Introduction to Information Theory

Meryem Benammar

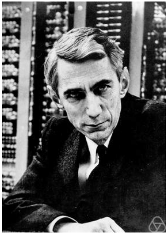
ISAE-Supaéro, DEOS

A bit of History

Measures of information

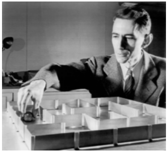
Shannon's main theorems

Claude Elwood Shannon

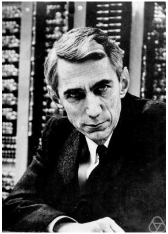


- Mathematician : "A mathematical theory of communication, Bell Labs, 1948"

$$\text{bit/sec}\ \left\{ R_{\max} = \overset{\text{Hz}}{B} \underbrace{\log_2\left(1+\frac{\overset{W}{P}}{\underset{W}{N}}\right)} \right. \qquad \text{bit/s/Hz}$$

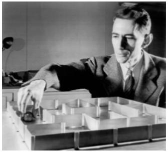


Claude Elwood Shannon



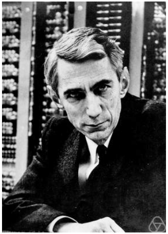
- Mathematician : "A mathematical theory of communication, Bell Labs, 1948"

$$R_{\max} = B \log_2 \left(1 + \frac{P}{N}\right)$$



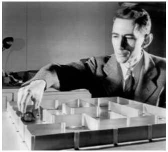
- Engineer : "Theseus : The maze mouse"
  `https://youtu.be/nS0luYZd4fs?t=59s`

Claude Elwood Shannon



- Mathematician : "A mathematical theory of communication, Bell Labs, 1948"

$$R_{\max} = B \log_2 \left(1 + \frac{P}{N}\right)$$



- Engineer : "Theseus : The maze mouse" `https://youtu.be/nS0luYZd4fs?t=59s`
- Inventor : "The ultimate machine" `https://www.youtube.com/watch?v=kt3csIz3hEk`,

Claude Elwood Shannon

- Juggler : "W.C. Fiels : The juggling machine"
  `https://www.youtube.com/watch?v=sBHGzRxfeJY`

Flying

$$(D + F)H = (D + V)N$$

Dwell hands vacancy balls

Claude Elwood Shannon

- Juggler : "W.C. Fiels : The juggling machine"
  `https://www.youtube.com/watch?v=sBHGzRxfeJY`

$$(D+F)H = (D+V)N$$

- Unicyclist

Claude Elwood Shannon

- Juggler : "W.C. Fiels : The juggling machine"
  `https://www.youtube.com/watch?v=sBHGzRxfeJY`

$$(D + F)H = (D + V)N$$

- Unicyclist
- Chess player : "Shannon number"

$$C = 10^{120}$$

Information Theory : Shannon's theory

The notion of digital information : bits (0, 1)

- A measure of the amount of information (bits, nats, ...)
- Based on the inherent uncertainty of a random process
- Different from semantic information : quantitative

Fundamental limits of manipulating digital information

- A communication channel in terms of bit rates (bits/sec)
- A lossless compression algorithm in terms of file sizes (Kbits, Kbytes)
- A lossy compression algorithm in terms of distortion (visual) and file size (Mbits)
- A cryptographic system in terms of key length (bits)

A bit of History

Measures of information

Shannon's main theorems

Entropy

- Let X be a random variable with pmf P_X
- Shannon's information content of a realization x is

$$c(x) \triangleq -\log_2 \left(P_X(x)\right)$$

Entropy

- Let X be a random variable with pmf P_X
- Shannon's information content of a realization x is

$$c(x) \triangleq -\log_2 \left(P_X(x)\right)$$

Shanon's entropy

$$H(X) \quad \triangleq \quad -\sum_{x \in \mathcal{X}} P_X(x) \log_2(P_X(x)) = \mathbb{E}\left(C(X)\right)$$

Entropy

- Let X be a random variable with pmf P_X
- Shannon's information content of a realization x is

$$c(x) \triangleq -\log_2 \left(P_X(x)\right)$$

Shanon's entropy

$$H(X) \quad \triangleq \quad -\sum_{x \in \mathcal{X}} P_X(x) \log_2(P_X(x)) = \mathbb{E}\left(C(X)\right)$$

- Measured in bits ($\log_2$) or in nats ($\log$)
- Positive for discrete support set $\mathcal{X}$
- Is maximized when X is uniform over $\mathcal{X}$ and its maximum value is $\log_2(|\mathcal{X}|)$
- Is minimized when X is deterministic and its minimum value is 0

Joint Entropy

$$H(x) = -\sum_{x} P_X(x) \log_2(P_X(x))$$

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The joint entropy of (X, Y) is defined by

$$H(X,Y) \triangleq -\sum_{(x,y)} P_{X,Y}(x,y) \log_2 \left(P_{X,Y}(x,y)\right)$$

Joint Entropy

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The joint entropy of (X, Y) is defined by

$$H(X, Y) \triangleq - \sum_{(x,y)} P_{X,Y}(x, y) \log_2 \left(P_{X,Y}(x, y)\right)$$

- Joint entropy is always positive
- Symmetric in X and Y
- Property $H(X, Y) \geq \max\{H(X), H(Y)\}$
- Is maximized when (X, Y) is uniform over $\mathcal{X} \times \mathcal{Y}$
- Is minimized when X and Y are deterministic

Conditional entropy

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The conditional entropy of X given Y is defined by :

$$H(X|Y) \quad \triangleq \quad -\sum_{(x,y)} P_{X,Y}(x, y) \log_2 \left(\frac{P_{X,Y}(x, y)}{P_Y(y)} \right)$$

Conditional entropy

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The conditional entropy of X given Y is defined by :

$$
\begin{aligned}
H(X|Y) &\triangleq -\sum_{(x,y)} P_{X,Y}(x,y)\log_2\left(\frac{P_{X,Y}(x,y)}{P_Y(y)}\right) \\
&= -\sum_{(x,y)} P_Y(y) P_{X|Y}(x|y)\log_2\left(P_{X|Y}(x|y)\right)
\end{aligned}
$$

$= P_{X|Y}(x|y)$

Conditional entropy

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The conditional entropy of X given Y is defined by :

$$
\begin{aligned}
H(X|Y) &\triangleq -\sum_{(x,y)} P_{X,Y}(x,y)\log_2\left(\frac{P_{X,Y}(x,y)}{P_Y(y)}\right) \\
&= -\sum_{(x,y)} P_Y(y)P_{X|Y}(x|y)\log_2\left(P_{X|Y}(x|y)\right) \\
&= \sum_y P_Y(y)H(X|Y=y)
\end{aligned}
$$

$\sum_x$

$Q^y_x(x)$

Conditional entropy

Let (X, Y) be a pair of random variables with pmf $P_{X,Y}$.

- The conditional entropy of X given Y is defined by :

$$
\begin{aligned}
H(X|Y) &\triangleq -\sum_{(x,y)} P_{X,Y}(x,y) \log_2 \left(\frac{P_{X,Y}(x,y)}{P_Y(y)} \right) \\
&= -\sum_{(x,y)} P_Y(y) P_{X|Y}(x|y) \log_2 \left(P_{X|Y}(x|y) \right) \\
&= \sum_{y} P_Y(y) H(X|Y=y)
\end{aligned}
$$

- Conditional entropy is positive (discrete)
- Asymmetric in X and Y
- Upper bound : $\underline{H(X|Y)} \leq \underline{H(X)}$
- Maximum when Y and X are independent
- Minimum when X is a function of Y

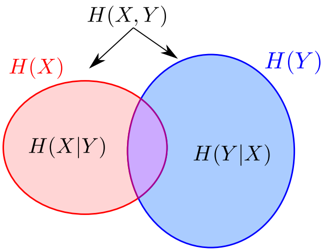
$H(X|Y) = H(X)$

$0 \leq H(X|Y) \leq H(X)$

Joint entropy and conditional entropy

Joint, individual and conditional entropies

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

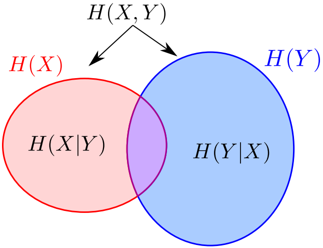


Wien diagram

Joint entropy and conditional entropy

Joint, individual and conditional entropies

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



Let $(X_1, \dots, X_n)$ be a set of random variables with pmf $P_{X_1,\dots,X_n}$.

$$H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \dots, X_{i-1})$$

$n = 2$

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

Continuous case

For a continuous variable with p.d.f $f_X(x)$, we define a differential entropy

$$H(X) \neq h(X) = -\int_{x\in\mathcal{X}} \underbrace{f_X(x)\log_2\left(f_X(x)\right)}\, dx$$

- Can be negative
- Might be undefined
- Not equivalent to the discrete entropy
- At a fixed variance $\mathbb{V}(X) = \sigma^2$, maximum for a Gaussian random variable $\mathcal{N}(0,\sigma^2)$

$$h(X) \leq \underbrace{\frac{1}{2}\log_2\left(2\pi e\sigma^2\right)}_{h_G(\sigma^2)}.$$

Continuous case

For a continuous variable with p.d.f $f_X(x)$, we define a differential entropy

$$h(X) = -\int_{x\in\mathcal{X}} f_X(x)\log_2\left(f_X(x)\right)\, dx$$

- Can be negative
- Might be undefined
- Not equivalent to the discrete entropy
- At a fixed variance $\mathbb{V}(X) = \sigma^2$, maximum for a Gaussian random variable $\mathcal{N}(0,\sigma^2)$

$$h(X) \leq \frac{1}{2}\log_2\left(2\pi e\sigma^2\right).$$

Conditional differential entropy is defined for a pair of random variables (X, Y) as

$$h(X|Y) = -\int_{x,y} f_{X,Y}(x,y)\log_2\left(\frac{f_{X,Y}(x,y)}{f_Y(y)}\right)\, dx\, dy$$

$h(X|Y) \leq h(X)$

Mutual information

Let (X, Y) be a pair of random variables. P_{xy}

- The mutual information between X and Y

$$I(X;Y) \triangleq \sum_{(x,y)} \underline{P_{X,Y}(x,y)} \log_2 \left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \right)$$

$P_{xy} \sim X \times Y$

$P_x P_y \sim X \times Y$

- Mutual information is related to KL-divergence

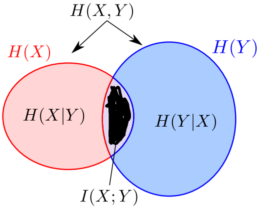
$$I(X;Y) = D_{KL}(\underline{P_{X,Y}} || \underline{P_X P_Y}) > 0$$

$P_{xy} = P_x P_y$

$\hookrightarrow I(X;Y) = 0$

- Symmetric in X and Y
- Satisfies the property

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - \underline{H(X,Y)} \end{aligned}$$



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

Auto-evaluation

At this point of the course, you should be able to

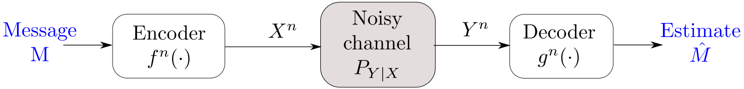
- Compute the expectation, variance, entropy based on the p.d.f / p.m.f of a random variable
- Compute conditional probability, marginal probabilities, conditional entropy and mutual information based on the joint p.d.f/p.m.f of a pair of random variables
- Relate entropy, mutual information, conditional entropy, and Kullback-Leibler divergence
- Compute mutual information for basic channels (BSC, BEC, Gaussian)

A bit of History

Measures of information

Shannon's main theorems

Channel coding theorem



- A message M of k bits $M \in [1 : 2^k]$
- Transmitted over n channel uses
- A rate defined by $R = \dfrac{k}{n}$
- An input alphabet $\mathcal{X}$ and sequence x^n
- An output alphabet $\mathcal{Y}$ and sequence y^n
- A memoryless channel $P_{Y|X}$
- An encoder $f^n \ : \ M \rightarrow X^n$
- A decoder $g^n \ : \ Y^n \rightarrow \hat{M}$

Is there a family of encoders-decoders (f^n, g^n) such that

$$\lim_{n\rightarrow\infty} \mathbb{P}(M \neq \hat{M}) = 0 \qquad ?$$

Channel coding theorem (Cont.)

Channel capacity

Such a family of pairs (f^n, g^n) exists if and only if (i.i.f)

$f(P_{XY})$

$$\frac{k}{n} = R \leq \max_{P_X} I(X;Y) = C(Y|X) \longrightarrow f(P_{Y|X})$$

- Characterized by Shannon in 1948
- $C(Y|X)$ is called the capacity of the channel $P_{Y|X}$
- Strong converse :

$$R > \max_{P_X} I(X;Y) \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}(M \neq \hat{M}) = 1$$

BSC, BEC, Gaussian

- Non-convex optimization in P_X
- Achieved by $P_X^\star$ called capacity achieving p.m.f (random coding argument)
- Valid only when $n \to \infty$
- Design of error correction codes (Turbo code, LDPC, Polar codes, BCH, Reed Solomon, Multi-Level Code, ...)

Lossless source coding

k bits

- An i.i.d source X^n of n symbols
- Conpressed in k bits ($m = 2$)
- A rate defined by $R = \dfrac{k}{n}$
- A compressor $f^n \ : \ X^n \to M$
- A decompressor $g^n \ : \ M \to \hat{X}^n$

Is there a family of compressors-decompressors (f^n, g^n) such that

$$\lim_{n\to\infty} \mathbb{P}(X^n \neq \hat{X}^n) = 0 \qquad ?$$

Lossless source coding

Lossless compression

Such a family of pairs (f^n, g^n) exists if and only if (i.i.f)

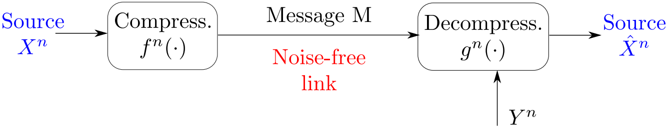
$$\frac{k}{n} = R \geq H(X) = f(P_X)$$

- Characterized by Shannon in 1948
- $H(X)$ is the minimum number of bits to compress a source X
- Strong converse :

$$R < H(X) \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}(X^n \neq \hat{X}^n) = 1$$

- Random binning argument
- Valid only when $n \to \infty$
- Design of lossless compressors : Audio (MPEG-4 ALS, DST) Image (PNG, JPEG-LS) Text (LZ-78, GZip)

Lossless source coding with side information



- An i.i.d source X^n of n symbols
- Conpressed in k bits
- A rate defined by $R = \dfrac{k}{n}$
- Receiver has a correlated sequence Y^n i.i.d $P_{Y|X}$
- A compressor $f^n \ : \ X^n \to M$
- A decompressor $g^n \ : \ (M, Y^n) \to \hat{X}^n$

? Is there a family of compressors-decompressors (f^n, g^n) such that

$$\lim_{n \to \infty} \mathbb{P}(X^n \neq \hat{X}^n) = 0 \qquad ?$$

Lossless source coding with side information

Slepian-Wolf Theorem

Such a family of pairs (f^n, g^n) exists if and only if (i.i.f)

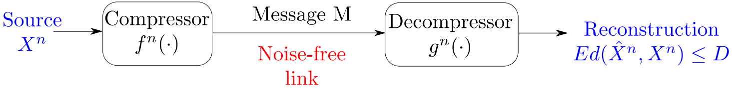
$$\frac{k}{n} = R \geq H(X|Y) \leq H(X)$$

- $H(X|Y)$ is the minimum number of bits to compress a source X, knowing apriori Y
- Entropy rate reduced by $H(X) - H(X|Y) = I(X;Y)$
- No need to know y^n at the transmission
- Random binning argument
- Strong converse :

$$R < H(X|Y) \quad \Rightarrow \quad \lim_{n \to \infty} \mathbb{P}(X^n \neq \hat{X}^n) = 1$$

- Valid only when $n \to \infty$
- DPCM based lossless video compression (MPEG-LS, H-LS series)

Lossy source coding



- Same source model as in lossless compression
- A distortion measure $d(\cdot,\cdot)$ defined by

$$\begin{array}{rcl} \mathcal{X} \times \hat{\mathcal{X}} & \to & \mathbb{R}^+ \\ (x, \hat{x}) & \to & d\left(x, \hat{x}\right) \end{array}$$

- Examples : Quadratic $d(x, \hat{x}) = |x - \hat{x}|^2$, Hamming $d(x, \hat{x}) = x \oplus \hat{x}$
- A compressor $f^n \;:\; X^n \to M$
- A decompressor $g^n \;:\; M \to \hat{X}^n$

Given a $D > 0$, is there a family of compressors-decompressors (f^n, g^n) such that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_{X^n, \hat{X}^n}} d(X_i, \hat{X}_i) \leq D \qquad ?$$

Lossy source coding

Lossy compression

Such a family of pairs (f^n, g^n) exists if and only if (i.i.f)

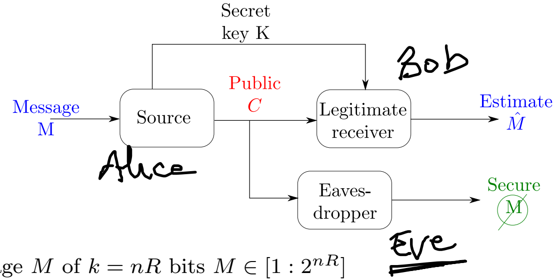
$$\frac{k}{n} = R(D) \geq \min_{\mathcal{P}} I(X; \hat{X})$$

where

$$\mathcal{P} = \{P_{\hat{X}|X}, \quad \mathbb{E}_{P_{X,\hat{X}}} d(X, \hat{X}) \leq D\}$$

- Rate distortion function $R(D)$
- Optimal compression p.d.f $P^{\star}_{\hat{X}|X}$ (random binning argument)
- $I(X; \hat{X}^*)$ minimum number of bits to compress X to a distortion level D
- Valid only when $n \to \infty$
- Property of $R(D)$: convex non-increasing in D
- Design of lossy compressors : Image (JPEG) Video (MPEG) Audio (Opus), Music (MP3)

Shannon's cipher system 1949



- A message M of $k = nR$ bits $M \in [1 : 2^{nR}]$
- A key K of $k_s = nR_s$ bits $K \in [1 : 2^{nR_s}]$
- An encoder $f^n \ : \ M \times K \to C$
- A decoder $g^n \ : \ C \times K \to \hat{M}$
- The security constraint writes as $I(M; C) \leq \epsilon_n$
- The link is noise free and accessed by the eavesdropper

Is there a family of encoders-decoders (f^n, g^n) such that

$$\lim_{n\to\infty} I(M; C) = 0 \text{ and } \lim_{n\to\infty} \mathbb{P}(\hat{M} \neq M) = 0?$$

Shannon's cipher system 1949 (Cont.)

One-time pad

Such a family of (f^n, g^n) exists if the message and key rates verify

$$R \leq R_s$$

- Encoding : $C = K \oplus (M, 0, \cdots, 0)$ (padded message M)
- Decoding : $\hat{M} = K \oplus C$
- The key can be used at most once : **one-time pad**
- Without a key, $R_s = 0$, no secure communication is possible
- Pessimistic result
- Base of cryptographic systems

Assymtery of information between users is crucial for secrecy

Auto-evaluation

At this point of the course, you should be able to

- List the main coding theorems of Information Theory
- Describe the optimal bounds given by the different theorems
- Infer the consequences of these theorems on practical schemes
- List practical algorithms for each theorem