Entropy properties

A.J. Han Vinck
University of Duisburg-Essen
April 2013

content

- Introduction
  - Entropy and some related properties
- Source coding
- Channel coding

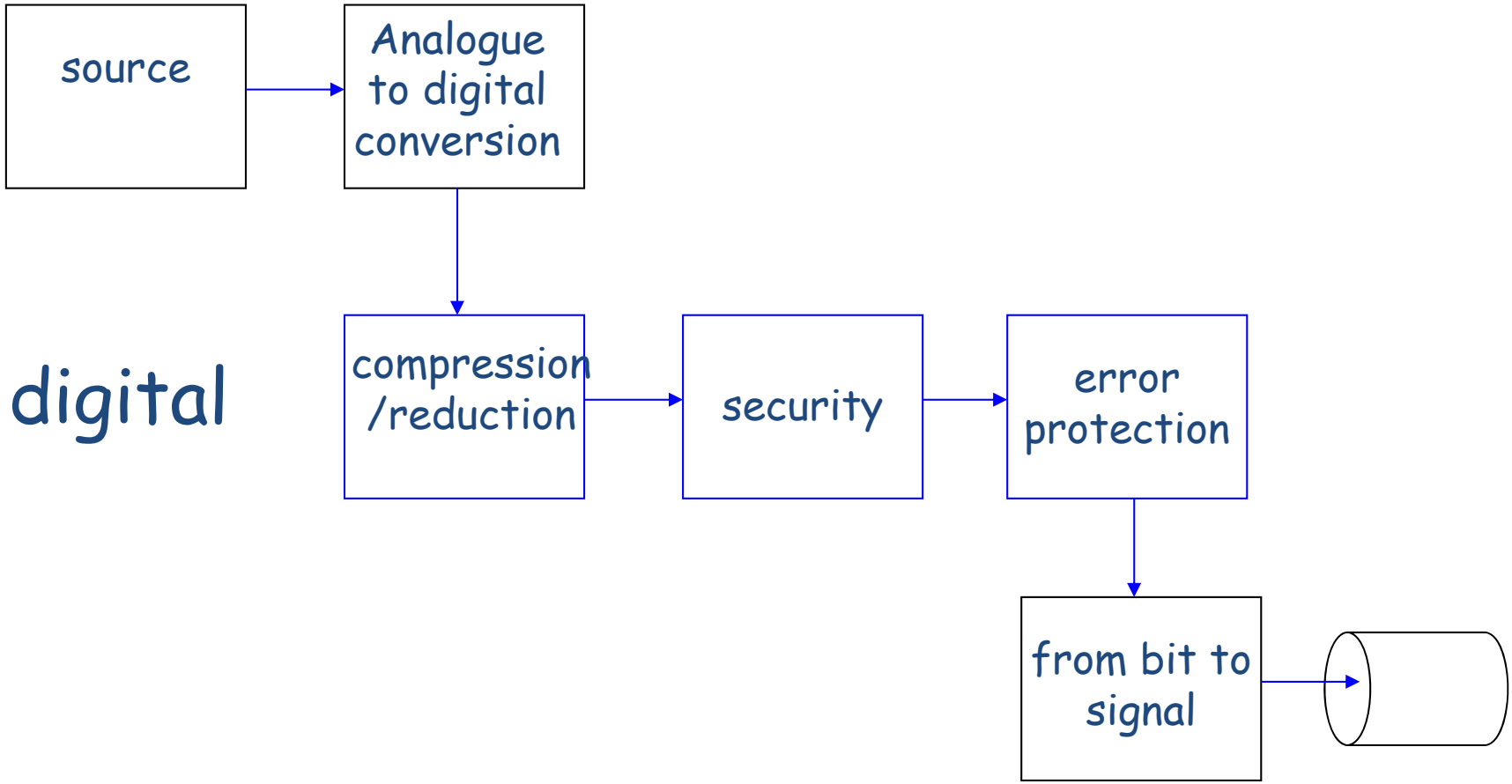
Field of Interest

Information theory deals with the problem of efficient and reliable transmission of information

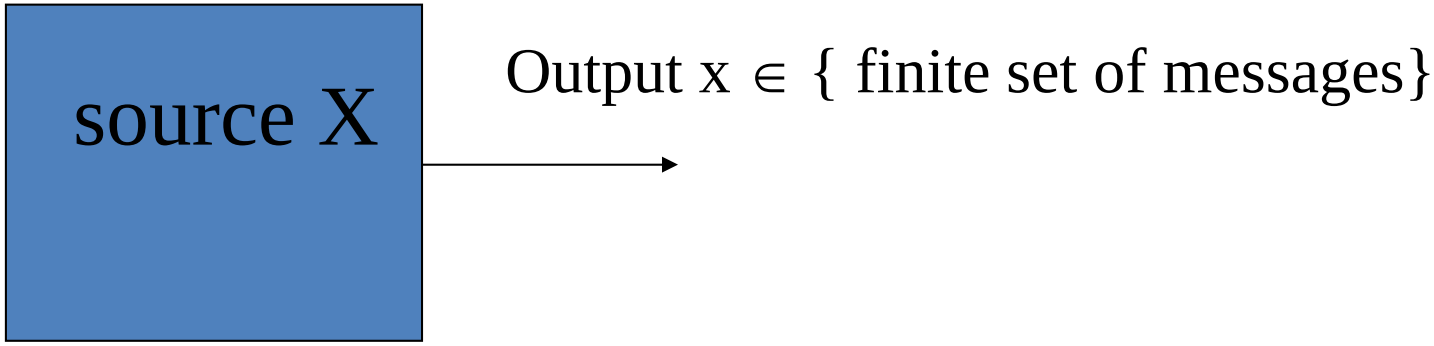
It specifically encompasses theoretical and applied aspects of

- *coding, communications and communications networks*
- *complexity and cryptography*
- *detection and estimation*
- *learning, Shannon theory, and stochastic processes*

Communication model



A generator of messages: the discrete source



Example:

binary source: $x \in \{ 0, 1 \}$ with $P(x = 0) = p$; $P(x = 1) = 1 - p$

M-ary source: $x \in \{1,2, \bullet\bullet\bullet, M\}$ with $\Sigma P_i = 1$.

Express everything in bits: 0 and 1

Discrete finite ensemble:

a,b,c,d $\Rightarrow$ 00, 01, 10, 11

k binary digits specify 2^k messages

M messages need $\lceil \log_2 M \rceil$ bits

The entropy of a source

a fundamental quantity in Information theory

entropy

The ***minimum*** average number of binary digits needed

to

specify a source output (message) uniquely is called

"SOURCE ENTROPY"

SHANNON (1948):

1) Source entropy:= $-\sum_{i=1}^{M} P(i) \log_2 P(i) \leq \sum_{i=1}^{M} P(i)\ell(i)$ = L

2) minimum can be obtained !

QUESTION: how to represent a source output in digital form?

QUESTION: what is the source entropy of text, music, pictures?

QUESTION: are there algorithms that achieve this entropy?

http://www.youtube.com/watch?v=z7bVw7lMtUg

Properties of entropy

A: For a source X with M different outputs: $\log_2 M \geq H(X) \geq 0$

the „worst" we can do is

just assign $\log_2 M$ bits to each source output

B: For a source X „related" to a source Y: $H(X) \geq H(X|Y)$

Y gives additional info about X

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

Joint Entropy: $H(X,Y) = H(X) + H(Y|X)$

- also $H(X,Y) = H(Y) + H(X|Y)$
  - intuition: first describe Y and then X given Y
  - from this: $H(X) - H(X|Y) = H(Y) - H(Y|X)$

- **Homework:** check the formel

Cont.

- As a formel:

$$H(X,Y) = -\sum_{X,Y} P(x,y)\log_2 P(x,y) = -\sum_{X,Y} P(x,y)\log_2 P(x)P(y\mid x)$$

$$= -\sum_{X}\sum_{Y} P(x)P(y\mid x)\log_2 P(x) - \sum_{X,Y} P(x,y)\log_2 P(y\mid x) = H(X) + H(Y\mid X)$$

$$H(X\mid Y) = -\sum_{X,Y} P(x,y)\log_2 P(x\mid y)$$

Entropy: Proof of A

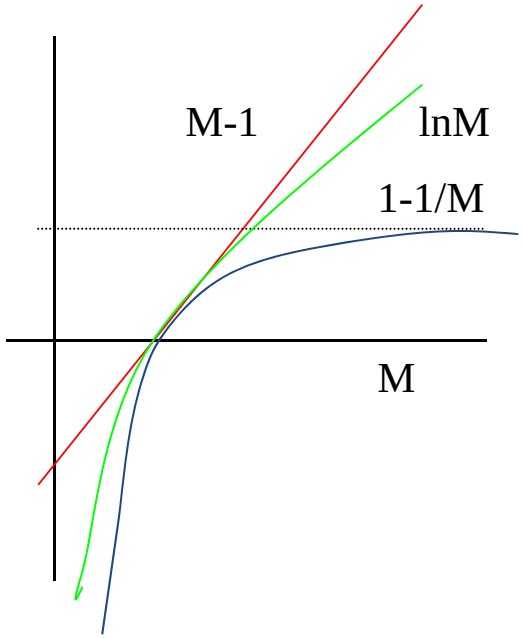
We use the following important inequalities

$\log_2 M = \ln M \log_2 e$

$\ln x = y \Rightarrow x = e^y$

$\log_2 x = y \log_2 e = \ln x \log_2 e$

$$1 - \frac{1}{M} \leq \ell n M \leq M - 1$$



Homework: draw the inequality

Entropy: Proof of A

$$H(X) - \log_2 M = \log_2 e \sum_x P(x) \ln \frac{1}{MP(x)}$$

$$\leq \log_2 e \sum_x P(x) \left(\frac{1}{MP(x)} - 1 \right)$$

$$\leq \log_2 e \left(\sum_x \frac{1}{M} - 1 \right) = 0$$

Entropy: Proof of B

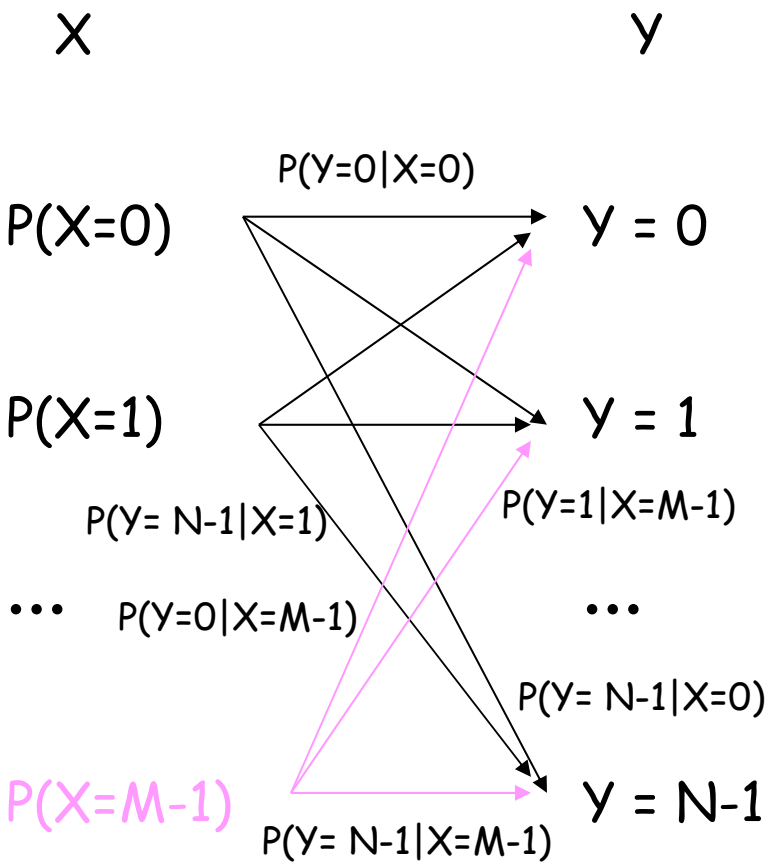
$$H(X) - H(X \mid Y) =$$

$$= -\log_2 e \sum_{x,y} P(x,y) \ln \frac{P(x)}{P(x \mid y)}$$

$$\geq -\log_2 e \sum_{x,y} P(x,y) \left(\frac{P(x)}{P(x \mid y)} - 1 \right)$$

$$\geq 0$$

The connection between X and Y



$$P(X=i, Y=j) = P(X=i)P(Y=j \mid X=i)$$
$$= P(Y=j)P(X=i \mid Y=j)$$

(Bayes)

$$P(Y=j) = \sum_{X=0}^{X=M-1} P(X=i)P(Y=j \mid X=i)$$

$$= \sum_{X=0}^{X=M-1} P(X=i, Y=j)$$

Entropy: corrolary

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

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

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

$$\leq H(X) + H(Y) + H(Z)$$

Binary entropy

$$\lim_{n\to\infty}\frac{1}{n}\log_2\binom{n}{pn}=h(p) \quad \Rightarrow \quad \binom{n}{pn}\to 2^{nh(p)}$$

interpretation:

let a **binary** sequence contain pn ones, then we can specify each sequence with

$\log_2 2^{nh(p)} = n\,h(p)$ bits

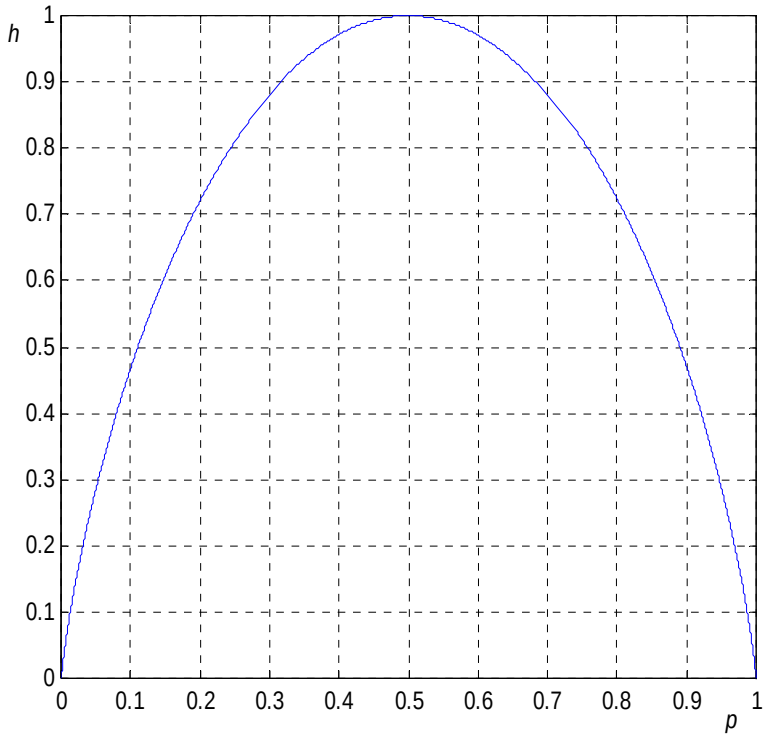
Homework: Prove the approximation using $\ln N! \sim N \ln N$ for N large.

Use also $\log_a x = y \rightarrow \log_b x = y \log_b a$

The Stirling approximation → $N! \to \sqrt{2\pi N}\, N^N e^{-N}$

The Binary Entropy:

$$h(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

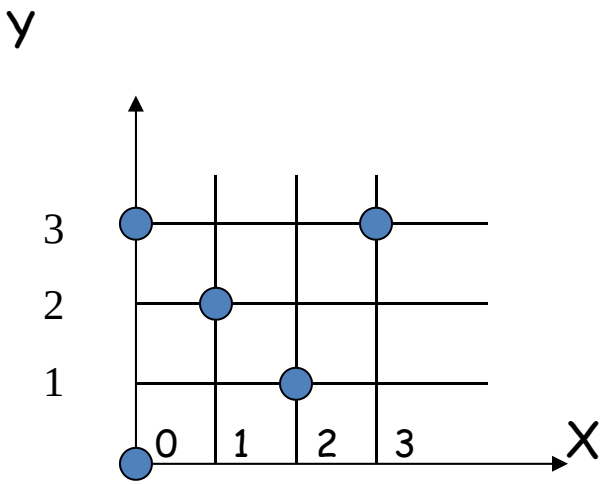


Note:

$h(p) = h(1-p)$

homework

- Consider the following figure



All points ● are equally likely. Calculate H(X), H(X|Y) and H(X,Y)

mutual information I(X;Y):=

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

$$= H(Y) - H(Y|X)$$ (homework: show this!)

i.e. the **reduction** in the description length of X given Y

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

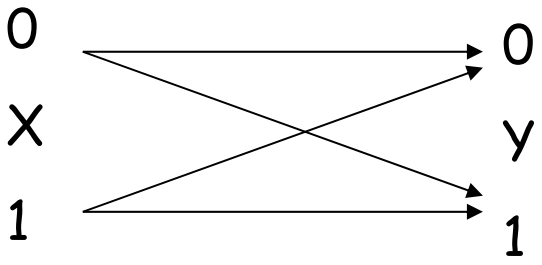
or: the amount of information that Y gives about X

equivalently: $I(X;Y|Z) = H(X|Z) - H(X|YZ)$

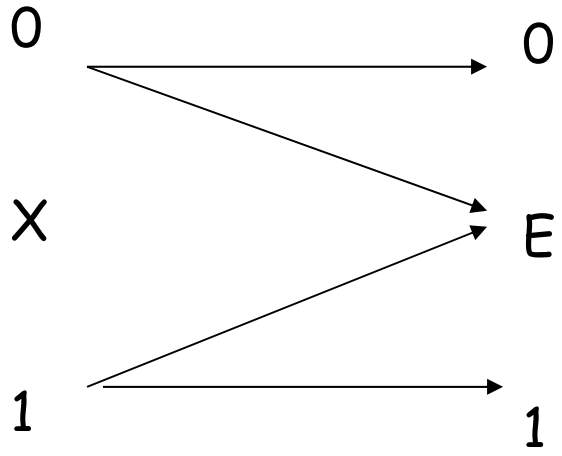
the amount of information that Y gives about X given Z

3 classical channels

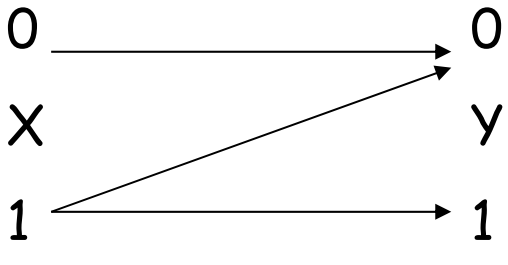
Binary symmetric (satellite)



erasure (network)



Z-channel (optical)



Homework:

find maximum H(X)-H(X|Y) and the corresponding input distribution

Example 1

- Suppose that X $\in$ { 000, 001, •••, 111 } with H(X) = 3 bits
- Channel: X Y = parity of X

- H(X|Y) = 2 bits: we transmitted H(X) – H(X|Y) = 1 bit of information!

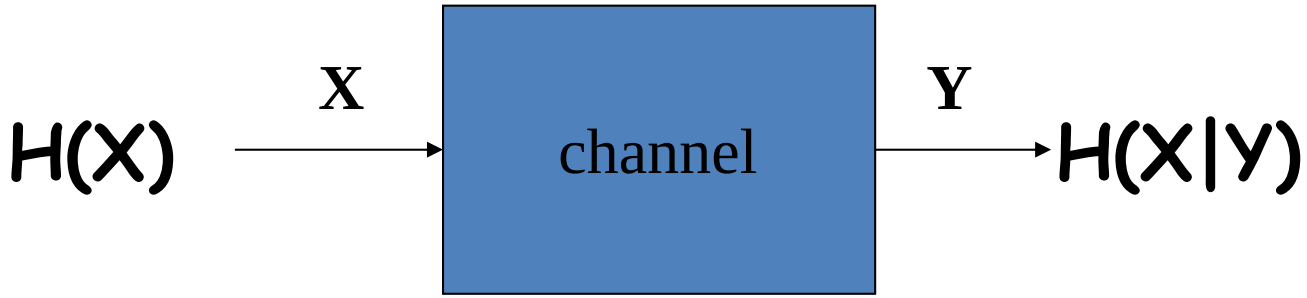
  We know that X|Y $\in$ { 000, 011, 101, 110 } or X|Y $\in$ { 001, 010, 001, 111 }

Homework: suppose the channel output gives the number of ones in X. What is then H(X) – H(X|Y)?

Transmission efficiency

I need on the average H(X) bits/source output to describe the source symbols X

After observing Y, I need H(X|Y) bits/source output



Reduction in description length is called the transmitted information

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

$= H(Y) - H(Y|X)$ from earlier calculations

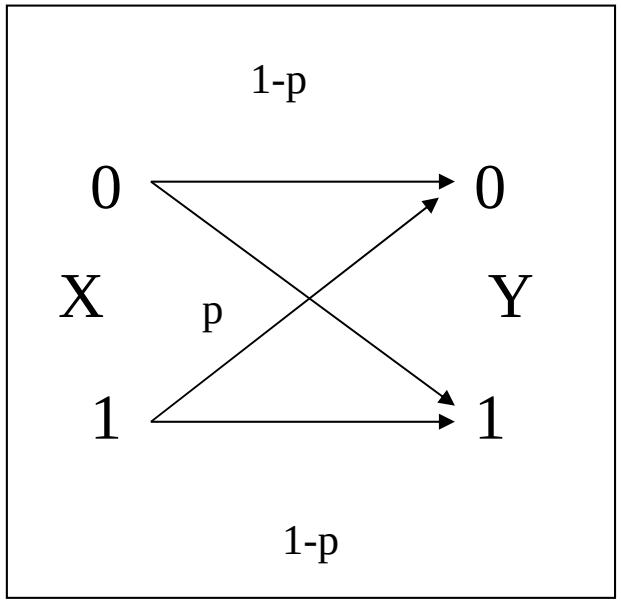
We can ***maximize R*** by changing the input probabilities.
The maximum is called CAPACITY (Shannon 1948)

Transmission efficiency

- Shannon shows that error correcting codes exist that have
  - An efficieny k/n $\leq$ Capacity
    - n channel uses for k information symbols
  - Decoding error probability $\approx$ 0
    - when n very large

- Problem: how to find these codes

channel capacity: the BSC



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

the maximum of $H(Y) = 1$

since Y is binary

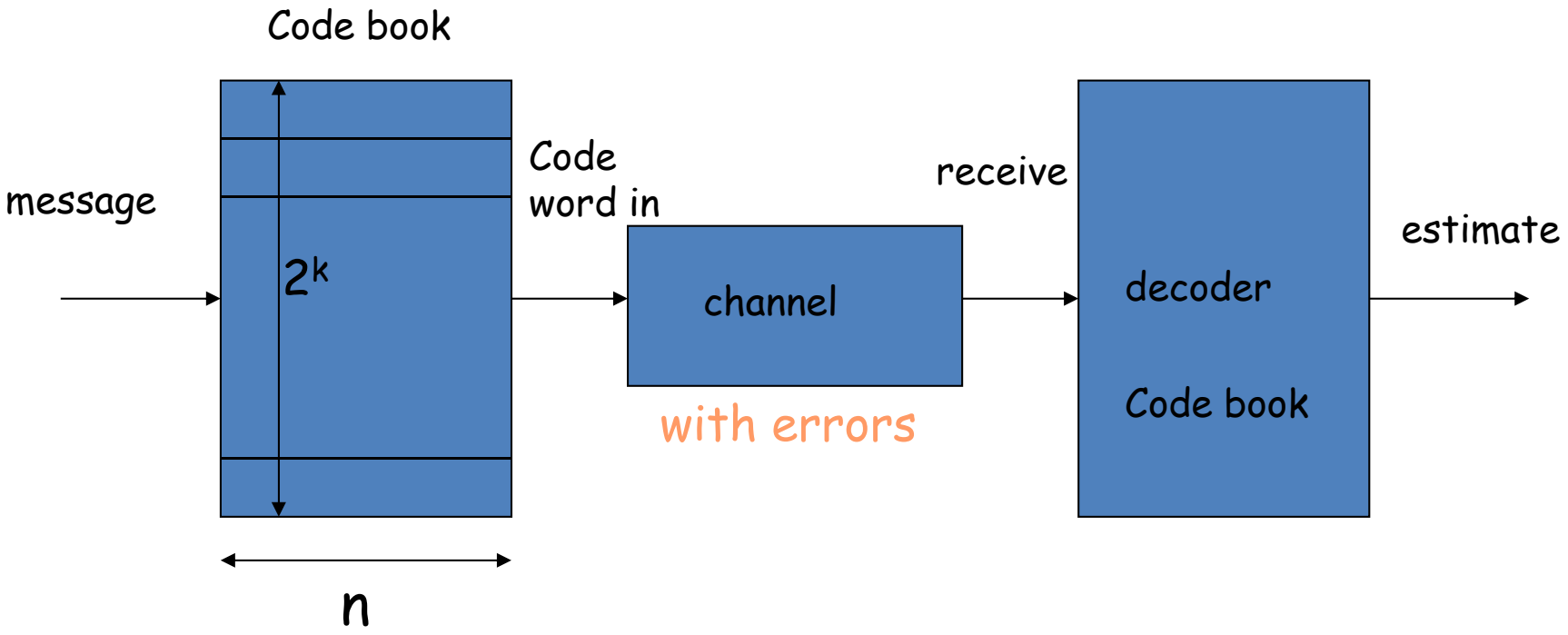
$H(Y|X) = h(p)$

$= P(X=0)h(p) + P(X=1)h(p)$

Conclusion: the capacity for the BSC $C_{BSC} = 1 - h(p)$

Homework: draw C_{BSC} , what happens for $p > \frac{1}{2}$

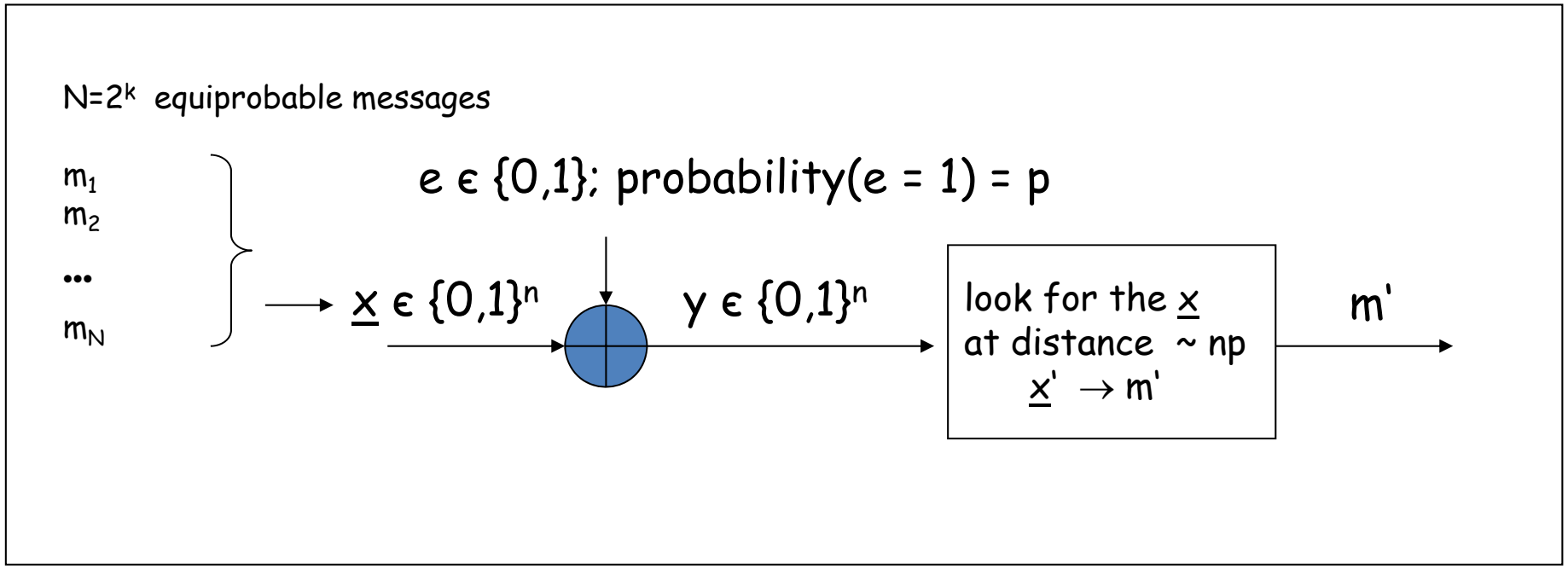
Practical communication system design



There are 2^k code words of length n

k is the number of information bits transmitted in n channel uses

A look at the binary symmetric channel



Conclusion:

if more than one x is connected with y, we have to accept a decision error

Encoding and decoding according to Shannon

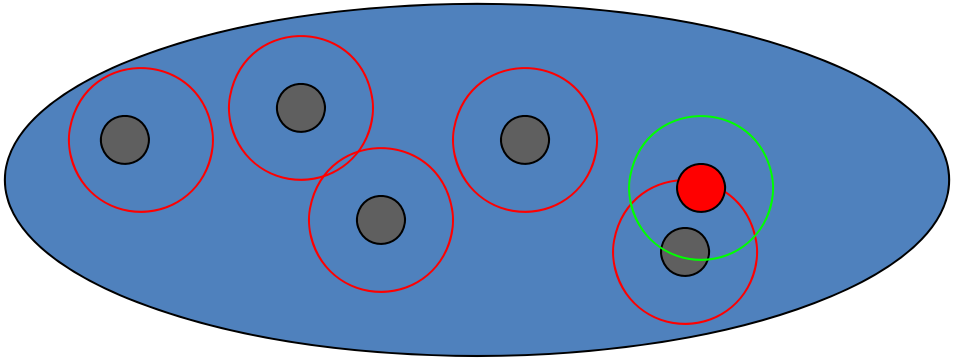
Code: 2^k binary codewords where $p(0) = P(1) = \frac{1}{2}$

Channel errors: $P(0 \rightarrow 1) = P(1 \rightarrow 0) = p$

i.e. # error sequences $\approx 2^{nh(p)}$

Decoder: search around received sequence for codeword

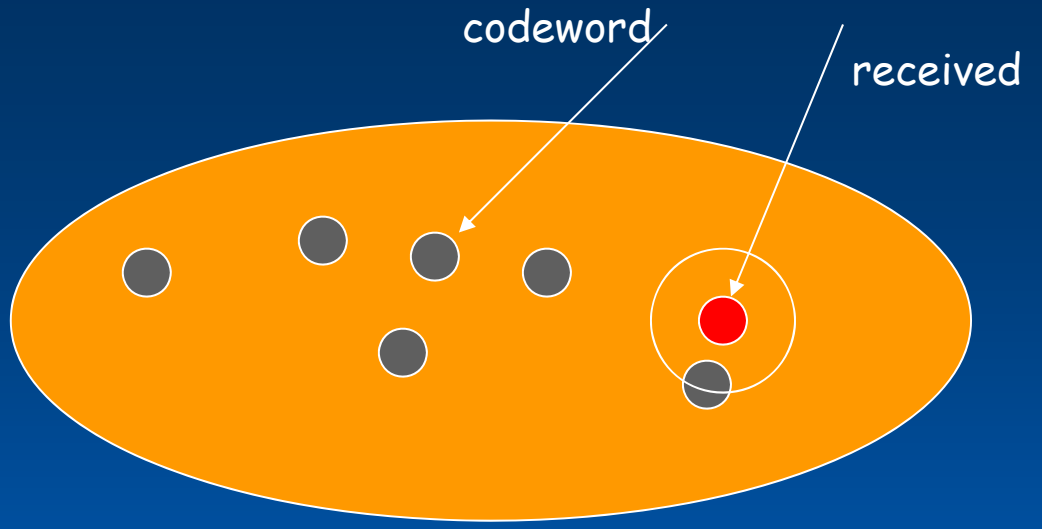
with $\approx np$ differences



space of 2^n binary sequences

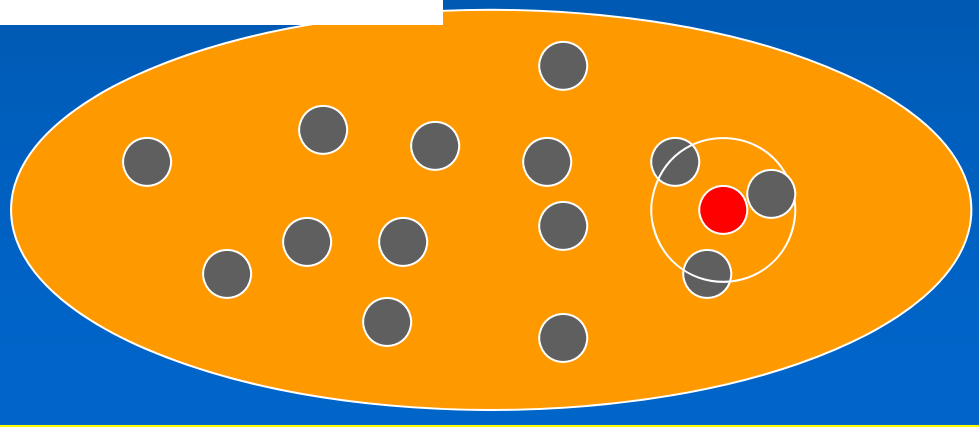
Illustration of the decoding error probability

1. For t errors: $\text{Prob}(|t/n-p|> \epsilon) \rightarrow 0$
   for $n \rightarrow \infty$ (law of large numbers)



2. Expected # of code word in region
   (codewords are random)

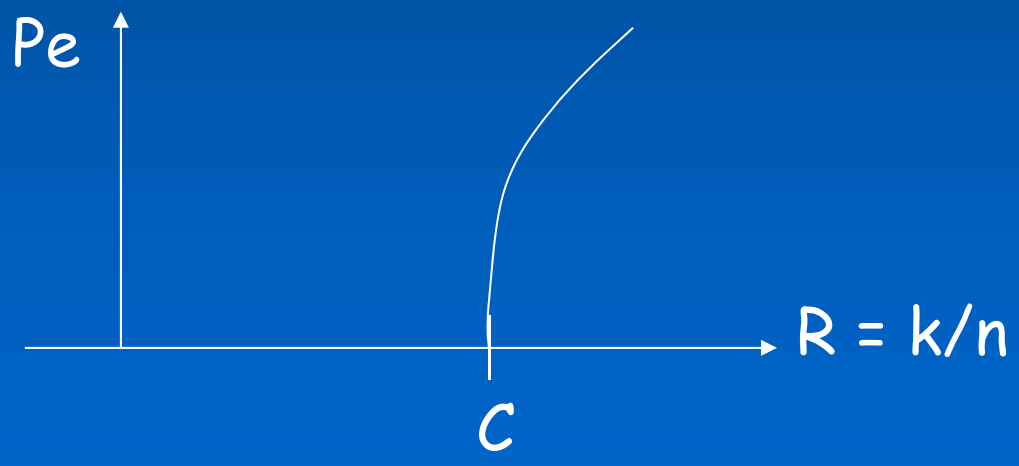
$$2^k \frac{2^{nh(p)}}{2^n} = 2^{n[\frac{k}{n}-(1-h(p)]} := 2^{n[R-C]} \text{ possible vectors } \underline{x}$$



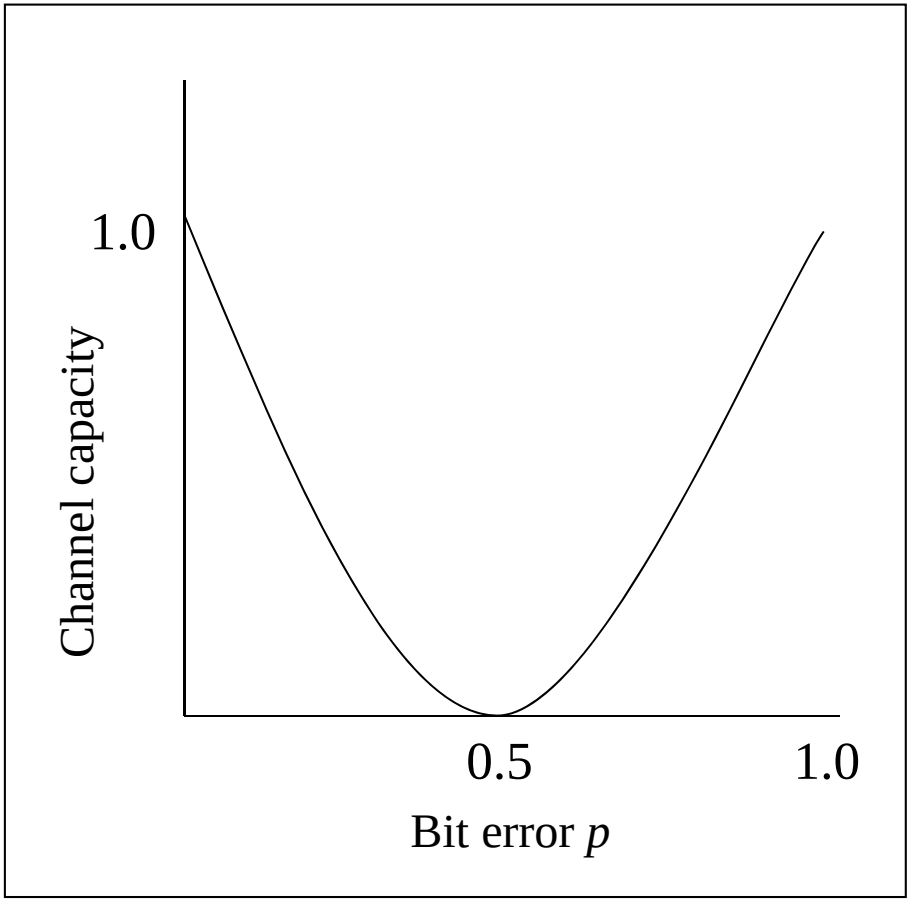
Conclusion: probability of a correct guess of $\underline{x}$ is $2^{-n(R-C)}$ => 0 for R > C

conclusion

For R > C the decoding error probability > 0



channel capacity: the BSC



Explain the behaviour!