UNIT 3: Information Theory

Gabriel Caffarena Fernández
3rd Year Biomedical Engineering Degree
EPS – Univ. San Pablo – CEU
(Based on slides by Carlos Oscar Sánchez Sorzano)

Entropy

Example: Temperature in a room

Source → $X \in \{-10, -9, ..., 28, 29, 30\}$

Scenario 1 (no information): 23, 23, 23, 23, 23, 23, …

Scenario 2 (little information): 23, 23, 23, 24, 23, 23, …

Scenario 3 (a lot of information): -10, 23, 30, -5, 0, 15, …

Example: Who wins a football match?

Fuente → $X \in \{A, B\}$

Scenario 1 (no information): A,A,A,A,A,A,A,A,A,A,…

Scenario 2 (little information): A,A,A,B,B,A,A,A,A,B,A,A,…

Scenario 3 (Lots of information): A,B,B,A,A,A,B,B,B,A,A,B,…

Entropy

- Given an information source $\boldsymbol{X} \in \{X_0, X_1, \ldots, X_{n-1}\}$ (e.g. experiment, random variables,etc.) the **Shannon information** of a particular outcome is defined as

$$h(X_i) = \log\left(\frac{1}{P(X=X_i)}\right) = -\log(P(X = X_i)$$

- The Shannon information is the amount of surprise that the outcome produces
- If the base of the logarithm is 2, then the **information is measured in bits**

Entropy

- **Example:**

$$X \in \{0,1,2,3\};$$

$$P(X=0)=0.25;\ P(X=1)=P(X=2)=0.125$$

$$P(X=3)=0.5$$

$$\boldsymbol{h(0)}=\log\left(\frac{1}{0.25}\right)=\log(4)=\boldsymbol{2\ bits}$$

$$\boldsymbol{h(1)}=\boldsymbol{h(2)}=\log\left(\frac{1}{0.125}\right)=\log(8)=\boldsymbol{3\ bits}$$

$$\boldsymbol{h(3)}=\log\left(\frac{1}{0.5}\right)=\log(2)=\boldsymbol{1\ bits}$$

$$\boldsymbol{h(3)}<\boldsymbol{h(0)}<\boldsymbol{h(1)}=\boldsymbol{h(2)}$$

Less surprise — More surprise

Entropy

- **Entropy** is the average of Shannon information of an information source X

$$H(X) = E[h(X)] = \sum_i h(X_i)P(h(Xi)) = \sum_i h(X_i)P(Xi)$$

$$= \sum_i P(Xi)\log\left(\frac{1}{P(Xi)}\right) = -\sum_i P(Xi)\log\left(P(X_i)\right)$$

Entropy

$$H(X)=\sum_i P(Xi)\log\left(\frac{1}{P(Xi)}\right)=-\sum_i P(Xi)\log\left(P(X_i)\right)$$

Example: Who wins a football match?

Scenario 1 (no information): A,A,A,A,A,A,A,A,A,A,…

$$H(X)=-p(A)\log p(A)-p(B)\log p(B)$$
$$=-1\log 1-0\log 0=0-0=0$$

The most probable event is the one contributing less

Scenario 2 (little information): A,A,A,B,B,A,A,A,A,B,A,A,…

$$H(X)=-\tfrac{3}{4}\log\tfrac{3}{4}-\tfrac{1}{4}\log\tfrac{1}{4}=-(-0.2158)-(-0.3466)=0.5624$$

Scenario 3 (a lot of information): A,B,B,A,A,A,B,B,B,A,A,B,…

$$H(X)=-\tfrac{1}{2}\log\tfrac{1}{2}-\tfrac{1}{2}\log\tfrac{1}{2}=-(-0.3466)-(-0.3466)=0.6932$$

$$H(X)=-\tfrac{1}{2}\log_2\tfrac{1}{2}-\tfrac{1}{2}\log_2\tfrac{1}{2}=-(-0.5)-(-0.5)=1(bits)$$

Entropy

Example:

Source → $X \in \{a,b,c,d\}$

$$p(a) = \tfrac{1}{2}$$
$$p(b) = \tfrac{1}{4}$$
$$p(c) = \tfrac{1}{8}$$
$$p(d) = \tfrac{1}{8}$$

$$H(X) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{8}\log\tfrac{1}{8} - \tfrac{1}{8}\log\tfrac{1}{8} = \tfrac{7}{4}\,bits$$

Entropy

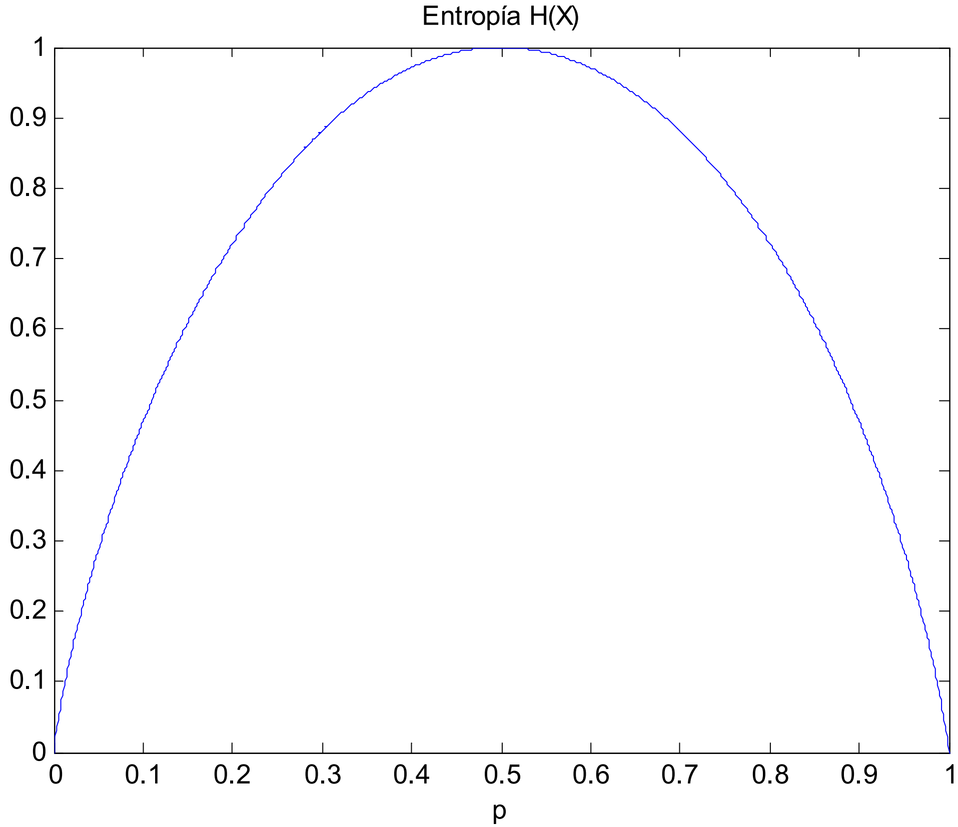
Example:

Source → $X \in \{0,1\}$

$p(0) = p$

$p(1) = 1 - p$

$$H(X) = -p\log p - (1-p)\log(1-p)$$



Entropy

Some comments on operations

$$\left.\begin{aligned} 0\log\tfrac{0}{q} &= 0 \\ p\log\tfrac{p}{0} &= \infty \end{aligned}\right\} \text{Consensus}$$

$$\log_a x = \frac{\log_b x}{\log_b a} \Rightarrow \log_2 x = \frac{\log_{10} x}{\log_{10} 2} \approx 3.32\log_{10} x$$

$$\log_a x = \log_a b \cdot \log_b x$$

Entropy

Properties

$$H(X) \geq 0$$

Proof:

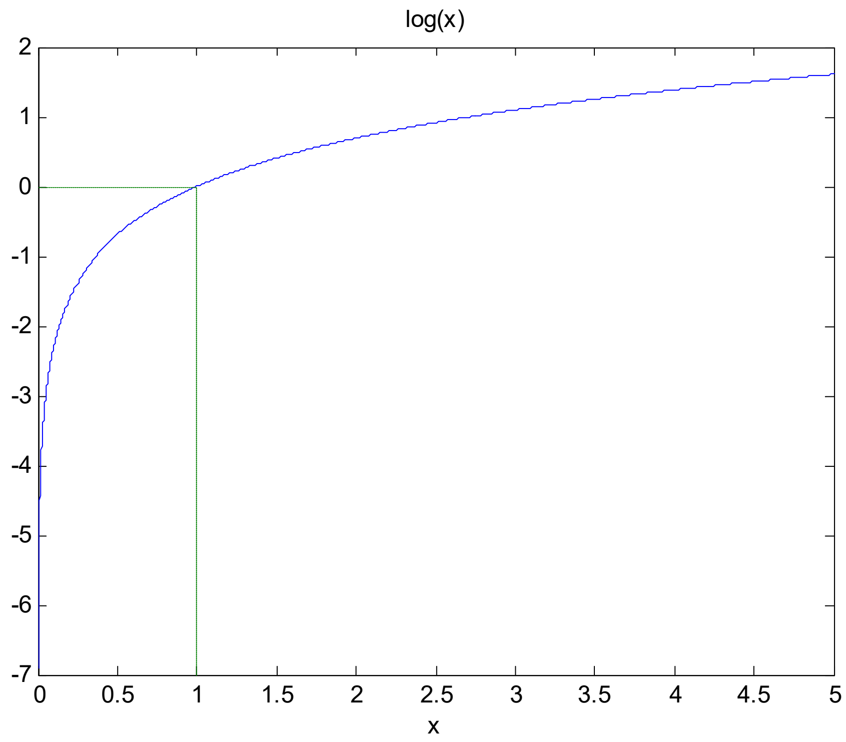
$$0 \leq p(x) \leq 1 \Rightarrow \frac{1}{p(x)} \geq 1 \Rightarrow \log \frac{1}{p(x)} \geq 0$$

$$\downarrow$$

$$H(X) = \sum_{x \in \Xi} p(x) \log \frac{1}{p(x)} \geq 0$$

$$\uparrow$$

$$p(x) \geq 0$$



Entropy

Properties

$$H_b(X) = \left(\log_b a\right) H_a(X)$$

Proof:

$$H_b(X) = E\left\{\log_b \frac{1}{p(x)}\right\} = E\left\{\frac{1}{\log_a b} \log_a \frac{1}{p(x)}\right\}$$

$$= E\left\{\log_b a \log_a \frac{1}{p(x)}\right\} = \left(\log_b a\right) H_a(X)$$

Joint entropy

- Given two information sources **X** and **Y** the **joint Shannon information** of a particular joint outcome is defined as

$$h(X_i, Y_j) = \log\left(\frac{1}{P(X=X_i, X=X_j)}\right) = -\log(P(X = X_i, X = X_j))$$

- The joint entropy is the average of the joint Shannon information

$$H(X,Y) = E[h(X,Y)] = \sum_i \sum_j h(X_i, Y_j) P(X_i, Y_j)$$

$$= \sum_{i,j} P(X_i, Y_j) \log\left(\frac{1}{P(X_i, Y_j)}\right) = -\sum_{i,j} P(X_i, Y_j) \log\left(P(X_i, Y_j)\right)$$

Joint Entropy

Joint Entropy $H(X,Y)=\sum_{i,j}P(X_i,Y_j)\log\left(\frac{1}{P(X_i,Y_j)}\right)=-\sum_{i,j}P(X_i,Y_j)\log\left(P(X_i,Y_j)\right)$

Example: Peter is bilingual (Spanish/English) and he reads "The Times" with probability 0.5 and "El País" with probability 0.5.

$$H(newspaper)=1\,bit$$

$$p(newspaper, language)$$

newspaper \ language	English	Spanish
The Times	0.5	0
El País	0	0.5

$$H(newspaper, language)=$$
$$=-\tfrac{1}{2}\log\tfrac{1}{2}-\tfrac{1}{2}\log\tfrac{1}{2}=1\,bit$$

Joint Entropy

Joint Entropy $H(X,Y) = \sum_{i,j} P(X_i,Y_j) \log\left(\frac{1}{P(X_i,Y_j)}\right) = -\sum_{i,j} P(X_i,Y_j) \log\left(P(X_i,Y_j)\right)$

Example: Peter watches the CNN and the BBC with the following probabilities

$p(TV, language)$

TV \ language	English	Spanish
BBC	0.5	0
CNN	0.25	0.25

$$H(TV) = 1\, bit$$

$$H(TV, language) =$$

$$= -\tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{2}\log\tfrac{1}{2}$$

$$= 1.5\, bits$$

Conditional Entropy

- Given two information sources **X** and **Y**
- The average information of **Y** given that $\mathbf{X = X_i}$ is

$$H(Y|X = X_i) = E[h(Y|X = X_i)] = \sum_j p\big(h(Y_j|X_i)\big)h(Y_j|X_i)$$

$$= \sum_j p(Y_j|X_i) \log\left(\frac{1}{p(Y_j|X_i)}\right)$$

So this is a **conditional entropy** for a <u>given outcome of X.</u>

Conditional Entropy

- **Conditional entropy** for a <u>given outcome of X.</u>

$$H(Y|X = X_i) = \sum_j p(Y_j|X_i) \log\left(\frac{1}{p(Y_j|X_i)}\right)$$

- The **conditional entropy** is the average of the conditional entropy for a given outcome of X, that has already averaged the Shannon information over Y, so it is going to be averaged **over X**.

$$H(Y|X) = E_{P(X_i)}[H(Y|X_i)] = \sum_i \boldsymbol{p(X_i)H(Y|X = X_i)}$$

$$= \sum_i P(X_i) \sum_j P(\boldsymbol{Y_j|X_i}) \log\left(\frac{1}{P(Y_i|X_j)}\right)$$

$$= \sum_i \sum_j P(X_i)P(\boldsymbol{Y_j|X_i}) \log\left(\frac{1}{P(Y_i|X_j)}\right)$$

$$= \sum_i \sum_j P(X_i, Y_j) \log\left(\frac{1}{P(Y_i|X_j)}\right) = \sum_i \sum_j \boldsymbol{P(X_i, Y_j) \log\left(\frac{P(X_i)}{P(X_i, Y_j)}\right)}$$

Conditional Entropy

Conditional Entropy $H(Y|X) = \sum_i p(X_i) H(Y|X = X_i)$

$$= \sum_i \sum_j P(X_i) P(Y_j|X_i) \log\left(\frac{P(X_i)}{P(X_i, Y_j)}\right) = \sum_i \sum_j P(X_i, Y_j) \log\left(\frac{P(X_i)}{P(X_i, Y_j)}\right)$$

Properties

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

$$\downarrow$$

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

$$H(X|Y) \neq H(Y|X)$$

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

If X and Y are independent, then $H(Y|X) = H(Y)$

Conditional Entropy

Example: Peter watches CNN and BBC

$p(TV, language)$

TV \ language	English	Spanish
BBC	0.5	0
CNN	0.25	0.25

$p(TV \mid language)$

TV \ language	English	Spanish
BBC	0.666	0
CNN	0.333	1

$p(language \mid TV)$

TV \ language	English	Spanish
BBC	1	0
CNN	0.5	0.5

Conditional Entropy

$p(TV, language)$

TV \ language	English	Spanish
BBC	0.5	0
CNN	0.25	0.25

→

$p(language \mid TV)$

TV \ idioma	Inglés	Español
BBC	1	0
CNN	0.5	0.5

TV \ language	English	Spanish
BBC	1	0

$$p(language \mid TV = BBC)$$

$$H(language \mid TV = BBC) = 0\,bits$$

TV \ language	English	Spanish
CNN	0.5	0.5

$$p(language \mid TV = CNN)$$

$$H(language \mid TV = CNN) = 1\,bits$$

$$H(language \mid TV) = 0.5\,bits$$

Differential or Relative Entropy (Kullback-Leibler distance)

Relative Entropy $D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = E_{p(x)}\left\{\log \frac{p(x)}{q(x)}\right\}$

<u>Properties</u>

$$D(p \| q) \geq 0$$
$$D(p \| p) = 0$$
$$D(p \| q) \neq D(q \| p)$$

Conditional relative entropy

$$D(p(y|x)||q(y|x)) = \sum_{x} p(x) \underbrace{\sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)}}_{D(p(y|X=x)\,\|\,q(y|X=x))}$$

$$= \sum_{x} \sum_{y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \qquad = E_{p(x,y)}\left\{\log \frac{p(y|x)}{q(y|x)}\right\}$$

Relative Entropy

Example: Let's asume that the actual probabilities of an information source are

Source $\longrightarrow$ $X \in \{0,1\}$

$p(0) = p$

$p(1) = 1 - p$

$$H_p(X) = -p \log p - (1-p)\log(1-p)$$

However, due to estimation errors, what we really have is

Source $\longrightarrow$ $X \in \{0,1\}$

$p(0) = q$

$p(1) = 1 - q$

$$H_q(X) = -q \log q - (1-q)\log(1-q)$$

$$\left.\begin{aligned} D(p \,\|\, q) &= p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \\ D(q \,\|\, p) &= q \log \frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \end{aligned}\right\} \xrightarrow[p=q]{} D(p \,\|\, q) = D(q \,\|\, p) = 0$$

Relative Entropy

Example: Given the following actual distribution of a set of symbols

Source → $X \in \{0,1\}$

$p(0) = \frac{1}{2}$

$p(1) = \frac{1}{2}$

$$H_p(X) = 1 bit$$

The estimated probabilities are

Source → $X \in \{0,1\}$

$p(0) = \frac{1}{3}$

$p(1) = \frac{2}{3}$

$$H_q(X) = 0.9183 bits$$

$$D(p \| q) = 0.0850 bits$$

$$D(q \| p) = 0.0817 bits$$

$$\longrightarrow H_p(X) = H_q(X) + D(q \| p)$$

Mutual Information

Mutual Information

$$I(X;Y) = \sum_{x\in\Xi, y\in\Psi} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

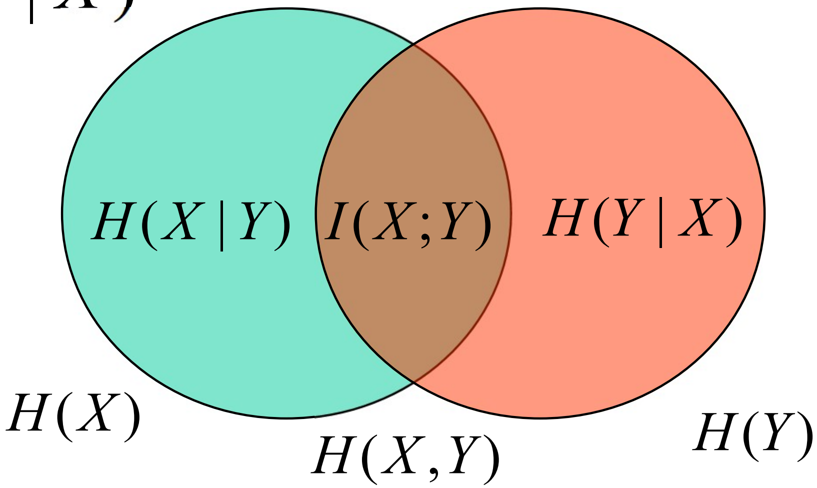
$$= E_{p(x,y)}\left\{\log\frac{p(x,y)}{p(x)p(y)}\right\} = D\big(p(x,y)\,\|\,p(x)p(y)\big)$$

Properties

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

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

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



Jensen’s inequality

Jensen’s inequality

Given the convex function *f(x)* and the r.v. X, then

$$E\{f(X)\} \geq f\left(E\{X\}\right)$$

Thanks to this inequality it can proved the following

$$D(p\|q) \geq 0$$

$$I(X;Y) \geq 0$$

$$H(X) \leq \log(\#X)$$

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

$$H(X_1,\ldots,X_N) \leq \sum_{i=1}^{N} H(X_i)$$

$$D(p\|q) = 0 \leftrightarrow p = q$$

$$I(X;Y|Z) \geq 0$$

$$H(X) = \log(\#X) \leftrightarrow p(X) = k$$

$$H(X|Y) = H(X) \leftrightarrow X, Y \text{ independent}$$

$$H(X_1,\ldots,X_N) = \sum_{i=1}^{N} H(X_i) \leftrightarrow X_i \text{ independent}$$

#X≡number of symbols (values) of X

Jensen's inequality

Highlights

- $H(X) \leq \log(\# X)$

  EXAMPLE: If X can have 4 values, its entropy cannot be greater than $\log_2(4)=2$ bits

- $H(X) = \log(\# X) \leftrightarrow p(X) = k$

An information source provides maximum information when its values (symbols) have equal probabilities

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

Any extra knowledge will never increase the information given by a source

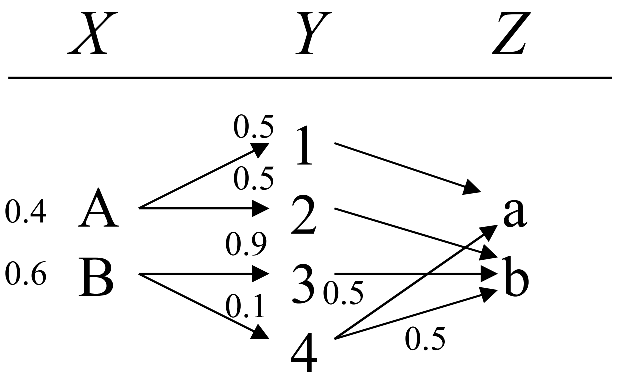
- $H(X|Y) = H(X) \leftrightarrow X, Y \textit{ independent}$

If extra knowledge about another information source Y does not vary the information provided by X, then X and Y are independent

Markov chains

Given the random variables X,Y,Z, these variables conform a Markov chain $X \rightarrow Y \rightarrow Z$ if $p(z \mid x, y) = p(z \mid y)$

Example:



Possible sequences:
B3b, A1a, A2b, etc.

$p(x = A) = 0.4$

$p(x = B) = 0.6$

$p(y = 1 \mid x = A) = 0.5$

$p(y = 2 \mid x = A) = 0.5$

$p(y = 3 \mid x = A) = p(y = 4 \mid x = A) = 0$

$p(y = 1 \mid x = B) = p(y = 2 \mid x = B) = 0$

$p(y = 3 \mid x = B) = 0.9$

$p(y = 4 \mid x = B) = 0.1$

$p(z = a \mid y = 1) = 1$

$p(z = b \mid y = 2) = 1$

$p(z = b \mid y = 3) = 1$

$p(z = a \mid y = 4) = 0.5$

$p(z = b \mid y = 4) = 0.5$

Signal processing inequality

$$X \rightarrow Y \rightarrow Z \Rightarrow I(X;Y) \geq I(X;Z)$$

The signal processing inequality leads to the assertion that if X is processed to generate Y, and Y is post-process to generate Z, then, X has more information on Y tan on Z. Thus, Z does not provides more information about X tan Y..

$$X \rightarrow Y \rightarrow f(Y) \Rightarrow I(X;Y) \geq I(X;f(Y))$$

If Y is generated from X, the information that Y contains regarding X, is not going to be increased if Y is processes by any signal processing algorithm.

$$X \rightarrow Y \rightarrow Z \Rightarrow I(X;Y\,|\,Z) \leq I(X;Y)$$

In a Markov chain, knowing the value of Z decreases (or just keeps) the dependency between X and Y.

Coding and compression: Block Code

Example:

x_i	$\Pr\{X = x_i\}$	$C(x_i)$
1	1/2	0
2	1/4	10
3	1/8	110
4	1/8	111

AVERAGE LENGTH OF THE CODE

$$L(C) = E\left[l\left(C(x_i)\right)\right]$$

Length of code C(xi)

$$H(X) = L(C) = 1.75 bits$$

Example:

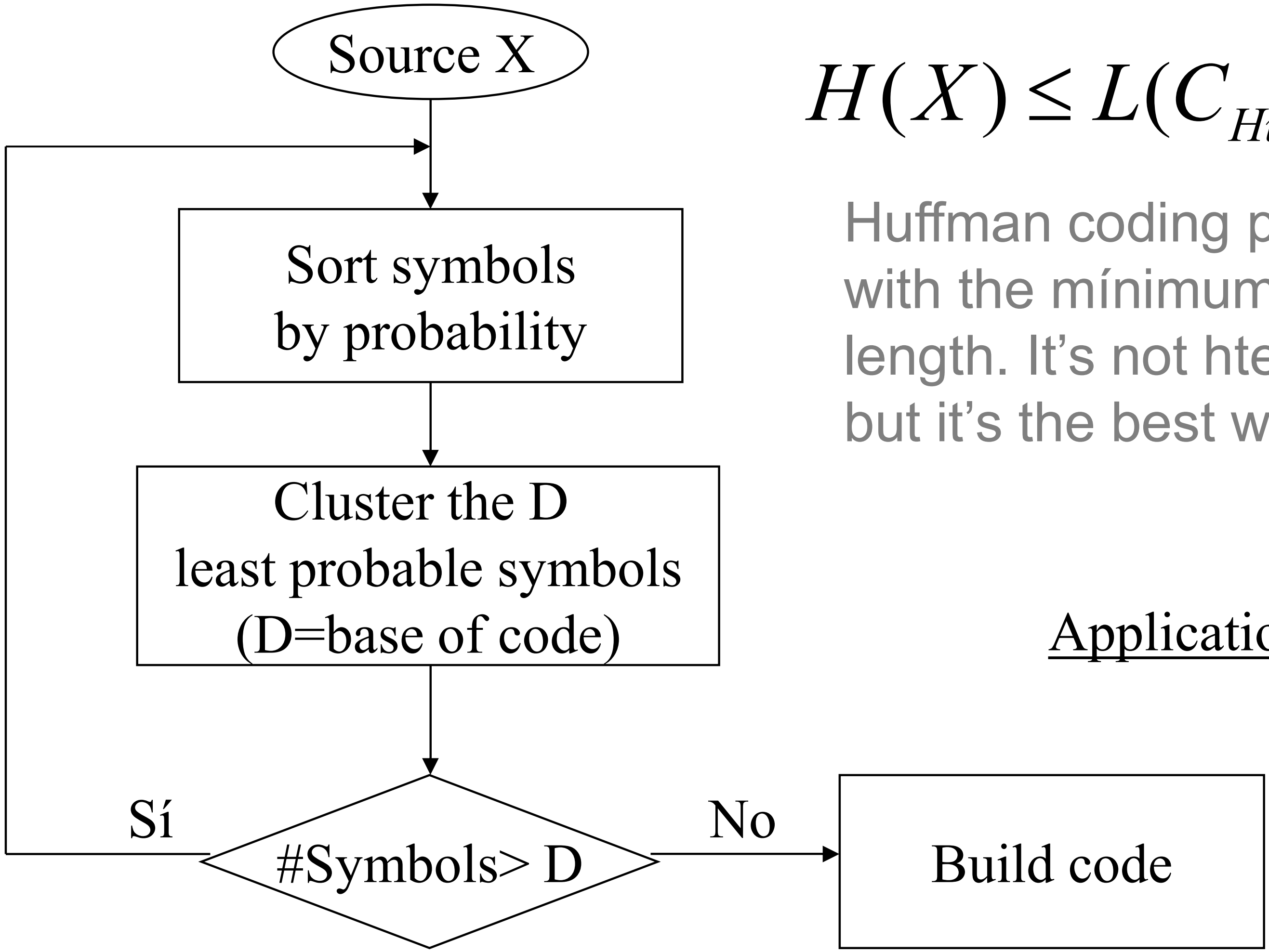
x_i	$\Pr\{X = x_i\}$	$C(x_i)$
1	1/3	0
2	1/3	10
3	1/3	11

$$H(X) = 1.58 bits$$
$$L(C) = 1.66 bits$$

H(X)≤L(C)

This is a fundamental property in statistical coding

Huffman coding



$$H(X) \leq L(C_{Huffman}) \leq L(C)$$

Huffman coding produces a code with the mínimum posible average length. It's not hte optimum code, but it's the best we can get.

Applications: JPEG, MP3, etc.

Huffman coding: D=2

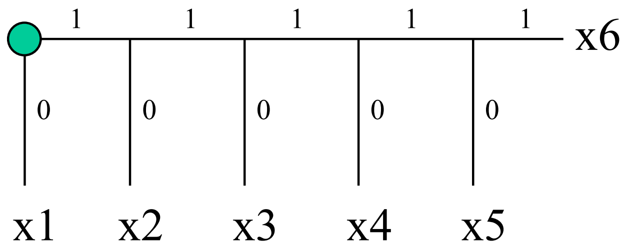
x_i	p_i
x1	0.4
x2	0.3
x3	0.1
x4	0.1
x5	0.06
x6	0.04

x_i	p_i
x1	0.4
x2	0.3
x3	0.1
x4	0.1
(x5x6)	0.1

x_i	p_i
x1	0.4
x2	0.3
x4(x5x6)	0.2
x3	0.1

x_i	p_i
x1	0.4
x2	0.3
x3(x4(x5x6))	0.3

x_i	p_i
x2(x3(x4(x5x6)))	0.6
x1	0.4



x_i	$C(x_i)$
x1	0
x2	10
x3	110
x4	1110
x5	11110
x6	11111

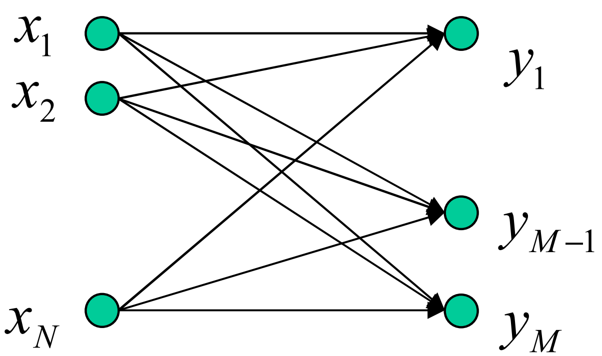
$$H(X) = 2.1435 bits$$

$$L(C_{Huffman}) = 2.2 bits$$

Channel characterization

Memoryless channel

$$P(y[n] \mid x[n], x[n-1], x[n-2], \ldots) = P(y[n] \mid x[n])$$



$$Q = \begin{pmatrix} P(y_1 \mid x_1) & P(y_2 \mid x_1) & \ldots & P(y_M \mid x_1) \\ P(y_1 \mid x_2) & P(y_2 \mid x_2) & \ldots & P(y_M \mid x_2) \\ \ldots & \ldots & \ldots & \ldots \\ P(y_1 \mid x_N) & P(y_2 \mid x_N) & \ldots & P(y_M \mid x_N) \end{pmatrix}$$

Channel with memory

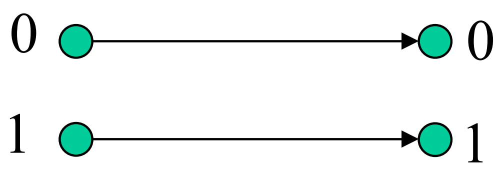
$$P(y[n] \mid x[n], x[n-1], x[n-2], \ldots) = P(y[n] \mid x[n]x[n-1])$$ Memory=1

$$P(y[n] \mid x[n], x[n-1], x[n-2], \ldots) = P(y[n] \mid x[n]x[n-1]x[n-2])$$ Memory=2

Channel capacity

$$C = \max_{p(X)} I(Y;X) \qquad 0 \le C \le \min(\log \# X, \log \# Y)$$

Example: Binary Channel with no noise



$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

P(X=0)=p
P(X=1)=1-p

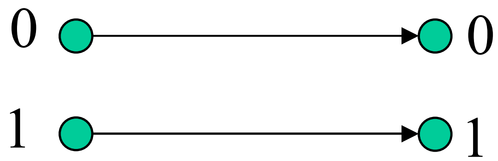
$$I(Y;X) = \sum_{x\in\Xi, y\in\Psi} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} =$$

$$= p(x=0,y=0)\log\frac{p(x=0,y=0)}{p(x=0)p(y=0)} + p(x=0,y=1)\log\frac{p(x=0,y=1)}{p(x=0)p(y=1)}$$

$$+ p(x=1,y=0)\log\frac{p(x=1,y=0)}{p(x=1)p(y=0)} + p(x=1,y=1)\log\frac{p(x=1,y=1)}{p(x=1)p(y=1)}$$

Channel capacity

Example: Binary channel with no noise



$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$ ← **Transition matrix**

$$p(X=x, Y=y) = p(X=x)\,p(Y=y \mid X=x) = \begin{cases} p(X=x) & x=y \\ 0 & x \neq y \end{cases}$$

$$p(Y=y) = \sum_x p(Y=y \mid X=x)\,p(X=x) = p(X=y)$$

$$I(Y;X) = p(x=0,y=0)\log\frac{p(x=0,y=0)}{p(x=0)\,p(y=0)} + p(x=0,y=1)\log\frac{p(x=0,y=1)}{p(x=0)\,p(y=1)}$$

$$+p(x=1,y=0)\log\frac{p(x=1,y=0)}{p(x=1)\,p(y=0)} + p(x=1,y=1)\log\frac{p(x=1,y=1)}{p(x=1)\,p(y=1)}$$

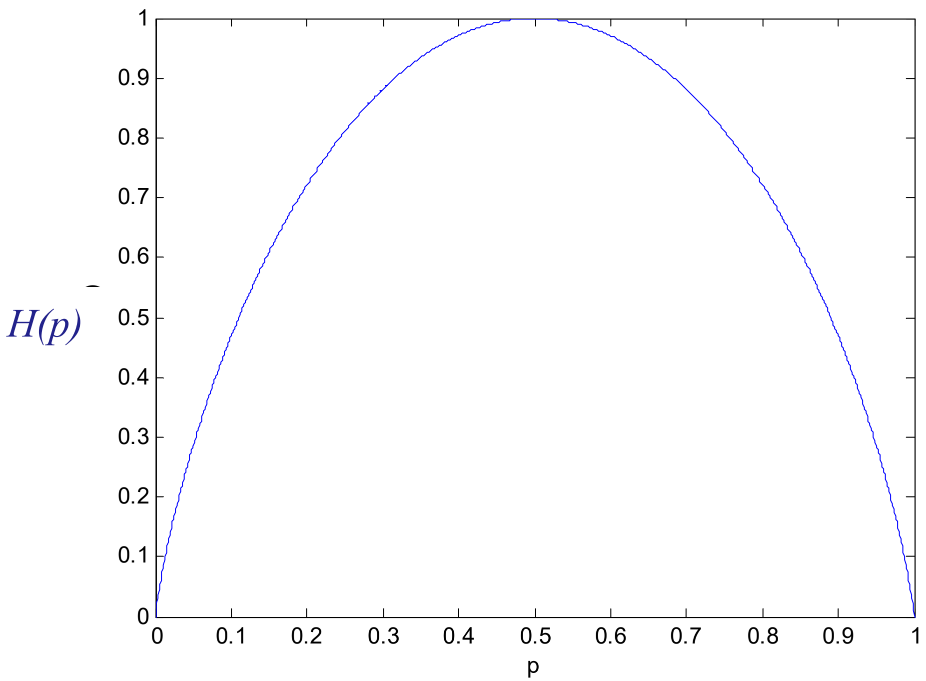
$$= p(x=0)\log\frac{p(x=0)}{p(x=0)\,p(x=0)} + p(x=1)\log\frac{p(x=1)}{p(x=1)\,p(x=1)}$$

$$= p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p} = H(X)$$

Channel capacity

Example: Binary channel with no noise

$$C = \max_{p(X)} I(Y;X) = \max_{p}\left\{ p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p} \right\}$$

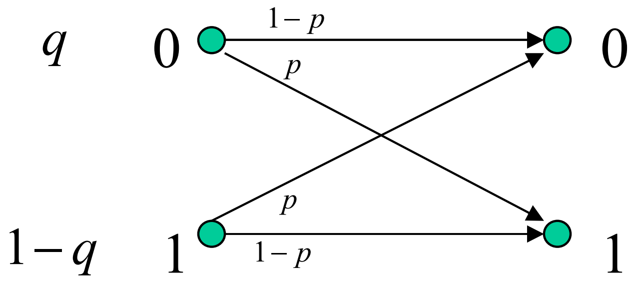


$$p = \frac{1}{2} \Rightarrow C = 1bit$$

(Remember: P(X=0)=*p*)

Channel capacity

Example: Binary Symetric Channel (BSC) – channel with noise



$$q=\frac{1}{2}\Rightarrow C=\max_q I(Y;X)=1-H(p)$$

$$H(Y\mid X=0)=p(Y=0\mid X=0)\log\frac{1}{p(Y=0\mid X=0)}+p(Y=1\mid X=0)\log\frac{1}{p(Y=1\mid X=0)}$$

$$=(1-p)\log\frac{1}{1-p}+p\log\frac{1}{p}=H(p)=H(Y\mid X=1)$$

$$P(Y=0)=P(X=0)P(Y=0\mid X=0)+P(X=1)P(Y=0\mid X=1)=q(1-p)+(1-q)p$$

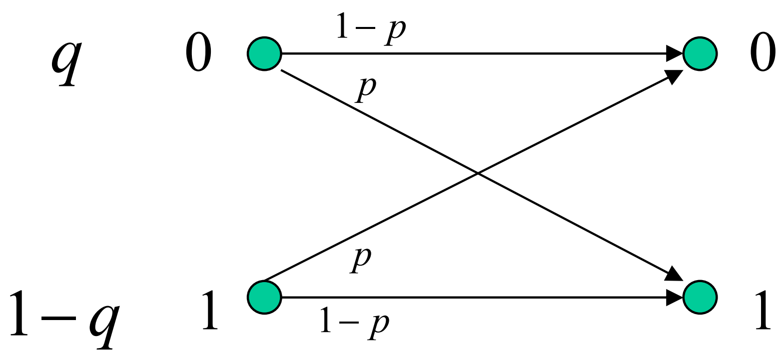
$$P(Y=1)=P(X=0)P(Y=1\mid X=0)+P(X=1)P(Y=1\mid X=1)=qp+(1-q)(1-p)$$

$$H(Y)=P(Y=0)\log\frac{1}{P(Y=0)}+P(Y=1)\log\frac{1}{P(Y=1)}\leq H(X)\leq 1bit$$

Only equal if Y is uniform

Channel capacity

Example: Binary Symetric Channel (BSC) – channel with noise



$$q = \frac{1}{2} \Rightarrow C = \max_q I(Y;X) = 1 - H(p)$$

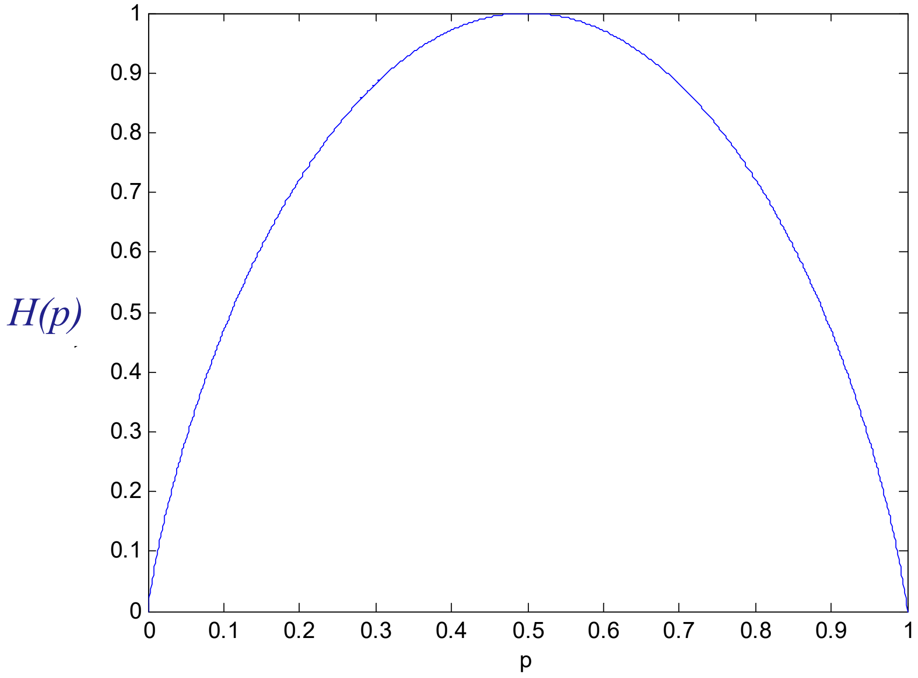
$$I(Y;X) = H(Y) - H(Y \mid X)$$
$$= H(Y) - \sum_x p(x) H(Y \mid X = x) \leq 1 - \sum_x p(x) H(p) = 1 - H(p) = C$$

If Y is uniform

Channel Capacity

Example: Binary channel with noise (BSC)

$$C = \max_{p(X)} I(Y;X) = 1 - H(p)$$

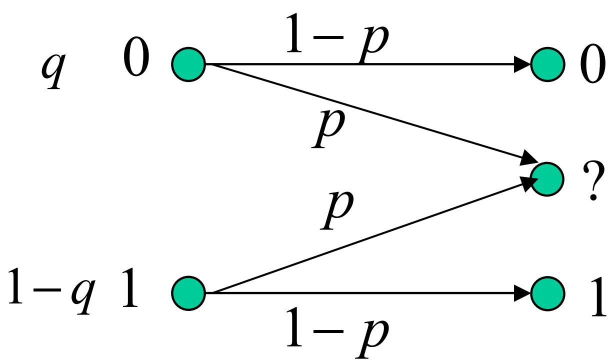


p=0 or p=1 → C=1 bit
p=0.5 → C=0 bits

(Remember: $P(Y=1|X=0)=P(Y=0|X=1)=p$)

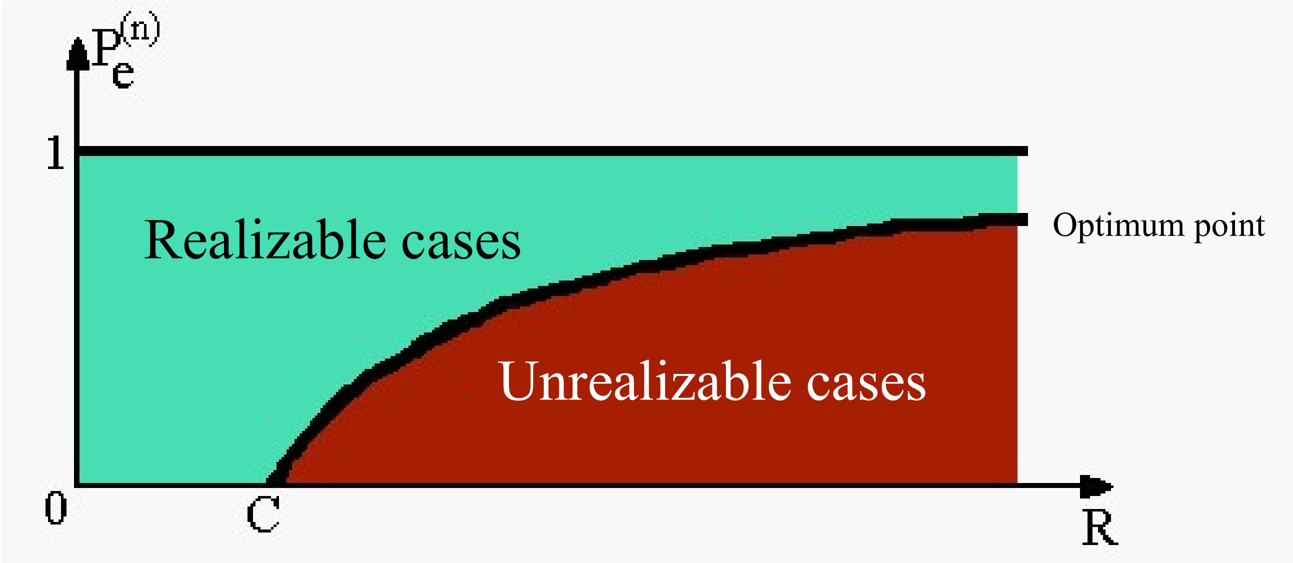
Challenge: Channel capacity

Given an "erasure channel", where the data is received corrected ("0" or "1") or
not received at all ("?") show that the capacity ***C=1-p***



$$Q = \begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix}$$

Channel coding theorem



It is also posible to express channel capacity in bits per second. Given C and the transmission rate R, it is proved that it is posible to achieve an error probability Pe=0 if R<C. Also, if R>C then the error probability increases.
It is posible to find codes that achieves Pe=0 and also that increase the transmission speed increasing the error rate, however the theorem does not indicate how are these codes.

Case study:
Are brains good at processing information?
(Introduction to Information Theory, J.V. Stone, 2018)

- Neurons communicate to each other continuosly
  - This communication must be performed efficiently

- Horace Bellow supports the *efficiency coding hypothesis*, where the sensory input must be encoded efficiently before being sent to the brain

Case study:
Are brains good at processing information?
(Introduction to Information Theory, J.V. Stone, 2018)

- Information in spiking neurons
  - The neurons propagate action potentials (AP, a.k.a. spikes)
  - If the spikes are encoded as 0's (no AP) and 1's (AP) it is possible to compute the entropy of the neuron (information source)
  - Considering an average firing rate of *r* spikes/s the capacity of the channel might be equal to r bits/s
  - However it can be proven that the information rate is bigger than that
  - The trick is that the neurons also use temporal information to encode information, so that one single spike encodes more than 1 bit

Case study:
Are brains good at processing information?
(Introduction to Information Theory, J.V. Stone, 2018)

- Mutual information between the input and output of a neuron
  - An experiment showed that a neuron (mechanical receptor of a cricket) generates 600 bits/s
  - 300 bits/s are related to the input, the rest is noise
  - Thinking that a neuron works with “packets” of 300 bits is out of line
  - There are theories that indicate that those 300 bits are actually divided in packets of 3 bits every 10 ms, providing continuous information about changes in the output (speed of an object)

Case study:
Are brains good at processing information?
(Introduction to Information Theory, J.V. Stone, 2018)

- Shannon optimal coding: maximizing entropy
  - In the human eye, the information provided by “Red” and “Green” receptors is very similar
  - Using two different nerve fibre per receptor is a waste of channel capacity (since there is a lot of redundancy over time)
  - However, it can be proved that the addition and substraction of the outputs of the R and G receptors leads to signals with uniform distributions
  - Also the substraction and summation are independent of each other
  - So:
    1. The use of two, instead of three, separate fibres is now justified
    2. The entropy is maximized so that the information rate is highly increased
  - Ganglion cells in the retina perform this operations, but they use several receptor outputs to **compress** information and reduce the necessary nerve fibres: 126 million receptors → 1 million nerve fibre

SUMMARY

- Entropy
- Mutual information
- Signal processing inequality
- Huffman coding
- Channel Capacity
- Channel coding theorem