Capacity of a channel – Shannon's second theorem

Outline

1. Memoryless channels, examples ;
2. Capacity ;
3. Symmetric channels ;
4. Channel Coding ;
5. Shannon's second theorem, proof.

1. Memoryless channels

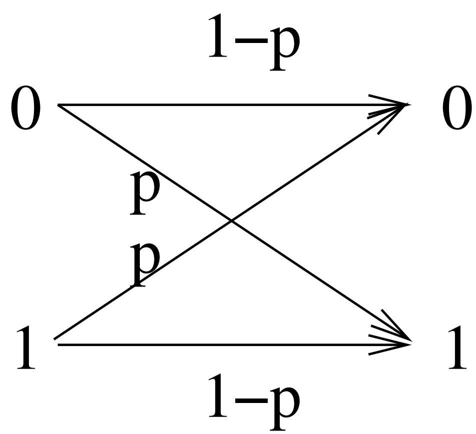
Definition A discrete channel is given by
- an input alphabet $X = \{a_1, \ldots, a_K\}$
- an output alphabet $Y = \{b_1, \ldots, b_J\}$
- transition probabilities $P_{Y|X}$, i.e. a stochastic matrix

$$\Pi = \begin{pmatrix} \mathbf{P}(b_1 \mid a_1) & \ldots & \mathbf{P}(b_J \mid a_1) \\ \vdots & \ddots & \vdots \\ \mathbf{P}(b_1 \mid a_K) & \ldots & \mathbf{P}(b_J \mid a_K) \end{pmatrix}$$

The channel is memoryless if for all transmitted $(x_1, \ldots, x_n)$ and all received $(y_1, \ldots, y_n)$, we have

$$\mathbf{P}(y_1, \ldots, y_n | x_1, \ldots, x_n) = \mathbf{P}(y_1|x_1) \ldots \mathbf{P}(y_n|x_n).$$

Example – Binary symmetric channel

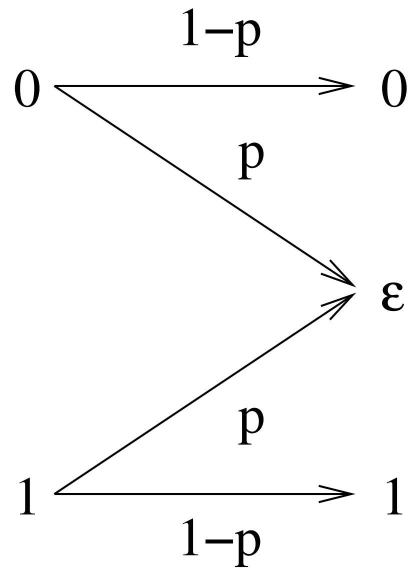


The stochastic matrix is

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

p is called the *crossover probability* of the channel.

Erasure channel



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

2. Capacity

The capacity of a channel is defined by the maximum mutual information between a random variable X taking its values on the input alphabet and the corresponding output Y of the channel

$$C \stackrel{\text{def}}{=} \sup_X I(X;Y) \text{ with}$$

$$X \stackrel{\text{channel}}{\rightsquigarrow} Y$$

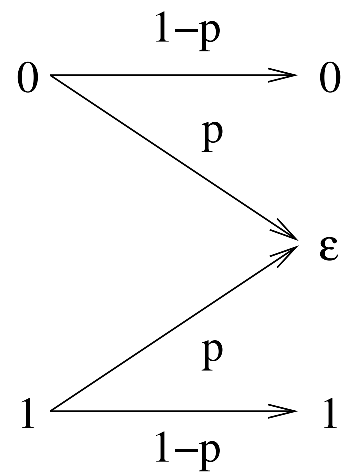
Capacity

It is useful to note that $I(X;Y)$ can be written as follows (using only the input distribution and the transition probabilities)

$$I(X;Y) = \sum_{x,y} \mathbf{P}(y \mid x)\mathbf{P}(x) \log_2 \frac{\mathbf{P}(y \mid x)}{\mathbf{P}(y)}$$

$$\mathbf{P}(y) = \sum_{x} \mathbf{P}(y \mid x)\mathbf{P}(x).$$

Capacity of a binary erasure channel



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

Observe that $H(Y|X) = \mathbf{P}(X=0)h(p) + \mathbf{P}(X=1)h(p) = h(p)$ with $h(p) \stackrel{\text{def}}{=} -p\log_2 p - (1-p)\log_2(1-p)$.

Capacity of the binary erasure channel (II)

Letting $a \stackrel{\text{def}}{=} \mathbf{P}(X = 1)$, we obtain :

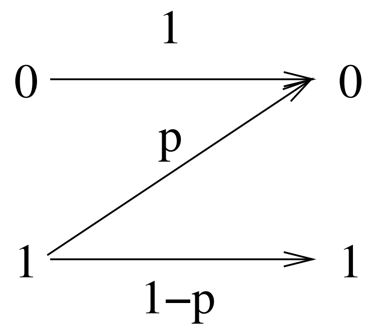
$$\begin{aligned}
\mathbf{P}(Y = 1) &= a(1-p) \\
\mathbf{P}(Y = 0) &= (1-a)(1-p) \\
\mathbf{P}(Y = \epsilon) &= ap + (1-a)p = p
\end{aligned}$$

$$\begin{aligned}
H(Y) &= -a(1-p)\log a(1-p) \\
&\quad -(1-a)(1-p)\log(1-a)(1-p) - p\log p \\
&= (1-p)h(a) + h(p)
\end{aligned}$$

Therefore

$$C = \max_a (1-p)h(a) = 1-p$$

Z-channel



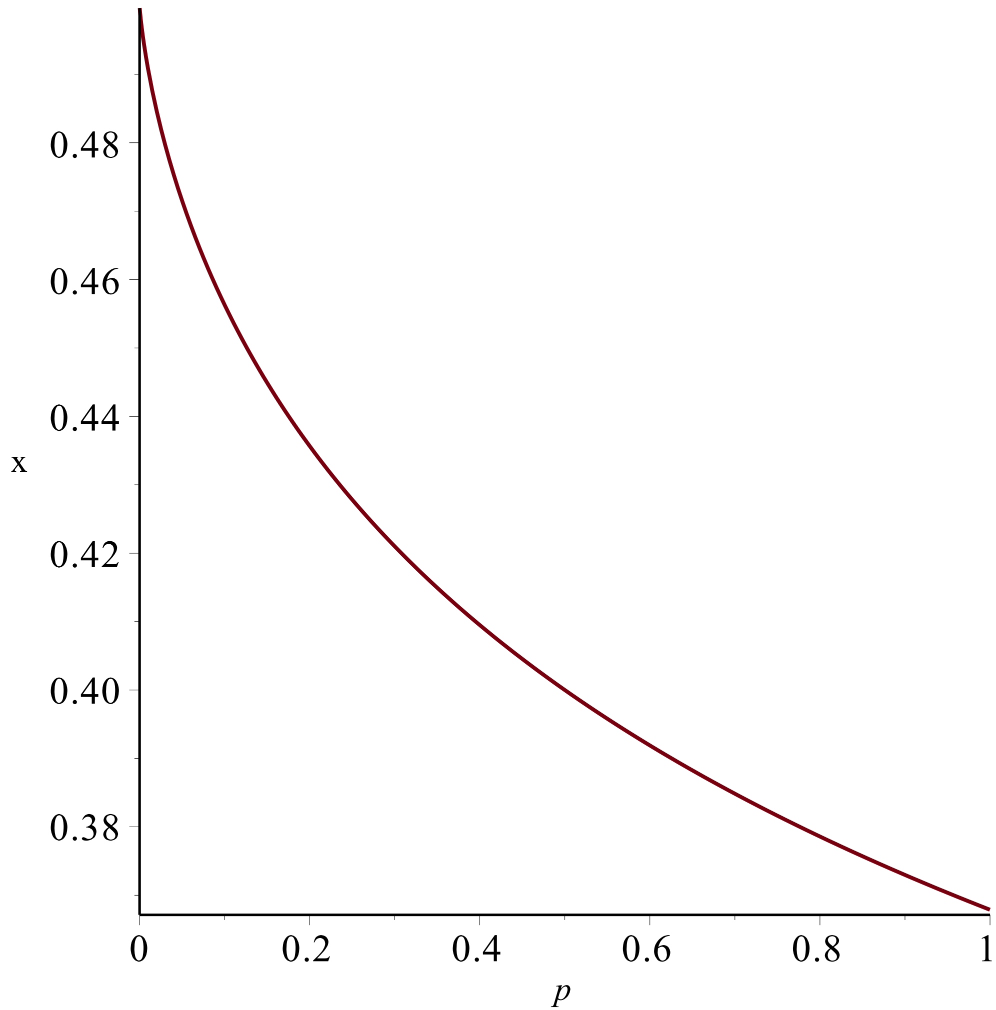
The stochastic matrix is

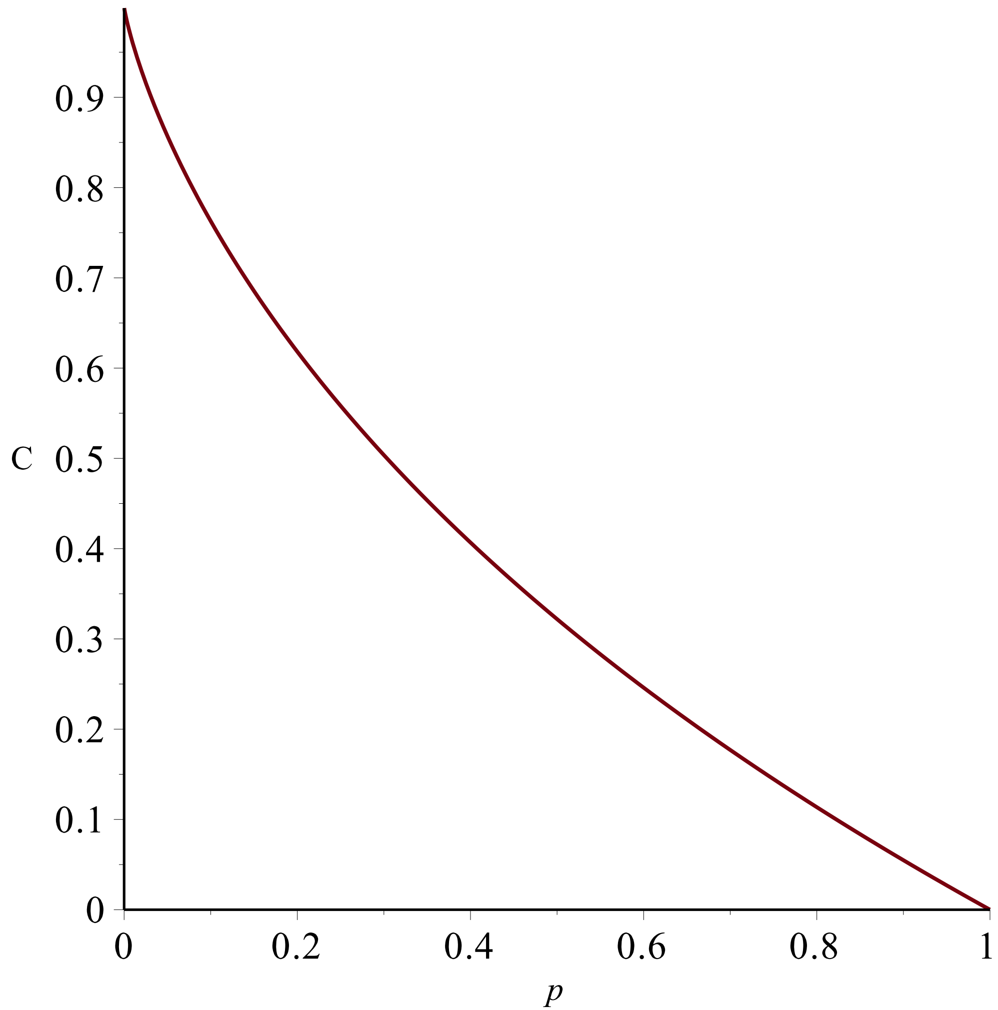
$$\begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}.$$

For a distribution $x \stackrel{\text{def}}{=} \mathbf{P}(X=1)$, we have

$$I(X;Y) = h(x(1-p)) - xh(p)$$

Maximum attained in $x = \left((1-p)\left(1+2^{\frac{h(p)}{1-p}}\right)\right)^{-1}$


0.48
0.46
0.44
x
0.42
0.40
0.38
0
0.2
0.4
0.6
0.8
1
p


C
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
p

Symmetric channels

Definition A discrete memoryless channel is *symmetric* if each row/column is a permutation of the first row/column.

Proposition 1. *In a symmetric channel $H(Y|X)$ does not depend on X.*

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

Proof

$$\begin{aligned} H(Y \mid X) &= -\sum_x P(x) \sum_y P(y \mid x) \log_2 P(y \mid x) \\ &= -\sum_x P(x) H(\Pi) = H(\Pi) \end{aligned}$$

where $H(\Pi) = -\sum_y P(y \mid x) \log_2 P(y \mid x)$ is independent of x.

Capacity of a symmetric channel

Hence

$$
\begin{aligned}
C &= \sup_X I(X;Y) \\
&= \sup(H(Y) - H(Y \mid X)) \\
&= \sup(H(Y)) - H(\Pi) \\
&\leq \log_2 |Y| - H(\Pi).
\end{aligned}
$$

The entropy is maximised when Y is uniform. Note that Y is uniform when X is uniform for a symmetric channel.

Capacity of a symmetric channel (II)

Proposition 2. *The capacity of a symmetric channel is attained for a uniform distribution on the inputs and is equal to*

$$C = \log_2 |Y| - H(\Pi)$$

Example

Capacity of a binary symmetric channel

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

To compar with the capacity of the binary erasure channel :

$$C = 1 - p$$

3. Channel coding

Let us consider a discrete memoryless channel $\mathcal{T} = (X, Y, \Pi)$

Definition A *block code* of *length* n and of *cardinality* M is a set of M sequences of n symbols of X. It is an (M, n)-code. Its elements are called codewords. The *code rate* is equal to

$$R = \frac{\log_2 M}{n}$$

Such a code allows to encode $\log_2 M$ bits per codeword transmission.

R is also equal to the number of transmitted bits per channel use.

An *encoder* is a procedure that maps a finite binary sequence to a finite sequence of elements of X.

Code performance – Decoding

Let $\mathcal{C}$ be an (M, n)-block code used over a discrete memoryless channel (X, Y, Π)

Definition A *decoding algorithm* for $\mathcal{C}$ is a procedure which maps any block of n symbols of Y to a codeword in $\mathcal{C}$.

The event "bad decoding" for a decoding algorithm is defined by :

> *A codeword* $\mathbf{x} \in \mathcal{C} \subset X^n$ *is transmitted through the channel, the word* $\mathbf{y} \in Y^n$ *is received and is decoded in* $\tilde{\mathbf{x}} \neq \mathbf{x}$*.*

Definition The *error rate of $\mathcal{C}$* (for a given channel and sent codeword $\mathbf{x}$) denoted by $P_e(\mathcal{C}, \mathbf{x})$ is the probability of bad decoding when $\mathbf{x}$ is transmitted.

Examples for the binary symmetric channel

Repetition code of length 3

$$C = \{000, 111\}$$

Single Parity-check code of length 4

$$C = \{0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111\}$$

Hamming code of length 7

$$
\begin{aligned}
C \quad = \quad & \{0000000, 1101000, 0110100, 0011010, \\
& 0001101, 1000110, 0100011, 1010001, \\
& 1111111, 0010111, 1001011, 1100101, \\
& 1110010, 0111001, 1011100, 0101110\}
\end{aligned}
$$

Decoding of the Hamming code

The Hamming distance $d(x, y)$ is equal to

$$d(x, y) = |\{i; x_i \neq y_i\}|$$

It can be verified that all codewords of the Hamming code are at distance at least 3 from each other. This implies that the balls of radius 1 centered around each codeword do not intersect.

Moreover any binary word of length 7 is at distance at most 1 from a Hamming codeword, since

$$16(1 + 7) = 2^4 \times 2^3 = 2^7.$$

Decoding algorithm : when y is received, return the codeword x at distance ≤ 1 from y.

Shannon's second theorem

Theorem 1. *Consider a discrete memoryless channel of capacity C. For all $R < C$, there exists a sequence of block codes $(\mathcal{C}_n(M, n))_{n>0}$ of rate R_n together with a decoding algorithm such that*

$$\lim_{n\to\infty} R_n = R \quad \textit{and} \quad \lim_{n\to\infty} \sup_{\mathbf{x}\in\mathcal{C}_n} P_e(\mathcal{C}_n, \mathbf{x}) = 0$$

Theorem 2. *Consider a discrete memoryless channel of capacity C. Any code $\mathcal{C}$ of rate $R > C$ satisfies $\frac{1}{M}\sum_{\mathbf{x}\in\mathcal{C}} P_e(\mathcal{C}, \mathbf{x}) > K(C, R)$, where $K(C, R) > 0$ depends on the channel and the rate but is independent of the code.*

Error exponent

There is even a stronger version of Shannon's theorem : there are block codes of rate R and length n for which

$$\sup_{x} P_e(\mathcal{C}, x) \approx \mathbf{e}^{-nE(R)}$$

where $E(R)$ is called the *error exponent*. It depends on the channel and the transmission rate and satisfies

$$E(R) > 0 \text{ if } R < C$$

Jointly typical sequences

Definition [Jointly typical set] Let $(X^{(n)}, Y^{(n)})$ be a pair of r.v. taking its values in a discrete set $\mathcal{A}^n \times \mathcal{B}^n$, $p(\mathbf{x}, \mathbf{y})$ be the joint probability distribution of $(X^{(n)}, Y^{(n)})$, and let $p(\mathbf{x})$ and $p(\mathbf{y})$ be the probability distribution of $X^{(n)}$ and $Y^{(n)}$ respectively. The set of jointly typical sequences $\mathcal{T}_\epsilon^{(n)}$ is given by

$$
\begin{aligned}
\mathcal{T}_\epsilon^{(n)} = \; & \{(\mathbf{x}, \mathbf{y}) \in \mathcal{A}^n \times \mathcal{B}^n : \\
& \left| \frac{1}{n} \left(-\log_2 p(\mathbf{x}) - H(X^{(n)}) \right) \right| < \epsilon \qquad (1) \\
& \left| \frac{1}{n} \left(-\log_2 p(\mathbf{y}) - H(Y^{(n)}) \right) \right| < \epsilon \qquad (2) \\
& \left| \frac{1}{n} \left(-\log_2 p(\mathbf{x}, \mathbf{y}) - H(X^{(n)}, Y^{(n)}) \right) \right| < \epsilon \Bigg\} \qquad (3)
\end{aligned}
$$

Theorem 3. *Let (X_i, Y_i) be a sequence of pairs of i.i.d. r.v. taking their values in $\mathcal{A} \times \mathcal{B}$ distributed as a fixed pair (X, Y). We define $\mathcal{T}_\epsilon^{(n)}$ from $(X^{(n)}, Y^{(n)})$ with $X^{(n)} \stackrel{\text{def}}{=} (X_1, X_2, \ldots, X_n)$ and $Y^{(n)} \stackrel{\text{def}}{=} (Y_1, Y_2, \ldots, Y_n)$. Then*

1. $\mathbf{Prob}(\mathcal{T}_\epsilon^{(n)}) > 1 - \epsilon$ *for n sufficiently large.*
2. $|\mathcal{T}_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *Let $\tilde{X}^{(n)}$ and $\tilde{Y}^{(n)}$ be 2 independent r.v. with $\tilde{X}^{(n)} \sim X^{(n)}$ and $\tilde{Y}^{(n)} \sim Y^{(n)}$. Then,*

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

Moreover, for n sufficiently large

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

Proof of Point 3.

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

Proof of Point 3. (II)

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

The direct part of Shannon's theorem

The crucial point : random choice of the code !

We begin by choosing a probability distribution $\mathbf{P}$ on the input symbols of the channel. Then we choose a code of length n and rate R by drawing 2^{nR} words in $\mathcal{A}^n$ randomly according to the distribution $\mathbf{P}^{(n)}$ on $\mathcal{A}^n$ given by

$$\mathbf{P}^{(n)}(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} \mathbf{P}(x_i).$$

Typical set decoding

$\mathbf{x}$ transmitted word, $\mathbf{y}$ received word. Let $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{2^{nR}}$ be the 2^{nR} codewords.

1. compute the 2^{nR} probabilities $\mathbf{P}$(received word $= \mathbf{y}$, transmitted word $= \mathbf{x}^s$) for $s \in \{1, \ldots, 2^{nR}\}$.
2. if more than one pair or no pair $(\mathbf{x}^i, \mathbf{y})$ is ϵ-jointly typical $\rightarrow$ "decoding failure".
3. otherwise output $\mathbf{x}_s$ such that $(\mathbf{x}_s, \mathbf{y})$ is jointly typical.

Analysis of the decoder

This decoder can fail for two reasons
- the right pair $(\mathbf{x}, \mathbf{y})$ is not jointly typical (event $\mathcal{E}_0$),
- At least one of the $2^{nR} - 1$ pairs $(\mathbf{x}^s, \mathbf{y})$ is jointly typical with $\mathbf{x}^s \neq \mathbf{x}$, (event $\mathcal{E}_1$).

$$\begin{aligned}\mathbf{Prob}(\text{decoding failure}) &= \mathbf{Prob}(\mathcal{E}_0 \cup \mathcal{E}_1) \\ &\leq \mathbf{Prob}(\mathcal{E}_0) + \mathbf{Prob}(\mathcal{E}_1)\end{aligned}$$

The probability of the first event

$$\begin{aligned}\mathbf{Prob}(\mathcal{E}_0) &= \mathbf{Prob}\left((X^{(n)}, Y^{(n)}) \text{ is not typical}\right) \\ &= 1 - \mathbf{Prob}(\mathcal{T}_\epsilon^{(n)})\end{aligned}$$

By using Point 1. of Theorem 3, we obtain $1 - \mathbf{Prob}(\mathcal{T}_\epsilon^{(n)}) \leq \epsilon$.

The probability of the second event

$$\begin{aligned}\mathbf{Prob}(\mathcal{E}_1) &= \mathbf{Prob}\left(\cup_{s:\mathbf{x}^s \neq \mathbf{x}} \{(\mathbf{x}^s, \mathbf{y}) \text{ is typical}\}\right) \\ &\leq \sum_{s:\mathbf{x}^s \neq \mathbf{x}} \mathbf{Prob}\left((\mathbf{x}^s, \mathbf{y}) \text{ is typical}\right)\end{aligned}$$

$(\tilde{X}^{(n)}, \tilde{Y}^{(n)})$ where $\tilde{X}^{(n)} \sim X^{(n)}$ $\tilde{Y}^{(n)} \sim Y^{(n)}$ and $(\tilde{X}^{(n)},\ \tilde{Y}^{(n)})$ independent

$$\mathbf{Prob}\left((\mathbf{x}^s, \mathbf{y}) \text{ is typical}\right) = \mathbf{Prob}\left((\tilde{X}^{(n)}, \tilde{Y}^{(n)}) \text{ is typical}\right)$$

The probability of the second event (II)

$$\mathbf{Prob}\left\{\left(\tilde{X}^{(n)}, \tilde{Y}^{(n)}\right) \text{ is typical}\right\} \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Therefore

$$\begin{aligned}\mathbf{Prob}(\mathcal{E}_1) &\leq (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq 2^{-n(I(X;Y)-R-3\epsilon)}.\end{aligned}$$

End of proof

$$\mathbf{Prob}(\text{decoding failure}) \leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}.$$

End : choosing X s.t. $C = I(X;Y)$.