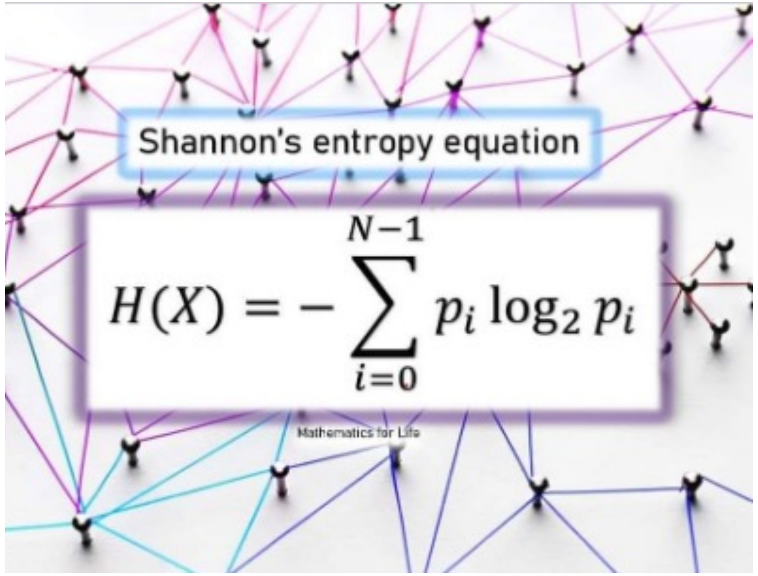
Intro to Information Theory

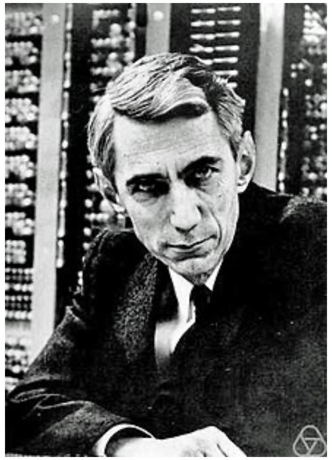
Statistics and data analysis

Ben Galili

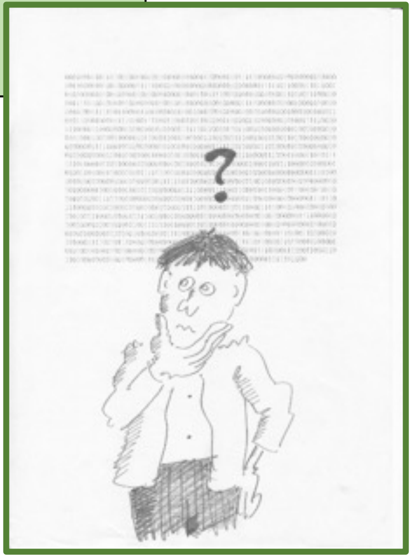
Zohar Yakhini

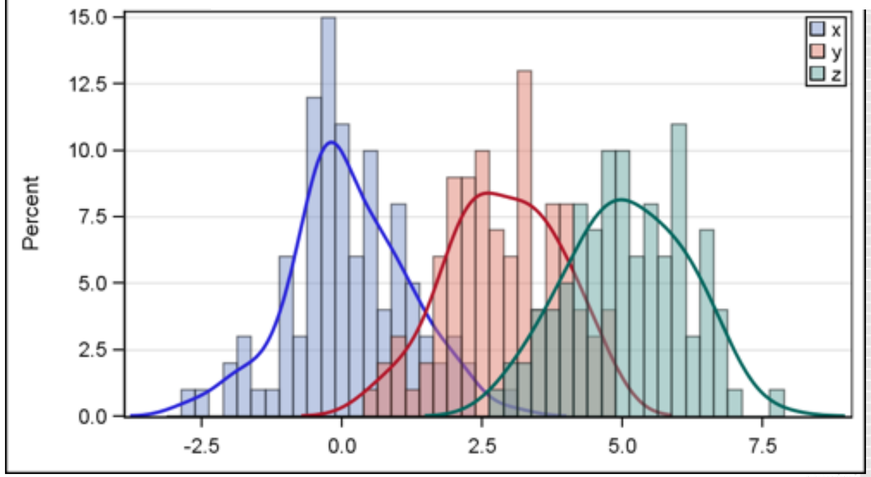
RUNI, Herzeliya

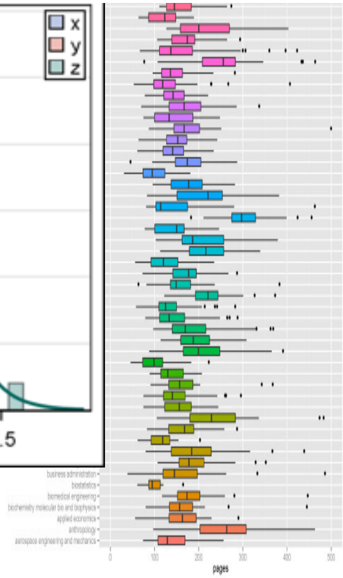




00100111010101001010101001001000010
1010100010101111101011010011001001
11101010100110100101100101101100010







Background

The paper ”A Mathematical Theory of Communication” was published by **Claude Shannon** in 1948.

This was the beginning of the information theory.

He defined the bit as an information measurement and the Entropy of information source as the minimal number of bits needed to code any message from the source

IDC HERZLIYA

The Idea

The amount of information in a result of an experiment (that has more than one possible outcome) increases when the
result is more surprising.

Example:

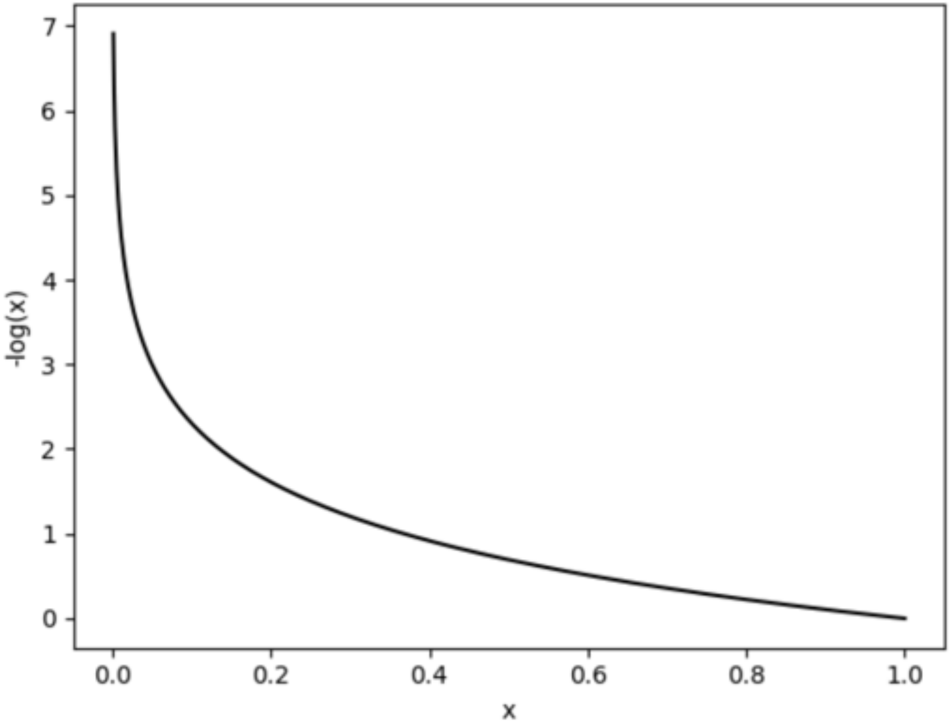
- We roll a fair die
- Where do we have more information? “the result is not 6” or “the result is 6”.

We want to measure the amount of information in a message/result

Information

$$I(p) = \log_2\left(\frac{1}{p}\right)$$

- Large p $\rightarrow$ small $I(p)$
- $I(p) \geq 0$
- $I(1) = 0$
- If the events are independent
  $I(p_1, p_2) = I(p_1) + I(p_2)$
  $* \log_2(p_1 p_2) = \log_2(p_1) + \log_2(p_2)$



Entropy – Definition

Let X be a discrete random variable with some PMF.

Let $\mathrm{H}(X)$ be the Entropy of X (=the amount of information in the experiment defined by X):

$$H(X) = \sum_{x \in X} P(x) I\big(P(x)\big) = \sum_{x \in X} P(x) \log_2 \left(\frac{1}{P(x)}\right) = E\left[\log_2 \left(\frac{1}{P(x)}\right)\right]$$

Entropy = uncertainty measurement

IDC HERZLIYA

Entropy – Comments

- Convention – $0\log_2 0 = 0$
- The entropy depends only on the probabilities and NOT on the possible values of the random variable
- $\mathrm{H}(X) \geq 0$

IDC HERZLIYA

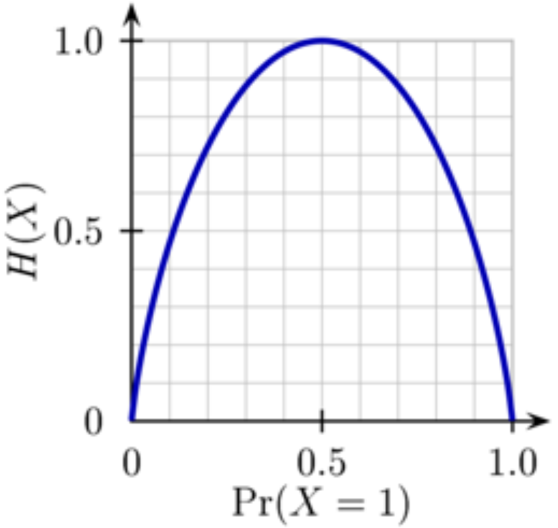
Entropy – Example 1

Let

$$X = \begin{cases} 0, & 1-p \\ 1, & p \end{cases}$$

- $\mathrm{H}(X) = p\log\left(\frac{1}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right) = -p\log p - (1-p)\log(1-p)$
- $\mathrm{H}(X) = \sum_x p\log\left(\frac{1}{p}\right) = -\sum_x p\log p$
- If $p = \frac{1}{2}$:

  $\mathrm{H}\left(\frac{1}{2}\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = \log 2 = 1$

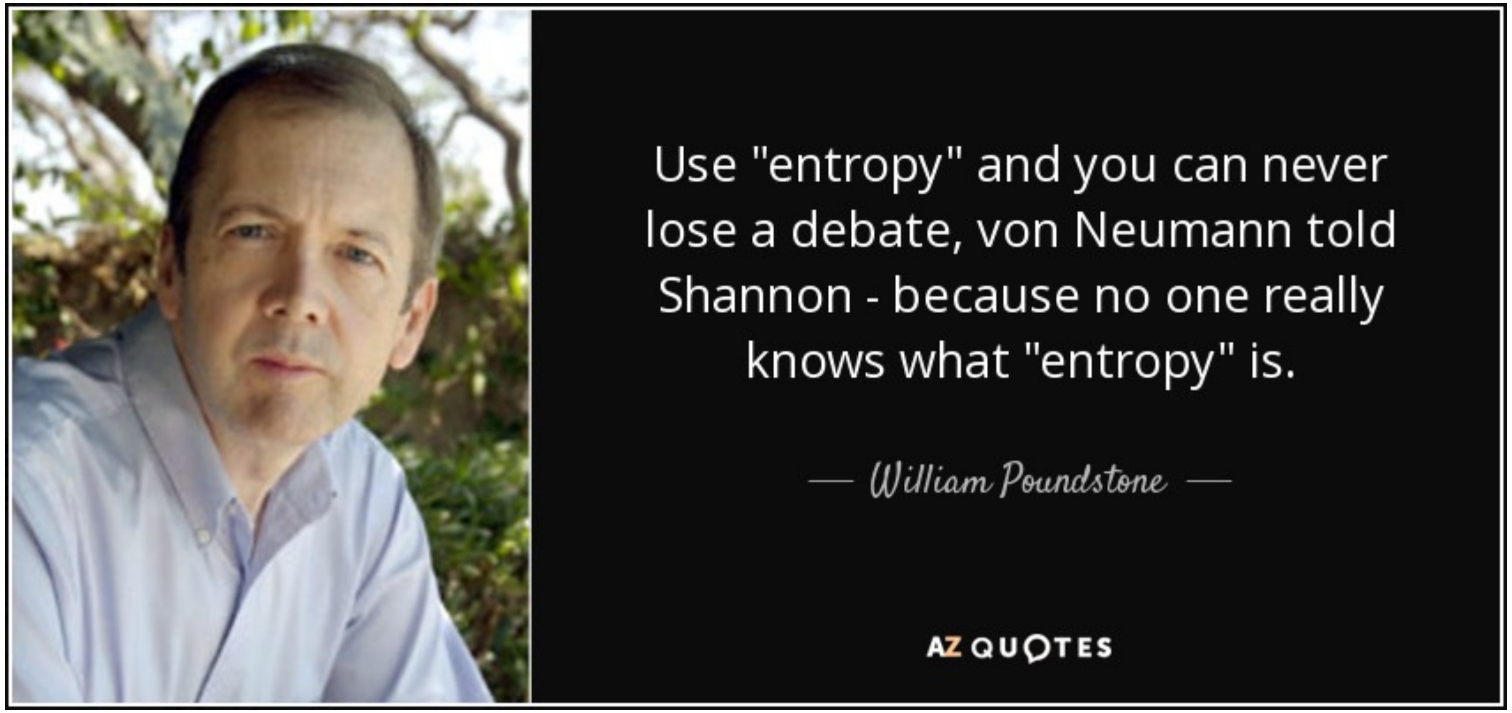


IDC HERZLIYA

Entropy – Example 2

Let

$$X = \begin{cases} a, & \frac{1}{2} \\ b, & \frac{1}{4} \\ c, & \frac{1}{8} \\ d, & \frac{1}{8} \end{cases}$$



$$\mathrm{H}(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8$$

$$= \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 = 1.75$$

Joint Entropy – Definition

Let X, Y be two discrete random variables.

Let $\mathrm{H}(X, Y)$ be the Joint Entropy:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(x,y)\big) = -E\big[\log\big(P(x,y)\big)\big]$$

IDC HERZLIYA

Conditional Entropy – Definition

Let X, Y be two discrete random variables.

Let $\mathrm{H}(Y|X)$ be the Conditional Entropy:

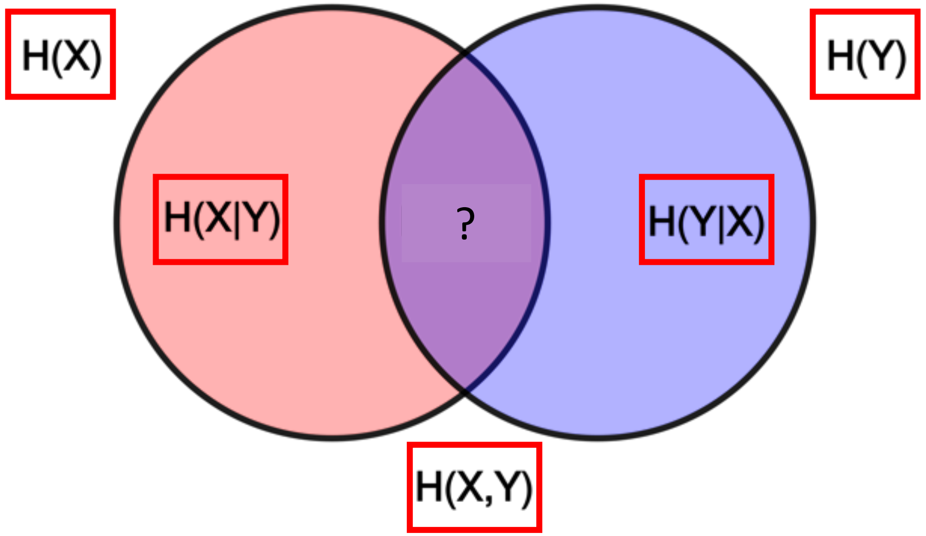
$$\mathrm{H}(Y|X) = \sum_{x \in X} P(x) H(Y|X = x) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log\big(P(y|x)\big)$$

$$= -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log\big(P(y|x)\big) = -E_{P(x,y)}\left[\log\big(P(y|x)\big)\right]$$

IDC HERZLIYA

Chain Rule

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



Chain Rule – Proof

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

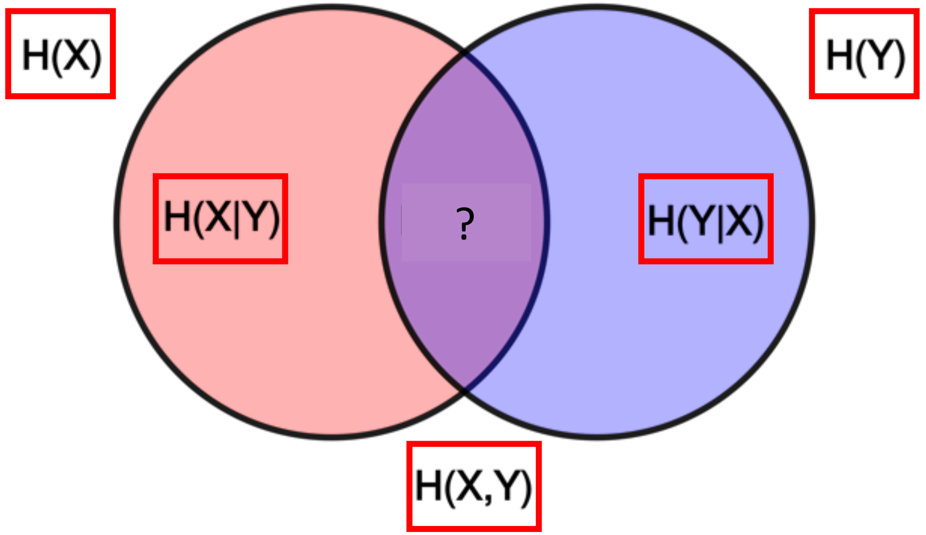
$$\begin{aligned}
\mathrm{H}(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(x,y)\big) \\
&= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(x)P(y|x)\big) \\
&= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(x)\big) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(y|x)\big) \\
&= -\sum_{x\in X} P(x)\log\big(P(x)\big) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\big(P(y|x)\big) \\
&= \mathrm{H}(X) + \mathrm{H}(Y|X)
\end{aligned}$$

IDC HERZLIYA

Chain Rule – Example

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

Y \ X	1	2	3	4	P(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
P(X)	1/2	1/4	1/8	1/8	

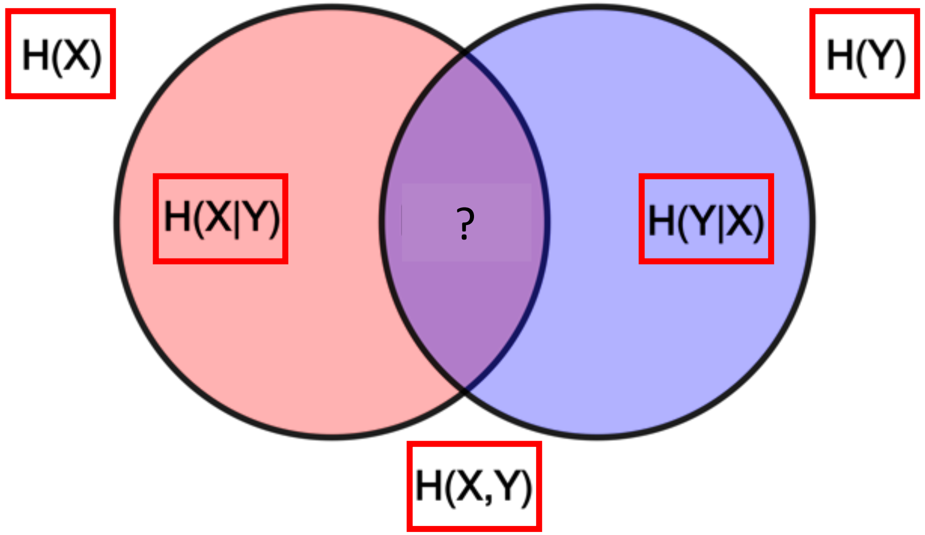


$$\mathrm{H}(X) = ?$$
$$\mathrm{H}(Y) = ?$$

Chain Rule – Example

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

Y \ X	1	2	3	4	P(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
P(X)	1/2	1/4	1/8	1/8	



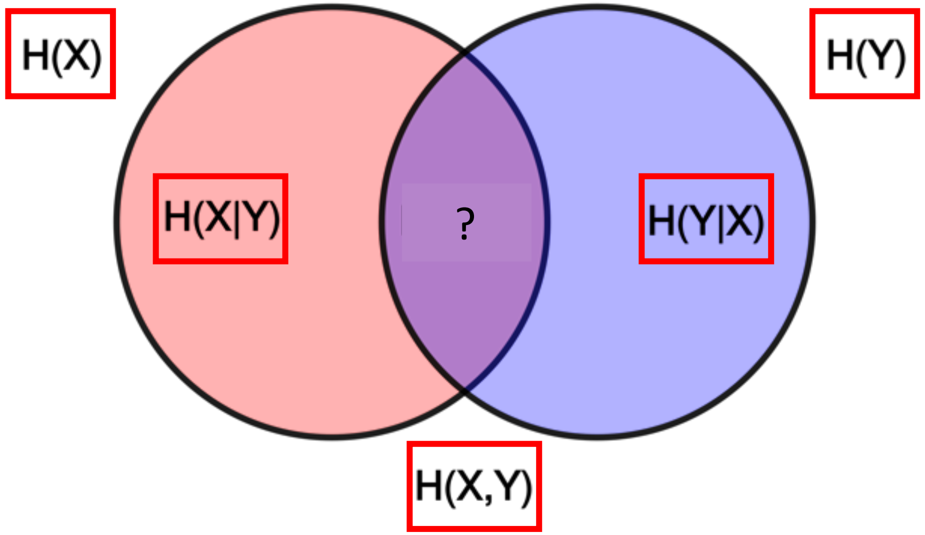
$$\mathrm{H}(X|Y) = ?$$
$$\mathrm{H}(Y|X) = ?$$

IDC HERZLIYA

Chain Rule – Example

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

Y \ X	1	2	3	4	P(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
P(X)	1/2	1/4	1/8	1/8	

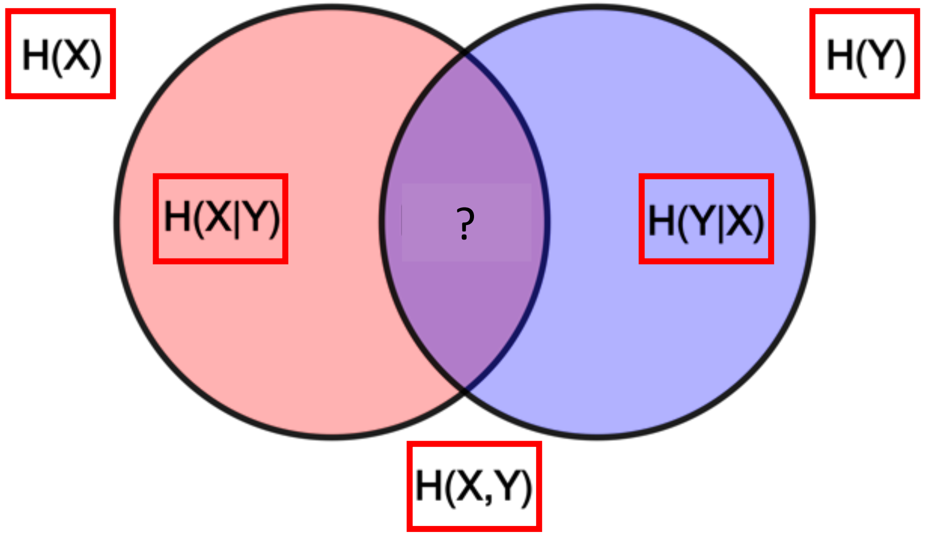


$$\mathrm{H}(X,Y) = ?$$

Chain Rule – Example

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

Y \ X	1	2	3	4	P(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
P(X)	1/2	1/4	1/8	1/8	



$$\mathrm{H}(X) - \mathrm{H}(X|Y) \stackrel{?}{=} \mathrm{H}(Y) - \mathrm{H}(Y|X)$$

We will define this term (the intersection) soon

IDC HERZLIYA

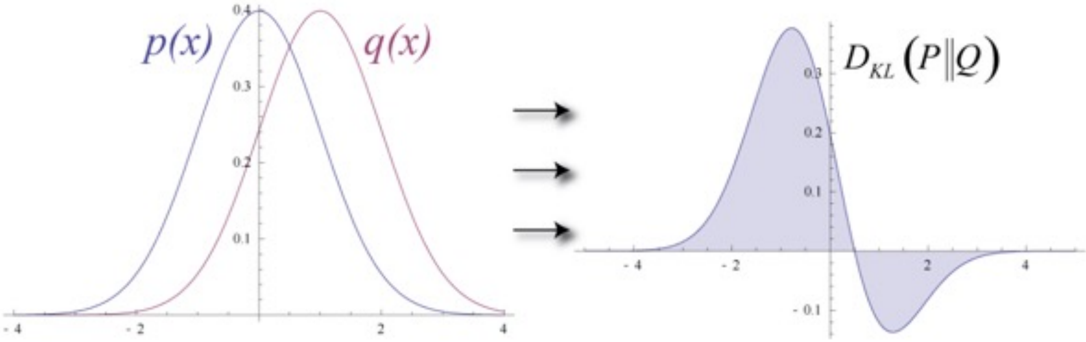
Relative Entropy = Kullback-Leibler divergence

$$D_{\mathrm{KL}}(P \parallel Q)$$

Measure the “distance” (difference) between the probability distribution P and the probability distribution Q.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in X} p(x) \log\left(\frac{p(x)}{q(x)}\right) = E_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$$

- $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$
- $\boxed{D_{\mathrm{KL}}(P \parallel Q) \geq 0}$
- $P = Q \Leftrightarrow D_{\mathrm{KL}}(P \parallel Q) = 0$
- $0 \log \frac{0}{0} = 0, 0 \log \frac{0}{q} = 0, 0 \log \frac{p}{0} = \infty$



Original Gaussian PDF’s

KL Area to be Integrated

IDC HERZLIYA

Jensen's Inequality

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}^2[X] \geq 0$$

Thus

$$\mathrm{E}[X^2] \geq \mathrm{E}^2[X]$$

If we define $g(x) = x^2$, we can write the above inequality as

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$$

The function $g(x) = x^2$ is an example of convex function.

Jensen's inequality states that, for any convex function g, we have

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$$

Johan Ludvig William Valdemar Jensen (1859-1925)

Jensen's Inequality

$g(x)$ is convex if and only if $-g(x)$ is concave.

We can state the definition for convex and concave functions in the following way:

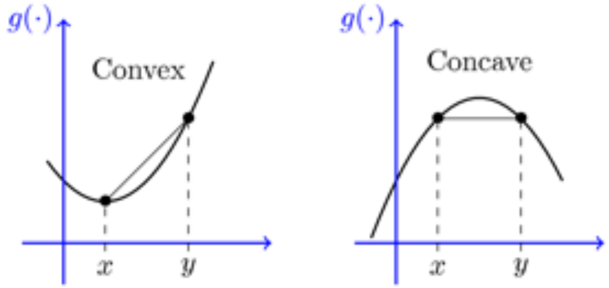
Consider a function $g\colon \mathrm{I} \to \mathbb{R}$, where I is an interval in $\mathbb{R}$.

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

$$g(\alpha x + (1-\alpha)y) \le \alpha g(x) + (1-\alpha)g(y)$$

We say that g is **concave** if

$$g(\alpha x + (1-\alpha)y) \ge \alpha g(x) + (1-\alpha)g(y)$$



Jensen's Inequality

More generally, for a convex function $g\colon \mathrm{I} \to \mathbb{R}$ and $x_1, x_2, \dots, x_n$ in I and nonnegative real numbers α_i such that $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$, we have

$$g(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \leq \alpha_1 g(x_1) + \alpha_2 g(x_1) + \cdots + \alpha_n g(x_n)$$

If $n = 2$, the above statement is the definition of convex functions.
We can extend it to higher values of n by induction.

IDC HERZLIYA

Jensen's Inequality

Now, consider a discrete random variable X with n possible values $x_1, x_2, \dots, x_n$.

In the previous equation,

$$g(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \leq \alpha_1 g(x_1) + \alpha_2 g(x_1) + \cdots + \alpha_n g(x_n)$$

we can choose $\alpha_i = P(X = x_i)$.

Then, the left-hand side becomes $g(\mathrm{E}[X])$ and the right-hand side becomes $\mathrm{E}[g(X)]$.

Jensen's Inequality:

If $g(x)$ is a convex function, and $\mathrm{E}[g(X)]$ and $g(\mathrm{E}[X])$ are finite, then

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$$

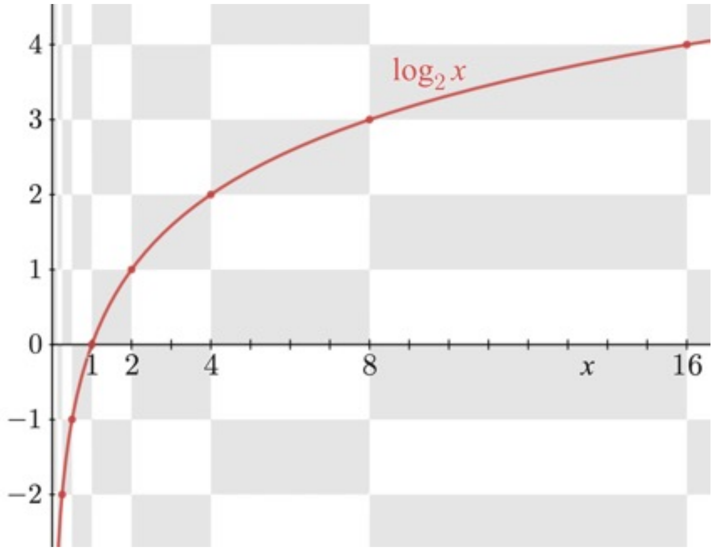
IDC HERZLIYA

Relative Entropy = Kullback-Leibler divergence

$$D_{\mathrm{KL}}(P \parallel Q) \geq 0$$

Proof:

$$\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in X} p(x) \log\left(\frac{p(x)}{q(x)}\right) \\
&= -\sum_{x \in X} p(x) \log\left(\frac{q(x)}{p(x)}\right) \\
&= -E_p\left[\log\left(\frac{q(x)}{p(x)}\right)\right] \\
&\geq -\log\left(E_p\left[\frac{q(x)}{p(x)}\right]\right) \text{ (by Jensen's Inequality for concave function log)} \\
&= -\log\left(\sum p(x)\frac{q(x)}{p(x)}\right) = -\log\left(\sum q(x)\right) = 0
\end{aligned}$$



IDC HERZLIYA

Entropy Max Value

Consider a discrete random variable X with k possible values $x_1, x_2, \dots, x_k$.

$$\mathrm{H}(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = E\left[\log \frac{1}{p(x)}\right]$$

(by Jensen's Inequality
for concave function $\log$)

$$\leq \log E\left[\frac{1}{p(x)}\right] = \log \sum_{x \in X} p(x) \frac{1}{p(x)} = \log k$$

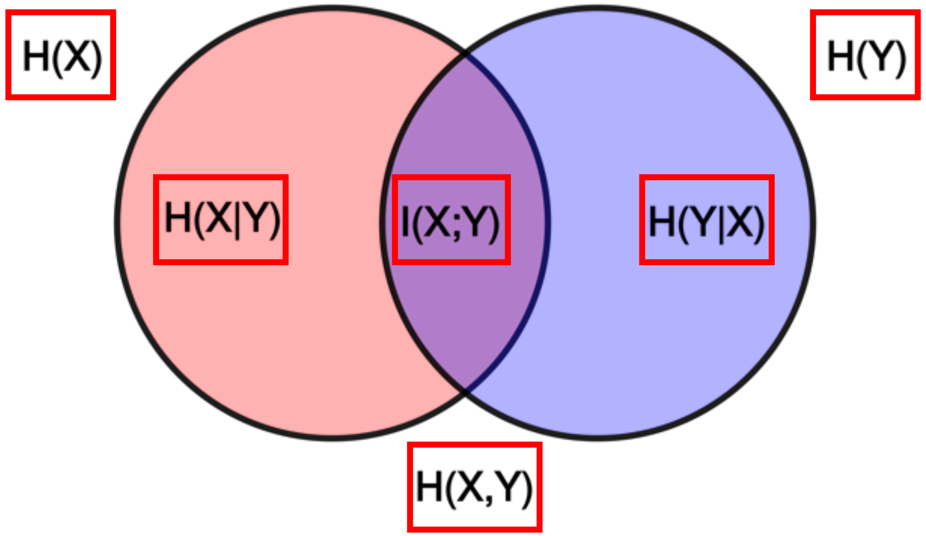
$$\mathrm{H}(X) \leq \log k$$

Mutual Information

Let X and Y be two random variables with probability distributions $P(X)$ and $P(Y)$ respectively, and a joint distribution $P(X,Y)$.

The mutual information $\mathrm{I}(X;Y)$ is the relative entropy between the joint distribution and the marginal distributions

$$\begin{aligned}\mathrm{I}(X;Y) &= D\big(P(x,y)\|P(x)P(y)\big) \\ &= \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right) \\ &= E_{p_{X,Y}}\left[\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\right]\end{aligned}$$



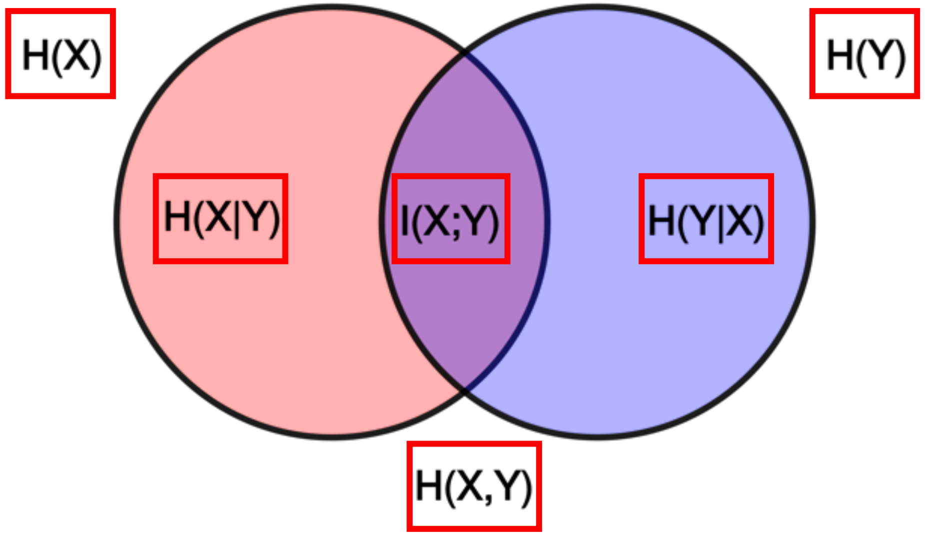
IDC HERZLIYA

Mutual Information

$$\mathrm{I}(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

$$= \sum_{x,y} P(x,y)\log\left(\frac{P(x|y)}{P(x)}\right)$$

$$= -\sum_{x,y} P(x,y)\log P(x) - -\sum_{x,y} P(x,y)\log P(x|y)$$

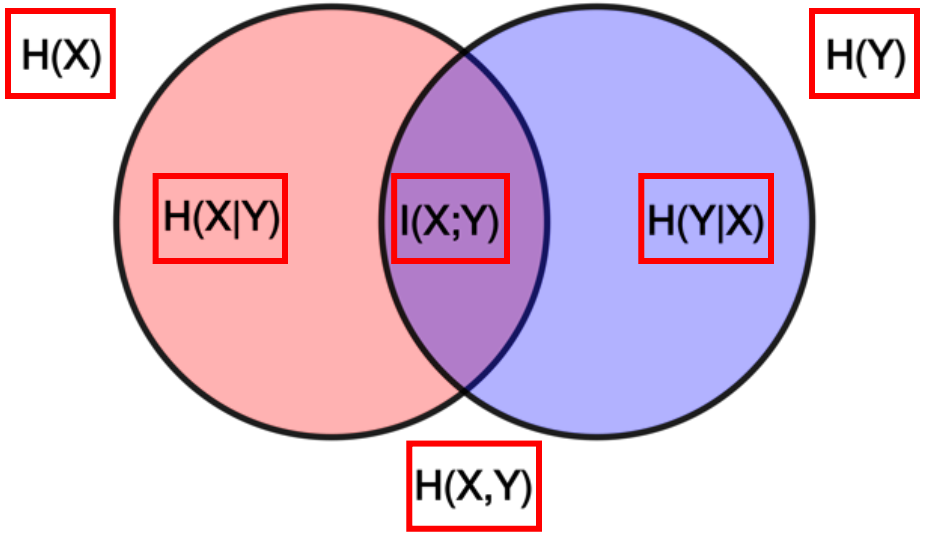


$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$

$$\mathrm{H}(X,Y) = \mathrm{H}(Y) + \mathrm{H}(X) - \mathrm{I}(X;Y)$$ (from the chain rule)

Mutual Information

- $\mathrm{I}(X;X) = H(X) - H(X|X) = H(X)$
  (Self information)
- $\mathrm{I}(X;Y) \geq 0$
  (Kullback-Leibler)
- $\mathrm{I}(X;Y) = 0$ iff $\log\left(\frac{P(x,y)}{P(x)P(y)}\right) = 0$
  iff X, Y are independent
- $\mathrm{H}(X) \geq \mathrm{H}(X|Y)$



Example

Y \ X	1	2	P(Y)
1	0	3/4	3/4
2	1/8	1/8	1/4
P(X)	1/8	7/8	

$$\mathrm{H}(X) = \mathrm{H}\left(\frac{1}{8},\frac{7}{8}\right) = 0.544$$

$$\mathrm{H}(Y) = \mathrm{H}\left(\frac{1}{4},\frac{3}{4}\right)$$

$$\mathrm{H}(X|Y=1) = 0$$

$$\mathrm{H}(X|Y=2) = 1$$

$$\mathrm{H}(X|Y) = P(Y=1)\mathrm{H}(X|Y=1) + P(Y=2)\mathrm{H}(X|Y=2) = \frac{3}{4}\cdot 0 + \frac{1}{4}\cdot 1 = 0.25$$

Entropy of a Function of a Random Variable

- Let X be a random variable and let $g(X)$ be a function on X

$$X \in \mathbb{R} \quad g: \mathbb{R} \to \mathbb{R}$$

$$X = \begin{cases} 1 & \frac{1}{3} \\ 0 & \frac{1}{3} \\ -1 & \frac{1}{3} \end{cases} \qquad g(x) = x^2 = \begin{cases} 1 & \frac{2}{3} \\ 0 & \frac{1}{3} \end{cases}$$

- The entropy of a variable can only decrease when the latter is passed through a function

IDC HERZLIYA

Entropy of a Function of a Random Variable

$$\mathrm{H}\big(X, G(X)\big) = \mathrm{H}(X) + \mathrm{H}(G(X)|X)$$
$$\mathrm{H}(G(X)|X) = 0 \ \rightarrow \mathrm{H}\big(X, G(X)\big) = \mathrm{H}(X)$$

$$\mathrm{H}\big(X, G(X)\big) = \mathrm{H}\big(G(X)\big) + \mathrm{H}(X|G(X)) \geq \mathrm{H}\big(G(X)\big)$$

$$\mathrm{H}(X) \geq \mathrm{H}\big(G(X)\big)$$

- מתי יהיה שוויון?

IDC HERZLIYA

Entropy of a Function of a Random Variable

- ChatGPT

b

When the entropy of a random variable is equal to the entropy of a function of the random variable

Inaccurate

Entropy is a measure of uncertainty or randomness in a random variable. The entropy of a random variable is equal to the entropy of a function of that random variable if the function does not reduce the amount of uncertainty or randomness in the variable. For example, if a random variable X has a known probability distribution, and we apply a function f to X to create a new random variable Y, the entropy of Y will be equal to the entropy of X if the function f does not change the probability distribution of X. In other words, if the function f is deterministic and does not introduce any additional randomness, the entropy of Y will be equal to the entropy of X.

IDC HERZLIYA

Entropy of a Function of a Random Variable

- ChatGPT

b

Write a python code that calculate the mutual information of two random variables

Copy code

```
def mutual_information(X, Y):
  # first we compute the joint probability distribution of X and Y
  joint_distribution = compute_joint_distribution(X, Y)

  # then we compute the marginal probability distributions of X and Y
  marginal_x = compute_marginal_distribution(X)
  marginal_y = compute_marginal_distribution(Y)

  # now we can compute the mutual information using the formula above
  mi = 0
  for x in marginal_x:
    for y in marginal_y:
      joint_prob = joint_distribution[(x, y)]
      marginal_prob = marginal_x[x] * marginal_y[y]
      if joint_prob > 0 and marginal_prob > 0:
        mi += joint_prob * log(joint_prob / marginal_prob)

  return mi
```

Summary

- Information
- Entropy
- Joint Entropy
- Conditional Entropy
- Relative Entropy
- Mutual Information