# Information Theory
# Degree in Data Science and Engineering
## Lesson 5: Capacity of discrete channels

Jordi Quer, Josep Vidal

Mathematics Department, Signal Theory and Communications Department
{jordi.quer, josep.vidal}@upc.edu

2018/19 - Q1

# Definition of communication

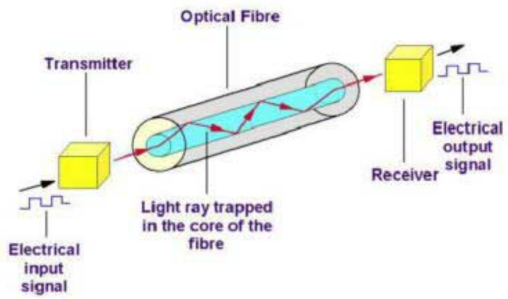
Communication between two points $A$ and $B$ is a procedure whereby physical acts in $A$ induce a desired state in $B$.

Succesful communication: when $A$ and $B$ agree on what was sent, in spite of the noise and imperfections of the signalling process that might induce errors.

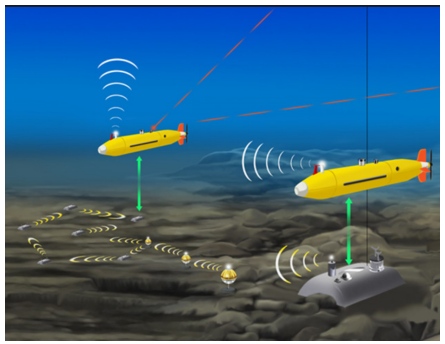
Data communication systems are strongly related to the medium and the transceiver design...

# Some real communication channels

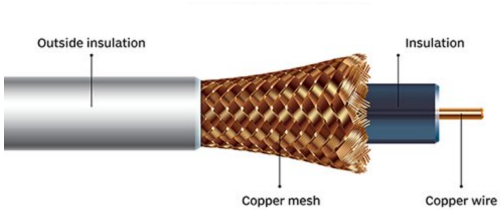
### Light on optical fiber



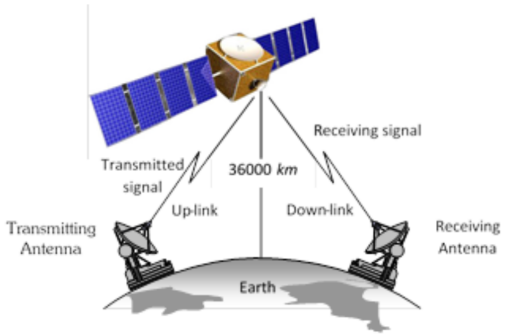

### Acoustic medium



### Electric signals on copper cable




### Electromagnetic waves




### Wireless optical

### Data storage devices

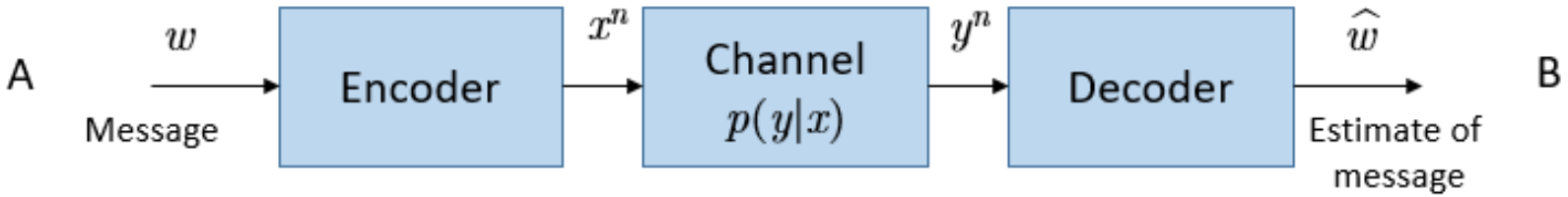

What is the origin of errors? Thermal noise in all electronic devices, and also: dust or scratches in HDD and DVD, solar wind in satellite comms, nearby transmissions in cellular comms, cosmic rays in solid state memories, propellers and biological noise in underwater comms, black body radiation,...

# Goals

We want to ellucidate...

- How to design distinguishable codewords in $B$ so that we can reconstruct the original sequence sent in $A$ with arbitrary low probability of error?
  Channel capacity theorem
- At which rate can these codewords transmit information?
  Channel capacity theorem
- Can feedback from the receiver improve the transmission rate?
  Feedback capacity theorem
- Is it efficient to separately design source and channel encoders?
  Joint source-channel coding theorem

# Formal definitions




- Discrete noisy channel. It consists of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a set of transition probabilities $p(y|x)$ accounting for the probability of observing the output symbol $y$ when $x$ was sent. Then, two different input sequences may give rise to the same output sequence.
- Memoryless channel. The current output $y_i$ only depends on the input $x_i$, and is independent of past inputs and past outputs:

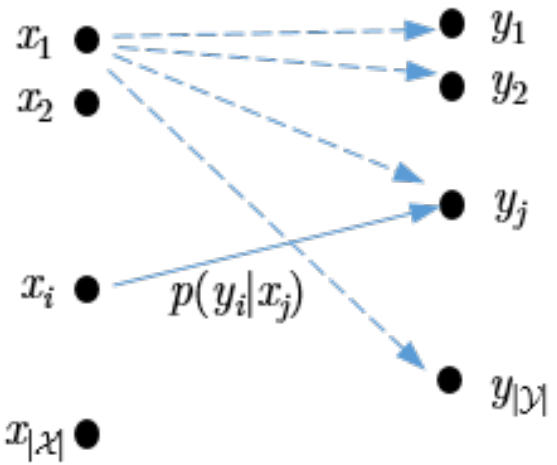
$$p(y_i|x_i, x_{i-1}, \ldots, x_1, y_{i-1}, \ldots, y_1) = p(y_i|x_i)$$

- Message. It is a random variable $W$ taking values from a set $\{1, 2, \ldots, M\}$. The encoder selects a codeword $X^n(W)$ which is received as $Y^n$. The decoder then guesses the index $W$ by an appropriate decoding rule $\hat{W} = g(Y^n)$. The receiver decides an error if $\hat{W} \neq W$.

# Formal definitions

- If $x^n \in \mathcal{X}^n$ is the transmitted codeword, the probability of receiving $y^n \in \mathcal{Y}^n$ is

$$p\left(y^n|x^n\right) = \prod_{i=1}^{n} p(y_i|x_i)$$




- The conditional symbol error probability given that index $i$ was sent is

$$\lambda_i = Pr\left(g(Y^n) \neq i|x^n(i)\right) = \sum_{y^n \in \mathcal{Y}^n} p\left(y^n|x^n(i)\right) I\left(g(y^n) \neq i\right)$$

# Formal definitions

- The maximal probability of error $\lambda^{(n)}$ for an $(M, n)$ code is

$$\lambda^{(n)} = \max_{i \in 1,2,\ldots,M} \lambda_i$$

- The average probability of error $P_e^{(n)}$ for an $(M, n)$ code is

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

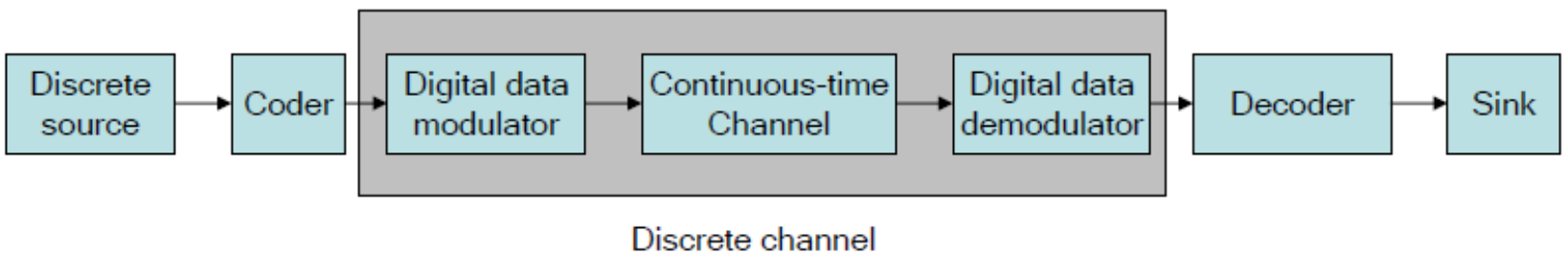
- The rate $R$ of an $(M, n)$ code is $R = \frac{\log M}{n}$ bits per transmission
- A rate $R$ is said to be achievable if there exists a sequence of $\left(\lceil 2^{nR} \rceil, n\right)$ codes such that $\lambda^{(n)} \to 0$ as $n \to \infty$.

# Range of application

The models used in the sequel do not include:

- Discrete-time continous amplitude input/output channels
- Channels with memory (i.e. frequency selective channels, where outputs depend on previous inputs)
- Continuous channels
- Multi-user transmissions (e.g. multiple-access channels, broadcast channels, interference channels)

But the model is useful for transmission schemes where modulator and demodulator are part of the channel:



# Channel capacity

The capacity of the channel is defined as

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

where the maximum is taken over all input distributions $p(x)$.

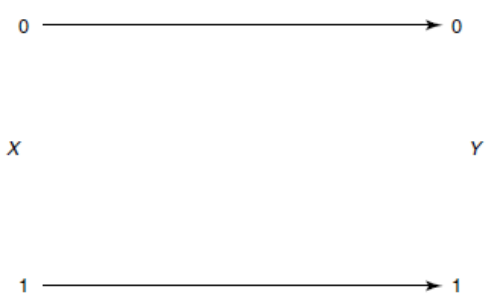
It is measured in bits/transmission, or bits/channel use.

The following duality should be noted:

- Data compression - use source encoder to remove redundancy and compress.
- Data transmission - use channel encoder to add redundancy and combat channel errors.

# Examples of discrete channels

- Noiseless channel

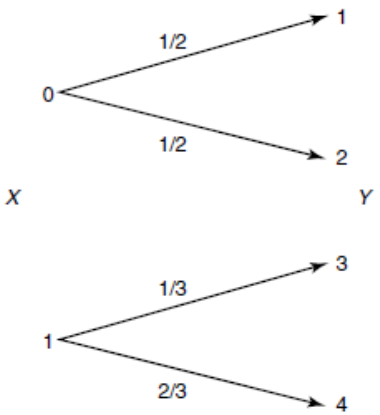



$$p(y_i|x_j) = \delta_{i,j}$$

$$I(X;Y) = H(X) - H(X|Y) = H(p) - 0$$

$$C = \max_{p(x)} I(X;Y) = 1 \text{ bit/tr}$$

- Noisy channel with non-overlapping outputs

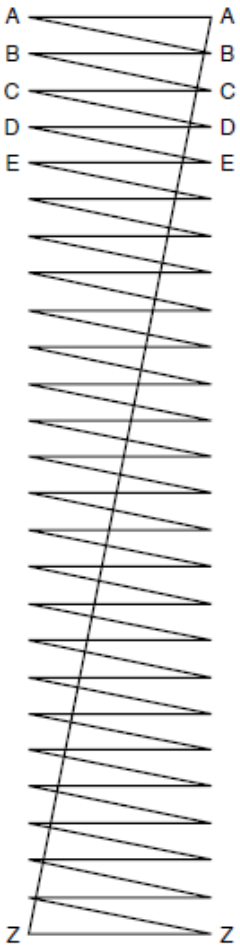



The channel is random, but the inputs can be reconstructed from the outputs without errors, so $H(X|Y) = 0$ and

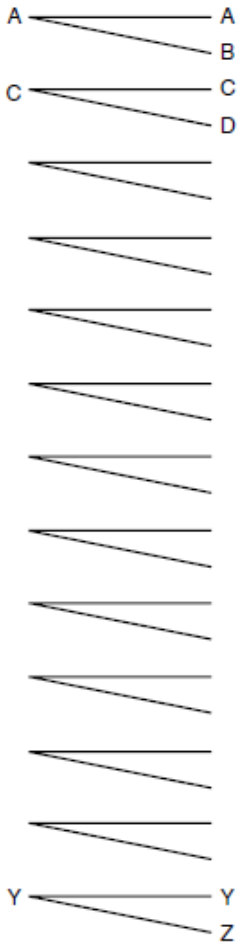
$$C = \max_{p(x)} I(X;Y) = 1 \text{ bit/tr}$$

# Examples of discrete channels

- Noisy typewritter



Noisy channel



Noiseless subset of inputs

Using only the alternate input symbols, we can transmit 13 symbols without errors and we are in the previous channel case, so $C = \log 13$.
Alternatively, we can compute capacity as:

$$p(x|y = \mathsf{B}) = \begin{cases} \frac{1}{2} & \text{if } x = \mathsf{A} \\ \frac{1}{2} & \text{if } x = \mathsf{B} \\ 0 & \text{otherwise} \end{cases}$$

$$H(X|Y) = \sum_{y=\mathsf{A}}^{\mathsf{Z}} p(y) H(X|Y = y) = 1 \text{ bit}$$

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

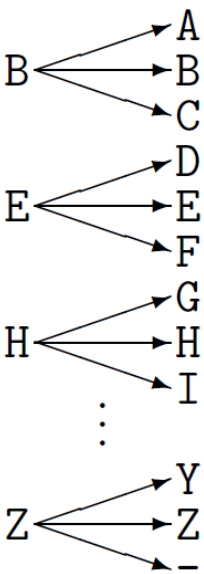
$$= \log 26 - 1 = \log 13 \text{ bits/tr}$$

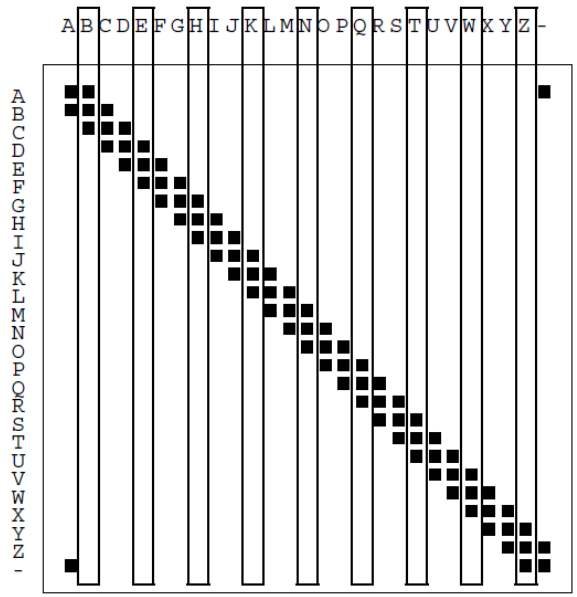
Note that if a uniform distribution is used for $X$, we have a uniform distribution for $Y$.

# Examples of discrete channels

- Noisy typewriter (cont.)

Non-confusable set of inputs for a three-outputs noisy typewriter channel:

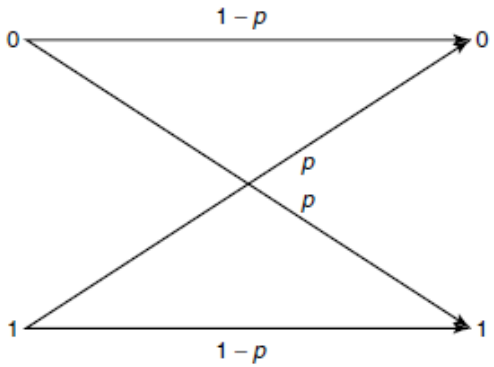



For large block lengths, every channel looks like the typewritter: any input is likely to produce a channel output in a small subset of the output alphabet. Then, capacity is obtained from the non-confusable subset of inputs that produce disjoint output sequences (will be discussed later in this lesson at the noisy-channel capacity theorem).

# Examples of discrete channels

- Binary symmetric channel




$$p(x|y) = \begin{cases} 1-p & \text{if } x = y \\ p & \text{if } x \neq y \end{cases}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) \\ &= H(Y) + p \log p + (1-p) \log(1-p) \\ &= H(Y) - H(p) \end{aligned}$$

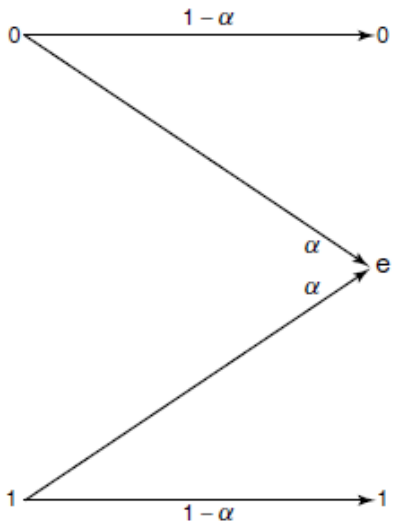
The maximum is achieved when $p(x=1) = \frac{1}{2}$, then $p(y=1) = \frac{1}{2}$, so

$$C = 1 - H(p) \text{ bits/tr}$$

Note that the rate at which we can transmit information is not $(1-p)$ bits per channel use, since the receiver does not know where the error ocurred. In fact, if $p = \frac{1}{2}$, we cannot transmit any information at all!

# Examples of discrete channels

- Binary erasure channel




Some bits are lost, rather than corrupted:

$$p(y|x=0) = \begin{cases} 1-\alpha & \text{if } y = 0 \\ \alpha & \text{if } y = \mathsf{e} \\ 0 & \text{if } y = 1 \end{cases}$$

$$p(y|x=1) = \begin{cases} 0 & \text{if } y = 0 \\ \alpha & \text{if } y = \mathsf{e} \\ 1-\alpha & \text{if } y = 1 \end{cases}$$

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \left(H(Y) - H(Y|X)\right) = \max_{p(x)} H(Y) - H(\alpha)$$

Let us compute

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log p(y) \text{ where } p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$$

and maximize it w.r.t. $p(x)$.

# Examples of discrete channels

- Binary erasure channel (cont.)

By computing the probability of the output it is clear that, for an arbitrary value of $\alpha$, we cannot make $Y$ symbols equiprobable, and $H(Y) < \log 3$:

$$H(Y) = H(\alpha) + (1-\alpha)H(\pi)$$

where $\pi = p(x=1)$. Therefore,

$$C = \max_{\pi}(1-\alpha)H(\pi)$$

so the maximum is achieved at $\pi = \frac{1}{2}$, and $C = (1-\alpha)$ bits/tr.

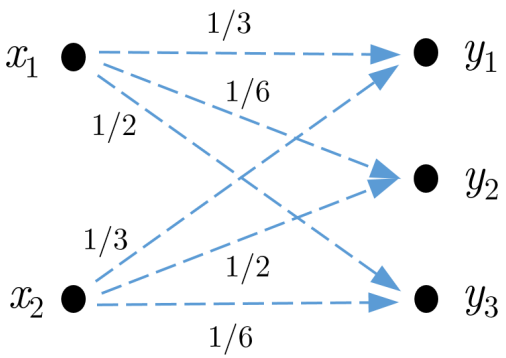
The intuition for the expression is the following: since a fraction $\alpha$ of symbols is lost, we can only transmit a fraction $(1-\alpha)$.

# Examples of discrete channels

- Symmetric channel

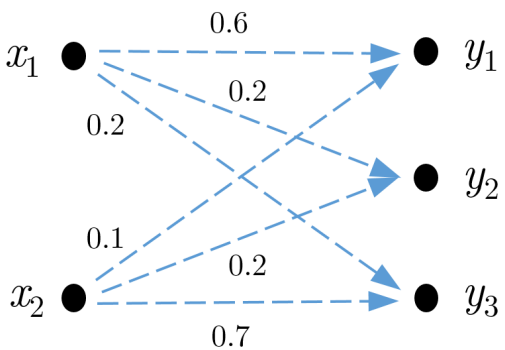
Weakly symmetric channel: in matrix $\mathbf{Q}$, that contains the transition probabilites, rows are permutations of other rows, and all the column sums are equal.
Symmetric channel: in matrix $\mathbf{Q}$, rows are permutations of other rows, and columns are permutations of other columns.

Examples:




**Weakly symmetric channel**

$$\mathbf{Q} = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$




**Non-symmetric channel**

$$\mathbf{Q} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$

$$Y = X + Z \bmod c$$

**Symmetric channel**

$\mathcal{X} = \mathcal{Z} = \{0, 1, ..., c-1\}$
$X$ and $Y$ are independent
$p(z)$ is arbitrary
Determine $\mathbf{Q}$ as an exercise

# Examples of discrete channels

- Symmetric channel (cont.)

Both for symmetric and weakly symmetric channels:

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) + \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x)$$

$$= H(Y) - H(\mathbf{q})$$

where $\mathbf{q}$ is a row of $\mathbf{Q}$. If $p(x)$ is uniform,

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} p(y|x) = \frac{c(y)}{|\mathcal{X}|}$$

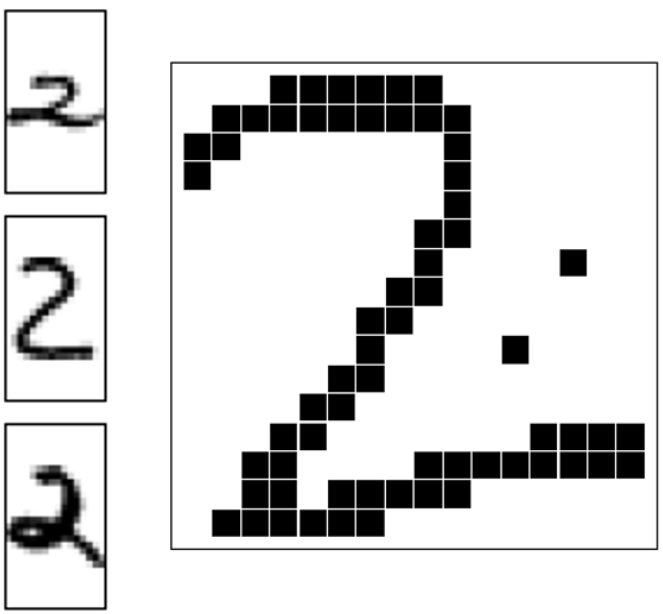
where $c(y)$ is the sum of elements in the $y$-th column of the transition matrix and it is constant for both symmetric and quasi-symmetric channels. Therefore the capacity is

$$C = \max_{p(x)} I(X;Y) = \log|\mathcal{Y}| - H(\mathbf{q})$$

# Examples of discrete channels

- Pattern recognition

Consider the problem of recognizing handwritten digits. In this case the input to the channel is a decimal digit $X \in \mathcal{X} = \{0, 1, 2, \ldots, 9\}$. What comes out is a pattern of ink on a paper that can be represented as a vector $\mathbf{y}$.



If the ink pattern is digitized to $16 \times 16$ binary pixels, the output of the channel is a vector random variable $\mathbf{Y} \in \{0, 1\}^{256}$ (unlike previous examples, where one scalar input produced one scalar output).

One strategy for pattern recognition (that is, decoding) is to build a model for $p(\mathbf{y}|x)$ and use it to infer $X$ given $\mathbf{Y}$ using Bayes' theorem:

$$\hat{x} = \operatorname*{argmax}_x p(x|\mathbf{y}) = \operatorname*{argmax}_x \frac{p(\mathbf{y}|x)p(x)}{p(\mathbf{y})}$$

# Examples of discrete channels

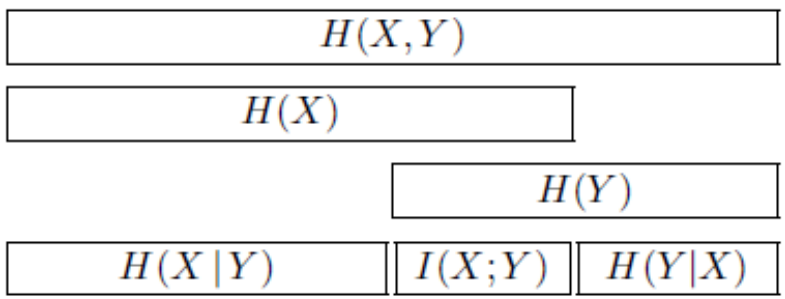
- Natural evolution

Natural evolution can be considered as a channel that models how information about the environment is transferred to the genome.

# Properties of channel capacity

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

1. $C > 0$ since $I(X;Y) \geq 0$.
2. $C \leq \log |\mathcal{X}|$ since $C = \max_{p(x)} I(X;Y) \leq \max_{p(x)} H(X) = \log |\mathcal{X}|$.
3. $C \leq \log |\mathcal{Y}|$ for the same reason.
4. $I(X;Y)$ is a continuous function of $p(x)$.
5. $I(X;Y)$ is a concave function of $p(x)$.
6. Reminder of relations between information and entropy in graphic form:



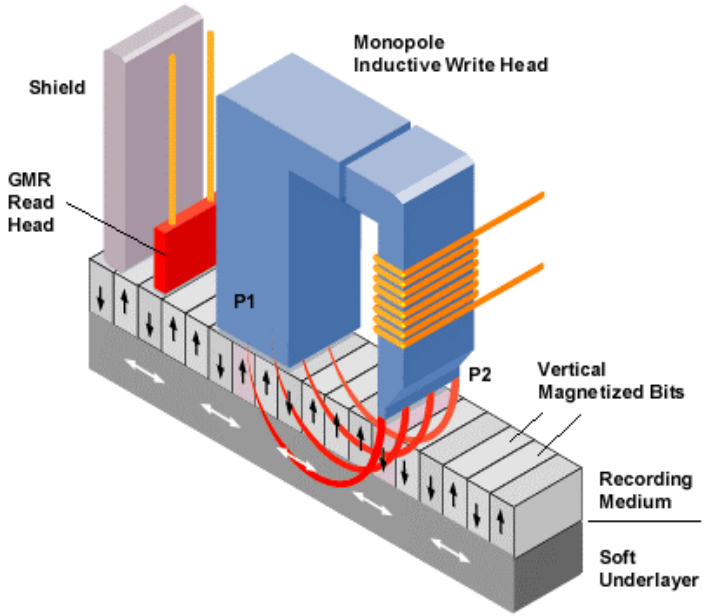

Rationale: The mutual information is the uncertainty at the channel input minus the remaining uncertainty when the channel output is observed.

# Example: A code for reliable storage

A *magnetic hard disk drive (HDD)* records data by magnetizing a thin film of ferromagnetic material in flat circular disks. Bits are stored by changing the direction of magnetization through a magnetic coil head. A reading head is used to detect the magnetization of the material underneath.

**Hard disk drive**

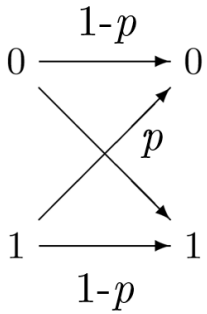
**Magnetic recoding read/write head**




*Watch this video on how an HDD works.*

# Example: A code for reliable storage

The reading/writing process can be modeled as a transmission of a sequence of bits through a BSC:



- Assume that the crossover probability of the BSC is $p = 0.1$.
- If we read/write 1 Gbyte of data per day for 10 years, the number of bits sent through the channel is $2,92 \times 10^{13}$.
- A useful HDD should not deliver any erroneous bit in its entire life.
- This requires a target bit error probability on the order of $10^{-15}$, or even smaller.

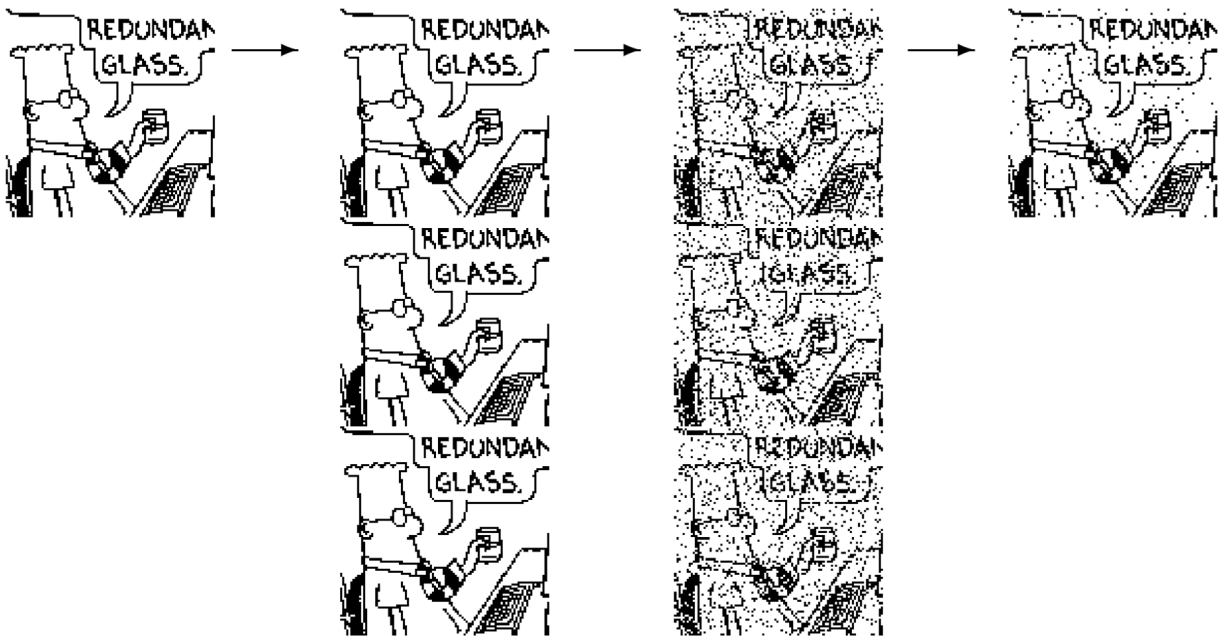
An error-correcting code (channel encoder/decoder) is needed.

# Example: A code for reliable storage

Let us try to reduce the probability of error in the BSC using a repetition code $\mathrm{R}_N$, for $N = 3$ parallel disks:

$$1\,0\,1\,1\,0\,0\ldots \rightarrow 111\,000\,111\,111\,000\,000\ldots$$


$p = 0.1$




The transmission rate is $R = 1/N$.

# Example: A code for reliable storage

The optimum decoding strategy is deciding the transmitted bit by majority voting among the $N$ received symbols. Then, the probability of error is:

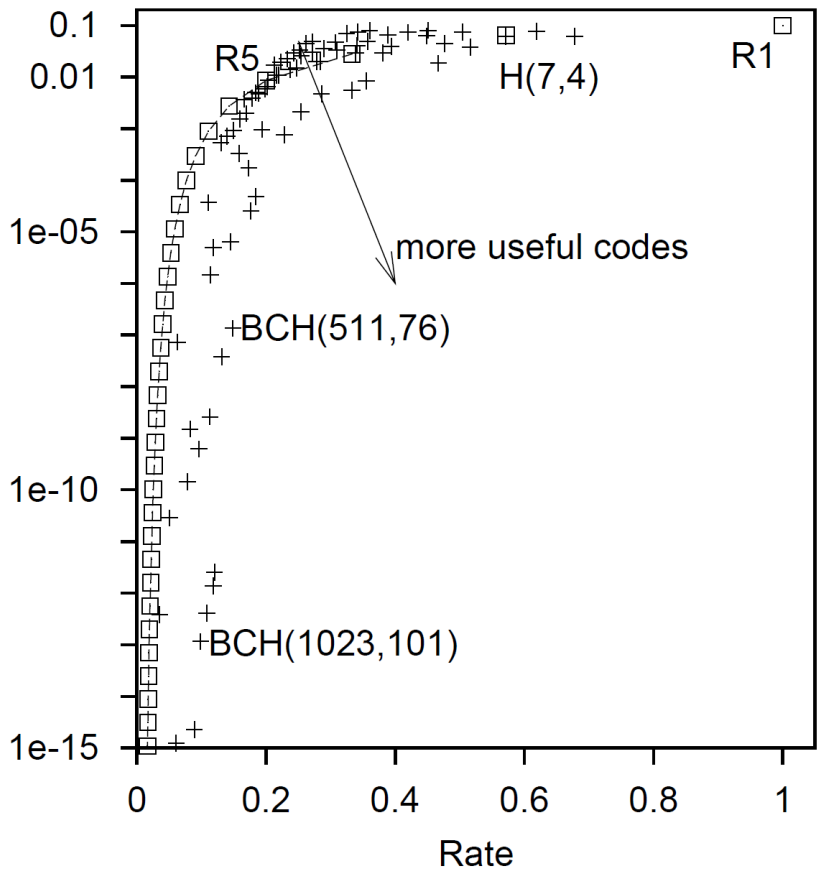
$$P_e^N = \sum_{n=(N+1)/2}^{N} \binom{N}{n} p^n (1-p)^{N-n} \approx (4p(1-p))^{\frac{N}{2}}$$

For $P_e^N = 10^{-15}$, $N$ must be at least $68$: we need so many parallel disks to achieve the target bit error probability!!

Clearly, better codes are needed.

# Example: A code for reliable storage

Plot $P_e^N$ vs bitrate for $p = 0.1$ and different codes




What is the tradeoff between redundancy and error probability?
Is it possible to transmit at $R > 0$ with $P_e^N = 0$?

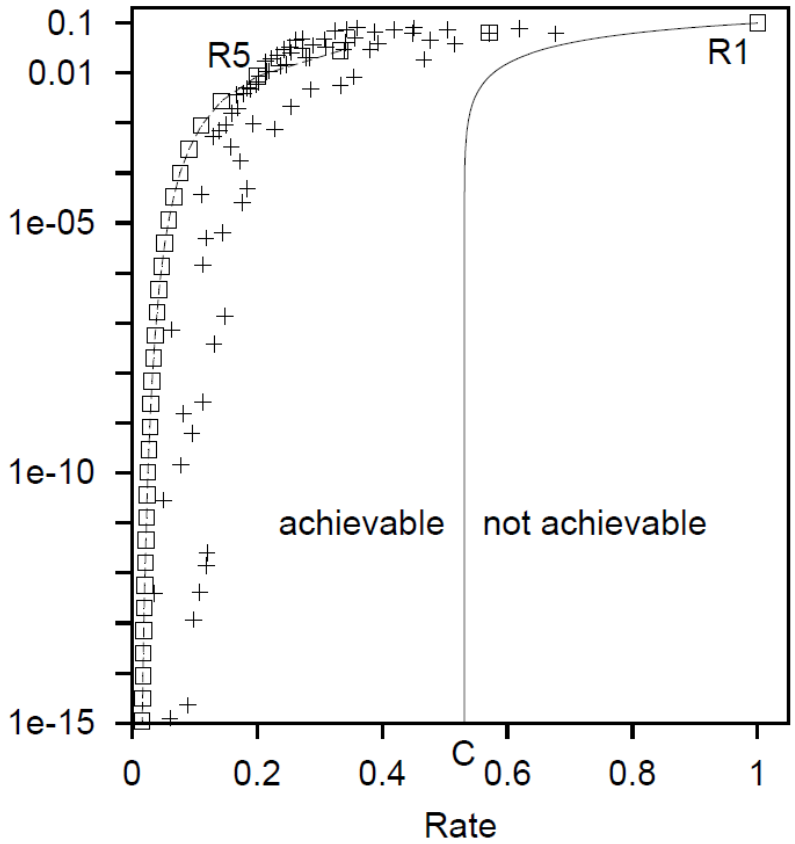
# Better channel codes and a teaser

Before Shannon announced his theorem, it was believed that zero-error communication implied zero-rate transmission.

Shannon stated the tradeoff between $P_e$ and $R$: there is a non-zero rate $R \leq C$ at which we can transmit information with $P_e = 0$.

The theorem proves that every channel behaves like a typewritter channel: a subset of inputs produce disjoint (and hence non-confusable) sequences at the output.

We need to evaluate how many of these sequences are possible and how to decode them.



# Jointly typical sequences

## Definition

The set $\mathcal{A}_\epsilon^{(n)}$ of jointly typical sequences $(x^n, y^n)$ is the set of $n$-sequences with empirical entropies $\epsilon$-close to the true ones, that is

$$\begin{aligned}\mathcal{A}_\epsilon^{(n)} =& \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ & \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon\}\end{aligned}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$

# Example

Let us evaluate the joint typicality of two sequences $n = 100$:
$x^n$ (with $p(x = 0) = 0.9$) was transmitted at the input of a binary symmetric channel (with $p(y|x) = 0.2$), and $y^n$ was received:

$x^n$ 1111111111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

$y^n$ 0011111111100000000000000000000000000000000000000000000000000000000000000000000000000000011111111111111111

From the probabilities above, we can easily compute

$$p(y) = \sum_{x=0}^{1} p(y|x)p(x) = \begin{cases} 0.74 & \text{if } y = 1 \\ 0.26 & \text{if } y = 0 \end{cases}$$

| $p(x, y)$ | 0 | 1 |
|---|---|---|
| 0 | 0.72 | 0.02 |
| 1 | 0.18 | 0.08 |

and check that

$$H(Y|X) = 1.8140, \;\; H(X, Y) = H(X) + H(Y|X) = 2.2830$$

which exactly matches the empirical entropies computed from the sequences above, so sequences are jointly typical. Now flip the last bit of $y^n$. To which tolerance $\epsilon$ both sequences are now jointly typical?

# Properties of the joint typical sequences

**Theorem (5.1)**

$$\Pr\left((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}\right) > 1 - \epsilon \text{ as } n \to \infty$$

**Theorem (5.2)**

*For sufficiently large* $n$, $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq \left|\mathcal{A}_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}$

**Theorem (5.3)**

*If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ *(they are independent with the same marginals as* $x^n$ *and* $y^n$*), then the probability of being jointly typical is upper bounded by*

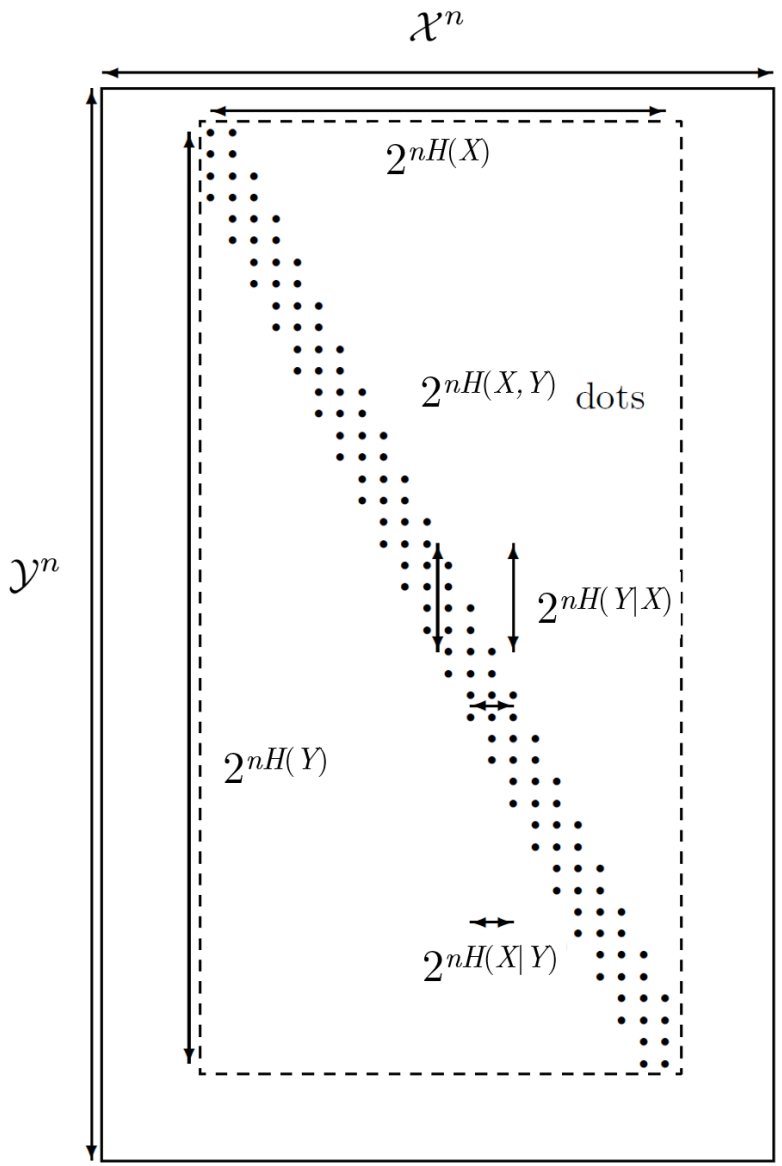
$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

*and for sufficiently large* $n$*, it is lower bounded by*

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Proofs. See annex.

# The size of the jointly typical set




The outer box represents all conceivable input/output pairs of sequences. Each dot represents a jointly-typical pair of sequences, whose total number is $2^{nH(X,Y)}$ (they can be computed offline, using exhaustive search).

Note that for very large block lengths, every channel looks like the typewritter: any input is very likely to produce a channel output in a small subset of the output alphabet.

So let us find a non-confusable subset of inputs that produce disjoint output sequences: this is the key idea behind the channel capacity theorem.

Receiver operation: we shall associate a channel output $y^n$ to the $i$-th message, if the transmitted codeword $x^n(i)$ is jointly typical with $y^n$.

# Noisy-channel coding theorem

**Theorem (C. E. Shannon, 1948)**

1. *Achievability. For every discrete memoryless channel, for $R < C$ and $n \to \infty$, there exist a code $(2^{nR}, n)$ and a decoding algorithm for which the maximum probability of error is $\lambda^{(n)} \to 0$.*
2. *Converse. For any $\lambda^{(n)} \to 0$, transmission rates greater than $C$ are not achievable.*

In the sequel we prove each part separately.

# 1. Achievability part

Proof of achievability. It is completed in two steps:

- Can we define a suitable encoding/decoding scheme?

**i**. A random code of $2^{nR}$ sequences of length $n$ is generated according to some pdf, such that $p(x^n) = \prod_{i=1}^{n} p(x_i)$, so the codebook can be described as

$$\mathcal{C} = \begin{bmatrix} x^n(1) \\ \vdots \\ x^n(2^{nR}) \end{bmatrix} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix},$$

and the probablity of the code is

$$\Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w)).$$

**ii**. The random code used is revealed to sender and receiver, who also knows the channel transition matrix $p(y|x)$.

# 1. Achievability part

**iii**. A message $W$ is selected according to a uniform distribution $Pr(W = w) = 2^{-nR}$, $w = 1, 2, \ldots, 2^{nR}$.

**iv**. The $w$-th codeword $x^n(w)$ is sent over the channel.

**v**. The receiver observes a sequence $y^n$ according to the distribution

$$p(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)).$$

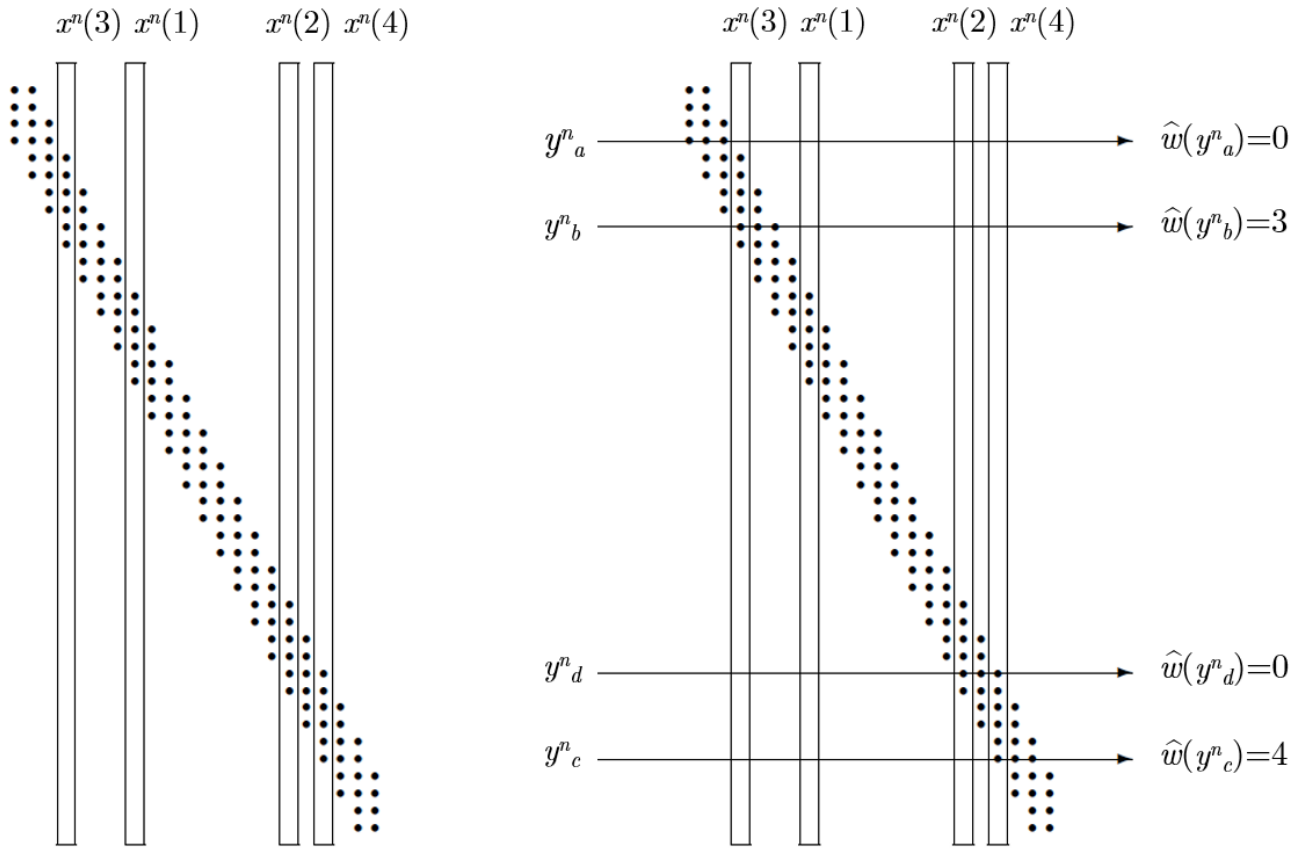
**vi**. The receiver guesses which message was sent according to the joint typicallity criterion. $\hat{w}$ is declared to have been sent if the following conditions are satisfied:

- $(x^n(\hat{w}), y^n)$ is jointly typical, and
- no other index $w' \neq \hat{w}$ such that $(x^n(w'), y^n) \in \mathcal{A}_\epsilon^n$

otherwise, an error is declared.

**vii**. There is an error $\mathcal{E}$ if $\hat{w} \neq w$.

# 1. Achievability part




Example of typical set decoding for a 4-symbols codebook: $y_a^n$ is not jointly typical with any codeword, $y_b^n$ is jointly typical with $x^n(3)$, $y_c^n$ is jointly typical with $x^n(4)$, $y_d^n$ is jointly typical with more than one codeword.

# 1. Achievability part

Proof (cont).

- Does the probability of error of this code reduces to zero?

**i**. The average probability of error for all codewords in all possible codebooks is:

$$\Pr(\mathcal{E}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C})$$
$$= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) = \Pr(\mathcal{E}|w=1)$$

**ii**. Let us define $E_i = \left\{(x^n(i), y^n) \in \mathcal{A}_\epsilon^{(n)}\right\}$ for $i \in \left\{1, 2, \ldots, 2^{nR}\right\}$. An error occurs if either:

- $E_1^c = \left\{(x^n(i), y^n) \notin \mathcal{A}_\epsilon^{(n)}\right\}$, or
- $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (a wrong codeword is jointly typical)

Therefore, we can write, using the union bound...

# 1. Achievability part

$$\Pr(\mathcal{E}|w=1) = \Pr(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|w=1)$$
$$\leq \Pr(E_1^c|w=1) + \sum_{i=2}^{2^{nR}} \Pr(E_i|w=1).$$

**iii**. By the joint AEP, $\Pr(E_1^c|w=1) \leq \epsilon$ for $n \to \infty$. By the code generation process, $x^n(1)$ and $x^n(i)$ are independent for $i \neq 1$ and so are $y^n$ and $x^n(i)$, and hence by theorem 5.3 of the joint AEP:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|w=1) \leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X,Y)-3\epsilon)}$$
$$= \epsilon + (2^{nR}-1)2^{-n(I(X,Y)-3\epsilon)}$$
$$\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X,Y)-R)}$$

A key point: we can make the last term in red lower than $\epsilon$ by increasing $n$, provided that $R < I(X,Y) - 3\epsilon$, and then $\Pr(\mathcal{E}) \leq 2\epsilon$.

# 1. Achievability part

**iv**. We can strengthen the conclusion by selecting the code in a smart way by:

a. Choosing $p(x)$ such that $I(X,Y)$ is maximal, and hence $R < C$.

b. Get rid of the average over codes and keep just the good one. Since averaging over all codes gives $\Pr(\mathcal{E}) \leq 2\epsilon$ there must be one code (found by exhaustive search over all $(2^{nR}, n)$ codes) such that

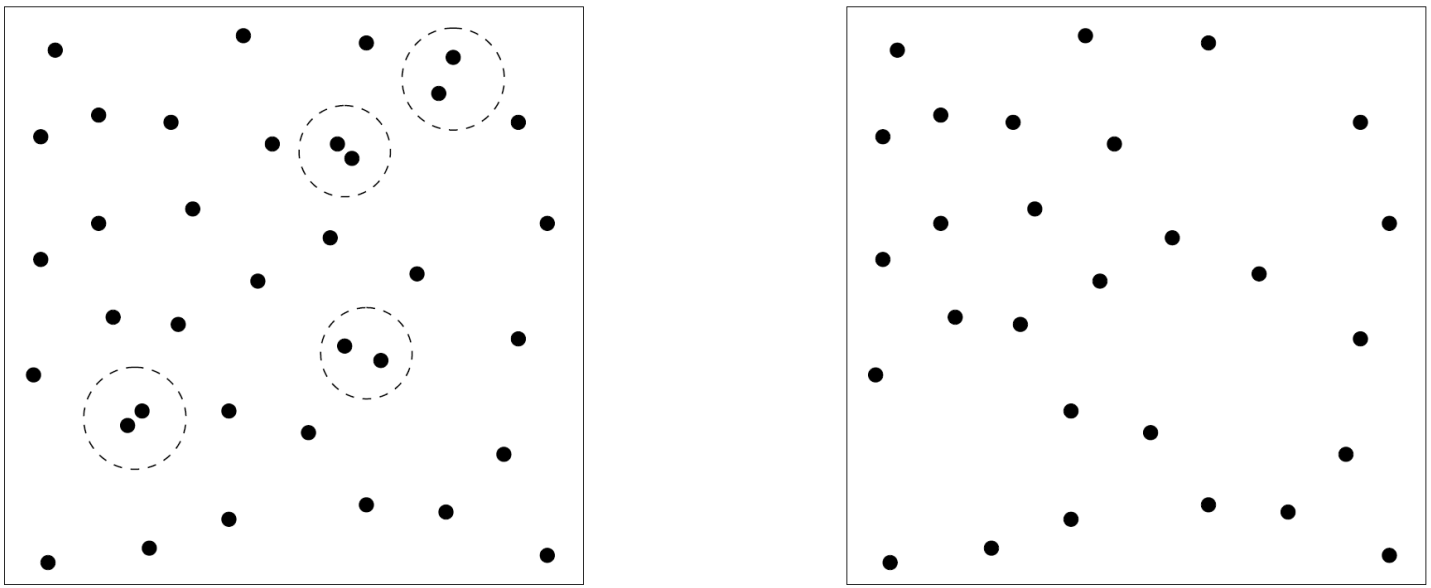
$$\Pr(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*) \leq 2\epsilon.$$

c. Throw away the worst half of codewords (so as we can minimize the maximal probability of error $\lambda_{max}(\mathcal{C}^*)$, not only the average):

$$\Pr(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2}\left[\frac{1}{2^{nR-1}} \sum_{worst\ i} \lambda_i(\mathcal{C}^*) + \frac{1}{2^{nR-1}} \sum_{best\ i} \lambda_i(\mathcal{C}^*)\right] \leq 2\epsilon$$

where the term in red is loosy upper bounded by $4\epsilon$.

# 1. Achievability part



A typical random code (left), where a small fraction of the codewords are sufficiently close to each other that the probability of error when either codeword is transmitted is not tiny. We obtain a new code by deleting all these confusable codewords (right). The resulting code has less codewords, so has a lower rate, and its maximal probability of error is greatly reduced.

# 1. Achievability part

The number of codewords has changed to $2^{nR-1}$ and therefore the rate is

$$\frac{1}{n}\log(2^{nR-1}) = \frac{1}{n}(nR-1) = R - \frac{1}{n}$$

In short: we have been able to turn a noisy channel into a noiseless channel, as long as the transmission rate is below the capacity, just by constructing a code of rate

$$R' = R - \frac{1}{n},$$

whose maximal probability of error is $\lambda^{(n)} \leq 4\epsilon$. □

# 2. Converse part

Proof of converse. Assume we have a code $(2^{nR}, n)$ with $\lambda^{(n)} \to 0$ (this implies $P_e^{(n)} \to 0$), an encoding rule $X^n(.)$ and a decoding rule $\hat{W} = g(Y^n)$ so we can construct the Markov chain

$$W \to X^n(W) \to Y^n \to \hat{W}.$$

If $W$ has a uniform distribution $\Pr(\hat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \sum_i \lambda_i$ and hence

$$\begin{aligned}
nR &= H(W) \\
&= H(W|\hat{W}) + I(W;\hat{W}) \quad \text{Entropy identity} \\
&\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \quad \text{Fano's theorem (upper bounding } H(P_e^{(n)}) < 1) \\
&\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) \quad \text{Data processing inequality} \\
&\leq 1 + P_e^{(n)} nR + nC \quad \text{Repeated use of channel does not increase capacity (see annex)}
\end{aligned}$$

Dividing by $n$ we obtain

$$R \leq P_e^{(n)} R + C + \frac{1}{n}$$

For large $n$, we can build codes that $P_e^{(n)} \to 0$ and hence $R \leq C$. □

# Communication above capacity

Can we achieve rates beyond capacity if some error is acceptable?

**Theorem (Rate-distorsion)**

*For a channel of capacity $C$, transmission rates up to*

$$R = \frac{C}{1 - H(P_e)}$$

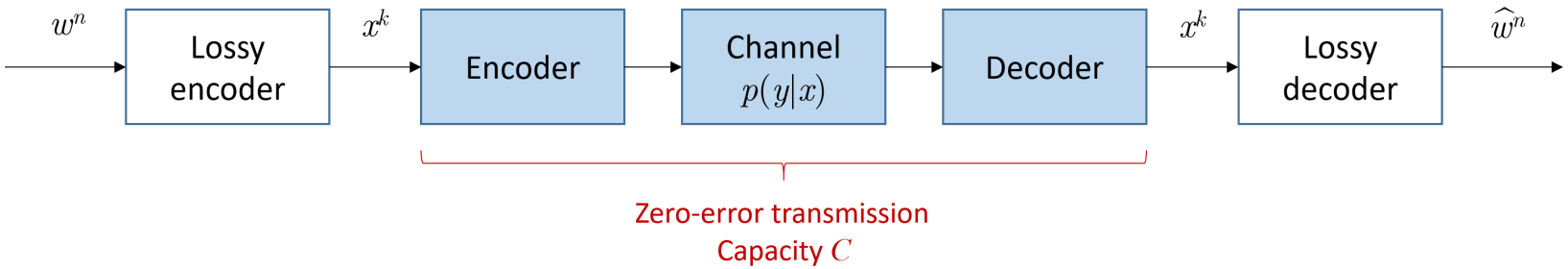
*can be achieved at a probability of bit error $P_e$.*

Proof. Take a channel of capacity $C$ obtained with its corresponding capacity-achieving encoder/decoder that allow zero-error transmission, and take the capacity-achieving encoder/decoder designed for a BSC whose transition probability is $q$.

Since we are interested in lossy transmission, let us reverse the use of encoder and decoder, i.e. the BSC-decoder is used as a lossy encoder, whose input is a sequence of $n$ symbols and the output is a codeword of length $k < n$, and its rate is $R' = n/k = 1/(1 - H(q))$.

# Communication above capacity

Proof (cont.) The lossy decoder will take a sequence of $k$ symbols and convert it to a sequence of $n$ symbols. Let us concatenate them with the $C$ capacity channel together with its own optimum encoder/decoder as follows...

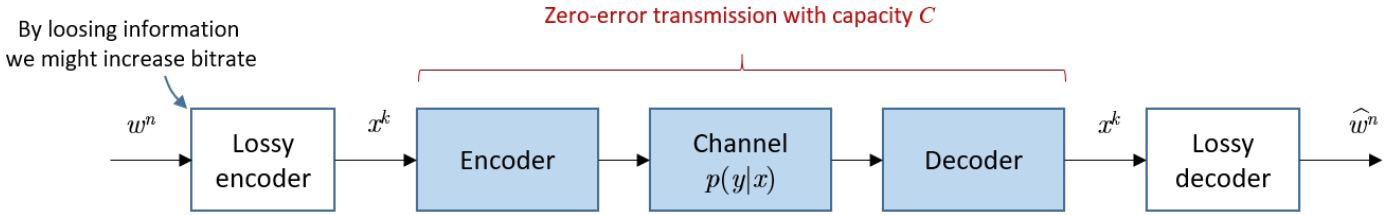


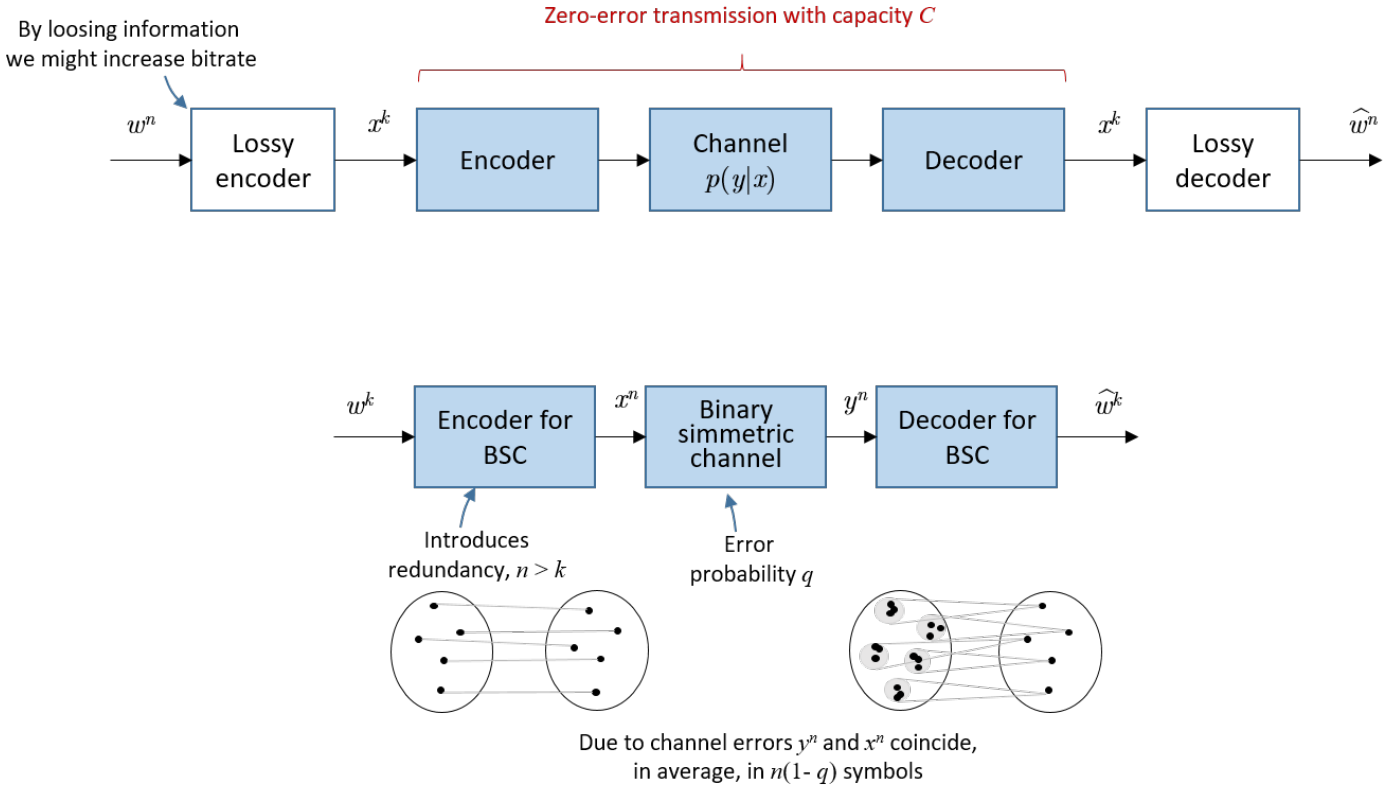

The lossy encoding is a *surjective* mapping and the lossy decoding is a *bijective* mapping. Both are designed using the joint typicality principle for the BSC channel, so $\hat{w}$ and $w$ will differ in $nq$ symbols, hence $P_e = q$.

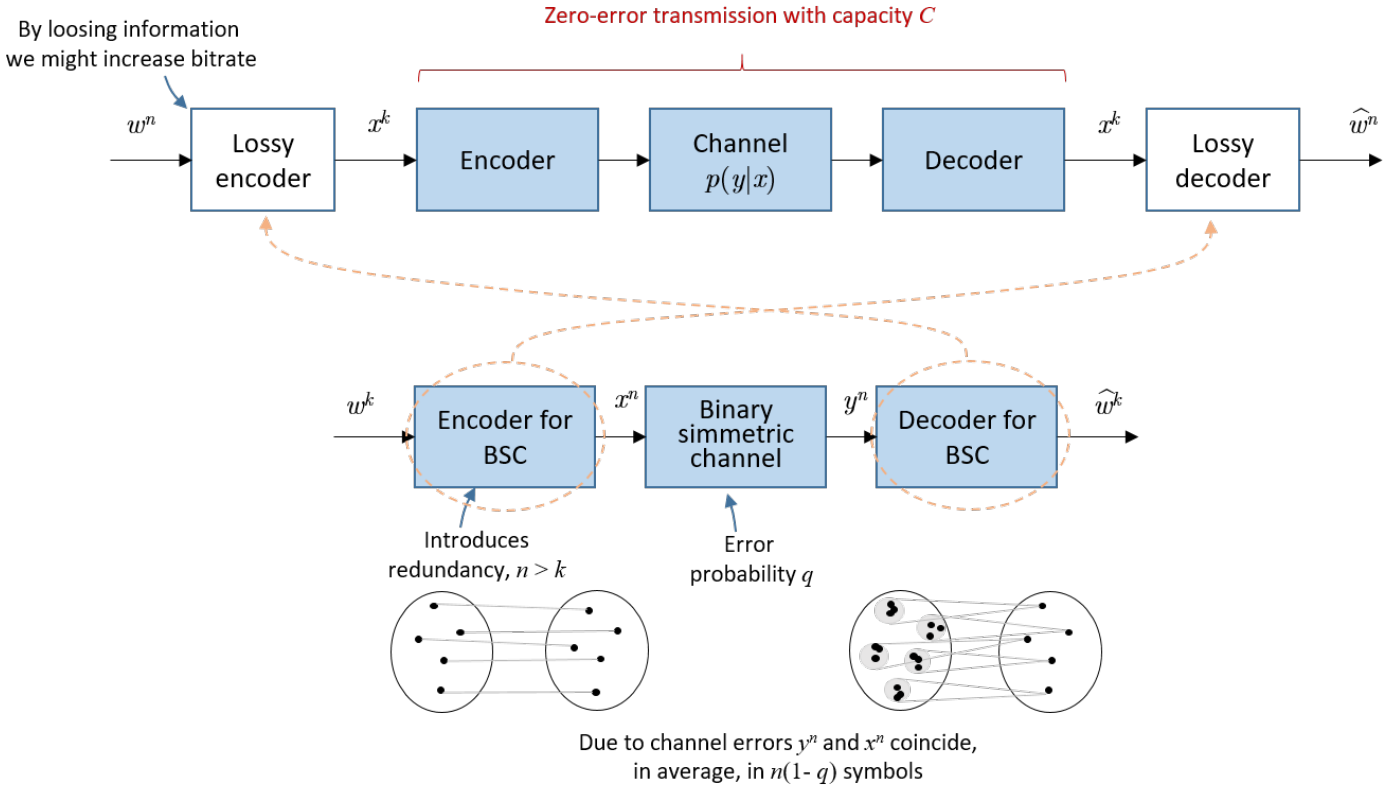
Now the rate of the transmission with errors is

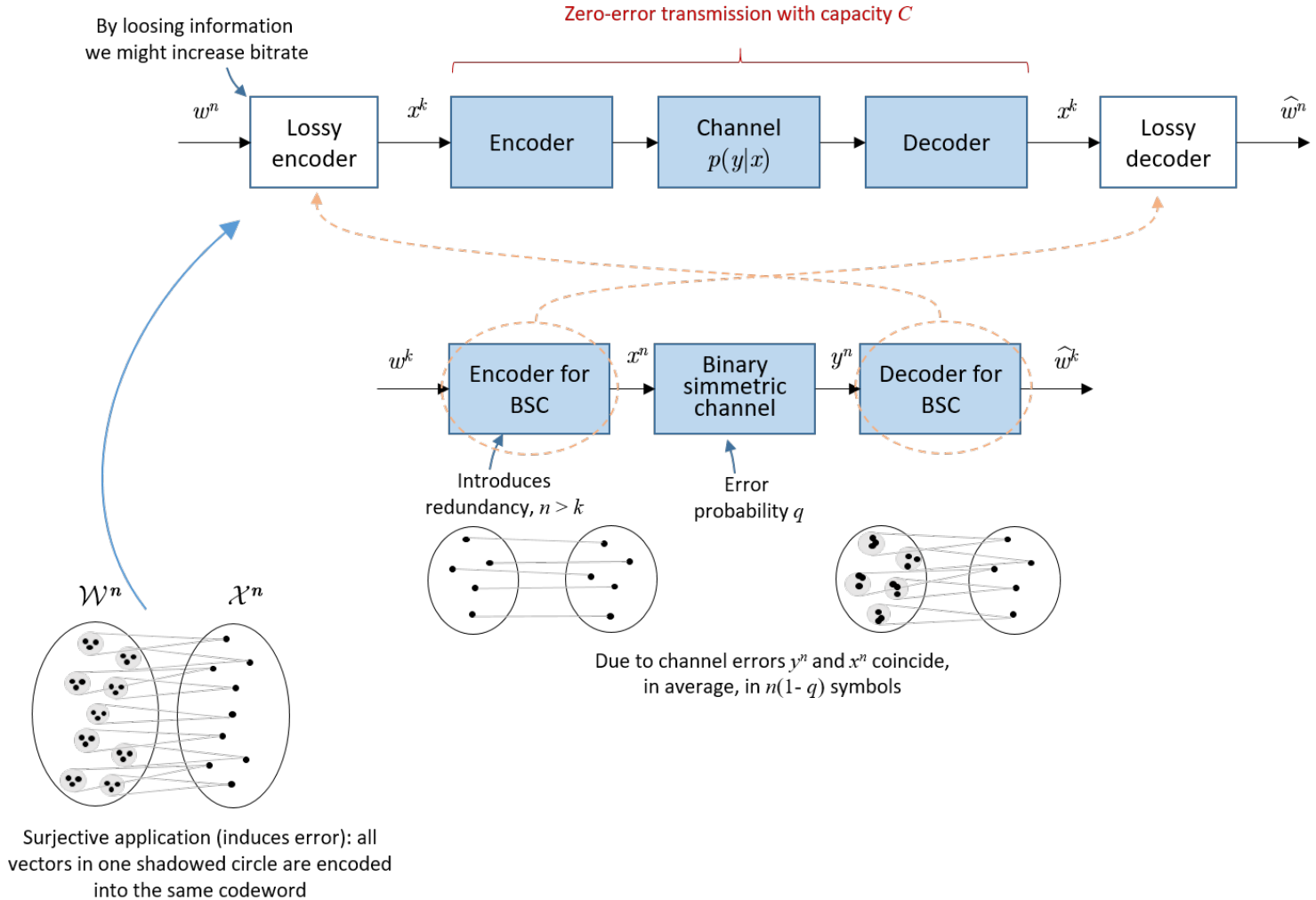
$$R = \frac{k}{\#\text{ transmissions}} \frac{n}{k} = C \cdot R' = C \frac{1}{1 - H(P_e)} \qquad \square$$

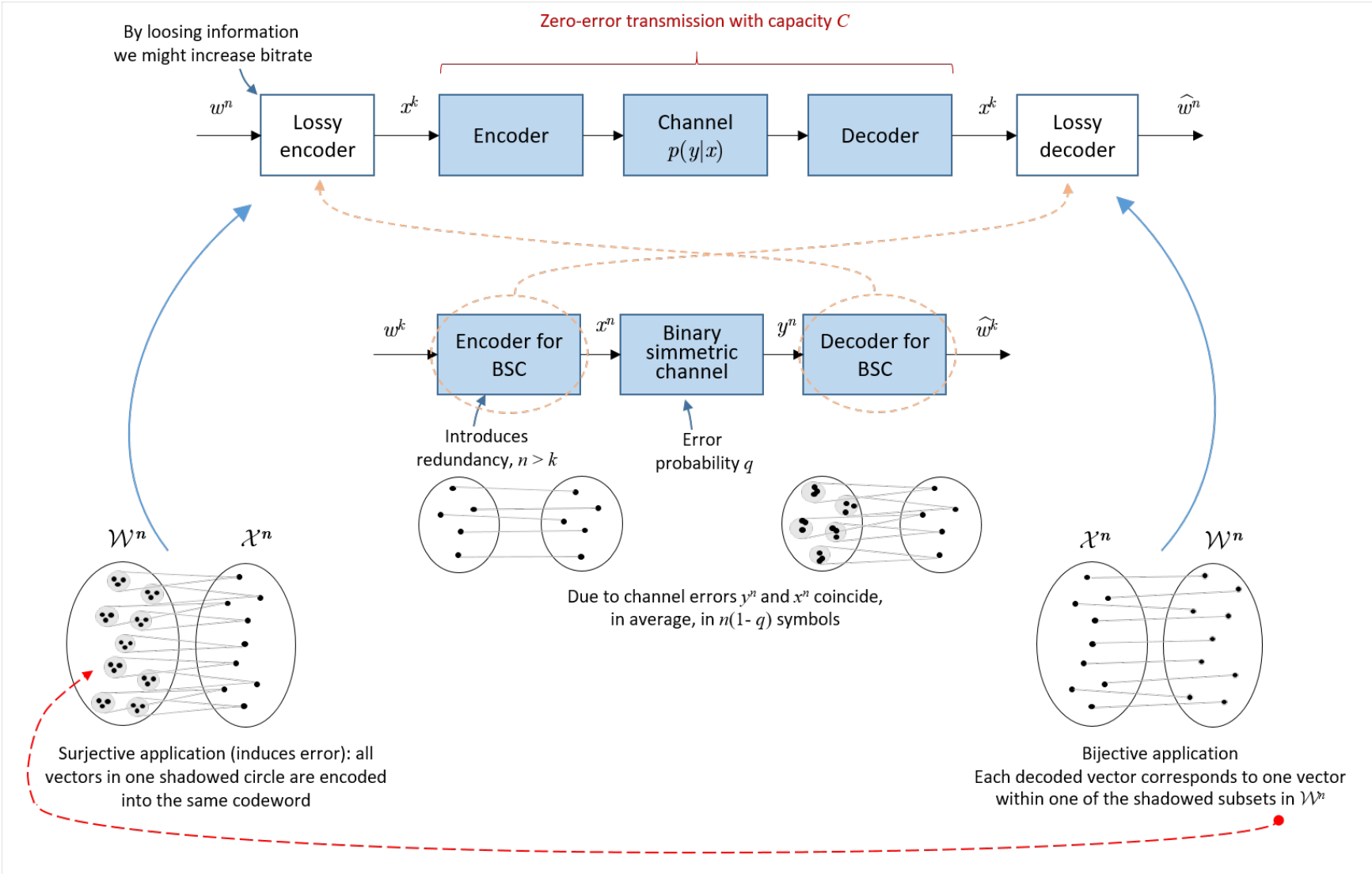
The following slides illustrate the theorem...


By loosing information
we might increase bitrate
Zero-error transmission with capacity $C$
$w^n$
Lossy encoder
$x^k$
Encoder
Channel $p(y|x)$
Decoder
$x^k$
Lossy decoder
$\widehat{w}^n$

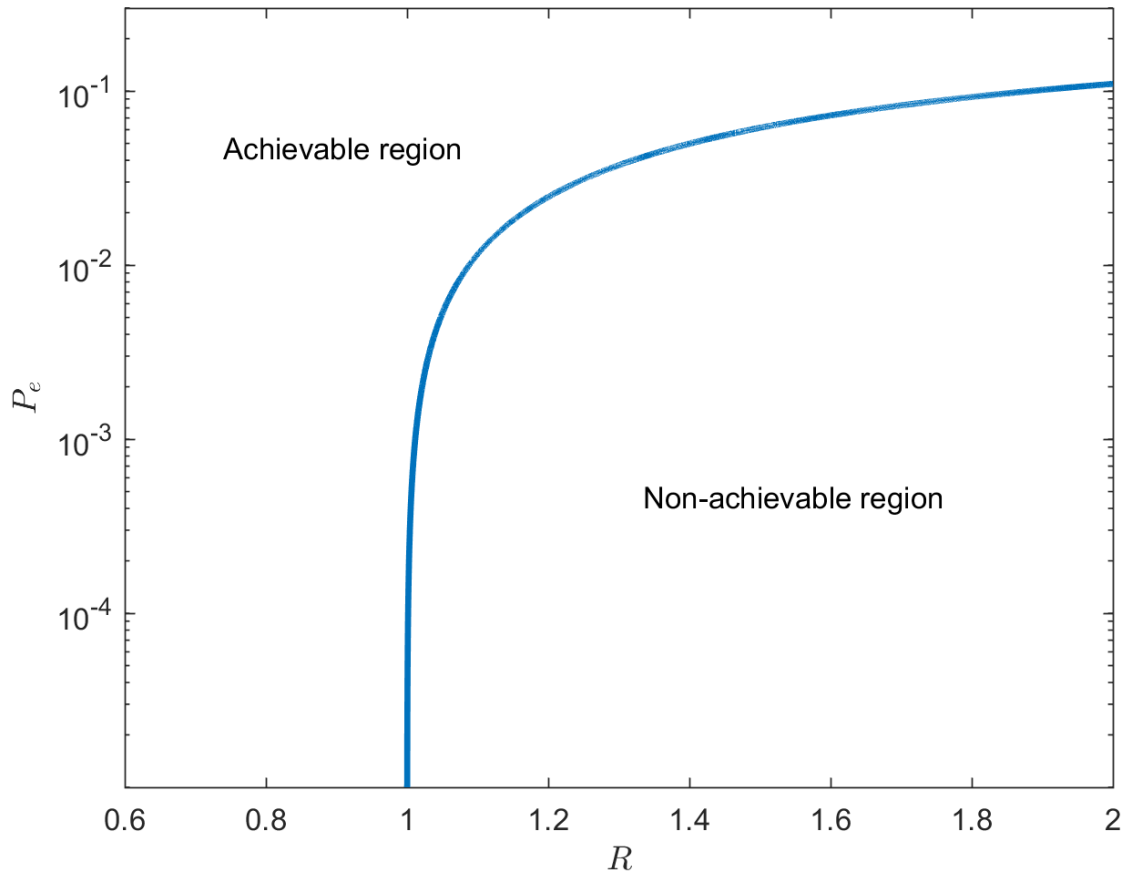

By loosing information
we might increase bitrate
Zero-error transmission with capacity $C$
$w^n$
Lossy encoder
$x^k$
Encoder
Channel $p(y|x)$
Decoder
$x^k$
Lossy decoder
$\widehat{w}^n$
$w^k$
Encoder for BSC
$x^n$
Binary simmetric channel
$y^n$
Decoder for BSC
$\widehat{w}^k$
Introduces redundancy, $n > k$
Error probability $q$
Due to channel errors $y^n$ and $x^n$ coincide,
in average, in $n(1\text{-}\,q)$ symbols


By loosing information
we might increase bitrate
Zero-error transmission with capacity $C$
$w^n$
Lossy encoder
$x^k$
Encoder
Channel $p(y|x)$
Decoder
$x^k$
Lossy decoder
$\widehat{w}^n$
$w^k$
Encoder for BSC
$x^n$
Binary simmetric channel
$y^n$
Decoder for BSC
$\widehat{w}^k$
Introduces redundancy, $n > k$
Error probability $q$
Due to channel errors $y^n$ and $x^n$ coincide,
in average, in $n(1- q)$ symbols


By loosing information
we might increase bitrate
Zero-error transmission with capacity $C$
$w^n$
Lossy encoder
$x^k$
Encoder
Channel $p(y|x)$
Decoder
$x^k$
Lossy decoder
$\widehat{w}^n$
$w^k$
Encoder for BSC
$x^n$
Binary simmetric channel
$y^n$
Decoder for BSC
$\widehat{w}^k$
Introduces redundancy, $n > k$
Error probability $q$
$\mathcal{W}^n$
$\mathcal{X}^n$
Surjective application (induces error): all vectors in one shadowed circle are encoded into the same codeword
Due to channel errors $y^n$ and $x^n$ coincide, in average, in $n(1-q)$ symbols


By loosing information
we might increase bitrate
Zero-error transmission with capacity $C$
$w^n$
Lossy encoder
$x^k$
Encoder
Channel $p(y|x)$
Decoder
$x^k$
Lossy decoder
$\widehat{w}^n$
$w^k$
Encoder for BSC
$x^n$
Binary simmetric channel
$y^n$
Decoder for BSC
$\widehat{w}^k$
Introduces redundancy, $n > k$
Error probability $q$
$\mathcal{W}^n$
$\mathcal{X}^n$
Due to channel errors $y^n$ and $x^n$ coincide,
in average, in $n(1- q)$ symbols
$\mathcal{X}^n$
$\mathcal{W}^n$
Surjective application (induces error): all vectors in one shadowed circle are encoded into the same codeword
Bijective application
Each decoded vector corresponds to one vector within one of the shadowed subsets in $\mathcal{W}^n$

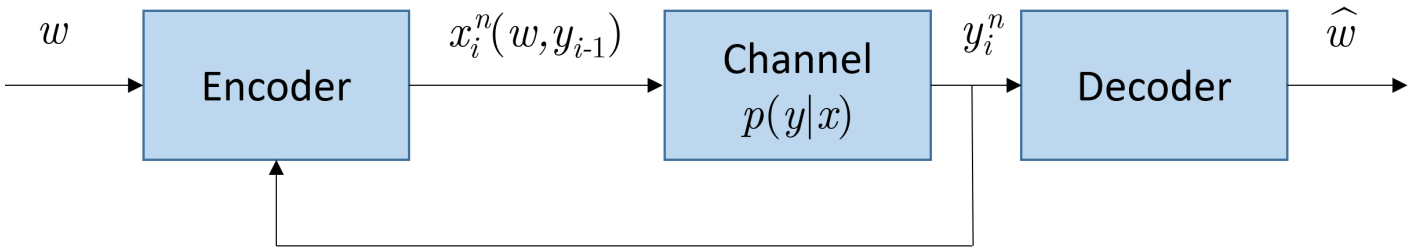
# Communication above capacity




Bit error probability versus transmission rate for a channel of capacity 1 bit/transmission (the non-achievability region is not proved).

# Capacity with feedback

Let us assume that the receiver can send back immediately and noiselessly the received symbols to the transmitter, which then can decide what to do next.
Can feedback from the receiver increase the channel capacity?




Let us define a $(2^{nR}, n)$ feedback code as a sequence of mappings $x_i\left(w, y_{i-1}\right)$ where the codeword is selected according to the input symbol and the past received symbols.

**Theorem (Feedback capacity)**

$$C_{FB} = C = \max_{p(x)} I(X;Y)$$

# Capacity with feedback

Proof. A non-feedback code is a special case, so $C_{FB} \geq C$. Let us prove that $C_{FB} \leq C$. Assume $W$ be uniformly distributed over $1, 2, ..., 2^{nR}$, and hence $Pr(W \neq \hat{W}) = P_e^{(n)}$. Then let us bound the rate as:

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \quad \text{Fano's inequality} \\ &\leq 1 + P_e^{(n)} nR + I(W;Y^n) \quad \text{Data processing inequality} \end{aligned}$$

The last term can be bounded as:

# Capacity with feedback

$$
\begin{aligned}
I(W;Y^n) &= H(Y^n) - H(Y^n|W) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W) \quad \text{Chain rule of } H \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W, X_i) \quad \text{Since } X_i = f\left(W, Y_{i-1}, \ldots, Y_1\right) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \leq nC \quad \text{Repeated use of channel does not increase capacity}
\end{aligned}
$$

Put altogether to obtain

$$nR \leq 1 + P_e^{(n)} nR + nC,$$

so when $n \to \infty$, $R \leq C$. □

# Example

Feedback cannot provide higher rates, but it helps in simplifying encoding and decoding in practical systems.

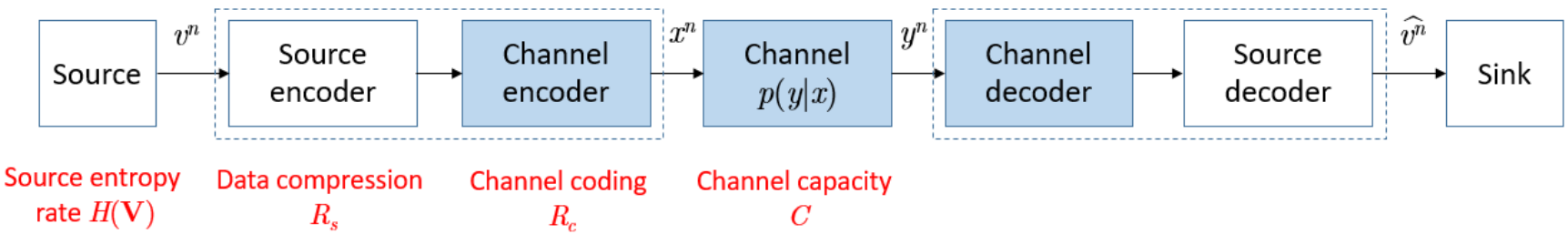
As an example, let us use feedback in the binary erasure channel: transmit the same bit through the channel until an erasure does not occur and the information bit is received correctly.

Can you compute the average number of uses of the channel it takes to transmit an information bit through this channel?

# Joint source-channel coding

Now we shall combine the fundamental results of channel coding $R_c \leq C$ and source coding $R_s \geq H$: is the condition $H \leq C$ necessary and sufficient?

If so, we can separately design a source encoder that encodes to $H$ bits/symbol and a channel encoder that encodes to $C$ bits/tr.



# Joint source-channel coding

**Theorem (Source-channel coding)**

*Let $\mathbf{V} = V_1, V_2, ..., V_n \in \mathcal{V}^n$ be a finite alphabet ergodic process that satisfies the AEP. Then*

1. *Achievability. If $H(\mathbf{V}) \leq C$ there exist a source-channel code with zero probability of error: $\Pr(\hat{v}^n \neq v^n) \to 0$.*
2. *Converse. If $H(\mathbf{V}) > C$ the probability of error is bounded away from zero.*

Proof of achievability. The ergodic source process $\mathbf{V}$ is mapped to codewords through an encoding rule $x^n(v^n)$ and sent over the channel. The receiver observes $y^n$ and makes a guess $\hat{v}^n$. An error is declared if $\hat{v}^n \neq v^n$. We shall follow three steps:

**i**. The size of the typical set $\mathcal{A}_\epsilon^n$ for $\mathbf{V}$ is $2^{n(H(\mathbf{V})+\epsilon)}$ so we need $n(H(\mathbf{V})+\epsilon)$ bits to encode it. Encoding non-typical sequences will entail an error $\epsilon \to 0$ (remember lecture 3).

# Joint source-channel coding

**ii**. We can transmit the index to the typical set with arbitrary low error if $H(\mathbf{V}) + \epsilon = R_s \leq C$ (according to the channel capacity theorem).

**iii**. The receiver can reconstruct $v^n$ by enumerating the typical set and choosing a sequence that matches the transmitted one with high probability:

$$\Pr\left(v^n \neq \hat{v}^n\right) \leq \Pr\left(v^n \notin \mathcal{A}_\epsilon^n\right) + \Pr\left(g(y^n) \neq v^n | v^n \in \mathcal{A}_\epsilon^n\right)$$

for large $n$. Therefore we can reconstruct the original sequence with low probability of error if $H(\mathbf{V}) \leq C$.

Proof of converse. Let us first identify the Markov's chain

$$\mathbf{V} \to X^n \to Y^n \to \hat{\mathbf{V}}$$

By Fano's inequality

$$H(\mathbf{V}|\hat{\mathbf{V}}) \leq 1 + \Pr\left(\hat{v}^n \neq v^n\right) \log |\mathcal{V}^n| = 1 + \Pr\left(\hat{v}^n \neq v^n\right) n \log |\mathcal{V}|$$

# Joint source-channel coding

Let us apply it to bound the error

$$
\begin{aligned}
H(\mathbf{V}) &\leq \frac{1}{n} H(V_1, ..., V_n) \\
&= \frac{1}{n} H(V^n|\hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n) \\
&\leq \frac{1}{n} \left(1 + \Pr\left(\hat{v}^n \neq v^n\right) n \log |\mathcal{V}|\right) + \frac{1}{n} I(V^n; \hat{V}^n) \quad \text{Fano's theorem} \\
&\leq \frac{1}{n} \left(1 + \Pr\left(\hat{v}^n \neq v^n\right) n \log |\mathcal{V}|\right) + \frac{1}{n} I(X^n; \hat{Y}^n) \quad \text{Data processing inequality} \\
&\leq \frac{1}{n} + \Pr\left(\hat{v}^n \neq v^n\right) \log |\mathcal{V}| + C \quad \text{Capacity of repeated use of channel (see annex)}
\end{aligned}
$$

By letting $n \to \infty$, if $\Pr\left(\hat{v}^n \neq v^n\right) \to 0$ then $H(\mathbf{V}) \leq C$. $\square$

# Conclusions

- The data compression theorem is based on the AEP: there exists a "small" subset (of size $2^{nH}$) of all possible source sequences that contain most of the probability and hence we can represent the source with a small probability of error using $H$ bits/symbol.
- The data transmission theorem is based on the joint AEP: for long block lengths the output sequence of the channel is very likely to be jointly typical with the input codeword, while any other codeword is jointly typical with probability $\simeq 2^{-nI}$. Hence, we can use about $2^{nI}$ codewords and still have negligible probability of error.
- The source–channel separation theorem shows that we can design the source code and the channel code separately and use them together to achieve optimal performance as long as $H \leq C$.

# Way through...

- Chapter 6 introduces practical channel codes that achieve low probability of error and aproach capacity.
- Other relevant aspects of capacity not considered in this course are:
  - How the probability of error decreases as a function of $n$ (i.e. error exponents, sphere packing bounds)
  - Optimization with respect to $p(x)$ (e.g. using Blahut-Arimoto algorithm)

# Proof of theorem 5.1

Proof. This proves that with high probability the sequences $(X^n, Y^n)$ of length $n$ are jointly typical. By the weak law of large numbers

$$-\frac{1}{n}\log p(X^n) \to -E\left[\log p(X)\right] = H(X)$$

Hence, given $\epsilon > 0$, there exist $n_1, n_2, n_3$ such that for all $n > n_1$, $n > n_2$, $n > n_3$

$$\Pr\left(\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \geq \epsilon\right) < \frac{\epsilon}{3}$$

$$\Pr\left(\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3}$$

$$\Pr\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3}$$

respectively. Then, by choosing $n > max(n_1, n_2, n_3)$, and using $\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B)$, the probability of the intersection of the three sets must be less than $\epsilon$. Hence, for $n$ sufficiently large, the probability of the set $\mathcal{A}_\epsilon^{(n)}$ is greater than $1 - \epsilon$. □

# Proof of theorem 5.2

Proof. For the upper bound,

$$1 = \sum_{(x^n,y^n)\in\{\mathcal{X}^n,\mathcal{Y}^n\}} p(x^n,y^n) \geq \sum_{(x^n,y^n)\in\mathcal{A}_\epsilon^{(n)}} p(x^n,y^n) \geq 2^{-n(H(X,Y)+\epsilon)} \left|\mathcal{A}_\epsilon^{(n)}\right|$$

For the lower bound, take theorem 5.1, and if $n$ is sufficiently large

$$1-\epsilon < \Pr\left(\mathcal{A}_\epsilon^{(n)}\right) \leq \sum_{(x^n,y^n)\in\mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X,Y)-\epsilon)} = 2^{-n(H(X,Y)-\epsilon)} \left|\mathcal{A}_\epsilon^{(n)}\right| \quad \square$$

The set may be small, its size depends on the joint entropy of $X$ and $Y$.

# Proof of theorem 5.3

Proof. Under the conditions stated in the theorem, and using theorem 5.2

$$\begin{aligned}\Pr\left((\tilde{X}^n,\tilde{Y}^n)\in\mathcal{A}_\epsilon^n\right) &= \sum_{(x^n,y^n)\in\mathcal{A}_\epsilon^{(n)}} p(x^n)p(y^n)\\ &\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}\\ &= 2^{-n(I(X;Y)-3)\epsilon}\end{aligned}$$

Similarly we can also prove using theorem 5.2

$$\begin{aligned}\Pr\left((\tilde{X}^n,\tilde{Y}^n)\in\mathcal{A}_\epsilon^n\right) &= \sum_{(x^n,y^n)\in\mathcal{A}_\epsilon^{(n)}} p(x^n)p(y^n)\\ &\geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)}\\ &= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}\end{aligned}$$

□

# Repeated use of channel

**Theorem (Capacity of repeated use of channel)**

*Let $Y^n$ be the result of passing $X^n$ through a discrete memoryless channel of capacity $C$. Then $I(X^n;Y^n) \leq nC$ for all $p(x^n)$.*

Proof.

$$
\begin{aligned}
I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X^n) \quad \text{by chain rule of entropy} \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{by definition of a memoryless channel} \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{by definition of joint entropy} \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \quad \text{by definition of mutual information} \\
&\leq nC
\end{aligned}
$$

□