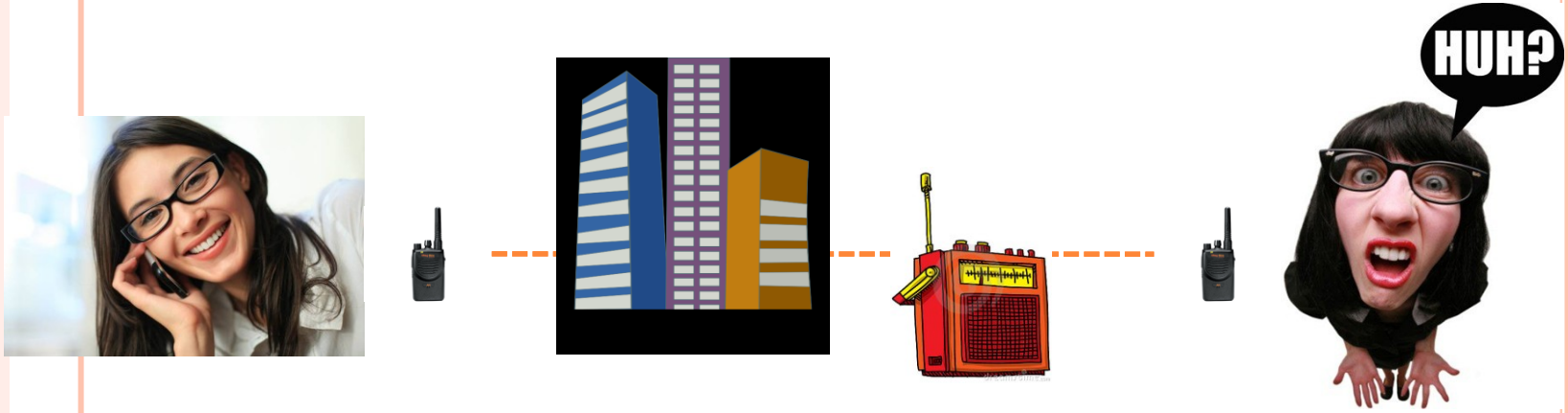
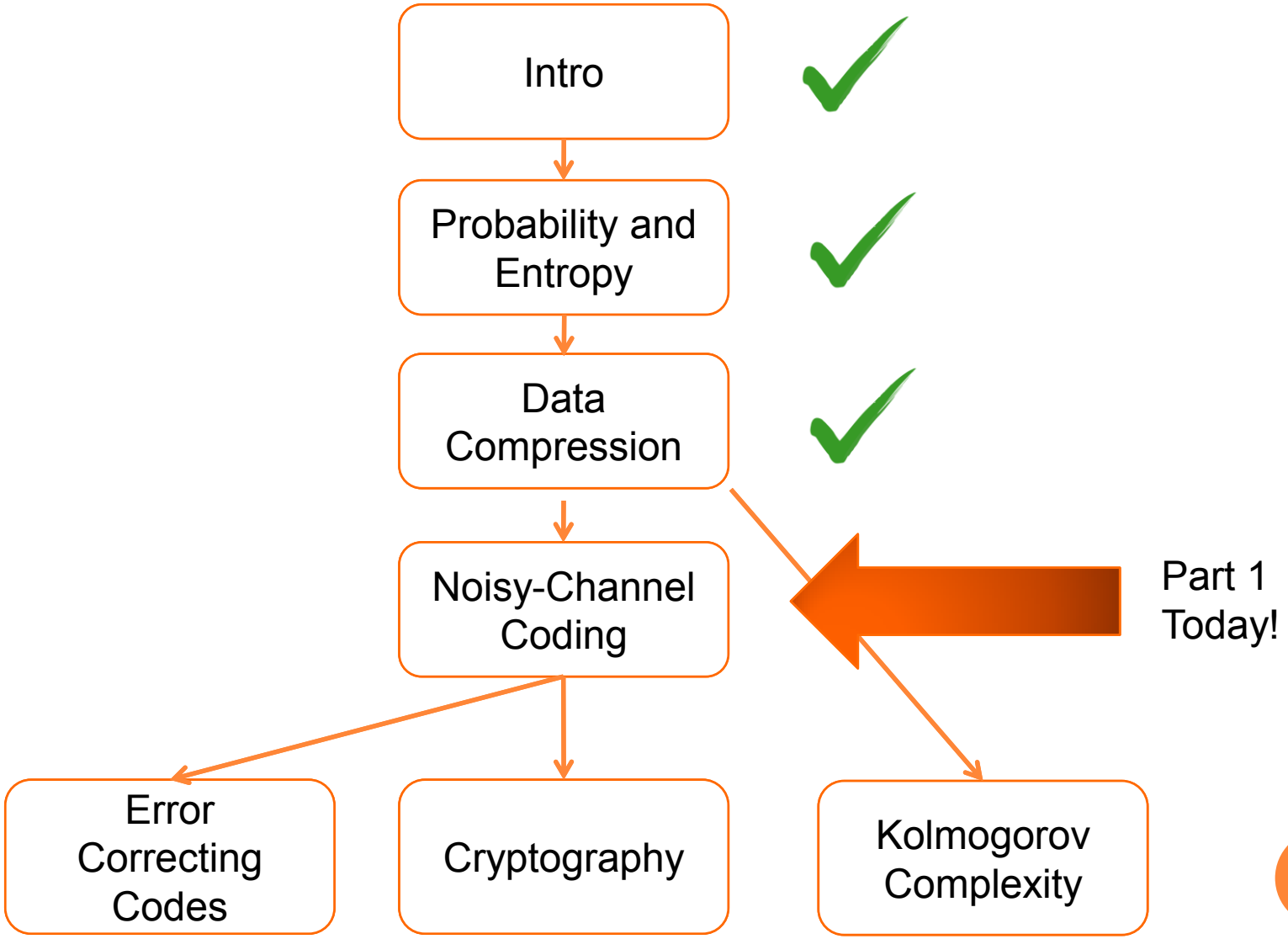


CS3236 Introduction to Information theory

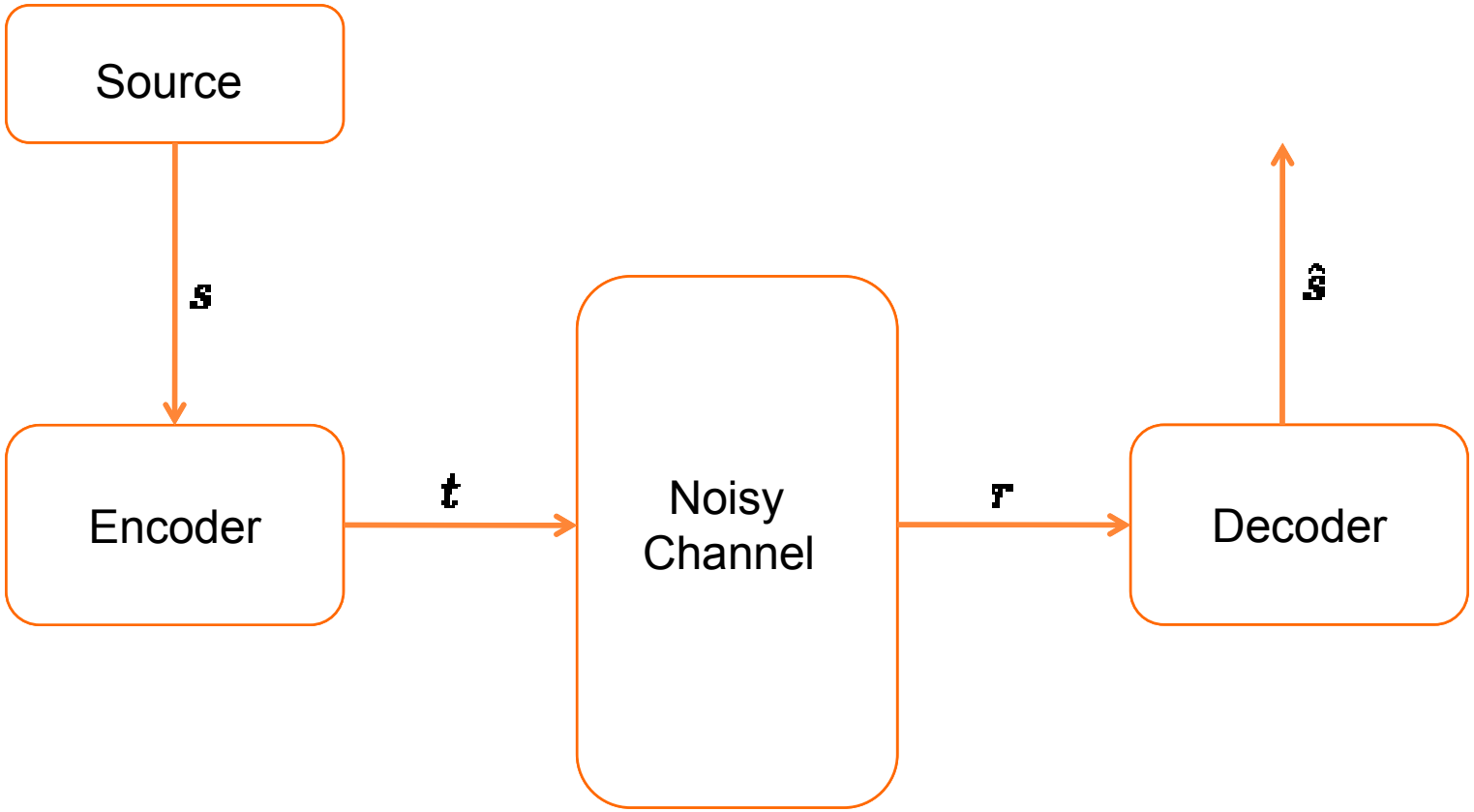
Lecture 6: Noisy Channel Coding

Course given by Pierre Senellart
Material by Stephanie Wehner, with additions by P. Senellart

Where do we go from here?



Sending information without error



Goal: Construct an encoder and decoder such that we can recover the information $\hat{s} = s$

What we'll do this time

- Noisy-channel coding
  - Conditional entropies
  - Mutual information
  - Capacity of a channel
  - Examples of noisy channels
  - Block error probability
- Statement of Shannon's channel coding theorem

Reminder: Entropies

- Shannon entropy

$$H(X) = -\sum_{x} \Pr_X(x) \log \Pr_X(x)$$

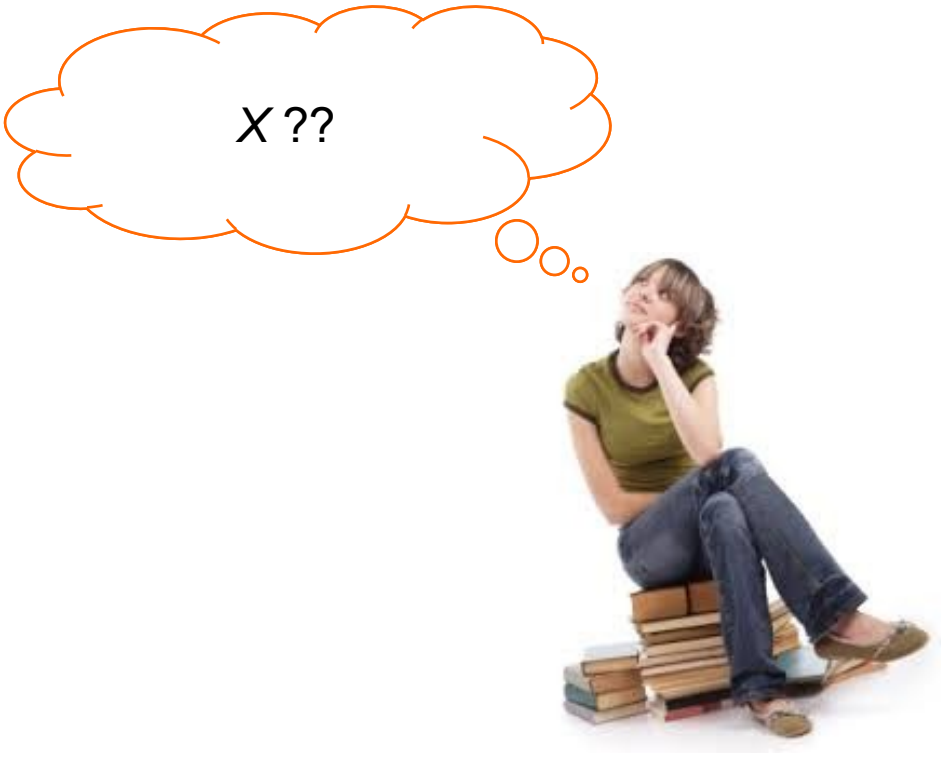
- Joint Shannon entropy

$$H(X,Y) = -\sum_{x,y} \Pr_{XY}(x,y) \log \Pr_{XY}(x,y)$$

- For a product distribution $\Pr_{XY}(x,y) = \Pr_X(x) \Pr_Y(y)$

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

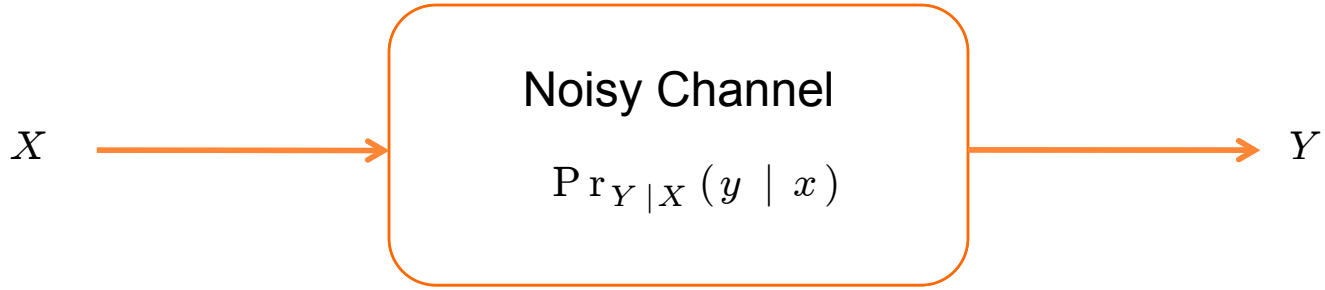
Conditional entropy $H(X|Y)$



Y

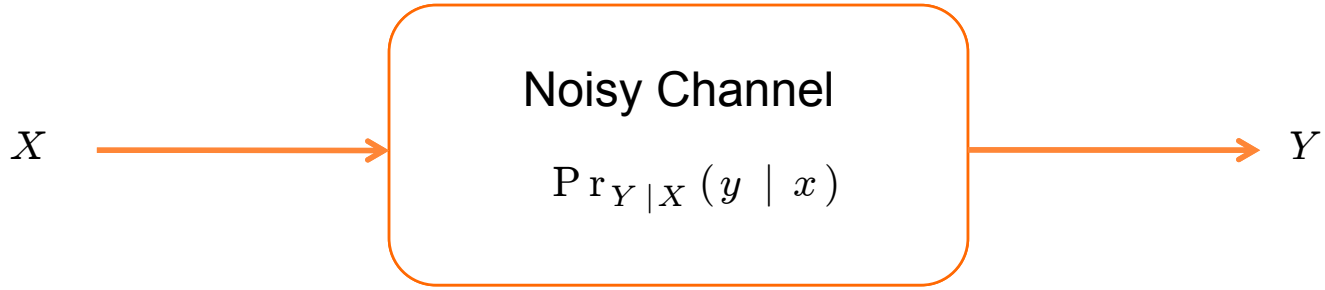
Motivation

- How can we quantify the entropy of X **given** Y?



Motivation

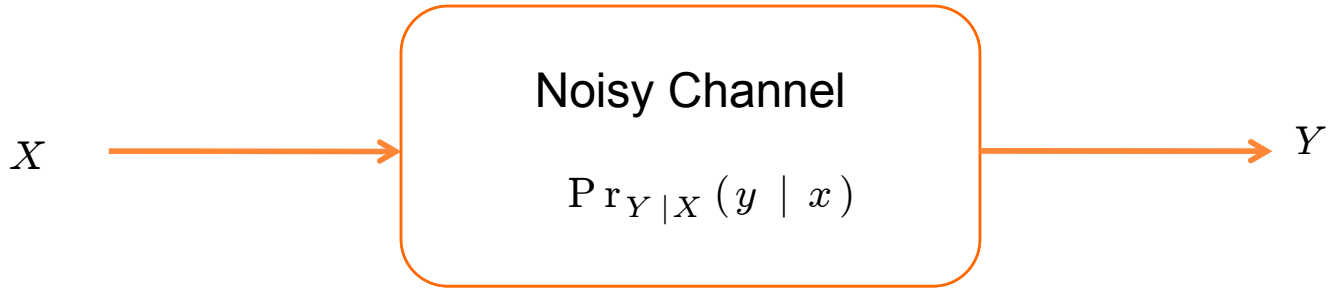
- How can we quantify the entropy of X **given** Y?



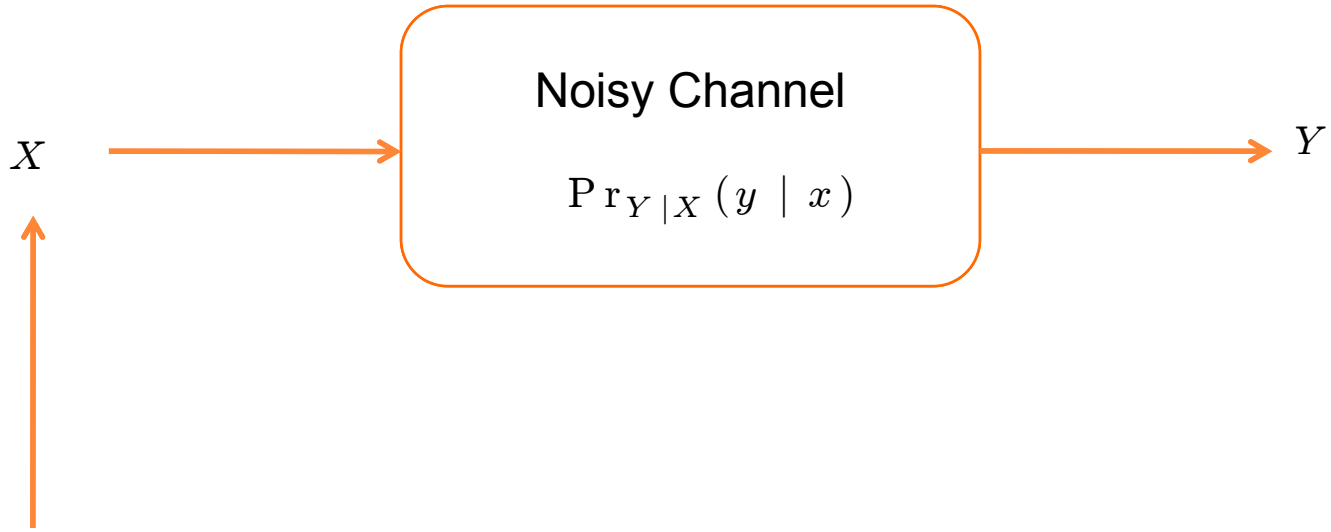
Example: Binary symmetric channel

$$\Pr_{Y|X}(y \mid x) = \begin{cases} 1-f & \text{if } y = x \in \{0,1\} \\ f & \text{otherwise} \end{cases}$$

Change of entropy



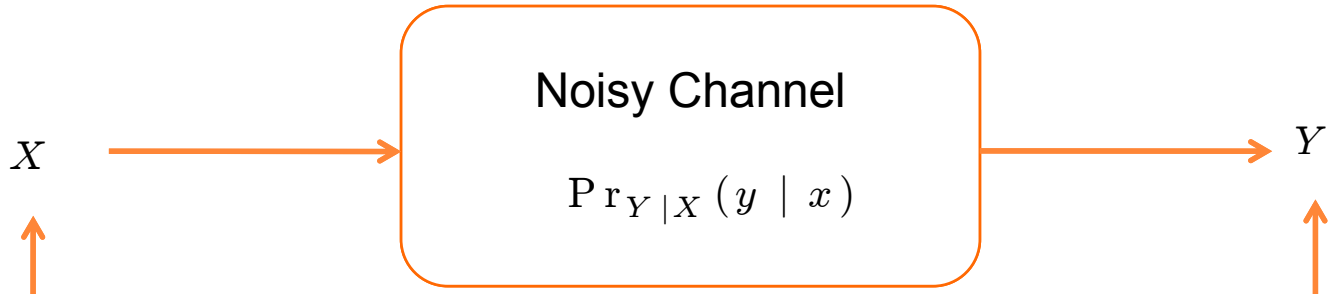
Change of entropy



Initial uncertainty
(without any observation!)
given by the entropy

$$H(X) = -\sum_{x} \Pr_X(x) \log \Pr_X(x)$$

Change of entropy



Initial uncertainty
(without any observation!)
given by the entropy

$$H(X) = -\sum_{x} \Pr_X(x) \log \Pr_X(x)$$

Now we observe *Y*
which could decrease
our uncertainty

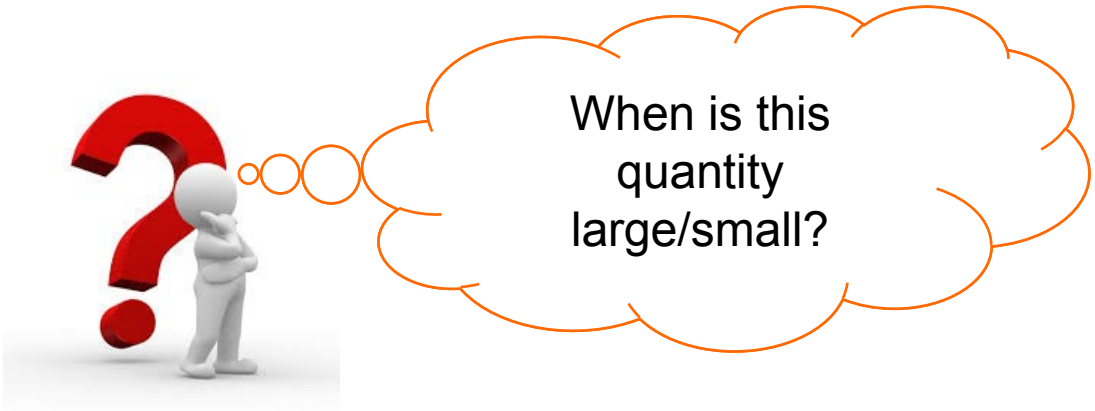
$$H(X \mid Y)$$

$$H(X \mid Y) = \sum_{y} \Pr_Y(y) H(X \mid Y = y)$$

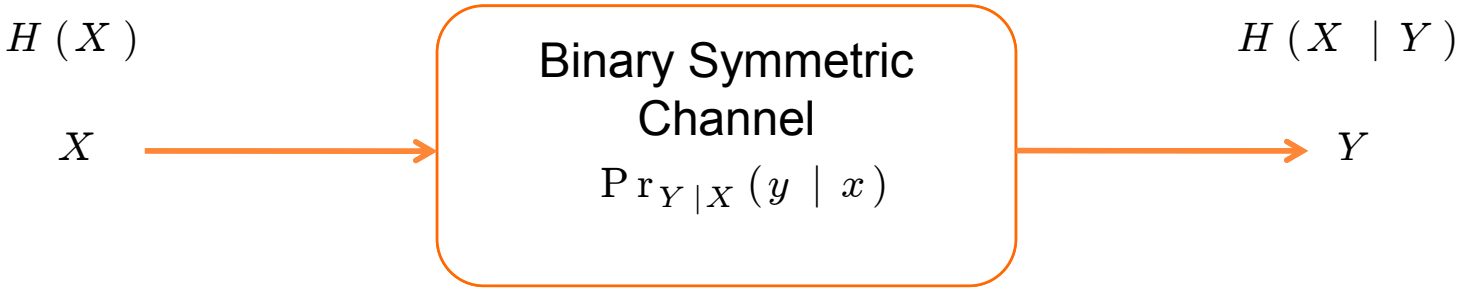
Conditional entropy

$$H(X \mid Y) = \sum_{y} \Pr_{Y}(y) H(X \mid Y = y)$$

$$H(X \mid Y = y) = -\sum_{x} \Pr_{X \mid Y}(x \mid y) \log \Pr_{X \mid Y}(x \mid y)$$



What do you expect?



$$\Pr_{Y|X}(y \mid x) = \begin{cases} 1 - f & \text{if } y = x \in \{0, 1\} \\ f & \text{otherwise} \end{cases}$$

Conditional entropy

- Entropy of X given Y

$$H(X \mid Y) = -\sum_{x,y} \Pr_Y(y) \Pr_{X|Y}(x \mid y) \log \Pr_{X|Y}(x \mid y) \geq 0$$

"Side information"

- Other examples
  - Compression: Y = C(X) is the compressed string
  - Cryptography: Y is some information that the adversary has gathered

Chain rule for information content

- We know that

$$\Pr_{X,Y}(x, y) = \Pr_{X|Y}(x \mid y) \Pr_Y(y) = \Pr_{Y|X}(y \mid x) \Pr_X(x)$$

- *Chain rule* for information content

$$\log \frac{1}{\Pr_{X,Y}(x, y)} = \log \frac{1}{\Pr_X(x)} + \log \frac{1}{\Pr_{Y|X}(y \mid x)}$$

A FEW WAYS TO WRITE THE CONDITIONAL ENTROPY: CHAIN RULE FOR ENTROPY

- Chain rule for information content

$$\log \frac{1}{\Pr_{X\,Y}(x, y)} = \log \frac{1}{\Pr_{X}(x)} + \log \frac{1}{\Pr_{Y\,|X}(y \mid x)}$$

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

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

A FEW WAYS TO WRITE THE CONDITIONAL ENTROPY: CHAIN RULE FOR ENTROPY

- Chain rule for information content

$$\log \frac{1}{\Pr_{X\,Y}(x, y)} = \log \frac{1}{\Pr_{X}(x)} + \log \frac{1}{\Pr_{Y\,|X}(y \mid x)}$$

Translated into entropies, this gives us the so-called *chain rule*

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

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

A FEW WAYS TO WRITE THE CONDITIONAL ENTROPY: CHAIN RULE FOR ENTROPY

- Chain rule for information content

$$\log \frac{1}{\Pr_{X,Y}(x,y)} = \log \frac{1}{\Pr_{X}(x)} + \log \frac{1}{\Pr_{Y|X}(y \mid x)}$$

Translated into entropies, this gives us the so-called *chain rule*

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

- *S*olving for $H(X \mid Y) = H(X,Y) - H(Y)$

A FEW WAYS TO WRITE THE CONDITIONAL ENTROPY: CHAIN RULE FOR ENTROPY

- Chain rule for information content

$$\log \frac{1}{\Pr_{X,Y}(x,y)} = \log \frac{1}{\Pr_{X}(x)} + \log \frac{1}{\Pr_{Y|X}(y \mid x)}$$

Translated into entropies, this gives us the so-called *chain rule*

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

- *S*olving for $H(X \mid Y) = H(X,Y) - H(Y)$

A FEW WAYS TO WRITE THE CONDITIONAL ENTROPY: CHAIN RULE FOR ENTROPY

- Chain rule for information content

$$\log \frac{1}{\Pr_{X,Y}(x,y)} = \log \frac{1}{\Pr_{X}(x)} + \log \frac{1}{\Pr_{Y|X}(y \mid x)}$$

Translated into entropies, this gives us the so-called *chain rule*

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

- *S*

olving for $H(X \mid Y) = H(X,Y) - H(Y)$

Intuitively, total entropy of *X* and *Y*, minus what we know, namely *Y,* gives *X*

How are they related?

- Would think that the entropy, i.e., uncertainty becomes less if we have Y then if we had nothing….
- Conditioning reduces entropy

$$H(X \mid Y) \cdot \quad H(X)$$

How small or large is this quantity?

- Conditional entropy $H(X \mid Y) = H(X, Y) - H(Y)$

How small or large is this quantity?

- Conditional entropy $H(X \mid Y) = H(X, Y) - H(Y)$
- Upper bound $H(X \mid Y) \cdot H(X) \cdot \log |X|$

  When is this attained?

How small or large is this quantity?

- Conditional entropy $H(X \mid Y) = H(X, Y) - H(Y)$
- Upper bound $H(X \mid Y) \cdot H(X) \cdot \log |X|$

When is this attained?
Y says nothing about X: product distribution

$$\Pr_{X Y}(x, y) = \Pr_X(x) \Pr_Y(y)$$

How small or large is this quantity?

- Conditional entropy $H(X \mid Y) = H(X, Y) - H(Y)$

- Upper bound $H(X \mid Y) \cdot H(X) \cdot \log |X|$

  When is this attained?
  Y says nothing about X: product distribution

  $$\Pr_{XY}(x, y) = \Pr_X(x) \Pr_Y(y)$$

Lower bound $0 \cdot H(X \mid Y)$

When is this attained?

How small or large is this quantity?

- Conditional entropy $H(X \mid Y) = H(X, Y) - H(Y)$

- Upper bound $H(X \mid Y) \cdot H(X) \cdot \log |X|$

  When is this attained?
  Y says nothing about X: product distribution

  $$\Pr_{XY}(x, y) = \Pr_X(x) \Pr_Y(y)$$

Lower bound $0 \cdot H(X \mid Y)$

When is this attained?
Y says everything about *X*

A WORD OF CAUTION

- Conditioning only reduces entropy *on average!*

$$H(X \mid Y) \cdot \quad H(X)$$

A WORD OF CAUTION

- Conditioning only reduces entropy *on average!*

$$H(X \mid Y) \cdot \quad H(X)$$

But it is not always the case that for all y

$$H(X \mid Y = y) = -\sum_{x} \Pr_{X|Y}(x \mid y) \log \Pr_{X|Y}(x \mid y) \not\cdot \quad H(X)$$

Example

- Conditioning reduces entropy but only on average

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

$H(X) = 7/4 = 14/8$ $\quad H(X \mid Y) = 11/8$

Example

- Conditioning reduces entropy but only on average

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

$H(X) = 7/4 = 14/8 \qquad H(X \mid Y) = 11/8$

$H(X \mid Y) < H(X)$

EXAMPLE

- Conditioning reduces entropy but only on average

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

$< H(X)$

$H(X) = 7/4 = 14/8 \qquad H(X \mid Y) = 11/8$

$H(X \mid Y) \cdot H(X)$

EXAMPLE

- Conditioning reduces entropy but only on average

$P(x \mid y)$		x				$H(X \mid y)$/bits	
		1	2	3	4		
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$	
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$	
	3	$1/4$	$1/4$	$1/4$	$1/4$	2	$> H(X)$
	4	1	0	0	0	0	$< H(X)$

$H(X) = 7/4 = 14/8 \qquad H(X \mid Y) = 11/8$

$H(X \mid Y) \cdot H(X)$

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

$H(X)$

Y

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

$H(X)$

$H(X|Y)$

Y

Mutual information

- Measures the reduction of uncertainty about *X* after we learn *Y*, i.e., the average amount of information that *Y* gives us about *X*

$$I(X;Y) = H(X) - H(X \mid Y)$$

Mutual information

- Measures the reduction of uncertainty about *X* after we learn *Y*, i.e., the average amount of information that *Y* gives us about *X*

$$I(X;Y) = H(X) - H(X \mid Y)$$

Is symmetric

$$I(X;Y) = H(Y) - H(Y \mid X)$$

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

Mutual information

- Measures the reduction of uncertainty about *X* after we learn *Y*, i.e., the average amount of information that *Y* gives us about *X*

$$I(X;Y) = H(X) - H(X \mid Y)$$

Is symmetric

$$I(X;Y) = H(Y) - H(Y \mid X)$$

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

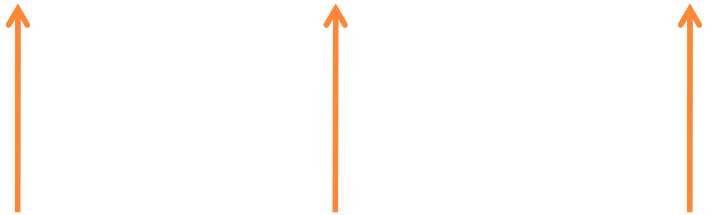
Conditioning reduces entropy implies that

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

Conditional mutual information

- What if we are already given some extra information, Z?

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$



Simply add the side information Z to all conditionings

How large or small can this get?

- Mutual information

$$I(X;Y) = H(X) - H(X \mid Y)$$

How large or small can this get?

- Mutual information

$$I(X;Y) = H(X) - H(X \mid Y)$$

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

When is this attained?

Product distribution $\Pr_{XY}(x,y) = \Pr_X(x)\Pr_Y(y)$

How large or small can this get?

- Mutual information

$$I(X;Y) = H(X) - H(X \mid Y)$$

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

When is this attained?

Product distribution $\Pr_{XY}(x,y) = \Pr_X(x) \Pr_Y(y)$

Upper bound $I(X;Y) \cdot H(X) \cdot \log |X|$

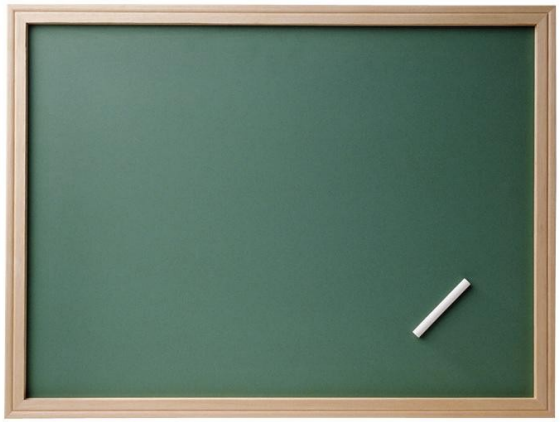
- When is this attained?
- Maximally far from product, *Y* tells us everything about *X*

Chain rule for the mutual information

$$
\begin{aligned}
I(X;Y\,Z) &= I(X;Y) + I(X;Z \mid Y) \\
&= I(X;Z) + I(X;Y \mid Z)
\end{aligned}
$$

Chain rule for the mutual information

$$
\begin{aligned}
I(X;Y\,Z) &= I(X;Y) + I(X;Z \mid Y) \\
&= I(X;Z) + I(X;Y \mid Z)
\end{aligned}
$$



Inbetween extremes

- A measure how far we are from the product distribution

- Can write the mutual information in terms of the relative entropy

$$I(X;Y) = D_{\mathrm{KL}}(\Pr_{XY} \| \Pr_X \Pr_Y)$$

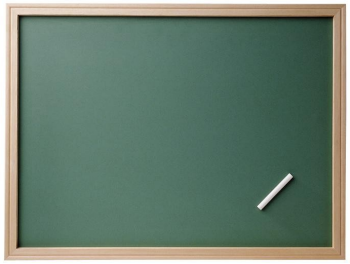
Intuively, measures how far we are from the product distribution

Inbetween extremes

- A measure how far we are from the product distribution

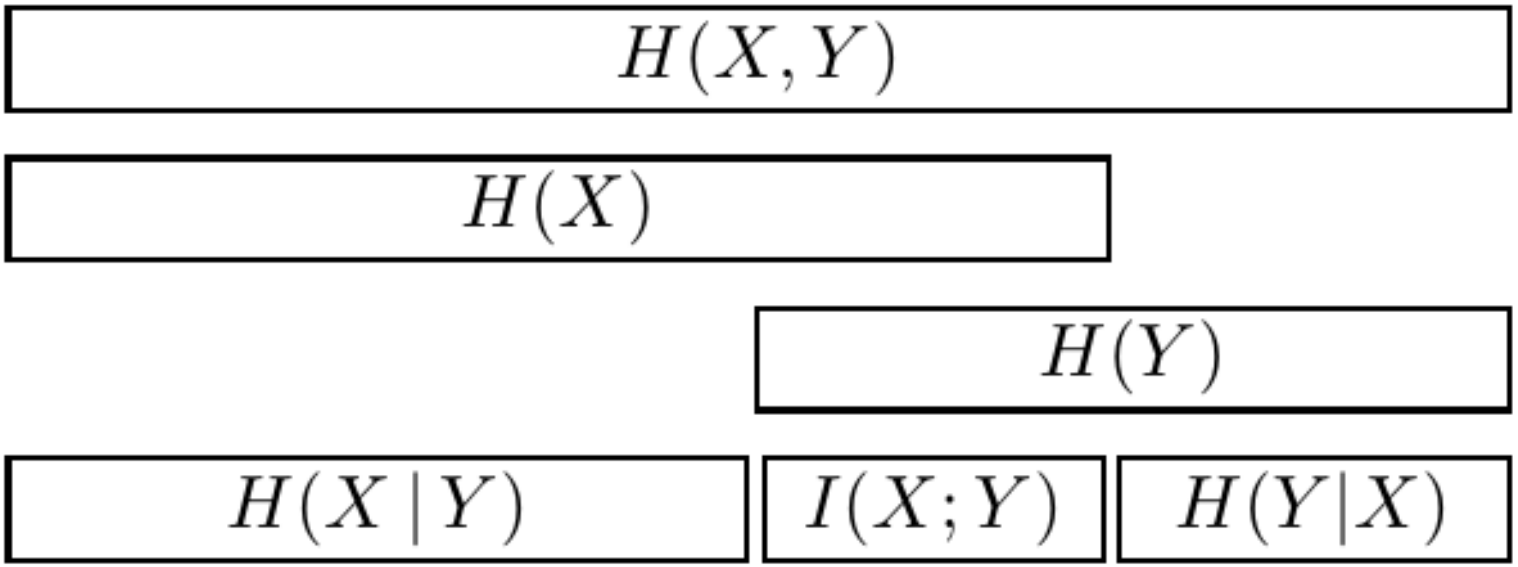
- Can write the mutual information in terms of the relative entropy

$$I(X;Y) = D_{\mathrm{KL}}(\Pr_{XY} \| \Pr_X \Pr_Y)$$



Intuively, measures how far we are from the product distribution

Relation among entropies



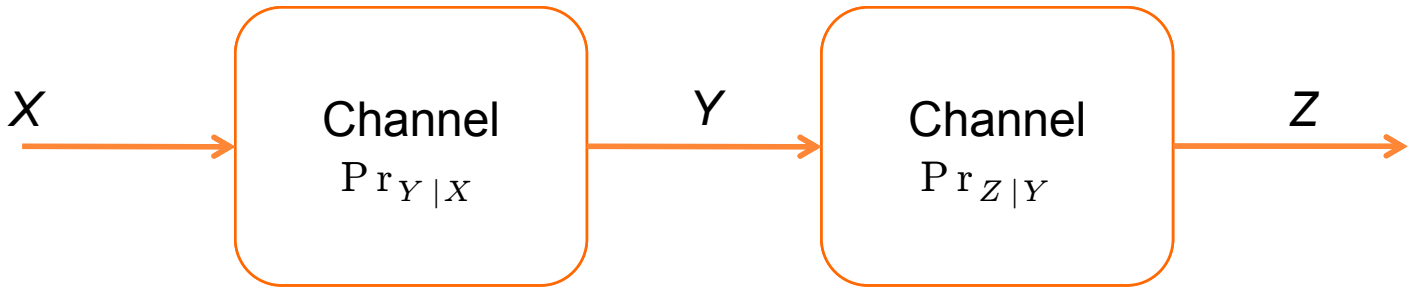
Markov Chain

- Random variables form a Markov Chain

$$X \rightarrow Y \rightarrow Z$$

if Z only depends on Y (but is conditionally independent of X)

$$\Pr_{X\,Y\,Z}(x, y, z) = \Pr_X(x)\,\Pr_{Y|X}(y \mid x)\,\Pr_{Z|Y}(z \mid y)$$



Properties of a Markov Chain

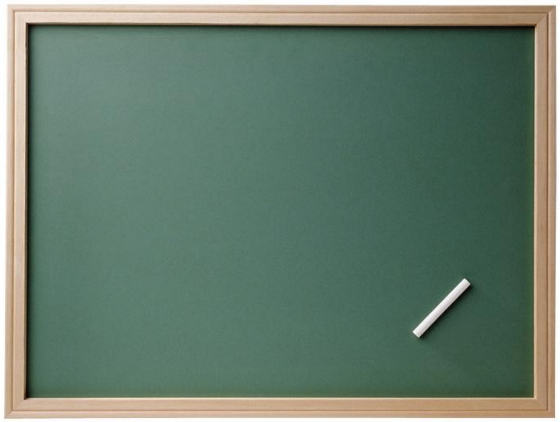
- X and Z are conditionally independent

$$\Pr_{X\,Z\,|Y}(x, z \mid y) = \Pr_{X\,|Y}(x \mid y)\Pr_{Z\,|Y}(z \mid y)$$

Properties of a Markov Chain

- X and Z are conditionally independent

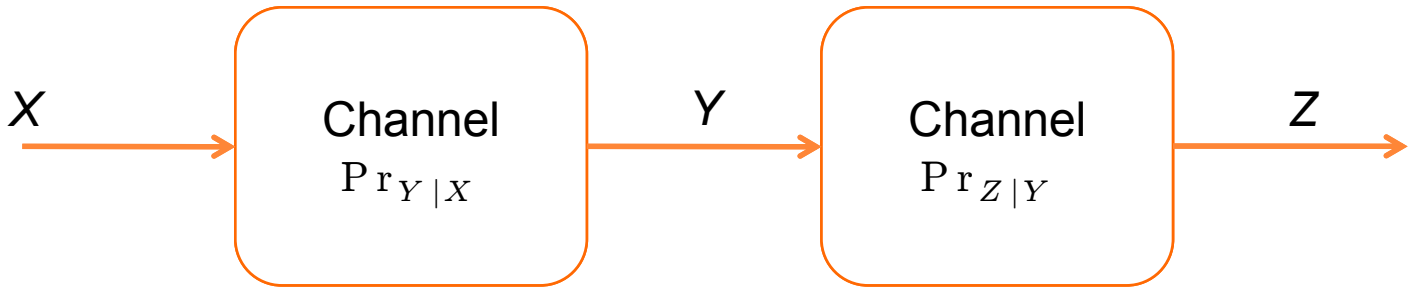
$$\Pr_{X\,Z\,|Y}(x, z \mid y) = \Pr_{X\,|Y}(x \mid y)\,\Pr_{Z\,|Y}(z \mid y)$$



Properties of a Markov chain

Symmetric: If $X \to Y \to Z$ then also $Z \to Y \to X$

If $Z = f(Y)$ then $X \to Y \to Z$



Data processing

- Suppose we have a Markov chain $X \to Y \to Z$
- What can we say about $I(X;Y)$ vs. $I(X;Z)$?

Data processing

- Suppose we have a Markov chain $X \rightarrow Y \rightarrow Z$
- What can we say about *I*(*X*;*Y*) vs. *I*(*X*;*Z*)?

Data processing

If *X*, *Y*, and *Z* form a Markov Chain, then

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

Data processing

- Suppose we have a Markov chain $X \rightarrow Y \rightarrow Z$
- What can we say about *I*(*X*;*Y*) vs. *I*(*X*;*Z*)?

Data processing

If *X*, *Y*, and *Z* form a Markov Chain, then

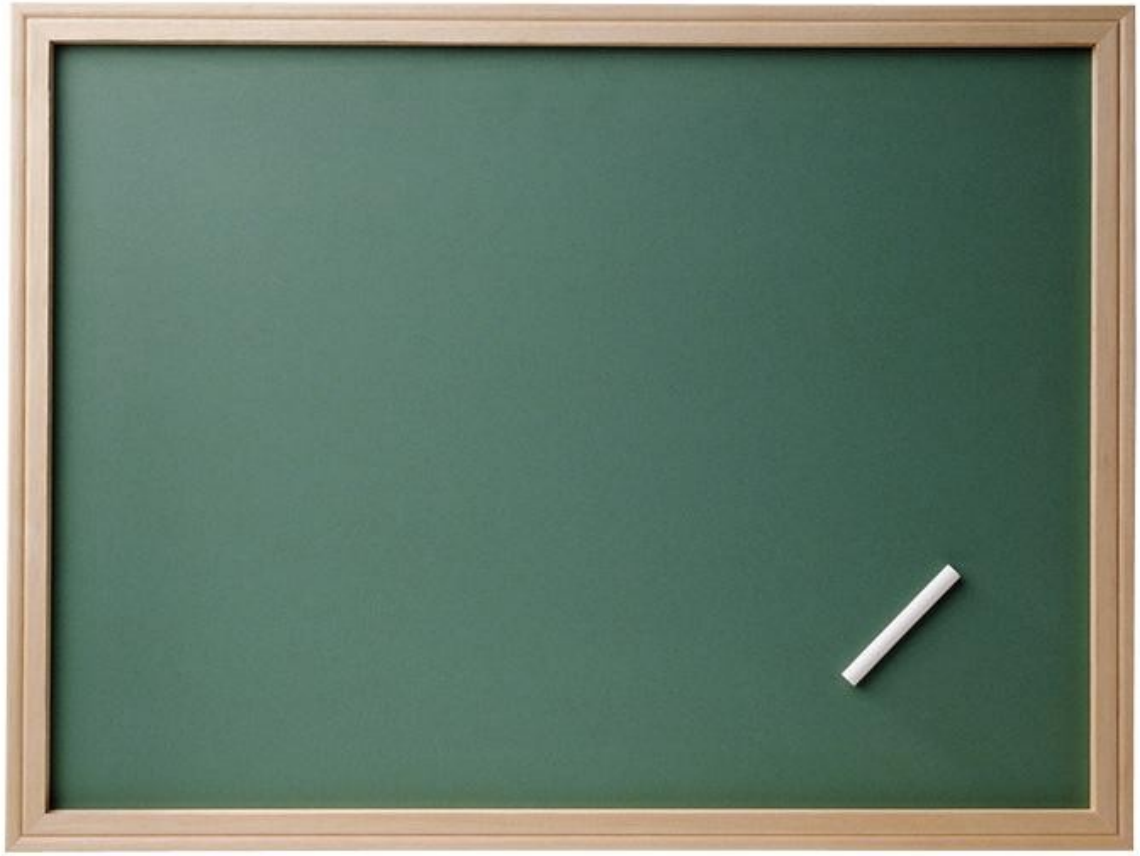
$$I(X;Y) \geq I(X;Z)$$

No processing of Y can improve our knowledge about X !

Proof

If X, Y and Z form a Markov Chain, then

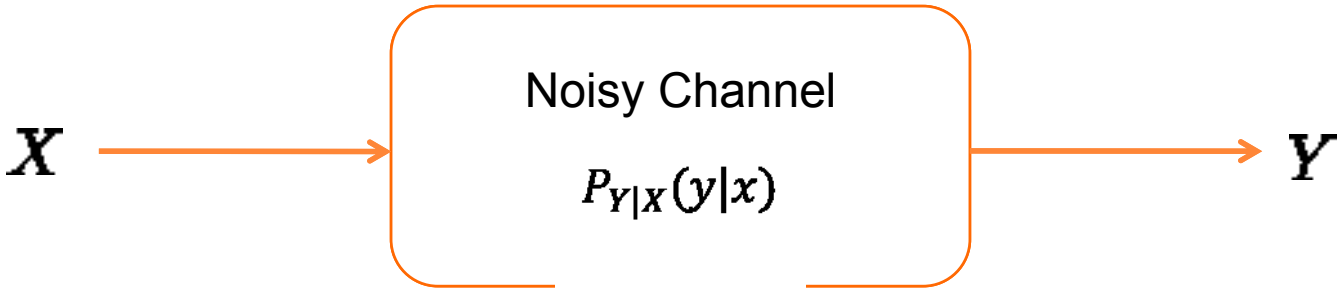
$$I(X:Y) \geq I(X:Z)$$



Summary

- Conditional entropy
  - Measures the uncertainty about *X* given *Y*
- Mutual information
  - Measures how much knowing *Y* reduces the uncertainty about *X* (on average)

How much information does the output give about the input?

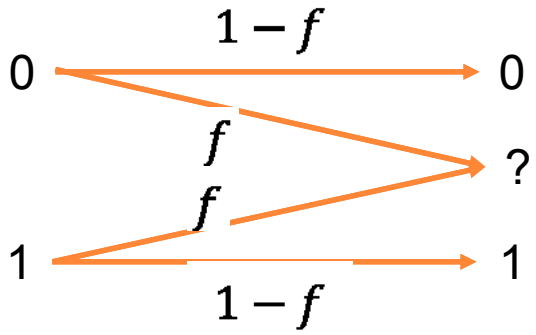


- Intuition: mutual information measures how much our uncertainty about *X* is reduced by learning *Y*

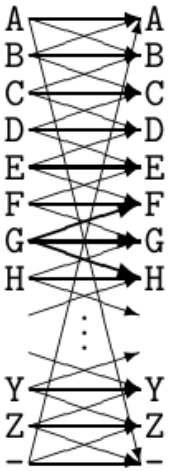
$$I(X;Y) = H(X) - H(X|Y)$$

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

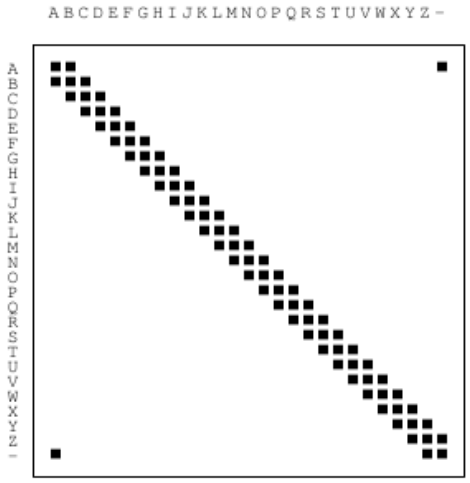
Example of channels: binary erasure channel



Example of channels: noisy-typewriter

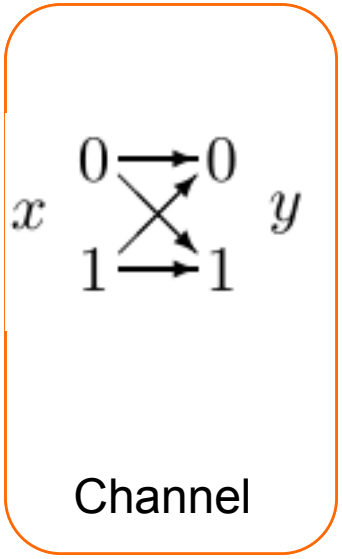


$$
\begin{aligned}
&\vdots \\
P(y=\mathtt{F}\,|\,x=\mathtt{G}) &= 1/3; \\
P(y=\mathtt{G}\,|\,x=\mathtt{G}) &= 1/3; \\
P(y=\mathtt{H}\,|\,x=\mathtt{G}) &= 1/3; \\
&\vdots
\end{aligned}
$$



Examples of channels: BSC

- Binary symmetric channel



$$\Pr_{Y|X}(y \mid x) = \begin{cases} 1-f & \text{if } y = x \in \{0,1\} \\ f & \text{otherwise} \end{cases}$$

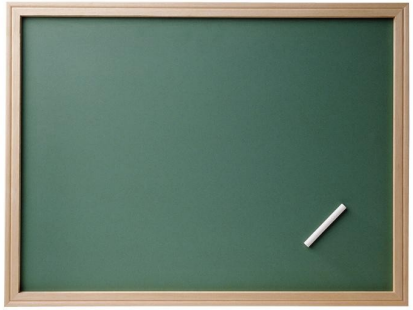
Example

$$f = 0.15$$

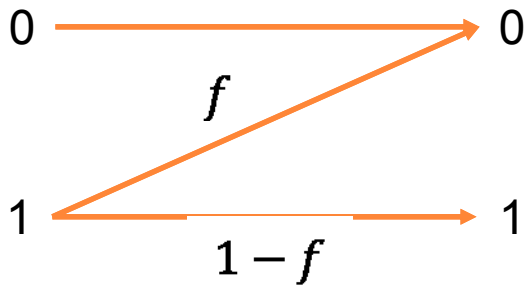
$$\Pr_X(0) = 0.9, \Pr_X(1) = 0.1$$

- Mutual information for

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



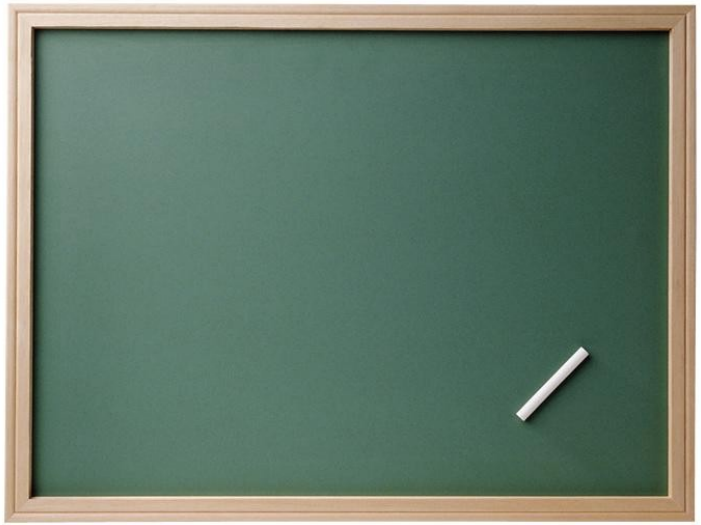
Example of channels: Z-Channel



$$f = 0.15$$

$$\Pr_X(0) = 0.9, \Pr_X(1) = 0.1$$

Mutual information?



Intuition

- How can we ensure we learn as much as possible from the output over the input?
- Remember: so far we don't "encode" we simply choose inputs *X* with some probability
- Mutual information measures how much the output reduces the uncertainty about the input

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

- Idea: Want to maximize the mutual information!

Capacity of a channel

- We define the capacity of a channel $\mathcal{N}$ by:

$$C(\mathcal{N}) = \max_{\Pr_X} I(X;Y)$$

- Determines how well we can send information.

For the binary symmetric channel

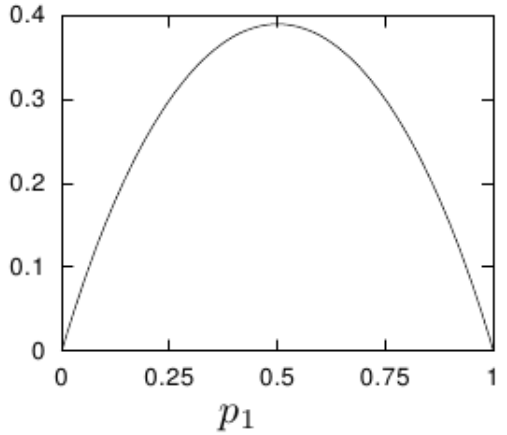
- What's $C(\mathcal{N}) = \max_{P_X} I(X;Y)$?
- Remember our example
- More generally

$$I(X;Y) = H_2((1-f)p_1 + f(1-p_1)) - H_2(f)$$

$I(X;Y)$

- Maximum at $p_1 = 1/2$
- Capacity

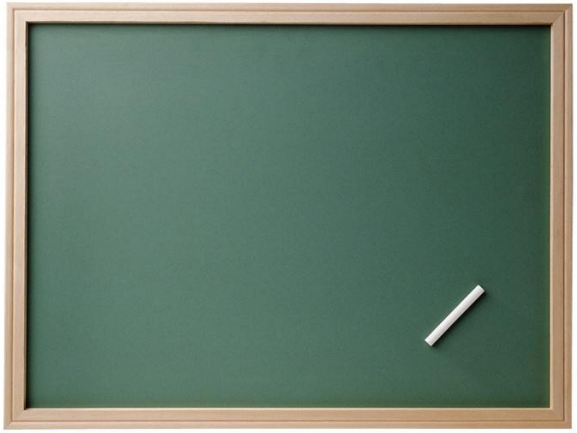
$$C(BSC) = 1 - H_2(f)$$



How about the Z-Channel?

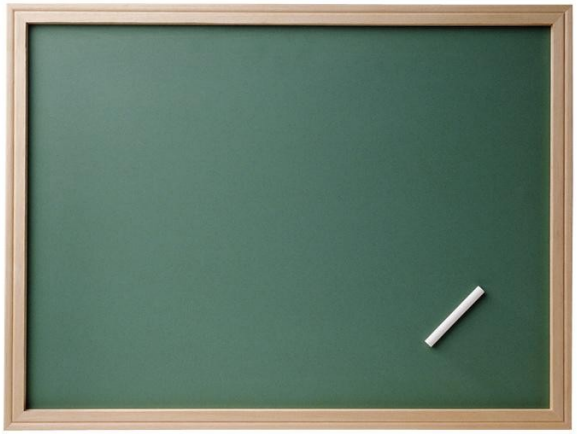
How about the Z-Channel?

- Computing the mutual information



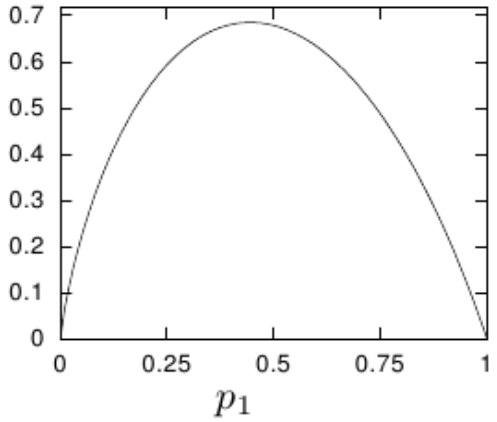
How about the Z-Channel?

- Computing the mutual information



For our example f = 0.15

$I(X;Y)$



How about the noisy-typewriter?

Terminology: block codes

- A (N,K) block code
  - Source symbols $s \in \{1, \ldots, 2^K\}$ ie K bits!
  - Codewords $x^{(s)}$ of length N bits!

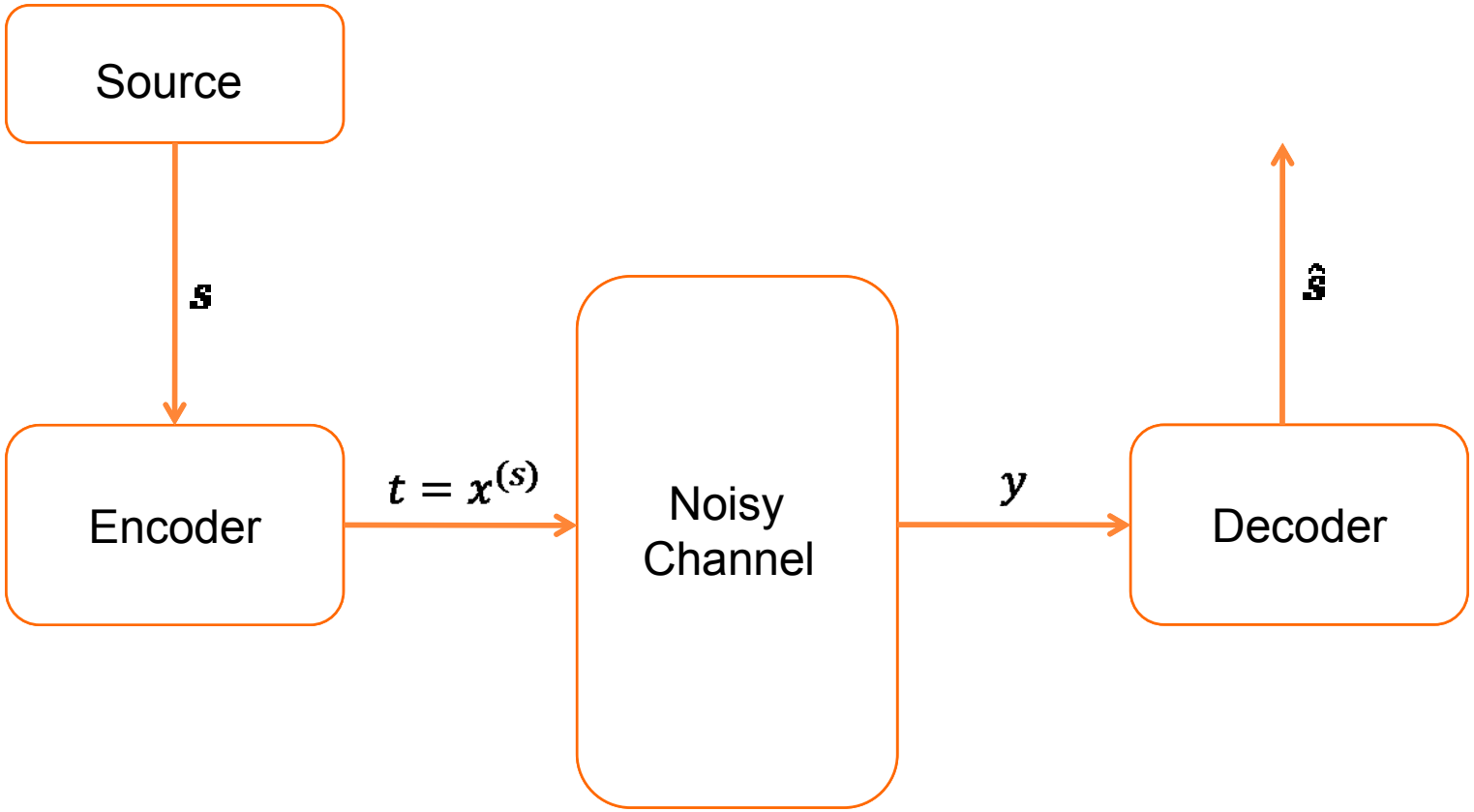
- List of possible codewords $\{x^{(1)}, \ldots, x^{(2^K)}\}$

- Can you name an example of block code? ☺

Terminology: rate of the code

- For an (N,K) block code has rate $R = \frac{K}{N}$
- What was the rate of the (7,4) Hamming code? ☺
- Rate measures the price of transmission: to send K bits of source signal we need to send $N = \frac{K}{R}$ bits
  - Lower rate means higher price
  - Higher rate means lower price

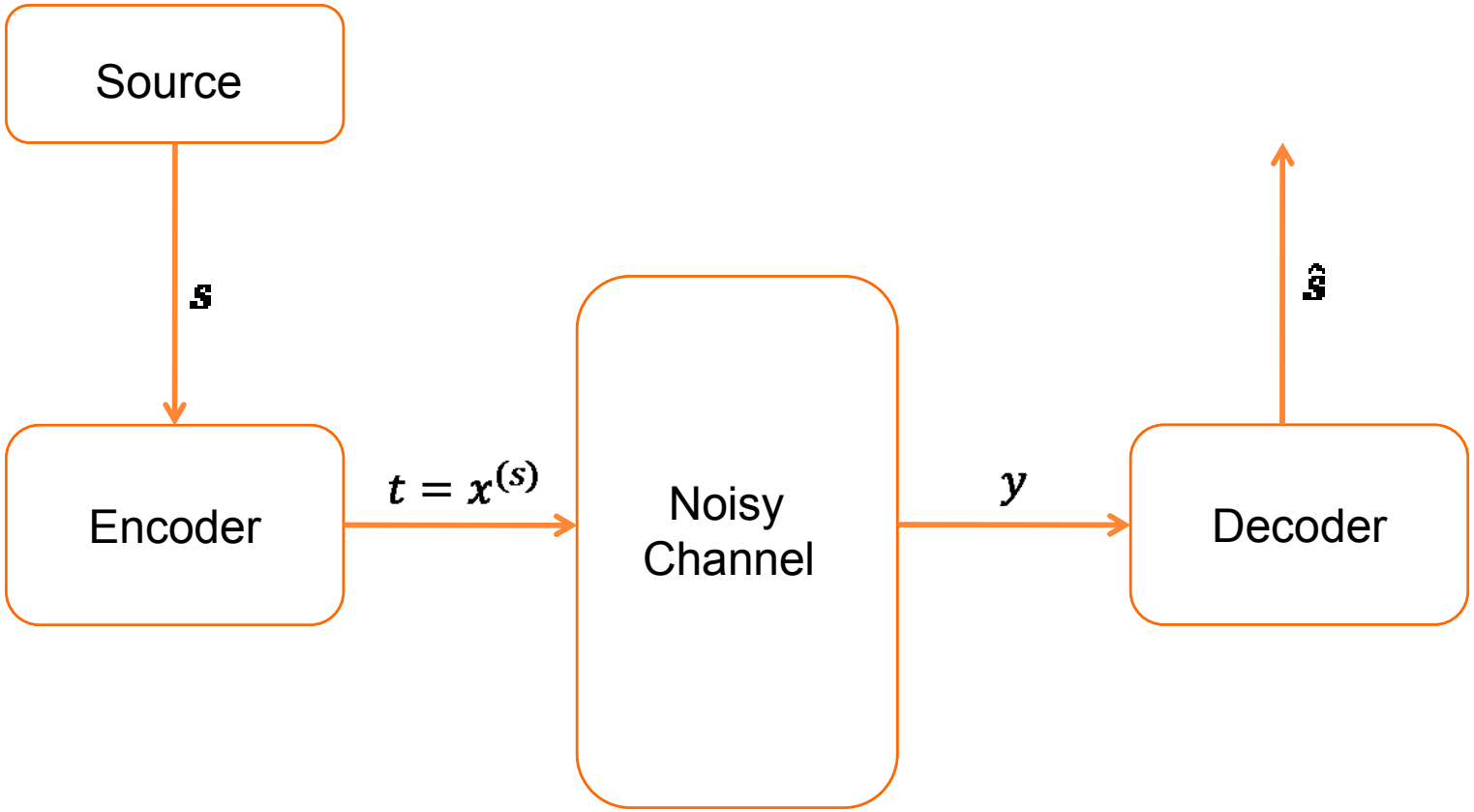
Terminology: encoder



Encoder: mapping from source symbols to channel inputs

Caution: channel inputs *x* are not the same as source symbols!

TERMINOLOGY : DECODER



Decoder: mapping from channel outputs to possible source symbols

Terminology: optimal decoder

- Most likely input symbol: $\hat{s} = \arg\max_s \Pr(s \mid y)$

Terminology: probability of block error

- Average probability of observing a wrong block

$$P_{\mathrm{Block}} = \sum_{s} \Pr(s) \Pr(\hat{s} \neq s \mid s)$$

- A block error occurs if a single input bit is decoded incorrectly!
- Do you remember what is a bit error in comparison?

Terminology: maximum probability of block error

- Worst case error

$$P_{\max} = \max_{s} \Pr(\hat{s} \neq s \mid s)$$

Shannon's noisy channel coding theorem (Part 1: Achievability)

- Associated with every discrete memoryless (IID) channel, there is a non-negative number C (the capacity) such that
  - For any error $\epsilon > 0$ and $R < C$ for large enough N, there exists a block code of length N and rate $\geq R$, and a decoding algorithm, such that the maximum probability of block error is $< \epsilon$
  - If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

$$R(p_b) = \frac{C}{1 - H_2(p_b)}$$

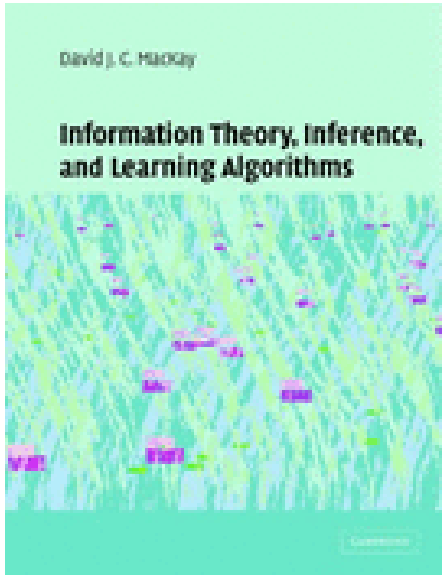
  - For any p_b, rates greater than $R(p_b)$ are not achievable

Next time

- Explanation and proof of noisy channel coding theorem!

Reading for this lecture

- Chapters 8, 9.1 to 9.5 in the book



Information Theory, Inference and Learning Algorithms
by David J. C. MacKay
Cambridge University Press, 2003

- Homework due by Monday 2pm two weeks from now