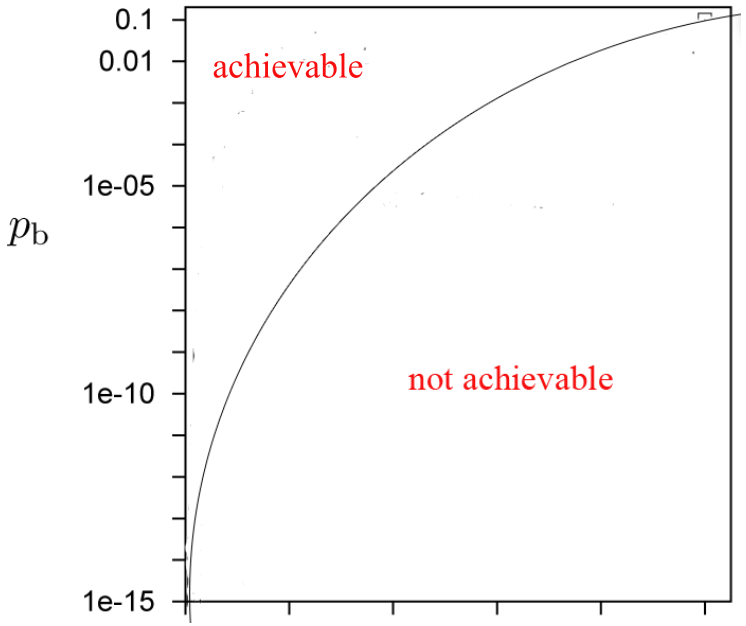
Information Theory
Communication over a noisy channel

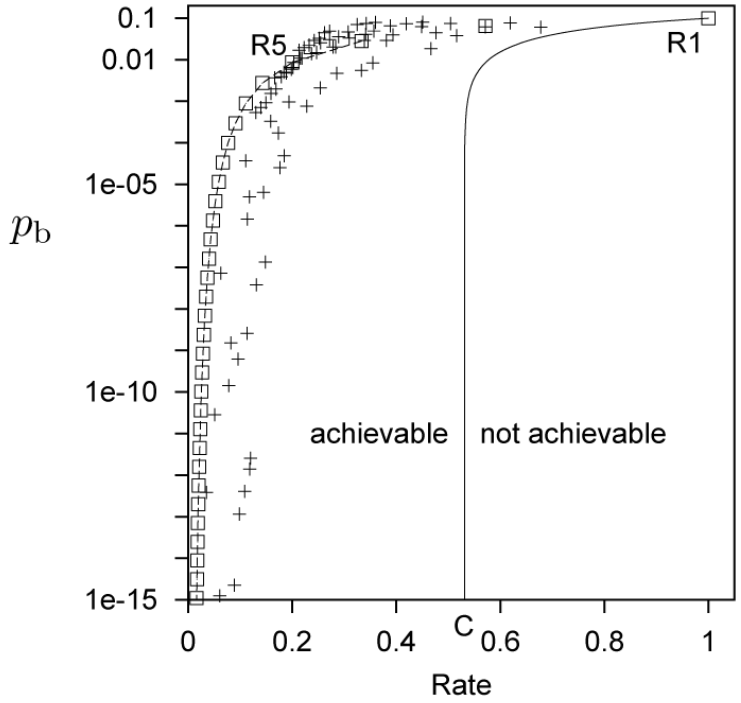
CoCoNut, 2016

Emmanuela Orsini

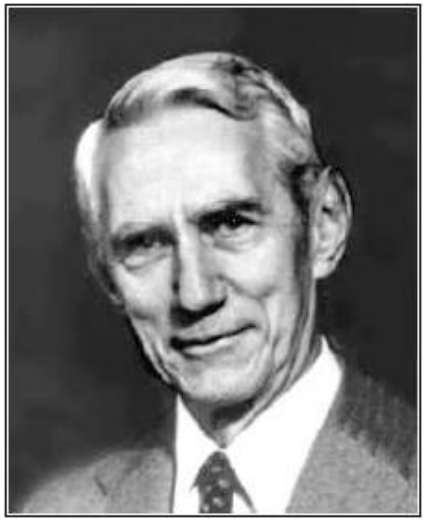
PREVIOUSLY...

Good codes exist! but. . .





Shannon proved that for every DMC with a finite number of inputs and outputs points, one may define the notion of **channel capacity**.



Claude Shannon 1916-2001

- If C is a code with rate $R > \mathcal{C}$, then the probability of error in decoding this code is bounded away from 0. (In other words, at any rate $R > \mathcal{C}$, reliable communication is not possible).
- For any information rate $R < \mathcal{C}$ and any $\delta > 0$, there exists a code C of length n_δ and rate R, such that the probability of error in maximum likelihood decoding of this code is at most δ.

  Proof: **Non-constructive!**

How can we find good code? Ingredients of Shannons proof:

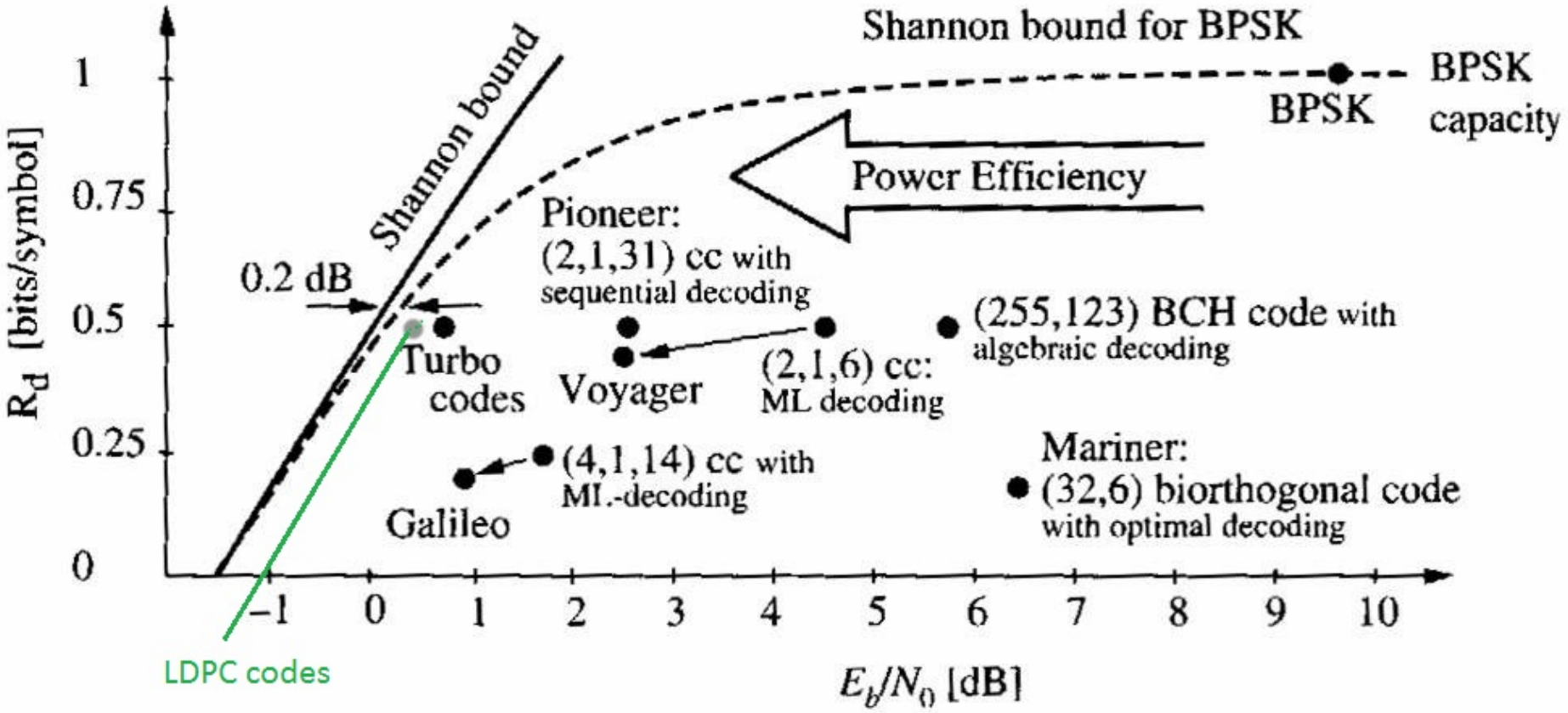
- Random code
- Large block length
- Optimal decoding

Solution:

- Long, structured, “pseudorandom” codes
- Practical, near-optimal decoding algorithms

State-of-art:

- Turbo codes and LDPC codes have brought Shannon limits to within reach on a wide range of channels


Shannon bound for BPSK
BPSK
BPSK capacity
Power Efficiency
Shannon bound
0.2 dB
Pioneer:
(2,1,31) cc with
sequential decoding
(255,123) BCH code with
algebraic decoding
Turbo
codes
Voyager
(2,1,6) cc:
ML decoding
(4,1,14) cc with
ML-decoding
Mariner:
(32,6) biorthogonal code
with optimal decoding
Galileo
LDPC codes
R_d [bits/symbol]
E_b/N_0 [dB]
1
0.75
0.5
0.25
0
−1
0
1
2
3
4
5
6
7
8
9
10

- LDPC codes are capacity-approaching codes
  - G.hn/G.9960 (ITU-T Standard for networking over power lines, phone lines and coaxial cable)
  - 802.3an (10 Giga-bit/s Ethernet over Twisted pair)
  - DVB-S2 / DVB-T2 / DVB-C2 (Digital video broadcasting, 2nd Generation) and DMB-T/H (Digital video broadcasting)
  - WiMAX (IEEE 802.16e standard for microwave communications)
  - IEEE 802.11n-2009 (Wi-Fi standard)

Binary LDPCC

A linear code C is LDPC if H is sparse.

Definition

An LDPC code is called (w_c, w_r)**-regular** if each its parity-check matrix contains a fixed number, w_c, of 1's per column, and a fixed number, w_r, of 1's per row.

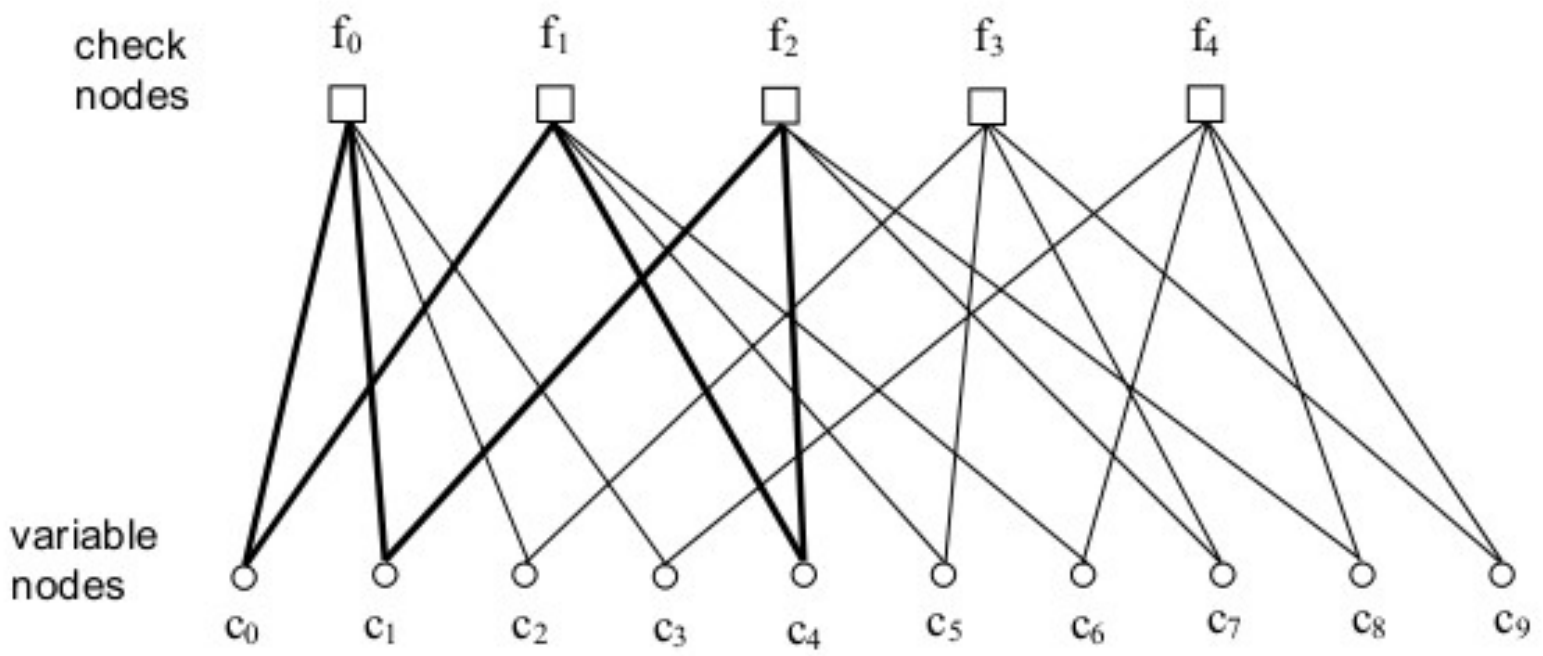
A *Tanner's graph of a code* is a graph with two types of nodes, the *variable nodes* and the *check nodes*.

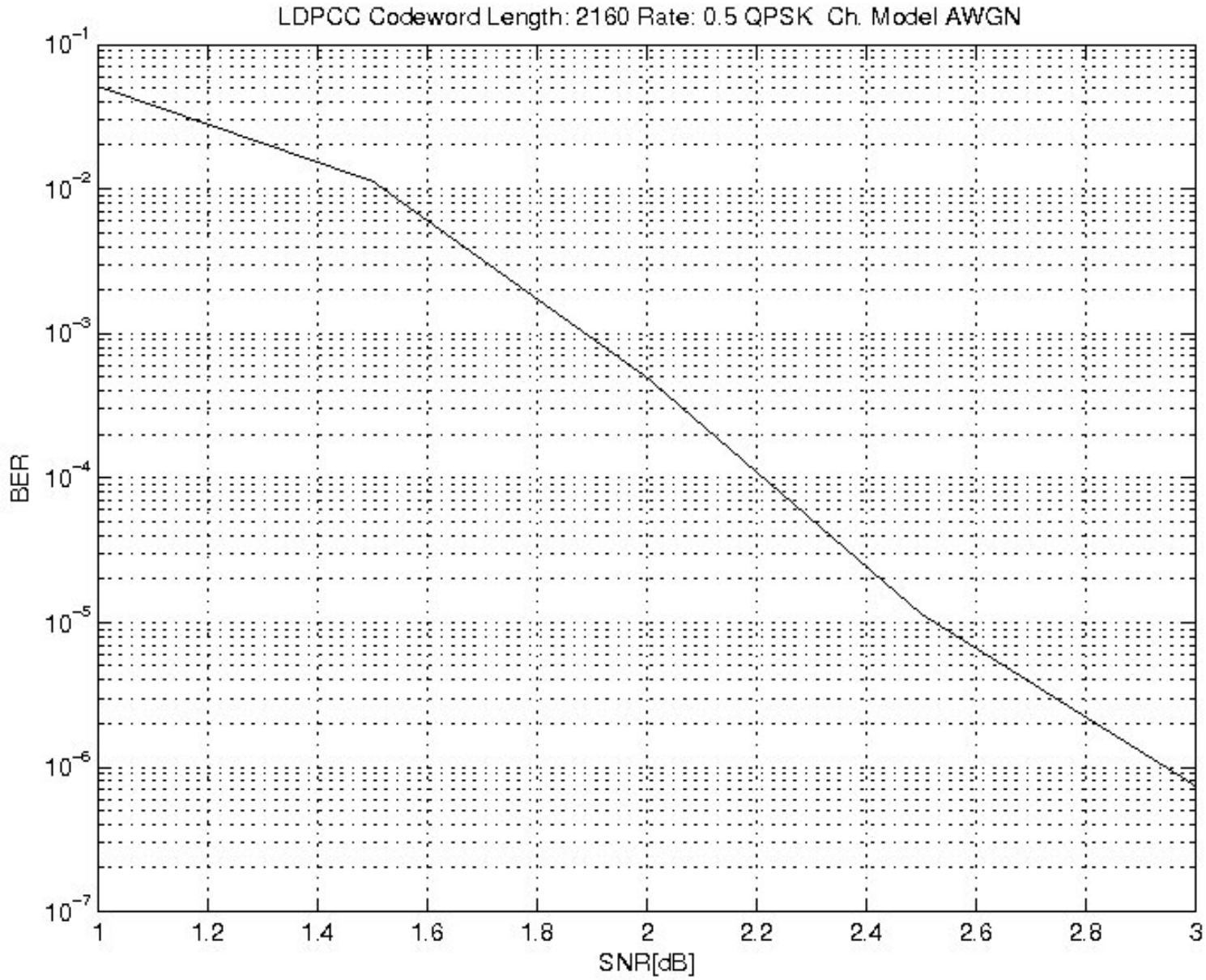
The Tanner graph of a code is drawn accordingly the following rule:

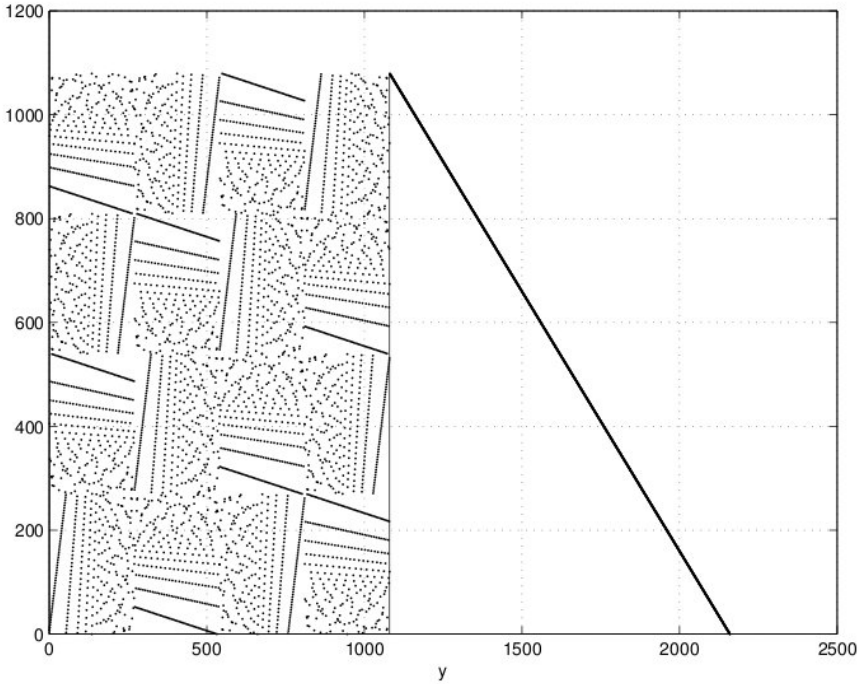
- check-node j is connected to variable node i whenever element h_{ij} in the parity-check matrix H is 1.

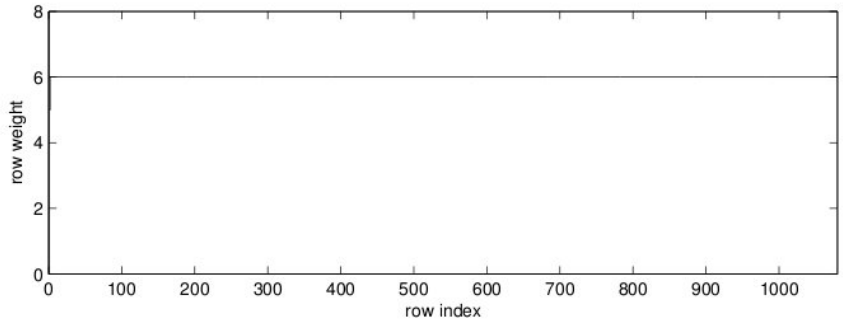
Consider a $(10, 5)$ linear code C with $w_c = 2$ and $w_r = 4$, with the following parity-check matrix

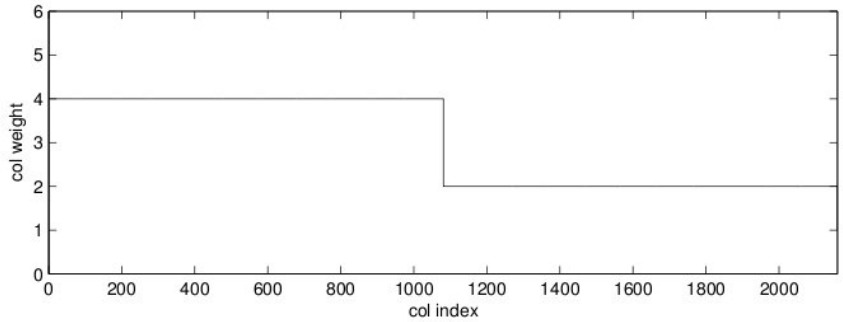
$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$




LDPCC Codeword Length: 2160 Rate: 0.5 QPSK Ch. Model AWGN
BER
SNR[dB]
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3


1200
1000
800
600
400
200
0
0
500
1000
1500
2000
2500
y


8
6
4
2
0
row weight
0
100
200
300
400
500
600
700
800
900
1000
row index


6
5
4
3
2
1
0
col weight
0
200
400
600
800
1000
1200
1400
1600
1800
2000
col index

The big picture

	Compression “SOURCE CODING”	Error correction “CHANNEL CODING”

The big picture

	Compression "SOURCE CODING"	Error correction "CHANNEL CODING"
Information theory		
Coding methods		

The big picture

	Compression "SOURCE CODING"	Error correction "CHANNEL CODING"
Information theory	Source coding theorem Kraft-McMillan ineq.	Channel coding theorem Channel capacity
Coding methods		

The big picture

	Compression "SOURCE CODING"	Error correction "CHANNEL CODING"
Information theory	Source coding theorem Kraft-McMillan ineq.	Channel coding theorem Channel capacity
Coding methods	Symbol codes Huffman codes	Hamming codes Reed Solomon codes LDPC codes

THIS LECTURE...

Informally

Shannon's definition represents a way to measure the amount of information that can potentially be gained when one learns of the outcome of a random process.

Example

- Suppose you need to uncover a certain English word of five letters.
- You manage to obtain one letter, namely an *e*.
- This is useful information, but the letter *e* is common in English, so this provides little information. If, on the other hand, the letter that you discover is *j* (the least common in English), the search has been more narrowed and you have obtained more information

The "uncertainty" of a random experiment is strictly related to the information carried by the result of this experiment.

We want to formalize this intuition: we want to quantify the amount of information that is carried by a random variable

- We consider a random variable X on a finite set
- The probability distribution of X is given by

$$p_X(x_i) = P(X = x_i), i = 1, \ldots, m \quad \text{and} \quad \sum_{i=1}^{m} p_X(x_i) = 1$$

We want to formalize this intuition: we want to quantify the amount of information that is carried by a random variable

- We consider a random variable X on a finite set
- The probability distribution of X is given by

$$p_X(x_i) = P(X = x_i), i = 1, \ldots, m \quad \text{and} \quad \sum_{i=1}^{m} p_X(x_i) = 1$$

How can we measure the information content of an outcome x_i?

◇ **Shannon information content**:

$$I(p_X(x_i)) = h(x_i) = -\log_2 p(x_i)$$

We want to formalize this intuition: we want to quantify the amount of information that is carried by a random variable

- We consider a random variable X on a finite set
- The probability distribution of X is given by

$$p_X(x_i) = P(X = x_i), i = 1, \ldots, m \quad \text{and} \quad \sum_{i=1}^{m} p_X(x_i) = 1$$

How can we measure the information content of an outcome x_i?

- **Shannon information content**:

$$I(p_X(x_i)) = h(x_i) = -\log_2 p(x_i)$$

- It is measured in bits.

Example. Consider a single toss of coin. How much information do we receive when we are told that the outcome is head?

We can consider different situations:

1. if $\Pr(H) = \Pr(T) = 1/2$:

$$I(\Pr(H)) = 1$$

2. if $\Pr(H) = 1$

$$I(\Pr(H)) = 0$$

3. if $\Pr(H) = 0.7$

$$I(\Pr(H)) \approx 0.511$$

We gain more information when something unexpected happens, and gain less information when something expected happens.

Entropy(Informally): average amount of information contained in each outcome we obtain.

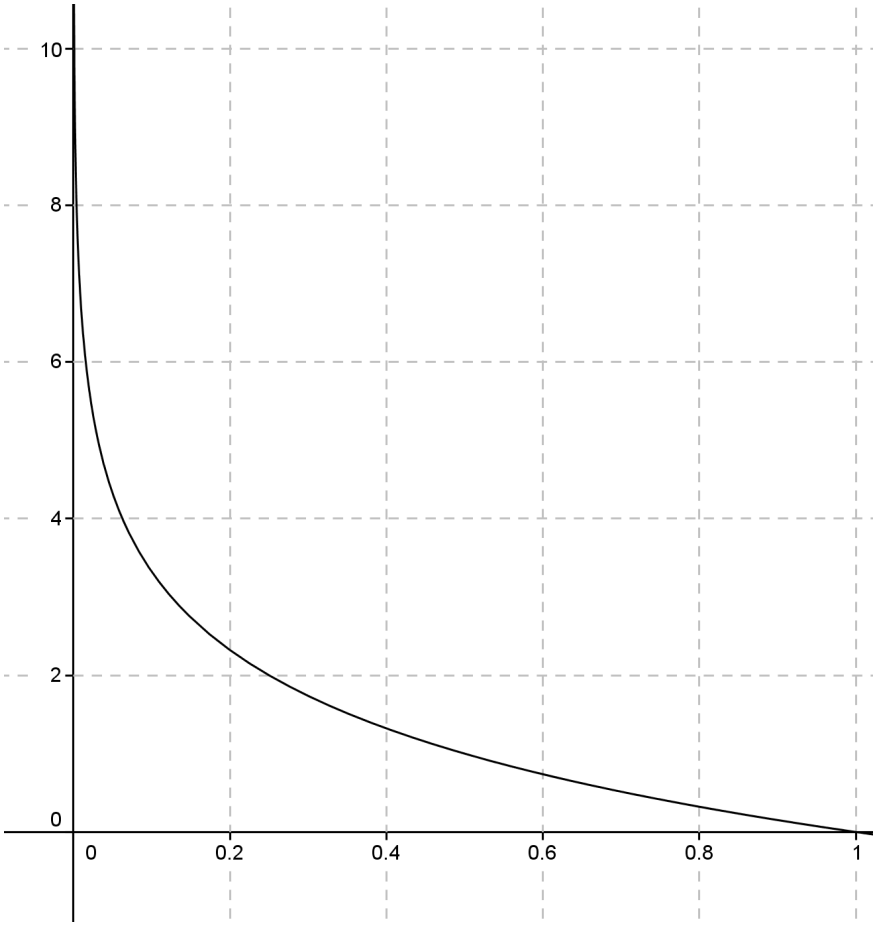
Definition (Entropy)

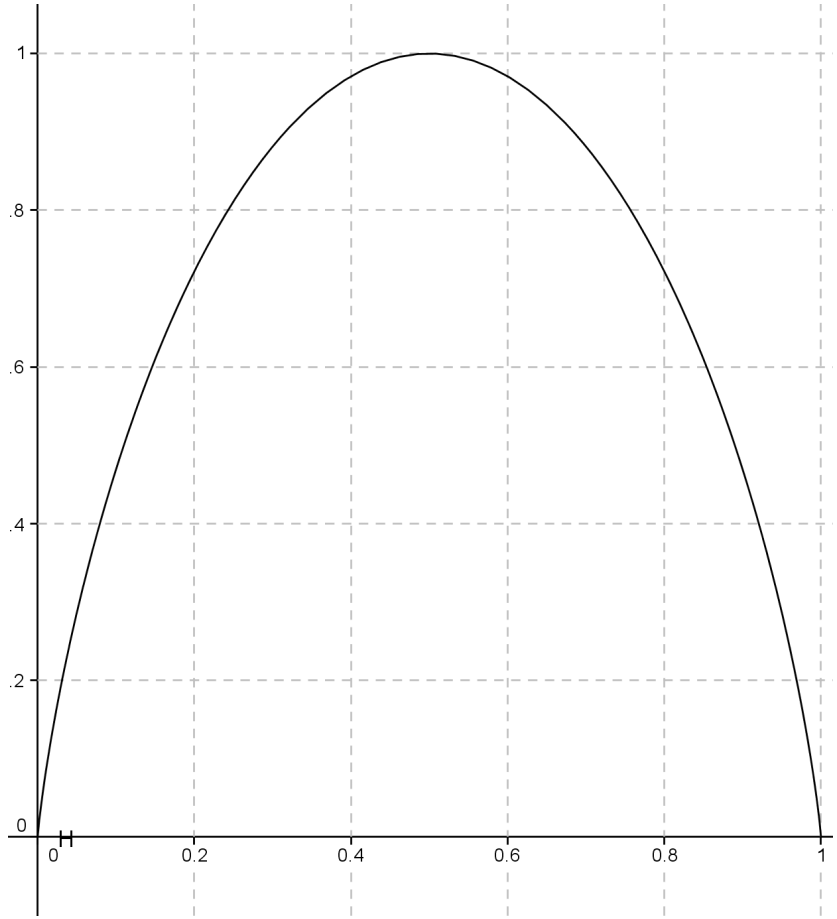
Given a discrete random variable X over a finite sized alphabet, the entropy of X is defined to be the expected value,

$$H(X) = \sum_{i=1}^{m} p(x_i) \log_2 \frac{1}{p(x_i)} = -\sum_{i=1}^{m} p(x_i) \log_2 p(x_i),$$

of the uncertainty in a single observation of X.

- We adopt the convention that $0 \log \frac{1}{0} = 0$.
- $0 \leq H(X) \leq \log_2 m$
- The *maximum entropy*, $H(X) = \log_2 m$, is reached when all the outcomes are equiprobable, i.e. $p(x_i) = 1/m$.
- If X is a r.v. such that $\Pr(X = 0) = p$ and $\Pr(X = 1) = 1 - p$, then the entropy of X is sometimes denoted by $H(p)$ and called **binary entropy function**.


10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1


1
.8
.6
.4
.2
0
H
0
0.2
0.4
0.6
0.8
1

- A **discrete memoryless channel** *Ch* is described by
  1. an input alphabet $\mathcal{X} = \{x_1, \ldots, x_m\}$
  2. an output alphabet $\mathcal{Y} = \{y_1, \ldots, y_{m'}\}$
  3. a set of conditional probabilities $p(y_j|x_i)$

  More precisely, the input source to a noisy channel is a random variable X over $\mathcal{X}$, the output source is a random variable Y over $\mathcal{Y}$, and
  $$p_{Y|X} = \Pr(Y = y_j|X = x_i)$$

- We can choose an input distribution p_X

- From the conditional distribution and the input distribution that we choose, we can deduce the **joint distribution**

$$p_{X,Y} = p_{Y|X} \cdot p_X$$

- We can obtain the **posterior distribution** using Bayes' theorem:

$$p_{X|Y} = \frac{p_{Y|X} \cdot p_X}{p_Y}$$

The input source to a noisy channel is a r. v. X over a, b, c, d. The output for this channel is a r. v. Y over the same alphabet. The joint probability of these two random variables is:

X \ Y	a	b	c	d	$\Pr_X(x)$
a	$1/8$	$1/16$	$1/16$	$1/4$	$1/2$
b	$1/16$	$1/8$	$1/16$	0	$1/4$
c	$1/32$	$1/32$	$1/16$	0	$1/8$
d	$1/32$	$1/32$	$1/16$	0	$1/8$
$\Pr_Y(y)$	$1/4$	$1/4$	$1/4$	$1/4$	1

We can compute $H(X), H(Y)$:

$$\begin{aligned} H(X) =& 1/2 \log_2 2 + 1/4 \log_2 4 \\ &+ 1/8 \log_2 8 + 1/8 \log_2 8 \\ =& 7/4 \text{ (bits)} \end{aligned}$$

$$H(Y) = 2 \text{ (bits)}$$

The input source to a noisy channel is a r. v. X over a, b, c, d. The output for this channel is a r. v. Y over the same alphabet. The joint probability of these two random variables is:

X \ Y	a	b	c	d	$\Pr_X(x)$
a	$1/8$	$1/16$	$1/16$	$1/4$	$1/2$
b	$1/16$	$1/8$	$1/16$	0	$1/4$
c	$1/32$	$1/32$	$1/16$	0	$1/8$
d	$1/32$	$1/32$	$1/16$	0	$1/8$
$\Pr_Y(y)$	$1/4$	$1/4$	$1/4$	$1/4$	1

We can compute $H(X), H(Y)$:

$$\begin{aligned} H(X) =& 1/2 \log_2 2 + 1/4 \log_2 4 \\ &+ 1/8 \log_2 8 + 1/8 \log_2 8 \\ =& \ 7/4 \text{ (bits)} \end{aligned}$$

$$H(Y) = 2 \text{ (bits)}$$

- We can compute the **joint entropy** of X and Y:

$$H(X, Y) = -\sum_{i=1}^{m} \sum_{j=1}^{m} p(x_i, y_j) \log_2(p(x_i, y_j)).$$

- We obtain $H(X, Y) = 27/8$ bits.

$P_{X\|Y}$		Y				
		a	b	c	d	
	a	$1/2$	$1/4$	$1/4$	1	
	b	$1/4$	$1/2$	$1/4$	0	
X	c	$1/8$	$1/8$	$1/4$	0	
	d	$1/8$	$1/8$	$1/4$	0	
$H(X\|y)$		$7/4$	$7/4$	2	0	

We can compute $H(X), H(Y)$:

$$\begin{aligned} H(X) =& 1/2 \log_2 2 + 1/4 \log_2 4 \\ &+ 1/8 \log_2 8 + 1/8 \log_2 8 \\ &= 7/4 \text{ (bits)} \\ H(Y) =& 2 \text{ (bits)} \end{aligned}$$

$P_{X\|Y}$		Y				
		a	b	c	d	
	a	$1/2$	$1/4$	$1/4$	1	
	b	$1/4$	$1/2$	$1/4$	0	
X	c	$1/8$	$1/8$	$1/4$	0	
	d	$1/8$	$1/8$	$1/4$	0	
$H(X\|y)$		$7/4$	$7/4$	2	0	

We can compute $H(X), H(Y)$:

$$\begin{aligned} H(X) =& 1/2 \log_2 2 + 1/4 \log_2 4 \\ &+ 1/8 \log_2 8 + 1/8 \log_2 8 \\ &= 7/4 \text{ (bits)} \\ H(Y) =& 2 \text{ (bits)} \end{aligned}$$

- We can compute the entropy of X conditioned to the events $Y = a$, $Y = b$, $Y = c$, $Y = d$, for example

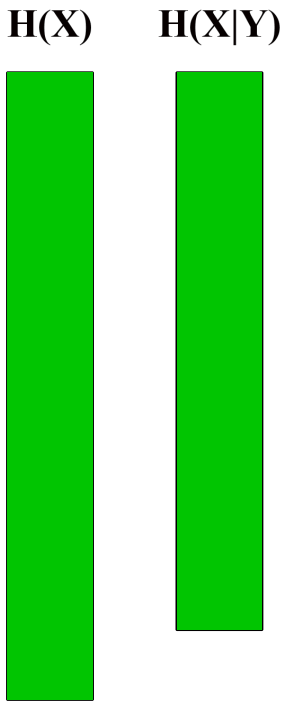
$$H(X|Y = a) = -\sum_{x \in \mathcal{X}} p(x|a) \log_2(p(x|a)).$$

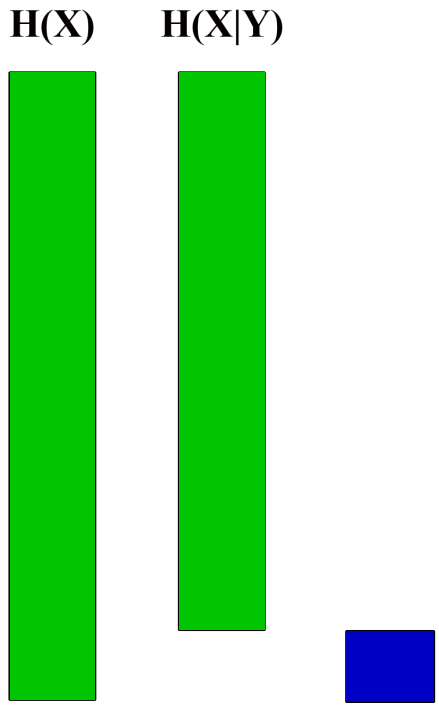
Conditional entropy (Informally): is the average over y_j of the conditional entropy of X given y_j

- We define the **conditional entropy** as the expected value:

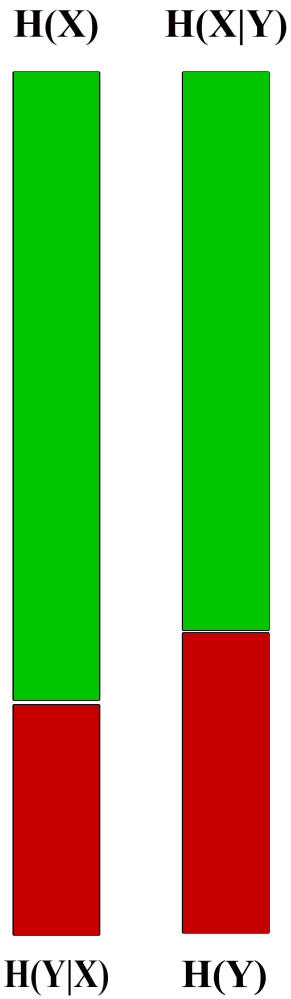
$$H(X|Y) = \sum_{k=1}^{m} p(y_k) H(X|Y = y_k)$$
$$= -\sum_{k=1}^{m}\sum_{i=1}^{m} p(x_i, y_k) \log_2(p(x_i|y_k)).$$

- $H(X|Y) = 11/8$ and $H(Y|X) = 13/8$ bits

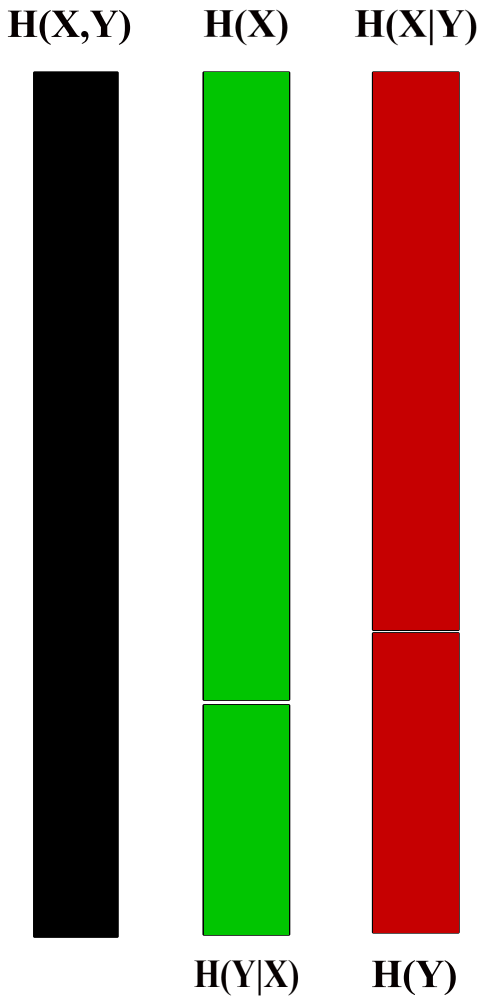

H(X)
H(X|Y)



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

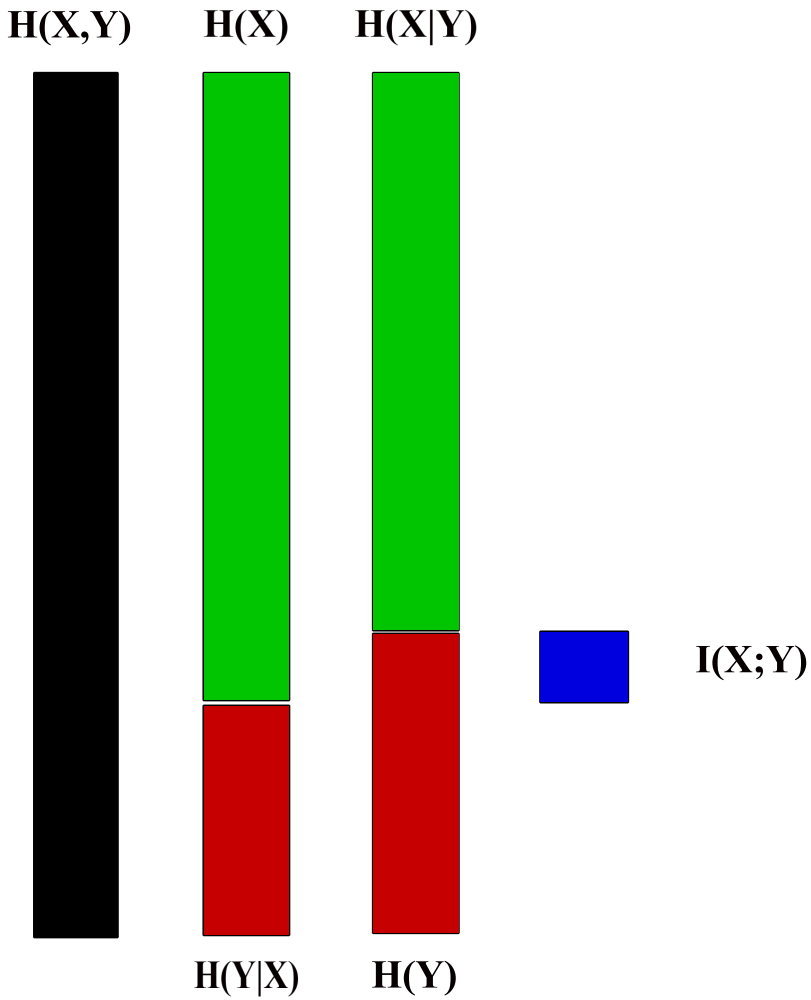


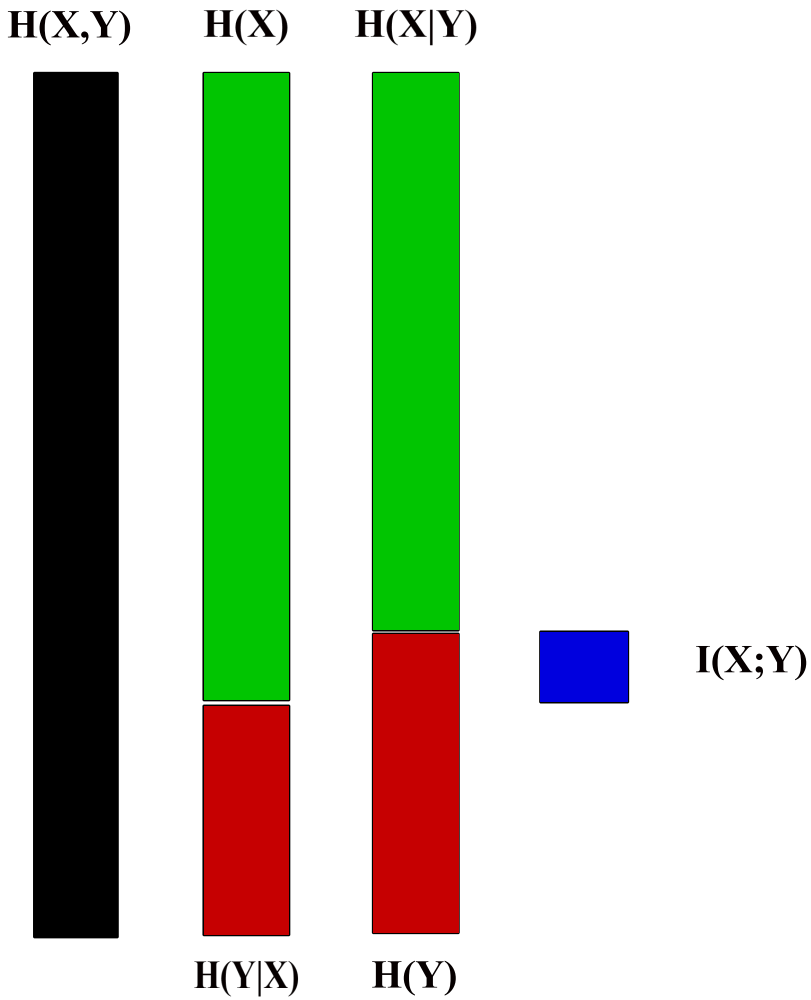
- $H(X|Y) \leq H(X)$



- $H(X|Y) \leq H(X)$
- **Chain rule for entropy**: The joint entropy, conditional entropy and marginal entropy are related by:

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


H(X,Y)
H(X)
H(X|Y)
I(X;Y)
H(Y|X)
H(Y)



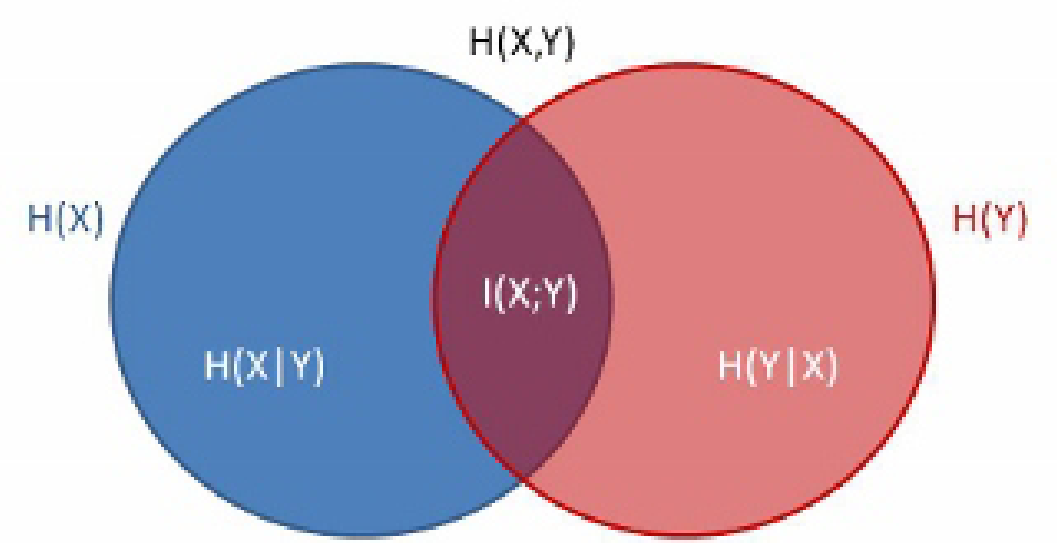
- **The mutual information**:

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

- (Informally) How much information the output convey about the input

The mutual information $I(X;Y)$ equals the reduction in the uncertainty of X due to the knowledge of Y, and similarly, it equals the reduction in the uncertainty of Y due to the knowledge of X.

Intuitively it is possible to represent the relation between entropy and mutual information using a Venn diagram.



$$I(X;Y) = H(X \cap Y)$$
$$H(X;Y) = H(X) + H(Y) - H(X \cap Y)$$

- We would like to determine the maximum information that can be sent over a channel.

- We would like to determine the maximum information that can be sent over a channel.
- At the beginning, when no symbol has been received, the uncertainty of X is its entropy $H(X)$.

- We would like to determine the maximum information that can be sent over a channel.
- At the beginning, when no symbol has been received, the uncertainty of X is its entropy $H(X)$.
- When a symbol is received, this uncertainty is reduced to $H(X|Y)$.

- We would like to determine the maximum information that can be sent over a channel.
- At the beginning, when no symbol has been received, the uncertainty of X is its entropy $H(X)$.
- When a symbol is received, this uncertainty is reduced to $H(X|Y)$.
- We can say that the information across the channel has been

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

- We would like to determine the maximum information that can be sent over a channel.
- At the beginning, when no symbol has been received, the uncertainty of X is its entropy $H(X)$.
- When a symbol is received, this uncertainty is reduced to $H(X|Y)$.
- We can say that the information across the channel has been

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

The **capacity** of a DMC is

$$C = max_{p(x)} I(X;Y).$$

- The capacity measures the information conveyed by the channel
  - The capacity measures the rate at which block of data can be communicated over the channel with arbitrarily small probability of error.
  - The capacity is the maximum rate at which we can send information over the channel and recover the information with a negligible error probability. C is only function of the probabilities defining the channel.

Theorem (Channel coding theorem)

The maximum rate R of information over a channel with arbitrarily low error probability is given by its channel capacity C.

This theorem, also known as **Noisy coding theorem** or **Shannon II Theorem**, states that as long as $R \leq C$, then the error probability can be made arbitrarily small. In other words it asserts that *good code exists*, but unfortunately its proof is not constructive and we do not know how to construct it.