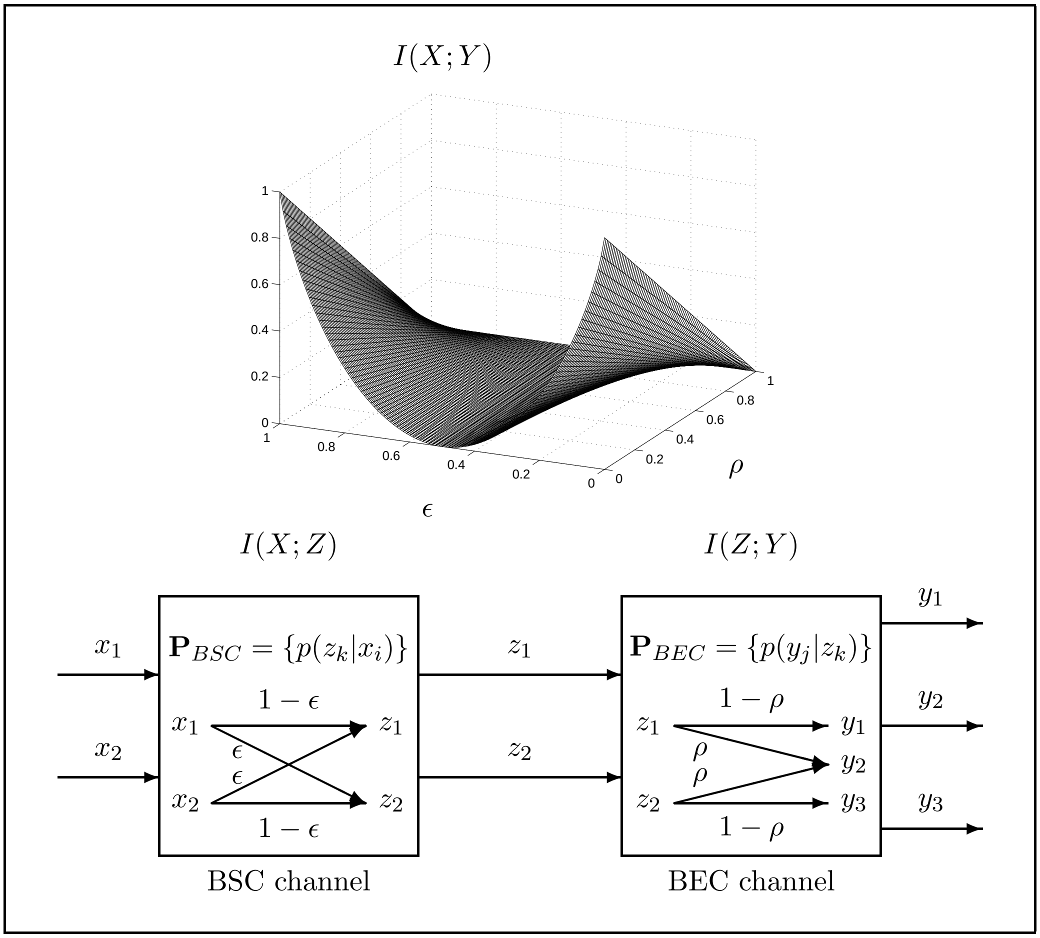
ELG 5170 Information Theory

Course Notes



by

Dr. Jean-Yves Chouinard

School of Information Technology and Engineering, University of Ottawa

April 2001

Contents

List of Figures

List of Tables

Chapter 1

Measure of Information

1.1 Self-information and source entropy

Source model:

Consider a source of information that generates at each instant a message x_i from a set $\mathcal{X} = \{x_i\}$, called the source alphabet:

$$\{x_i\} = \{x_1, x_2, \ldots, x_N\}$$

where N is the alphabet size.

Example 1 *(binary source):*

$$x_i \in \{0, 1\} \Longrightarrow \text{alphabet size } N = 2$$

Example 2 *(standard English alphabet):*

$$x_i \in \{a, b, c, \ldots, x, y, z\} \Longrightarrow \text{alphabet size } N = 26$$

These single letters are also called monograms.

Example 3 *(digrams):*

$$x_i \in \{aa, ab, ac, \ldots, zx, zy, zz\} \Longrightarrow \text{alphabet size } N = 26^2$$

Example 4 *(trigrams):*

$$x_i \in \{aaa, aab, aac, \ldots, zzx, zzy, zzz\} \Longrightarrow \text{alphabet size } N = 26^3$$

Example 5 *(US ASCII):*

$$x_i \in \{(000\ 0000), (000\ 0001), \ldots, (111\ 1111)\} \Longrightarrow \text{alphabet size } N = 128$$

USA Standard Code for Information Exchange (USASCII)								
b_4 b_3 b_2 b_1	b_7 b_6 b_5 0 0 0	b_7 b_6 b_5 0 0 1	b_7 b_6 b_5 0 1 0	b_7 b_6 b_5 0 1 1	b_7 b_6 b_5 1 0 0	b_7 b_6 b_5 1 0 1	b_7 b_6 b_5 1 1 0	b_7 b_6 b_5 1 1 1
0 0 0 0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	STX	DC2	″	2	B	R	b	r
0 0 1 1	ETX	DC3	#	3	C	S	c	s
0 1 0 0	EOT	DC4	$	4	D	T	d	t
0 1 0 1	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	ACK	SYN	&	6	F	V	f	v
0 1 1 1	BEL	ETB	´	7	G	W	g	w
1 0 0 0	BS	CAN	(	8	H	X	h	x
1 0 0 1	HT	EM	)	9	I	Y	i	y
1 0 1 0	LF	SUB	*	:	J	Z	j	z
1 0 1 1	VT	ESC	+	;	K	[	k	{
1 1 0 0	FF	FS	,	<	L	\	l	|
1 1 0 1	CR	GS	-	=	M	]	m	}
1 1 1 0	SO	RS	.	>	N	∧	n	∼
1 1 1 1	SI	US	/	?	O	_	o	DEL

Example 6 *(DES 64-bit blocks):*

$$x_i \in \{(000\ldots00), (000\ldots01), \ldots, (111\ldots11)\} \Longrightarrow \text{alphabet size } N = 2^{64}$$

This corresponds to the plaintext alphabet for the Data Encryption Standard (DES).

The less likely an event is expected to occur, the more information one obtains when this particular event happens. With each message x_i is associated a probability of occurrence $p(x_i)$. The amount of uncertainty of a particular message x_i is termed *self-information* and is defined as:

$$I_X(x_i) = \log_b \frac{1}{p(x_i)} \qquad (1.1)$$

$$\boxed{I_X(x_i) = -\log_b p(x_i)}$$

Note: The choice of the logarithmic base, i.e. b, determines the units of the information measure:

$b = 2$ *Sh* (*shannons*) or *bit* (*binary digit*)

$b = e$ *logons* or *nats* (*natural units*)

$b = 10$ *hartleys* (in honor of R.V.L. Hartley, pioneer in communication and information theory)

Whether base $b = 2, e$ or10, the information quantities are obviously the same. The conversion between the different bases is given by:

$$\begin{array}{rcccccr} 1\ \textit{logon} & = & \frac{1}{\ln 2} & = & \log_2 e & = & 1.443\ \textit{Sh} \\ 1\ \textit{hartley} & = & \frac{1}{\log_{10} 2} & = & \log_2 10 & = & 3.322\ \textit{Sh} \\ 1\ \textit{hartley} & = & \frac{1}{\log_{10} e} & = & \ln 10 & = & 2.303\ \textit{logons} \end{array} \tag{1.2}$$

A message x_i can take only one of the N possible values from the set of messages, or sample space $\mathcal{X}$, defined as the source alphabet:

$$\mathcal{X} \equiv \{x_1, x_2, \ldots, x_N\} \tag{1.3}$$

and the sum of the probabilities of occurrence of all the messages is equal to unity:

$$\sum_{i=1}^{N} p(x_i) = 1 \tag{1.4}$$

Definition *(Entropy):*

The entropy $H(X)$ of a source of information X (i.e. random variable) is defined as the weighted sum (or average) of the self-information of each message x_i from that source:

$$H(X) = \sum_{i=1}^{N} p(x_i)\ I_X(x_i) \tag{1.5}$$

$$\boxed{H(X) = -\sum_{i=1}^{N} p(x_i)\ \log_b p(x_i)}$$

Example *(quaternary source distribution):*

As an example, lets consider the distribution of quaternary sources. Let the distribution of a quaternary source be: $p(x_1) = \frac{1}{2}$, $p(x_2) = p(x_4) = \frac{1}{8}$, and $p(x_3) = \frac{1}{4}$. The self-information of each message is then:

$$\begin{array}{llll} I_X(x_1) & = - \ \log_2 \frac{1}{2} & = 1 \ Sh & \text{(shannon)} \\ I_X(x_2) = I_X(x_4) & = - \ \log_2 \frac{1}{8} & = 3 \ Sh & \\ I_X(x_3) & = - \ \log_2 \frac{1}{4} & = 2 \ Sh & \end{array}$$

As can be seen from the above equation, the more likely the event (i.e. the message), the less uncertainty its occurrence resolves. In other words, as the probability of an event increases, its corresponding self-information decreases. The entropy of this source of information is obtained by averaging the self-information over the set of messages:

$$H(X) = \sum_{i=1}^{4} p(x_i) \ I_X(x_i) = 1.75 \ Sh$$

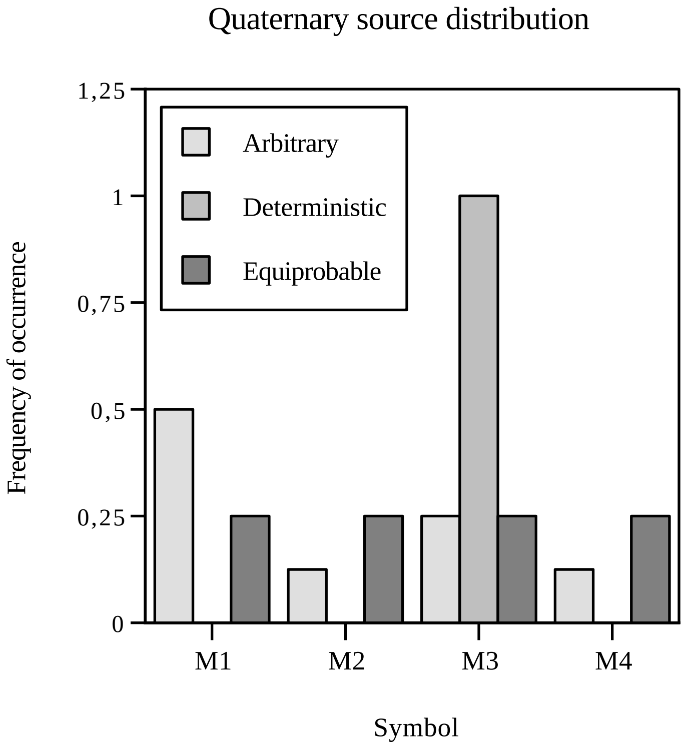


Figure 1.1: Example of quaternary source distributions (arbitrary, deterministic, and equiprobable).

Now, suppose that the quaternary source distribution has changed to the following: $p(x_1) = p(x_2) = p(x_4) = 0$, and $p(x_3) = 1$. This constitutes the special case of a deterministic source where it is *certain* that the third symbol x_3 will always occurs while x_1, x_2 and x_4 never occur. The self-information of symbol x_3 is simply $I_X(x_3) = - \ \log_2 1 = 0 \ Sh$ and thus the entropy $H(X)$ is also equal to 0 Sh. The observation of that source does not provide any additional information.

Finally, let the quaternary source have an equiprobable distribution, that is each symbol are produced with the same probability: $p(x_1) = p(x_2) = p(x_3) = p(x_4) = \frac{1}{4}$. the self-information is the same for the four symbols

$$I_X(x_1) = I_X(x_2) = I_X(x_3) = I_X(x_4) = -\ \log_2 \frac{1}{4} = 2\ Sh$$

and the entropy is simply:

$$H(X) = 2\ Sh$$

An equiprobable source is the source of information which provides the most uncertainty. This result is important in cryptography: the security of a cryptosystem is increased if the choice of encryption keys is equiprobable.

Example*(Standard alphabet distribution):*

The entropy of English language can be determined from the frequency of occurrence of the individual letters, as a first approximation. Table 1.1 indicates the relative frequencies of letters in English and French languages including the space character (represented by □).

$$H_{English}(X) = -\sum_{i=1}^{27} p(x_i)\ \log_2 p(x_i) = 4.0755\ Sh$$

By comparison, French language has a slightly lower entropy (it has slightly less uniform letter distribution).

$$H_{French}(X) = -\sum_{i=1}^{27} p(x_i)\ \log_2 p(x_i) = 3.9568\ Sh$$

Suppose that there exists a 27-letter language for which each letter x_i is equally probable, that is $P(x_i) = \frac{1}{27}$ for $1 \leq i \leq 27$. Then this new source's entropy is given by:

$$H_{Equiprobable}(X) = -\sum_{i=1}^{27} \left(\frac{1}{27}\right)\ \log_2 \left(\frac{1}{27}\right) = -\ \log_2 \left(\frac{1}{27}\right) = 4.7549\ Sh$$

which is the highest achievable entropy for a 27-letter alphabet.

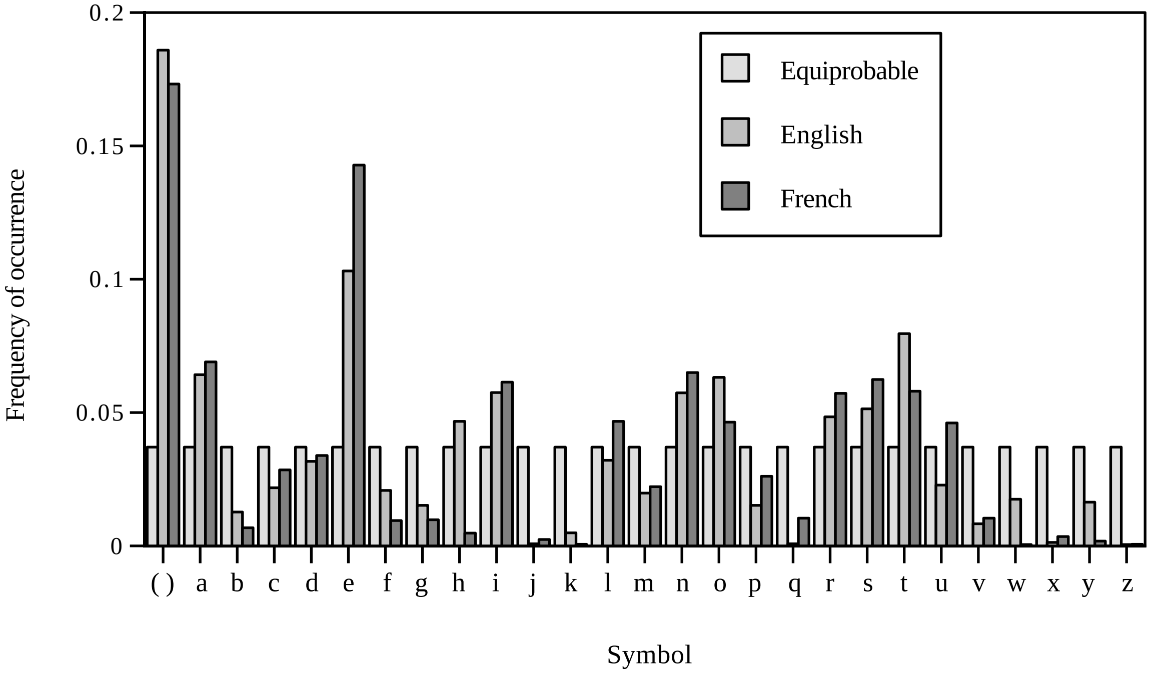


Figure 1.2: Letter distribution of standard alphabet (equiprobable, English, and French).

Table 1.1: Relative frequencies of letters for an equiprobable source, and English and French languages (alphabet size = 27).

Letter	Equiprobable		English language		French language	
x_i	$p(x_i) = \frac{1}{27}$	$-\log_2 \frac{1}{27}$	$p(x_i)$	$-\log_2 p(x_i)$	$p(x_i)$	$-\log_2 p(x_i)$
□	0.0370	4.7549	0.1859	2.4274	0.1732	2.5295
a	0.0370	4.7549	0.0642	3.9613	0.0690	3.8573
b	0.0370	4.7549	0.0127	6.2990	0.0068	7.2002
c	0.0370	4.7549	0.0218	5.5195	0.0285	5.1329
d	0.0370	4.7549	0.0317	4.9794	0.0339	4.8826
e	0.0370	4.7549	0.1031	3.2779	0.1428	2.8079
f	0.0370	4.7549	0.0208	5.5873	0.0095	6.7179
g	0.0370	4.7549	0.0152	6.0398	0.0098	6.6730
h	0.0370	4.7549	0.0467	4.4204	0.0048	7.7027
i	0.0370	4.7549	0.0575	4.1203	0.0614	4.0256
j	0.0370	4.7549	0.0008	10.2877	0.0024	8.7027
k	0.0370	4.7549	0.0049	7.6730	0.0006	10.7027
l	0.0370	4.7549	0.0321	4.9613	0.0467	4.4204
m	0.0370	4.7549	0.0198	5.6584	0.0222	5.4933
n	0.0370	4.7549	0.0574	4.1228	0.0650	3.9434
o	0.0370	4.7549	0.0632	3.9839	0.0464	4.4297
p	0.0370	4.7549	0.0152	6.0398	0.0261	5.2598
q	0.0370	4.7549	0.0008	10.2877	0.0104	6.5873
r	0.0370	4.7549	0.0484	4.3688	0.0572	4.1278
s	0.0370	4.7549	0.0514	4.2821	0.0624	4.0023
t	0.0370	4.7549	0.0796	3.6511	0.0580	4.1078
u	0.0370	4.7549	0.0228	5.4548	0.0461	4.4391
v	0.0370	4.7549	0.0083	6.9127	0.0104	6.5873
w	0.0370	4.7549	0.0175	5.8365	0.0005	10.9658
x	0.0370	4.7549	0.0013	9.5873	0.0035	8.1584
y	0.0370	4.7549	0.0164	5.9302	0.0018	9.1178
z	0.0370	4.7549	0.0005	10.9658	0.0006	10.7027

Example *(Binary source):*

Consider a binary source X with $0 < p(x_1) < 1$ and $p(x_2) = 1 - p(x_1)$. The entropy $H(X)$ of the source is then given by (see figure 1.3):

$$H(X) = -\left[p(x_1)\log_b p(x_1) + (1 - p(x_1))\log_b (1 - p(x_1))\right] \tag{1.6}$$

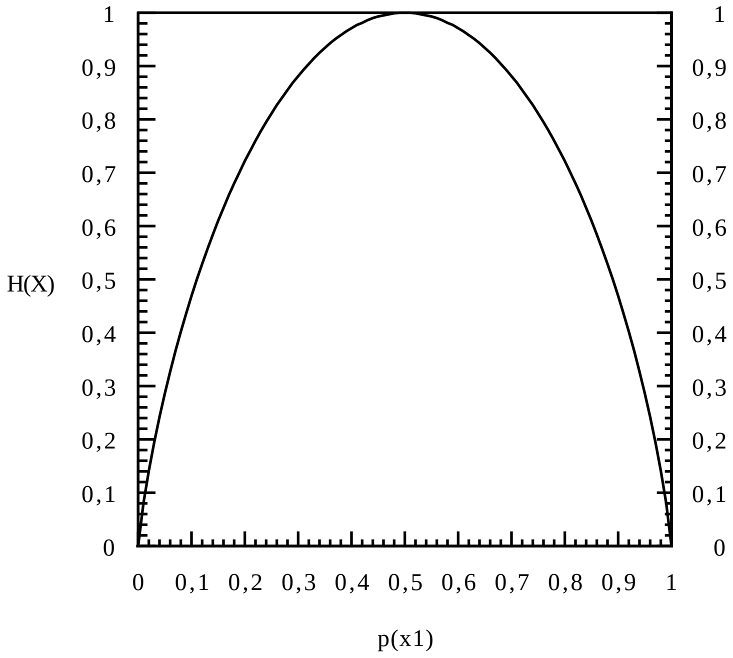


Figure 1.3: Entropy of a binary source as a function of its distribution.

1.2 Joint entropy and equivocation

Consider now two random variables X and Y having a joint probability density function (pdf) $p(x, y)$ (note that X and Y may happen to be independent). The *joint entropy* $H(XY)$ of X and Y is defined as:

$$H(XY) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log_b p(x,y) \tag{1.7}$$

$$\boxed{H(XY) = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i,y_j) \log_b p(x_i,y_j)} \tag{1.8}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the sample spaces of X and Y respectively.

The *conditional entropy* $H(X|Y)$, or *equivocation*, of a source X, given the observation of Y is defined as:

$$\boxed{H(X|Y) = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i,y_j) \log_b p(x_i|y_j)} \tag{1.9}$$

The equivocation $H(X|Y)$, or equivalently $H_Y(X)$, represents the *remaining* amount of uncertainty (or information) about X after the observation of Y.

Theorem *(Chain Rule):*

The joint entropy $H(XY)$ of a pair of random variables X and Y is equal to the sum of the entropy of X, that is $H(X)$, and the conditional entropy (or remaining uncertainty) of Y, given the observation of X.

$$\boxed{H(XY) = H(X) + H(Y|X)} \tag{1.10}$$

Proof:

Consider the sum $H(X) + H(Y|X)$:

$$\begin{aligned}
H(X) + H(Y|X) &= -\sum_{i=1}^{N} p(x_i) \log_b p(x_i) \\
&\quad -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(y_j|x_i) && (1.11) \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i) \\
&\quad -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(y_j|x_i) && (1.12) \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \left[\log_b p(x_i) + \log_b p(y_j|x_i)\right] && (1.13) \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b \left[p(x_i) p(y_j|x_i)\right] && (1.14) \\
H(X) + H(Y|X) &= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i, y_j) && (1.15)
\end{aligned}$$

Therefore

$$\boxed{H(XY) = H(X) + H(Y|X)} \tag{1.16}$$

QED

Theorem *(Chain Rule Generalization):*

Let $\bar{X} = X_1, X_2, \ldots, X_N$ be a *random vector* then the chain rule can be expressed as:

$$\boxed{H(X_1, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_N|X_1, \ldots, X_{N-1})} \tag{1.17}$$

1.3 Mutual information

Let X and Y be two random variables defined over a joint sample space $\mathcal{X}Y$:

$$\begin{aligned} \mathcal{X} &= \{x_1, \ldots, x_N\} && (1.18) \\ \mathcal{Y} &= \{y_1, \ldots, y_M\} && (1.19) \end{aligned}$$

For instance, x_i can be a symbol at the input of a communication channel while y_j represents the outcome from the channel, or the output symbol. The joint probability of both events: *"input symbol is x_i"* and *"output symbol is y_j"* is the probability of the joint event (x_i, y_j). One may raise the following question: *"How much information does the observation of a particular output symbol $y_j$from the channel provide about the input symbol x_i generated by the source?"*.

Before the observation of y_j, the probability of occurrence of the symbol x_i is simply $p(x_i)$ which is called the *"a priori"* probability. Upon reception of y_j, the probability that the symbol x_i was transmitted, given y_j, is the *"a posteriori"* probability $p(x_i|y_j)$.

This conditional probability $p(x_i|y_j)$ is also called the *"backward transition probability"* of the channel (for input symbol x_i and output symbol y_j). The *additional information* provided about the event x_i by the observation of the output symbol is given by:

$$\begin{aligned} I(x_i; y_j) &= I(x_i) - I(x_i|y_j) && (1.20) \\ I(x_i; y_j) &= -\log_b p(x_i) - [-\log_b p(x_i|y_j)] && (1.21) \end{aligned}$$

$$\boxed{I(x_i; y_j) = \log_b \frac{p(x_i|y_j)}{p(x_i)}} \qquad (1.22)$$

Then $I(x_i; y_j)$ is the *additional information* provided by the occurrence of y_j about x_i. Consider now the inverse case where one wants to find the additional information about the outcome y_j given that the specific symbol x_i has been transmitted through the channel.

$$\begin{aligned} I(y_j; x_i) &= I(y_j) - I(y_j|x_i) && (1.23) \\ I(y_j; x_i) &= -\log_b p(y_j) - [-\log_b p(y_j|x_i)] && (1.24) \end{aligned}$$

Therefore

$$I(y_j; x_i) = \log_b \frac{p(y_j|x_i)}{p(y_j)} \qquad (1.25)$$

Note that since the joint probability of (x_i, y_j) can be expressed as:

$$p(x_i, y_j) = p(x_i)\ p(y_j|x_i) = p(y_j)\ p(x_i|y_j) \tag{1.26}$$

then

$$\begin{aligned} I(y_j; x_i) &= \log_b \frac{p(y_j|x_i)}{p(y_j)} && (1.27) \\ I(y_j; x_i) &= \log_b \frac{p(x_i, y_j)}{p(x_i)\ p(y_j)} && (1.28) \\ I(y_j; x_i) &= \log_b \frac{p(x_i|y_j)}{p(x_i)} && (1.29) \end{aligned}$$

Therefore the expression for the additional information:

$$\boxed{I(x_i; y_j) = I(y_j; x_i)} \tag{1.30}$$

is called *mutual information* between events x_i and y_j, due to its symmetrical behavior.

1.4 Average mutual information

Definition *(Average mutual information):*

The *average mutual information* is defined as the weighted sum of the mutual information between each pair of input and output events x_i and y_j:

$$\boxed{I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j)\ I(x_i; y_j)} \tag{1.31}$$

or equivalently:

$$\boxed{I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j)\ \log_b \frac{p(x_i|y_j)}{p(x_i)}} \tag{1.32}$$

The average mutual information is a measure of the interdependence between the two random variables X and Y. Note that we can express the average mutual information as a function of the sets of joint probabilities $p(x_i, y_j)$ and marginal probabilities $p(x_i)$ and $p(y_j)$:

$$\boxed{I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j)\ \log_b \frac{p(x_i, y_j)}{p(x_i)\ p(y_j)}} \tag{1.33}$$

1.5 Relationship between the entropy and the (average) mutual information

The entropy of a source X, or its uncertainty, is denoted as:

$$H(X) = -\sum_{i=1}^{N} p(x_i)\ \log_b p(x_i) \tag{1.34}$$

where N is the size of the sample space $\mathcal{X}$. As seen previously, $H(X)$ represents the entropy of the source of information X *prior to* any observation.

On the other hand, the conditional entropy $H(X|Y)$, or equivocation, of this same source X given the observation of Y (e.g. output from a communication channel) is defined as:

$$H(X|Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j) \tag{1.35}$$

which indicates the remaining amount of information about the source X *after* the observation of Y. Consider the difference between these two entropy measures: the entropy $H(X)$ and the equivocation $H(X|Y)$.

$$\begin{aligned}
H(X) - H(X|Y) &= -\sum_{i=1}^{N} p(x_i)\ \log_b p(x_i) - \left[-\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j)\right] && (1.36)\\
H(X) - H(X|Y) &= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j)\ \log_b p(x_i) + \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j) && (1.37)\\
H(X) - H(X|Y) &= \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \left[\log_b p(x_i|y_j) - \log_b p(x_i)\right] && (1.38)\\
H(X) - H(X|Y) &= \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b \frac{p(x_i|y_j)}{p(x_i)} && (1.39)\\
H(X) - H(X|Y) &= I(X;Y) && (1.40)
\end{aligned}$$

Therefore, the mutual information $I(X;Y)$ between the two random variables X and Y is equal to the entropy $H(X)$ minus the equivocation (or remaining information in X given Y) $H(X|Y)$.

Theorem *((Average) Mutual Information):*

Let $\mathcal{XY}$ be a joint sample space. The (average) mutual information $I(X;Y)$ between the two random variables X and Y satisfies:

$$\boxed{I(X;Y) \geq 0} \tag{1.41}$$

with equality, if and only if, the X and Y are statistically independent.

Proof:

Consider the inequality $I(X;Y) \geq 0$, or equivalently, $-I(X;Y) \leq 0$. By definition of the (average) mutual information,

$$-I(X;Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \log_b \frac{p(x_i|y_j)}{p(x_i)} \tag{1.42}$$

$$-I(X;Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \log_b \frac{p(x_i)}{p(x_i|y_j)} \tag{1.43}$$

$$-I(X;Y) = (\log_b e) \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \ln \frac{p(x_i)}{p(x_i|y_j)} \tag{1.44}$$

If we consider only the events which have a non-zero probability of occurrence (i.e., the *probable events*), then:

$$p(x_i) > 0 \quad \text{and} \quad p(x_i|y_j) > 0 \quad \text{and thus} \quad p(x_i, y_j) > 0; \qquad \forall i, j \tag{1.45}$$

and therefore:

$$\frac{p(x_i)}{p(x_i|y_j)} > 0; \qquad \forall i, j \tag{1.46}$$

Since the natural logarithm $\ln x \leq (x - 1)$, for $x > 0$, then for each pair (x_i, y_j),

$$\ln \frac{p(x_i)}{p(x_i|y_j)} \leq \left[\frac{p(x_i)}{p(x_i|y_j)} - 1 \right] \tag{1.47}$$

a) If the random variables X and Y are independent then $p(x_i|y_j) = p(x_i)$, $\forall i, j$ which implies that:

$$-I(X;Y) = (\log_b e) \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \ln \frac{p(x_i)}{p(x_i|y_j)} \tag{1.48}$$

$$-I(X;Y) = (\log_b e) \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \ln \frac{p(x_i)}{p(x_i)} \tag{1.49}$$

$$-I(X;Y) = (\log_b e) \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \ \ln 1 \tag{1.50}$$

$$-I(X;Y) = 0 \tag{1.51}$$

and therefore:

$$\boxed{I(X;Y) = 0 \qquad \text{if and only if } X \text{ and } Y \text{ are independent}} \tag{1.52}$$

$$I(X;Y) = 0 \tag{1.53}$$

b) If X and Y are dependent from each other, then

$$\begin{aligned} -I(X;Y) &= (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \ \ln \frac{p(x_i)}{p(x_i|y_j)} && (1.54) \\ -I(X;Y) &< (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \ \left[\frac{p(x_i)}{p(x_i|y_j)} - 1 \right] && (1.55) \end{aligned}$$

but since the joint probability $p(x_i, y_j) = p(y_j)p(x_i|y_j)$, the above inequality can be expressed as:

$$\begin{aligned} -I(X;Y) &< (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \ \left[\frac{p(x_i)}{p(x_i|y_j)} - 1 \right] && (1.56) \\ -I(X;Y) &< (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} p(y_j)p(x_i|y_j) \ \left[\frac{p(x_i)}{p(x_i|y_j)} - 1 \right] && (1.57) \\ -I(X;Y) &< (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} \ \left[\frac{p(x_i)p(y_j)p(x_i|y_j)}{p(x_i|y_j)} - p(y_j)p(x_i|y_j) \right] && (1.58) \\ -I(X;Y) &< (\log_b e) \sum_{i=1}^{N} \sum_{j=1}^{M} \ [p(x_i)p(y_j) - p(x_i, y_j)] && (1.59) \\ -I(X;Y) &< (\log_b e) \left[\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i)p(y_j) \ - \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \right] && (1.60) \\ -I(X;Y) &< (\log_b e) \left[\sum_{i=1}^{N} p(x_i) \sum_{j=1}^{M} p(y_j) \ - \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \right] && (1.61) \\ -I(X;Y) &< (\log_b e) \left[(1 \times 1) - 1 \right] = 0 && (1.62) \end{aligned}$$

Therefore:

$$\boxed{I(X;Y) > 0 \qquad \text{when } X \text{ and } Y \text{ are dependent}} \tag{1.63}$$

QED

The average mutual information $I(X;Y)$ is equal to the difference between the entropy $H(X)$ and the equivocation $H(X|Y)$:

$$I(X;Y) = H(X) - H(X|Y) \tag{1.64}$$

Note that, since $I(X;Y)$ is always positive or equal to 0:

$$H(X) - H(X|Y) \geq 0 \tag{1.65}$$

which results in:

$$\boxed{H(X) \geq H(X|Y) \qquad \text{with equality when } X \text{ and } Y \text{ are independent}} \tag{1.66}$$

Then, the entropy of X, $H(X)$, is always larger or equal to the equivocation of X given Y, $H(X|Y)$.

1.6 Inequalities concerning the entropy and (average) mutual information

Theorem *(Entropy upper bound):*

Let $\mathcal{X}$ be a sample space consisting of N possible outcomes: $\{x_1, \ldots, x_N\}$, then:

$$\boxed{H(X) \leq \log_b N} \tag{1.67}$$

with equality, if and only if, the outcomes are equiprobable.

Proof:

Consider the difference $H(X) - \log_b N$:

$$H(X) - \log_b N = \sum_{i=1}^{N} p(x_i) \log_b \frac{1}{p(x_i)} - \log_b N \tag{1.68}$$

since $\sum_{i=1}^{N} p(x_i) = 1$ and the term $\log_b N$ is constant, the above expression can be rewritten as:

$$\begin{aligned} H(X) - \log_b N &= -\sum_{i=1}^{N} p(x_i) \log_b p(x_i) - \sum_{i=1}^{N} p(x_i) \log_b N && (1.69) \\ H(X) - \log_b N &= \sum_{i=1}^{N} p(x_i) \log_b \frac{1}{N p(x_i)} && (1.70) \end{aligned}$$

Converting to a natural logarithm;

$$H(X) - \log_b N = (\log_b e) \sum_{i=1}^{N} p(x_i) \ln \frac{1}{N p(x_i)} \tag{1.71}$$

The natural logarithm can be expanded as:

$$\ln x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \ldots \tag{1.72}$$

For $x > 0$, $\ln x \leq (x-1)$ with equality if $x = 1$ (see Figure 1.4).

Therefore, since

$$\ln \frac{1}{N p(x_i)} \leq \left[\frac{1}{N p(x_i)} - 1 \right] \tag{1.73}$$

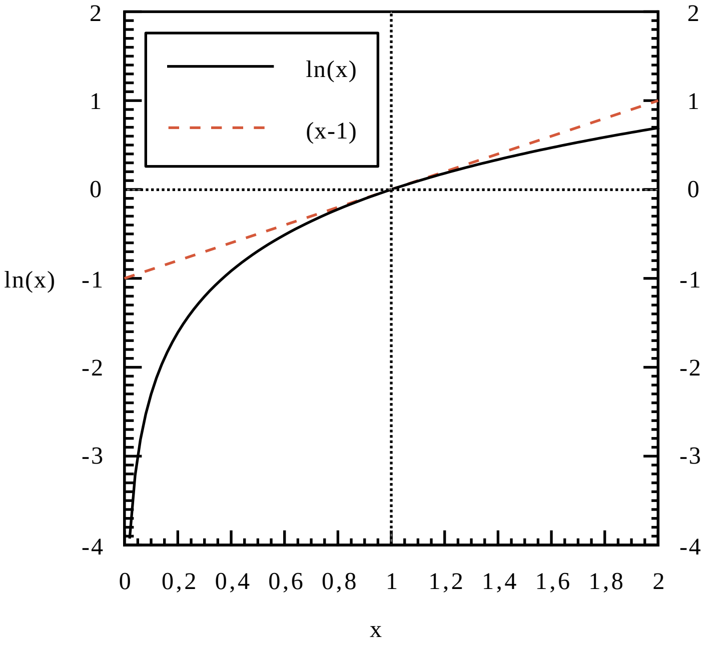


Figure 1.4: Natural logarithm ($\log(x) \leq x - 1 \quad \text{for } x > 0$).

then

$$
\begin{aligned}
H(X) - \log_b N &= (\log_b e) \sum_{i=1}^{N} p(x_i) \ln \frac{1}{N p(x_i)} && (1.74)\\
&\leq (\log_b e) \sum_{i=1}^{N} p(x_i) \left[\frac{1}{N p(x_i)} - 1 \right] && (1.75)\\
&\leq (\log_b e) \left[\sum_{i=1}^{N} \frac{1}{N} - \sum_{i=1}^{N} p(x_i) \right] && (1.76)\\
&\leq (\log_b e) \left[1 - 1 \right] && (1.77)\\
H(X) - \log_b N &\leq 0 && (1.78)
\end{aligned}
$$

or

$$\boxed{H(X) \leq \log_b N} \qquad (1.79)$$

QED

Note that if the source is equiprobable, then $p(x_i) = \frac{1}{N}$, for all i, and therefore:

$$\ln \frac{1}{Np(x_i)} = \ln 1 = (x - 1) = 0 \tag{1.80}$$

which implies that:

$$\boxed{H(X) = \log_b N} \tag{1.81}$$

The entropy $H(X)$ of a source can be increased by increasing the probability of an unlikely outcome x_i at the expense of a more probable outcome x_j.

1.7 Conditional and joint (average) mutual information

1.7.1 Conditional (average) mutual information

Let x_i, y_j and z_k be a set of specific outcomes in a joint sample space $\mathcal{XYZ}$. Then the conditional mutual information $I(x_i; y_j|z_k)$ between the events x_i and y_j, given z_k, is defined as:

$$\boxed{I(x_i; y_j|z_k) \equiv \log_b \frac{p(x_i|y_j, z_k)}{p(x_i|z_k)}} \tag{1.82}$$

Note: The condition of occurrence of event z_k affects *both* outcomes x_i and y_j. Thus the probability $p(x_i)$ becomes $p(x_i|z_k)$ while the conditional probability $p(x_i|y_j)$ now becomes $p(x_i|y_j, z_k)$. Also, the conditional mutual information $I(x_i; y_j|z_k)$ can be expressed as the difference between $I(x_i|z_k)$, the conditional self-information of x_i given z_k *before* the occurrence of y_j, and $I(x_i|y_j, z_k)$ which denotes the conditional self-information of x_i (still given z_k) *after* the occurrence of event y_j:

$$\boxed{I(x_i; y_j|z_k) = I(x_i|z_k) - I(x_i|y_j, z_k)} \tag{1.83}$$

The above result is demonstrated as follows:

$$\begin{aligned} I(x_i|z_k) - I(x_i|y_j, z_k) &= -\log_b p(x_i|z_k) - [-\log_b p(x_i|y_j, z_k)] && (1.84)\\ &= \log_b \frac{p(x_i|y_j, z_k)}{p(x_i|z_k)} && (1.85)\\ I(x_i|z_k) - I(x_i|y_j, z_k) &= I(x_i; y_j|z_k) && (1.86) \end{aligned}$$

Theorem *((Average) Conditional Mutual Information):*

Let $\mathcal{XYZ}$ be a joint sample space. Then the average conditional mutual information $I(X; Y|Z)$ between the X and Y random variables, given Z, is greater or equal to zero:

$$\boxed{I(X; Y|Z) \geq 0} \tag{1.87}$$

with equality, if and only if, conditional on each outcome z_k, X and Y are statistically independent, that is if:

$$p(x_i, y_j|z_k) = p(x_i|z_k)p(y_j|z_k), \qquad \text{for all } i, j, k \tag{1.88}$$

Note that all $p(z_k) > 0$. The proof of this theorem can be demonstrated in a similar manner than for theorem 2, by adding the conditioning on Z.

Consider the (average) conditional mutual information:

$$I(X;Y|Z) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) \log_b \left[\frac{p(x_i|y_j, z_k)}{p(x_i|z_k)}\right] \tag{1.89}$$

$$\begin{aligned} I(X;Y|Z) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) \log_b p(x_i|y_j, z_k) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{L} p(x_i, z_k) \log_b p(x_i|z_k) \end{aligned} \tag{1.90}$$

Then, the (average) conditional mutual information $I(X;Y|Z)$ can be expressed as the difference of 2 equivocations (i.e. $H(X|Z)$ and $H(X|YZ)$) of X:

$$\boxed{I(X;Y|Z) = H(X|Z) - H(X|YZ)} \tag{1.91}$$

and since $I(X;Y|Z) \geq 0$, this implies that:

$$\boxed{H(X|Z) \geq H(X|YZ)} \tag{1.92}$$

with equality if $I(X;Y|Z) = 0$, that is if, conditionally on Z, the random variables X and Y are statistically independent. Conditioning of X over the joint sample space $\mathcal{YZ}$ instead of $\mathcal{Z}$ alone reduces the uncertainty about X.

Also, by averaging the conditional mutual information $I(x_i, y_j|z_k)$ over the $\mathcal{XYZ}$ joint sample space, one obtains the average conditional mutual information $I(X;Y|Z)$:

$$I(X;Y|Z) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) I(x_i, y_j|z_k) \tag{1.93}$$

$$\begin{aligned} I(X;Y|Z) &= \sum_{i=1}^{N}\sum_{k=1}^{L} p(x_i, z_k) I(x_i|z_k) \\ &- \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) I(x_i|y_j, z_k) \end{aligned} \tag{1.94}$$

or once again:

$$\boxed{I(X;Y|Z) = H(X|Z) - H(X|Y, Z)} \tag{1.95}$$

Remark:

Even if both entropies (or equivocations) $H(X|Z)$ and $H(X|Y,Z)$ as well as the average conditional mutual information $I(X;Y|Z)$ are expressed using the same units, they have different meanings:

$H(X|Z)$: average uncertainty remaining in X after the observation in Z

$H(X|YZ)$: average uncertainty remaining in X after the observation in both Z and Y

$I(X;Y|Z)$: average amount of uncertainty in X *resolved* by the observation in Y

1.7.2 Joint (average) mutual information

Now lets go back to the $\mathcal{XYZ}$ joint sample space where $x_i \in \mathcal{X}$, $y_j \in \mathcal{Y}$ and $z_k \in \mathcal{Z}$.

Theorem *((Average) Mutual Information (over three sets)):*

The mutual information $I(x_i; y_j, z_k)$ between the event $x_i \in \mathcal{X}$ and the pair of events $(y_j, z_k) \in \mathcal{YZ}$ is equal to the sum of the mutual information $I(x_i; y_j)$ between the events x_i and y_j and the conditional mutual information $I(x_i; z_k|y_j)$ between x_i and z_k, given that y_j has occurred:

$$\boxed{I(x_i; y_j, z_k) = I(x_i; y_j) + I(x_i; z_k|y_j)} \tag{1.96}$$

or, equivalently:

$$\boxed{I(x_i; y_j, z_k) = I(x_i; z_k) + I(x_i; y_j|z_k)} \tag{1.97}$$

Proof:

Write the expressions for the (average) mutual information terms in the sum:

$$I(x_i; y_j) = \log_b \frac{p(x_i|y_j)}{p(x_i)} \quad \text{and} \tag{1.98}$$

$$I(x_i; z_k|y_j) = \log_b \frac{p(x_i|y_j, z_k)}{p(x_i|y_j)} \tag{1.99}$$

Therefore, using the properties of the logarithms, the sum becomes:

$$\begin{aligned}
I(x_i; y_j) + I(x_i; z_k|y_j) &= \log_b \frac{p(x_i|y_j)}{p(x_i)} + \log_b \frac{p(x_i|y_j, z_k)}{p(x_i|y_j)} && (1.100)\\
I(x_i; y_j) + I(x_i; z_k|y_j) &= \log_b \frac{p(x_i|y_j)}{p(x_i)} \frac{p(x_i|y_j, z_k)}{p(x_i|y_j)} && (1.101)\\
I(x_i; y_j) + I(x_i; z_k|y_j) &= \log_b \frac{p(x_i|y_j, z_k)}{p(x_i)} && (1.102)\\
I(x_i; y_j) + I(x_i; z_k|y_j) &= I(x_i; y_j, z_k) && (1.103)
\end{aligned}$$

QED

The average of the mutual information $I(x_i; y_j, z_k)$ over the entire joint sample space $\mathcal{XYZ}$ results in the average mutual information $I(X; YZ)$ between the single sample space $\mathcal{X}$ and the joint sample space $\mathcal{YZ}$:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) I(x_i; y_j, z_k) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) I(x_i; y_j) + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) I(x_i; z_k|y_j) \tag{1.104}$$

or

$$\boxed{I(X; YZ) = I(X; Y) + I(X; Z|Y)} \tag{1.105}$$

where $I(X; YZ)$ is the average mutual information between X and YZ, $I(X; Y)$ the average mutual information between X and Y, and $I(X; Z|Y)$ is the additional average mutual information between X and Z given Y.

We know that the conditional average mutual information $I(X; Z|Y)$ is always greater or equal to zero (with equality if and only if, conditional on Y, X and Z are independent). The average mutual information $I(X; YZ)$ should then be greater or at least equal to $I(X; Y)$.

For instance, one may consider a broadcast network for which the channel consists in a single source of information, e.g. X, and a number of receivers, say Y and Z. The message content from the source may have a common message intended for both receivers and some specific messages intended solely to each user independently of the other. The average mutual information term $I(X; YZ)$ represent the overall average mutual information between the source and the two receivers, whereas the $I(X; Y)$ term represent the average mutual information between X and Y, that is the common message and the specific message for this specific link. Finally, the remaining conditional average mutual information $I(X; Z|Y)$ represents only the information contents that is specific to the second receiver Z regardless of the common message sent to both users.

1.7.3 Average mutual information for cascaded channels

Another interesting example for the computation of joint (average) mutual information is for a cascaded channels for which the output of a channel in the chain depends uniquely on the preceeding regardless of the previous channels in that channel chain. A simple cascaded channel of only two channels is depicted in figure 1.5).

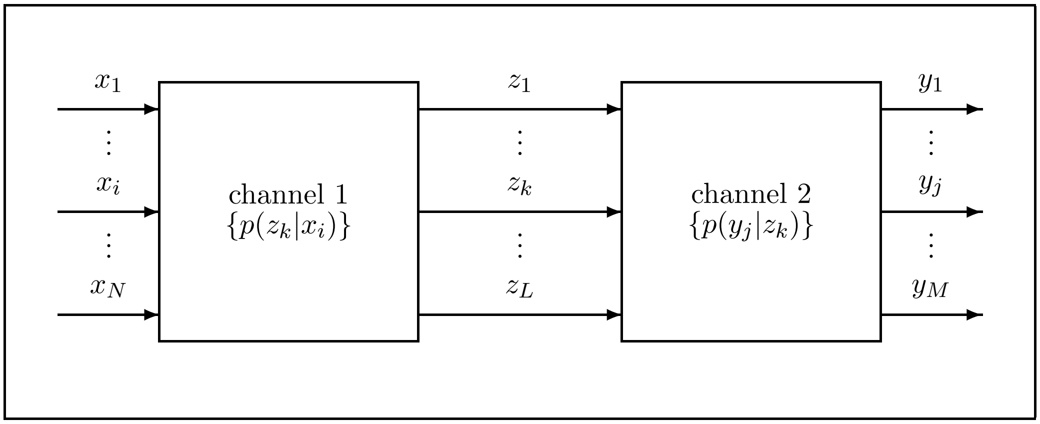


Figure 1.5: Transmission through two cascaded channels.

The output Y of the second channel depends entirely on its input Z, which itself depends only on its input X. Then for all i, j, and k, the conditional probability of the output symbol y_j given both inputs x_i and z_k, $p(y_j|x_i,z_k) = p(y_j|z_k)$. Multiplying both sides of the equality by $p(x_i|z_k)$ leads to:

$$\begin{aligned} p(y_j|x_i,z_k) &= p(y_j|z_k) && (1.106)\\ p(y_j|x_i,z_k)p(x_i|z_k) &= p(x_i|z_k)p(y_j|z_k) && (1.107)\\ p(x_i,y_j|z_k) &= p(x_i|z_k)p(y_j|z_k) && (1.108) \end{aligned}$$

which implies that, *conditionnally on the outcome* z_k, each input x_i and output y_j are statistically independent. Then, by the Theorem on (Average) Conditional Mutual Information, we have that the average conditional mutual information between X and Y, given Z is equal to zero:

$$I(X;Y|Z) = 0$$

Intuitively, one may expect that the average mutual information $I(X;Y)$ between the input X and output Y of the cascaded channels can not be greater than through either channel, that is $I(X;Z)$ or $I(Z;Y)$. Consider the average mutual information $I(X;YZ)$ between the input X and the two channels' outputs YZ.

$$\begin{aligned} I(X;YZ) &= I(X;Y) + I(X;Z|Y) \qquad \text{or} && (1.109)\\ I(X;YZ) &= I(X;Z) + I(X;Y|Z) && (1.110) \end{aligned}$$

Since, conditionnally on the output of the first channel Z, X and Y are independent we have that $I(X;Y|Z) = 0$ *but* the term $I(X;Z|Y) \geq 0$ and

$$\begin{aligned} I(X;Y) + I(X;Z|Y) &= I(X;Z) + I(X;Y|Z) && (1.111)\\ I(X;Y) + I(X;Z|Y) &= I(X;Z) && (1.112)\\ I(X;Y) &= I(X;Z) - I(X;Z|Y) && (1.113)\\ I(X;Y) &\leq I(X;Z) && (1.114) \end{aligned}$$

Or considering $I(Y;XZ)$, we obtain:

$$\begin{aligned} I(Y;XZ) &= I(Y;X) + I(Y;Z|X) \qquad \text{or} && (1.115)\\ I(Y;XZ) &= I(Y;Z) + I(Y;X|Z) \qquad \text{where } I(Y;X|Z) = 0 && (1.116)\\ I(Y;X) + I(Y;Z|X) &= I(Y;Z) && (1.117)\\ I(Y;X) &= I(Y;Z) - I(Y;Z|X) \qquad \text{where } I(Y;Z|X) \geq 0 && (1.118)\\ I(Y;X) &\leq I(Y;Z) && (1.119) \end{aligned}$$

By the *"commutativity property"* of the average *mutual* information:

$$\boxed{I(X;Y) \leq I(X;Z) \qquad \text{and} \qquad I(X;Y) \leq I(Z;Y)} \qquad (1.120)$$

In terms of entropies, the above can be restated as:

$$\begin{aligned} I(X;Y) &\leq I(X;Z) && (1.121)\\ H(X) - H(X|Y) &\leq H(X) - H(X|Z) && (1.122)\\ -H(X|Y) &\leq -H(X|Z) && (1.123)\\ H(X|Y) &\geq H(X|Z) && (1.124) \end{aligned}$$

The uncertainty (unresolved information) about the source X given the observation of the output of the first channel, i.e. $H(X|Z)$ is smaller than given the observation of the second channel (cascaded channel) output Y.

$$\boxed{H(X|Z) \leq H(X|Y)} \tag{1.125}$$

The conclusion is that the remaining uncertainty about the source X never decreases as we go further from the input through a series of cascaded channels. In our example, the second channel (it could even be a data processor such as an error correction decoder, the first channel being the noisy channel) cannot increases the average mutual information $I(X;Y)$ between the input X and cascaded channel output Y! Note however, that even if the mutual information decreases, the second channel can represent the available information at the output of the first channel (e.g. noisy channel), i.e. Z, in a more useful *format* at the output of the second channel and hence increasing the overall reliability of the cascaded channels without increasing the average mutual information.

1.8 Generalization over N sample spaces

1.8.1 Entropy of a random vector

Consider N sample spaces forming the joint sample space (or joint ensemble) $(\mathcal{X}_1, \ldots, \mathcal{X}_N)$; in other words, we consider an N-dimensional random vector $(X_1, \ldots, X_N)$. The probability of occurrence of a particular *string* of events $(x_1, \ldots, x_N)$ is given by:

$$p(x_1, \ldots, x_N) = p(x_1)\ p(x_2|x_1)\ p(x_3|x_1, x_2)\ \ldots\ p(x_N|x_1, \ldots, x_{N-1}) \tag{1.126}$$

The amount of self-information of this specific sequence of events $(x_1, \ldots, x_N)$ is then equal to:

$$\begin{aligned}
I(x_1, \ldots, x_N) &= -\log_b p(x_1, \ldots, x_N) &(1.127)\\
I(x_1, \ldots, x_N) &= -\log_b \left[p(x_1)\ p(x_2|x_1)\ p(x_3|x_1, x_2)\ \ldots\ p(x_N|x_1, \ldots, x_{N-1})\right] &(1.128)\\
I(x_1, \ldots, x_N) &= -\log_b p(x_1)\ -\log_b p(x_2|x_1)\ -\log_b p(x_3|x_1, x_2)\ \ldots\ -\log_b p(x_N|x_1, \ldots, x_{N-1}) &(1.129)
\end{aligned}$$

Then,

$$\boxed{I(x_1, \ldots, x_N) = I(x_1) + I(x_2|x_1) + I(x_3|x_1, x_2) + \ldots + I(x_N|x_1, \ldots, x_{N-1})}$$

This result indicates that the self-information of a string $(x_1, \ldots, x_N)$ is equal to the sum of the self-information of the first symbol in the string, namely x_1, the conditional self-information of the second symbol x_2, given symbol x_1, and so on up to the conditional self-information of the last symbol x_N, given the previous *substring* of events (or symbols) $(x_1, \ldots, x_{N-1})$.

The average of the self-information over all possible strings or symbols in the joint sample space $(\mathcal{X}_1, \ldots, \mathcal{X}_N)$, results in the entropy of the random vectors source:

$$\boxed{H(X_1, \ldots, X_N) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \ldots \sum_{k_N=1}^{K_N} p(x_{k_1}, x_{k_2}, \ldots, x_{k_N})\ I(x_{k_1}, x_{k_2}, \ldots, x_{k_N})}$$

or equivalently,

$$\boxed{H(X_1, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_N|X_1, \ldots, X_{N-1})}$$

This result is also known as the *chain rule* for the entropy of a random vector.

1.8.2 (Average) mutual information between two random vectors

Consider now two random vectors of dimension N and M respectively: $(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_M)$ defined on two joint ensembles $(\mathcal{X}_1, \ldots, \mathcal{X}_N)$ and $(\mathcal{Y}_1, \ldots, \mathcal{Y}_M)$. The (average) mutual information $I(X_1, \ldots, X_N; Y_1, \ldots, Y_M)$ between these two random vectors is:

$$\boxed{I(X_1, \ldots, X_N; Y_1, \ldots, Y_M) = H(X_1, \ldots, X_N) - H(X_1, \ldots, X_N | Y_1, \ldots, Y_M)}$$

which says that the (average) mutual information $I(X_1, \ldots, X_N; Y_1, \ldots, Y_M)$ is the difference between the joint entropy of the source $H(X_1, \ldots, X_N)$ and the equivocation of the source given the observation of the output random vector $H(X_1, \ldots, X_N | Y_1, \ldots, Y_M)$. But the joint entropy $H(X_1, \ldots, X_N)$ can be expressed as:

$$H(X_1, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_N|X_1, \ldots, X_{N-1}) \quad (1.130)$$

while the equivocation $H(X_1, \ldots, X_N | Y_1, \ldots, Y_M)$ is given by:

$$\begin{aligned} H(X_1, \ldots, X_N | Y_1, \ldots, Y_M) &= H(X_1|Y_1, \ldots, Y_M) + H(X_2|X_1, Y_1, \ldots, Y_M) + H(X_3|X_1, X_2, Y_1, \ldots, Y_M) + \ldots \\ &\quad + H(X_N|X_1, \ldots, X_{N-1}, Y_1, \ldots, Y_M) \qquad (1.131) \end{aligned}$$

The difference between the two terms is the (average) mutual information $I(X_1, \ldots, X_N; Y_1, \ldots, Y_M)$:

$$\begin{aligned} I(X_1, \ldots, X_N; Y_1, \ldots, Y_M) &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_N|X_1, \ldots, X_{N-1}) \\ &\quad - H(X_1|Y_1, \ldots, Y_M) - H(X_2|X_1, Y_1, \ldots, Y_M) - H(X_3|X_1, X_2, Y_1, \ldots, Y_M) \\ &\quad - \ldots - H(X_N|X_1, \ldots, X_{N-1}, Y_1, \ldots, Y_M) \qquad (1.132) \end{aligned}$$

Or in (average) mutual information terms:

$$\boxed{I(X_1, \ldots, X_N; Y_1, \ldots, Y_M) = I(X_1; Y_1, \ldots, Y_M) + I(X_2; Y_1, \ldots, Y_M | X_1) + \ldots + I(X_N; Y_1, \ldots, Y_M | X_1, \ldots, X_{N-1})}$$

The above result is known as the chain rule for the (average) mutual information between these two random vectors $I(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_M)$.

1.9 Relative entropy

Consider a sample space $\mathcal{X}$ and a random variable X with two different distributions $p(X) = \{p(x_i)\}$ and $q(X) = \{q(x_i)\}$ for $i = 1, \ldots, N$

Definition *(Relative entropy):*

The relative entropy $D\left[p(X)\|q(X)\right]$ between two distributions $p(X)$ and $q(X)$ is defined as the expectation of the logarithm of the ratio of the distributions:

$$D\left[p(X)\|q(X)\right] = \mathcal{E}\left[\log_b \frac{p(X)}{q(X)}\right] \tag{1.133}$$

In terms of probabilities of each outcome, the relative entropy $D\left[p(X)\|q(X)\right]$ becomes:

$$\boxed{D\left[p(X)\|q(X)\right] = \sum_{i=1}^{N} p(x_i)\ \log_b \left[\frac{p(x_i)}{q(x_i)}\right]}$$

The relative entropy is *a measure of the distance* between the two distributions (or probability mass function) $p(X)$ and $q(X)$. It is also known as the *Kullback-Leibler distance.* The concept of relative entropy is used in the *Maximum a Posteriori (MAP)* decoding techniques.

Note that if the two distributions are identical, i.e. $p(X) = q(X)$, or $p(x_i) = q(x_i) \quad \forall i$, then the relative entropy $D\left[p(X)\|q(X)\right] = 0$, since the term $\log_b \left[\frac{p(x_i)}{q(x_i)}\right] = \log_b 1 = 0$ for all i.

Example 1 *(relative entropy):*

Consider the following quaternary distributions $p(X)$ and $q(X)$ of the random variable X: $p(x_1) = p(x_2) = p(x_3) = p(x_4) = \frac{1}{4}$, and $q(x_1) = \frac{1}{2}$, $q(x_2) = \frac{1}{4}$, and $q(x_3) = q(x_4) = \frac{1}{8}$. The relative entropy (in *Sh*) between those distributions is:

$$\begin{aligned}
D\left[p(X)\|q(X)\right] &= \sum_{i=1}^{4} p(x_i)\ \log_2 \left[\frac{p(x_i)}{q(x_i)}\right] \\
D\left[p(X)\|q(X)\right] &= \frac{1}{4}\ \log_2 \left[\frac{\frac{1}{4}}{\frac{1}{2}}\right] + \frac{1}{4}\ \log_2 \left[\frac{\frac{1}{4}}{\frac{1}{4}}\right] + \frac{1}{4}\ \log_2 \left[\frac{\frac{1}{4}}{\frac{1}{8}}\right] + \frac{1}{4}\ \log_2 \left[\frac{\frac{1}{4}}{\frac{1}{8}}\right]
\end{aligned}$$

$$\begin{aligned}
D\left[p(X)\|q(X)\right] &= \frac{1}{4}\left[\log_2\frac{1}{2}+\log_2 1+\log_2 2+\log_2 2\right] \\
D\left[p(X)\|q(X)\right] &= \frac{1}{4}\ Sh
\end{aligned}$$

Now let's consider the relative entropy $D\left[q(X)\|p(X)\right]$ between $q(X)$ and $p(X)$:

$$\begin{aligned}
D\left[q(X)\|p(X)\right] &= \sum_{i=1}^{4} q(x_i)\ \log_2\left[\frac{q(x_i)}{p(x_i)}\right] \\
D\left[q(X)\|p(X)\right] &= \frac{1}{2}\ \log_2\left[\frac{\frac{1}{2}}{\frac{1}{4}}\right]+\frac{1}{4}\ \log_2\left[\frac{\frac{1}{4}}{\frac{1}{4}}\right]+\frac{1}{8}\ \log_2\left[\frac{\frac{1}{8}}{\frac{1}{4}}\right]+\frac{1}{8}\ \log_2\left[\frac{\frac{1}{8}}{\frac{1}{4}}\right] \\
D\left[q(X)\|p(X)\right] &= \frac{1}{2}\ \log_2 2+\frac{1}{4}\ \log_2 1+\frac{1}{8}\ \log_2\frac{1}{2}+\frac{1}{8}\ \log_2\frac{1}{2} \\
D\left[q(X)\|p(X)\right] &= \frac{1}{4}\ Sh
\end{aligned}$$

Note that here $D\left[p(X)\|q(X)\right]=D\left[q(X)\|p(X)\right]=0.250\ Sh$.

Example 2_(relative entropy):_

For this second example, $p(X)$ and $q(X)$ are two distributions of a binary random variable X where: $p(x_1)=p(x_2)=\frac{1}{2}$ and $q(x_1)=\frac{1}{4}$ and $q(x_2)=\frac{3}{4}$. The relative entropy $D\left[p(X)\|q(X)\right]$ is:

$$\begin{aligned}
D\left[p(X)\|q(X)\right] &= \sum_{i=1}^{2} p(x_i)\ \log_2\left[\frac{p(x_i)}{q(x_i)}\right] \\
D\left[p(X)\|q(X)\right] &= \frac{1}{2}\ \log_2\left[\frac{\frac{1}{2}}{\frac{1}{4}}\right]+\frac{1}{2}\ \log_2\left[\frac{\frac{1}{2}}{\frac{3}{4}}\right]=\frac{1}{2}\log_2 2+\log_2\frac{2}{3} \\
D\left[p(X)\|q(X)\right] &= 0.208\ Sh
\end{aligned}$$

whereas:

$$\begin{aligned}
D\left[q(X)\|p(X)\right] &= \sum_{i=1}^{2} q(x_i)\ \log_2\left[\frac{q(x_i)}{p(x_i)}\right] \\
D\left[q(X)\|p(X)\right] &= \frac{1}{4}\ \log_2\left[\frac{\frac{1}{4}}{\frac{1}{2}}\right]+\frac{3}{4}\ \log_2\left[\frac{\frac{3}{4}}{\frac{1}{2}}\right]=\frac{1}{4}\ \log_2\frac{1}{2}+\frac{3}{4}\ \log_2\frac{3}{2} \\
D\left[q(X)\|p(X)\right] &= 0.189\ Sh
\end{aligned}$$

In general, as in this second example, $D\left[p(X)\|q(X)\right]\neq D\left[q(X)\|p(X)\right]$.

The relative entropy $D\left[p(XY)\|p(X)p(Y)\right]$ between the joint distribution $p(XY)$ of two random variables X and Y and the product of their marginal distributions $p(X)$ and $p(Y)$ gives the (average) mutual information $I(X;Y)$ between the two random variables:

$$\boxed{D\left[p(XY)\|p(X)p(Y)\right] = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j)\ \log_b \left[\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right] = I(X;Y)}$$

1.10 Problems

Problem 1.1: A dishonest gambler has a loaded die which turns up the number 1 with a probability of 0.4 and the numbers 2 to 6 with a probability of 0.12 each. Unfortunately (or fortunately) he left the loaded die in a box with 2 honest dice and could not tell them apart. He picks at random one die from the box, rolls it once, and the number 1 appears.

a) What is the probability that he picked up the loaded die?

b) He rolls the same die once more and it comes up 1 again. What is the probability that he has picked the loaded die after the second rolling?

c) Repeat parts a) and b) but assuming this time that the first outcome was a 4 and the second outcome was 1.

Problem 1.2: A source of information produces letters from a three-symbol alphabet $X = \{x_0, x_1, x_2\}$ with a probability assignment $p(x_0) = p(x_1) = 1/4$ and $p(x_2) = 1/2$. Each source letter x_i is directly transmitted through two different channels simultaneously with outputs y_j and z_k for which the transition probabilities $p(y_j|x_i)$ and $p(z_k|x_i)$ are as indicated in figure 1.6 shown below. Note that this could be considered as a single channel with output (y_j, z_k).

a) Write the channel transition matrix for each channel.

b) Calculate the following entropies: $H(X)$, $H(Y)$, $H(Z)$ and $H(YZ)$.

c) Calculate the mutual information expressions: $I(X;Y)$, $I(X;Z)$, $I(X;Y|Z)$ and $I(X;YZ)$.

d) Interpret the mutual information expressions.

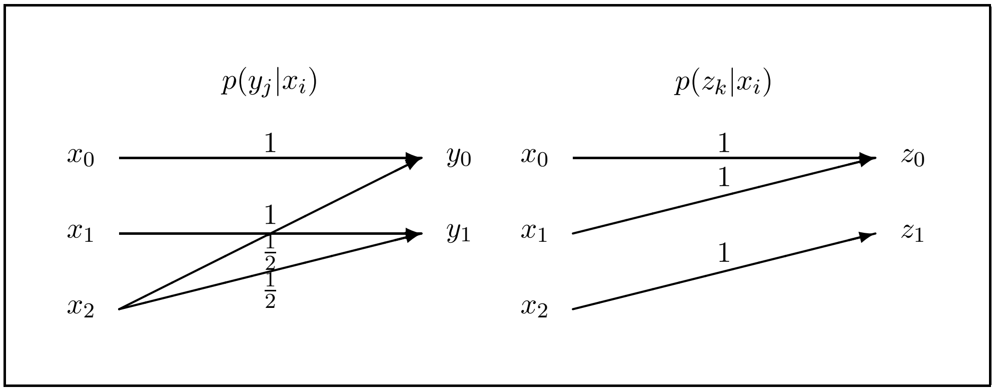


Figure 1.6: Simultaneous transmission through two channels.

Problem 1.3: A ternary source of information $X \equiv \{x_0, x_1, x_2\}$ is to be transmitted through a noisy communication channel. The channel probability transition matrix $\mathbf{P}$ is given by:

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/4 & 0 & 1/4 \\ 0 & 1/2 & 1/4 & 1/4 \\ 1/4 & 0 & 1/2 & 1/4 \end{pmatrix}$$

If the source letters are generated with the probabilities $p(x_0) = p(x_2) = \frac{1}{4}$ and $p(x_1) = \frac{1}{2}$, find the output letter probabilities $p(y_j)$ and the average mutual information $I(X;Y)$.

Problem 1.4: A source of information generates the symbols $\{a_0, \cdots, a_k, \cdots, a_7\}$ with the following probability:

$$p(a_k) = \binom{7}{k} \eta^k \, (1-\eta)^{7-k}$$

a) Find the source entropy $H(X)$ for $\eta = 1/4$.

b) If η is changed to 1/2, what is the new value the entropy $H(X)$?

Problem 1.5: Let U, V, W, X, Y, and Z be random variables.

a) Show that

$$I(XY;UVW) = I(XY;U|VW) + I(XY;V|W) + I(XY;W)$$

.

b) Do the conditions, $I(V;YZ|UX) = 0$ and $I(X;UZ|VY) = 0$, imply that $I(Z;XV|UY) \overset{?}{=} 0$? Justify your answer.

Problem 1.6: *(derived from Gallager)*

In Ottawa, a radio station weatherman's record is as follows: out of the 15% of the time when it actually rains, the weatherman predicts *"rain"* 12% of the time and *"no rain"* 3% of the time. The remaining 85% of the time, when it doesn't rain, the weatherman's prediction are *"no rain"* 64% of the time and *"rain"* 21% of the time.

A clever ELG-5170 Information Theory graduate student notices that the weatherman's predictions are correct 76% of the time. However, by predicting *"no rain"* all the time, he (or she) can achieve a higher success rate of 85%! The graduate student explains the situation to the weatherman's boss and applies for the job. However, the weatherman's boss, who is also an information theorist, decides to not hire the graduate student. Why?

Problem 1.7: Consider two statistical experiments represented by the random variables X and Y, where the sample space of X is (x_1, x_2, x_3, x_4) and the sample space of Y is (y_1, y_2, y_3). The joint probability matrix $\mathbf{P} = \{p(x_i, y_j)\}_{\substack{i=1,2,3,4\\ j=1,2,3}}$ for these 2 experiments is:

$$\mathbf{P} = \left[\begin{array}{cccc} p(x_1,y_1) & p(x_2,y_1) & p(x_3,y_1) & p(x_4,y_1) \\ p(x_1,y_2) & p(x_2,y_2) & p(x_3,y_2) & p(x_4,y_2) \\ p(x_1,y_3) & p(x_2,y_3) & p(x_3,y_3) & p(x_4,y_3) \end{array}\right] = \left[\begin{array}{cccc} \frac{3}{32} & \frac{1}{32} & \frac{1}{32} & \frac{7}{32} \\ \frac{1}{32} & \frac{3}{32} & \frac{3}{32} & \frac{1}{32} \\ \frac{7}{32} & \frac{1}{32} & \frac{1}{32} & \frac{3}{32} \end{array}\right]$$

a) How much information do we receive if someone tells us the outcome resulting from X and Y?

b) How much information do we receive if someone tells us the outcome of Y?

c) How much information do we receive if someone tells us the outcome of X if we already know the outcome of Y?

Chapter 2

Distortionless Source Coding

2.1 Tchebycheff Inequality and the weak law of large numbers

2.1.1 Tchebycheff inequality

Consider a random variable X having the input symbol distribution $\{p(x_k)\}$, an expectation η_X and a variance σ_X^2, that is:

$$\eta_X \equiv E[X] = \sum_{k=1}^{K} p(x_k) \times x_k \tag{2.1}$$

$$\sigma_X^2 \equiv E[(X - \eta_X)^2] = \sum_{k=1}^{K} p(x_k) \times (x_k - \eta_X)^2 \tag{2.2}$$

Definition *(Tchebycheff inequality):*

The Tchebycheff inequality states that:

$$\boxed{Pr\{|X - \eta_X| \geq \delta\} \leq \frac{\sigma_X^2}{\delta^2}}$$

2.1.2 Weak law of large numbers

Let $\overline{X} = X_1, \ldots, X_n, \ldots, X_N$ be a sequence of independent, identically distributed (i.i.d.) random variables with expectation η_X and variance σ_X^2. Define a new random variable Y_N which is the *sample mean* of the random sequence $\overline{X}$:

$$Y_N \equiv \frac{1}{N} \sum_{n=1}^{N} X_n \tag{2.3}$$

The mean η_{Y_N} of this new random variable is then given by:

$$\begin{aligned}
\eta_{Y_N} &= E[Y_N] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] \qquad (2.4)\\
&= \frac{1}{N} E\left[\sum_{n=1}^{N} X_n\right]\\
&= \frac{1}{N}\sum_{n=1}^{N} E\left[X_n\right]\\
&= \frac{1}{N}\sum_{n=1}^{N} \eta_X\\
\eta_{Y_N} &= \eta_X
\end{aligned}$$

The variance $\sigma^2_{Y_N}$ of the sample average Y_N is equal to the expectation of $(Y_N - \eta_{Y_N})^2$:

$$\begin{aligned}
\sigma^2_{Y_N} &= E\left[(Y_N - \eta_{Y_N})^2\right] \qquad (2.5)\\
&= E\left[\left(\frac{1}{N}\sum_{n=1}^{N} X_n - \eta_X\right)^2\right]\\
&= \frac{1}{N^2} E\left[\sum_{n=1}^{N} (X_n - \eta_X)^2\right]\\
\sigma^2_{Y_N} &= \frac{\sigma^2_X}{N}
\end{aligned}$$

Applying the Tchebycheff inequality:

$$Pr\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] - \eta_X\right| \geq \delta\right\} \leq \frac{\sigma^2_X}{N\delta^2} \tag{2.6}$$

As N tends towards infinity, the right side of the above inequality approaches zero.

Definition *(Weak Law of Large Numbers):*

The weak law of large numbers stipulates that the sample average or sample mean of the random sequence $\overline{X}$ approaches the statistical mean η_X with high probability:

$$\boxed{\lim_{N\to\infty} Pr\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] - \eta_X\right| \geq \delta\right\} = 0}$$

or equivalently:

$$\boxed{\lim_{N\to\infty} Pr\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] - \eta_X\right| < \delta\right\} = 1}$$

2.2 Typical and atypical sequences

Definition *(Typical and atypical sequences):*

Consider a memoryless source X having the input symbol distribution $\{p(x_k)\}$, $k = 1, \ldots, K$, and an entropy $H(X)$. Let $\mathbf{x}$ be a vector of blocklength N: $\mathbf{x} = (x_{k1}, \cdots, x_{kN})$. For any number $\delta > 0$, the set $\mathcal{T}_X(\delta)$ of typical sequences of blocklength N is defined as:

$$\boxed{\mathcal{T}_X(\delta) \equiv \{\mathbf{x} \text{ such that: } |-\frac{1}{N} \log_b p(\mathbf{x}) - \mathbf{H}(\mathbf{X})| < \delta\}}$$

The remaining vectors of length N form a complementary set; the set $\mathcal{T}_X^c(\delta)$ of atypical sequences:

$$\boxed{\mathcal{T}_X^c(\delta) \equiv \{\mathbf{x} : |-\frac{1}{N} \log_b p(\mathbf{x}) - \mathbf{H}(\mathbf{X})| \geq \delta\}}$$

Example *(Typical sequences:):*

Consider a binary source, or random variable, $X = \{x_i\}$ with the probabilities $p(x_1) = 1/4$ and $p(x_2) = 3/4$. The source entropy $H(X)$, expressed in Shannons (or bits), is then equal to:

$$\begin{aligned}
H(X) &= -\sum_{i=1}^{2} p(x_i) \;\; \log_2 p(x_i) && (2.7)\\
&= -\left[(1/4) \;\; \log_2 (1/4) \;\; + \;\; (3/4) \;\; \log_2 (3/4)\right]\\
H(X) &= 0.811 \; Sh
\end{aligned}$$

Now if the experiment is repeated twice, that is if the source generates two binary symbols; the outcomes will be all possible pairs $X_1, X_2 = \{(x_i, x_j)\}$. Since the random variables X_1 and X_2 are independent and also identically distributed (i.i.d.) then the probability of each pair $p(x_i, x_j)$ is equal to the product of the marginal probabilities:

$$\begin{aligned}
p(x_1, x_1) &= p(x_1)p(x_1) = 1/4 \times 1/4 = 1/16 && (2.8)\\
p(x_1, x_2) &= p(x_1)p(x_2) = 1/4 \times 3/4 = 3/16\\
p(x_2, x_1) &= p(x_2)p(x_1) = 3/4 \times 1/4 = 3/16\\
p(x_2, x_2) &= p(x_2)p(x_2) = 3/4 \times 3/4 = 9/16
\end{aligned}$$

For $N = 3$ (i.e. considering sequences of information of length 3), the probabilities of each sequence $p(x_i, x_j, x_k)$ is:

$$\begin{aligned}
p(x_1,x_1,x_1) &= p(x_1)p(x_1)p(x_1) = 1/64 \\
p(x_1,x_1,x_2) &= p(x_1)p(x_1)p(x_2) = 3/64 \\
p(x_1,x_2,x_1) &= p(x_1)p(x_2)p(x_1) = 3/64 \\
p(x_1,x_2,x_2) &= p(x_1)p(x_2)p(x_2) = 9/64 \\
p(x_2,x_1,x_1) &= p(x_2)p(x_1)p(x_1) = 3/64 \\
p(x_2,x_1,x_2) &= p(x_2)p(x_1)p(x_2) = 9/64 \\
p(x_2,x_2,x_1) &= p(x_2)p(x_2)p(x_1) = 9/64 \\
p(x_2,x_2,x_2) &= p(x_2)p(x_2)p(x_2) = 27/64
\end{aligned} \tag{2.9}$$

Note that six sequences of symbols have a probability $p(x_i, x_j, x_k) = 3/64$ or $9/64$, out of the $2^N = 8$ possible sequences of length 3. For sequences to be termed typical sequences, their probability of occurrence must be in the following range:

$$b^{-N[H(X)+\delta]} \leq p(\mathbf{x}) \leq \mathbf{b^{-N[H(X)-\delta]}}$$

where $N = 3$, $b = 2$, $H(X)$ is the source entropy (per symbol), δ an arbitrarily small positive number, and $\mathbf{x}$ is a specific sequence of length N:

$$\begin{aligned}
2^{-3[H(X)+\delta]} \leq & \quad p(x_i, x_j, x_k) \leq 2^{-3[H(X)-\delta]} \\
2^{-3[0.811+\delta]} \leq & \quad p(x_i, x_j, x_k) \leq 2^{-3[0.811-\delta]}
\end{aligned} \tag{2.10}$$

Writing the sequences' probabilities in the form of $b^{-N[H(X)\pm\delta]}$:

$$\begin{aligned}
p(x_1,x_1,x_1) &= 2^{-3\times 2.000} = 1/64 \\
p(x_1,x_1,x_2) &= 2^{-3\times 1.472} = 3/64 \\
p(x_1,x_2,x_1) &= 2^{-3\times 1.472} = 3/64 \\
p(x_1,x_2,x_2) &= 2^{-3\times 0.943} = 9/64 \\
p(x_2,x_1,x_1) &= 2^{-3\times 1.472} = 3/64 \\
p(x_2,x_1,x_2) &= 2^{-3\times 0.943} = 9/64 \\
p(x_2,x_2,x_1) &= 2^{-3\times 0.943} = 9/64 \\
p(x_2,x_2,x_2) &= 2^{-3\times 0.415} = 27/64
\end{aligned} \tag{2.11}$$

There are thus three sequences ((x_1, x_2, x_2), (x_2, x_1, x_2) and (x_2, x_2, x_1)) that have a probability of occurrence close to $b^{-N[H(X)\pm\delta]}$. These can be considered (depending on the value of δ, that we choose to be arbitrarily small), as typical sequences.

For $N = 20$, there is a single sequence having only the symbol x_1 in it, 20 sequences with one occurrence of x_2, $\binom{N}{2}$ sequences with two occurences of x_2, and so on.

$$\binom{N}{n} \equiv \frac{N!}{(N-n)!n!}$$

The probability of each sequence of length N depends on the number n of occurrences of each symbol x_1:

$$p(\mathbf{x}) = \mathbf{p(x_1)^n \; p(x_2)^{N-n}}$$

Table 1 indicates the number of sequences as a function of the number n of occurrences of symbol x_1, along with the probability of each of these sequences and and total probability of all sequences having symbol x_1 n times.

The total probability of occurrence of all typical sequences is high; for instance for $2 \leq n \leq 8$, the exponent in the probability expression of the individual sequences ranges from 0.573 to 1.049 while $H(X) = 0.811$. Thus for $\delta \leq .238$, the total probability of the occurrence of typical sequences is close to 94% (i.e. total probability is equal to 0.93478 for $2 \leq n \leq 8$).

Note also that for $n = 5$, the probability of each sequence consisting of $n = 5$ occurrences of the binary symbol x_1 is equal to $2^{-20 \times 0.811}$, which is exactly equal to $b^{-N[H(X)]}$. That is, for $\delta = 0$, the total probability of all sequences with $n = 5$ is already 20% (i.e. 0.20233 for $n = 5$). For these sequences, there are $n = 5$ occurrences of the x_1 in the 20-symbol vector, which represent the actual distribution of each individual symbol: $\{p(x_1) = 1/4, \, p(x_2) = 3/4\}$.

Table 2.1: Typical (binary) sequences of length $N = 20$: $p(x_1) = \frac{1}{4}$, $p(x_2) = \frac{3}{4}$

Occurrences of x_1 n	Number of sequences $\binom{N}{n}$	Probability of each sequence $p(x_1)^n \; p(x_2)^{N-n}$	Probability of all sequences $\binom{N}{n} \; p(x_1)^n \; p(x_2)^{N-n}$
0	1	$3,171 \times 10^{-3} = 2^{-20\times 0,415}$	0,003171
1	20	$1,057 \times 10^{-3} = 2^{-20\times 0,494}$	0,021141
2	190	$3,524 \times 10^{-4} = 2^{-20\times 0,574}$	0,066948
3	1140	$1,175 \times 10^{-4} = 2^{-20\times 0,653}$	0,133896
4	4845	$3,915 \times 10^{-5} = 2^{-20\times 0,732}$	0,189685
5	15504	$1,305 \times 10^{-5} = 2^{-20\times 0,811}$	0,202331
6	38760	$4,350 \times 10^{-6} = 2^{-20\times 0,891}$	0,168609
7	77520	$1,450 \times 10^{-6} = 2^{-20\times 0,970}$	0,112406
8	125970	$4,833 \times 10^{-7} = 2^{-20\times 1,049}$	0,060887
9	167960	$1,611 \times 10^{-7} = 2^{-20\times 1,128}$	0,027061
10	184756	$5,370 \times 10^{-8} = 2^{-20\times 1,208}$	0,009922
11	167960	$1,790 \times 10^{-8} = 2^{-20\times 1,287}$	0,003007
12	125970	$5,967 \times 10^{-9} = 2^{-20\times 1,366}$	0,000752
13	77520	$1,989 \times 10^{-9} = 2^{-20\times 1,445}$	0,000154
14	38760	$6,630 \times 10^{-10} = 2^{-20\times 1,525}$	0,000026
15	15504	$2,210 \times 10^{-10} = 2^{-20\times 1,604}$	0,000003
16	4845	$7,367 \times 10^{-11} = 2^{-20\times 1,683}$	0,000000
17	1140	$2,456 \times 10^{-11} = 2^{-20\times 1,762}$	0,000000
18	190	$8,185 \times 10^{-12} = 2^{-20\times 1,842}$	0,000000
19	20	$2,728 \times 10^{-12} = 2^{-20\times 1,921}$	0,000000
20	1	$9,095 \times 10^{-13} = 2^{-20\times 2,000}$	0,000000

2.3 Shannon-McMillan theorem

Theorem *(Shannon-McMillan theorem for typical sequences:):*

Given a memoryless source of entropy $H(X)$ and an arbitrary positive number δ, a blocklength $N \geq N_0$ can be choosen sufficiently large such that the set of all K^N possible vectors $\{\mathbf{x}\}$ can be partitioned into a set of typical (or likely) sequences $\mathcal{T}_X(\delta)$, and a complementary set of atypical (or unlikely) sequences $\mathcal{T}_X^c(\delta)$ having the following properties:

a) The probability that a particular sequence $\mathbf{x}$ of blocklength N belongs to the set of atypical sequences $\mathcal{T}_X^c(\delta)$ is upperbounded as:

$$\boxed{Pr[\mathbf{x} \in \mathcal{T}_X^c(\delta)] < \epsilon}$$

b) If a sequence $\mathbf{x}$ is in the set of typical sequences $\mathcal{T}_X(\delta)$ then its probability of occurrence $p(\mathbf{x})$ is approximately equal to $b^{-NH(X)}$, that is:

$$\boxed{b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}}$$

c) The number of typical, or likely, sequences $\|\mathcal{T}_X(\delta)\|$ is bounded by:

$$\boxed{(1-\epsilon)b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]}}$$

Proof:

a) $P\left[\mathbf{x} \in \mathcal{T}_X^c(\delta)\right] < \epsilon$:

By the *Asymptotic Equipartition Property* (AEP), the set of atypical sequences $\in \mathcal{T}_X^c(\delta)$ is upperbound as:

$$\begin{aligned} P\left[\mathbf{x} \in \mathcal{T}_X^c(\delta)\right] &= P\left[\left|-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\right| > \delta\right] \qquad (2.12) \\ P\left[\left|-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\right| > \delta\right] &< \frac{\sigma_X^2}{N\delta^2} \\ P\left[\left|-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\right| > \delta\right] &< \epsilon \\ P\left[\mathbf{x} \in \mathcal{T}_X^c(\delta)\right] &< \epsilon \end{aligned}$$

b) $b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}$:

If $\mathbf{x}$ is in the set of typical sequences $\mathcal{T}_X(\delta)$, we have by definition that:

$$\left| -\frac{1}{N} \log_b p(\mathbf{x}) - H(X) \right| < \delta \tag{2.13}$$

or, for N sufficiently large (i.e., $N \geq N_0$):

$$-\delta < -\frac{1}{N} \log_b p(\mathbf{x}) - H(X) < \delta \tag{2.14}$$

or, adding $H(X)$ everywhere:

$$H(X) - \delta < -\frac{1}{N} \log_b p(\mathbf{x}) < H(X) + \delta \tag{2.15}$$

Multiplying by $-N$ (and changing the inequality signs accordingly):

$$-N[H(X) - \delta] > \log_b p(\mathbf{x}) > -N[H(X) + \delta] \tag{2.16}$$

Raising to the power b (i.e., logarithmic base used for computing the entropy):

$$b^{-N[H(X)-\delta]} > b^{\log_b p(\mathbf{x})} = p(\mathbf{x}) > b^{-N[H(X)+\delta]} \tag{2.17}$$

Therefore,

$$\boxed{b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}} \tag{2.18}$$

for $\mathbf{x} \in \mathcal{T}_X(\delta)$, which happens with a probability greater or equal to $(1 - \epsilon)$, for $N \geq N_0$.

c) Number of typical sequences, $\|\mathcal{T}_X(\delta)\|$:

$$(1 - \epsilon) b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]}$$

i) The sum of probabilities of typical sequences is less than 1 (definition of a probability space):

$$\sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} b^{-N[H(X)+\delta]} < \sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} p(\mathbf{x}) \leq \sum_{i=1}^{K^N} p(\mathbf{x}) = 1 \tag{2.19}$$

where $b^{-N[H(X)+\delta]}$ is the minimum probability of occurrence that a typical sequence can have. Since the term $b^{-N[H(X)+\delta]}$ is constant then:

$$\sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} b^{-N[H(X)+\delta]} = \|\mathcal{T}_X(\delta)\| \; b^{-N[H(X)+\delta]} < 1 \tag{2.20}$$

which implies that:

$$\|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]} \tag{2.21}$$

ii) The sum of probabilities of all typical sequences is also lowerbounded by $(1-\epsilon)$ (definition of typical sequences and Asymptotic Equipartition Property):

$$(1-\epsilon) < \sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} p(\mathbf{x}) < \sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} b^{-N[H(X)-\delta]} \tag{2.22}$$

since $b^{-N[H(X)-\delta]}$ is the highest probability of occurrence of a typical sequence. Then

$$(1-\epsilon) < \sum_{\mathbf{x} \in \mathcal{T}_X(\delta)} b^{-N[H(X)-\delta]} = \|\mathcal{T}_X(\delta)\| b^{-N[H(X)-\delta]} \tag{2.23}$$

Therefore, the number of typical sequences, $\|\mathcal{T}_X(\delta)\|$, is lowerbounded as:

$$\frac{(1-\epsilon)}{b^{-N[H(X)-\delta]}} < \|\mathcal{T}_X(\delta)\| \tag{2.24}$$

$$\|\mathcal{T}_X(\delta)\| > (1-\epsilon)\ b^{N[H(X)-\delta]} \tag{2.25}$$

Combining the upper and lower bounds:

$$(1-\epsilon)\ b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]} \tag{2.26}$$

QED

2.4 Variable length codes (source coding)

Consider a source code $\mathcal{C}$ which encodes each different source symbol (or sourceword) with a *unique* codeword. To be able to retrieve the original information at the receiver (i.e. information sink), all codewords should be uniquely decodable. It is desirable to minimize the average codeword length.

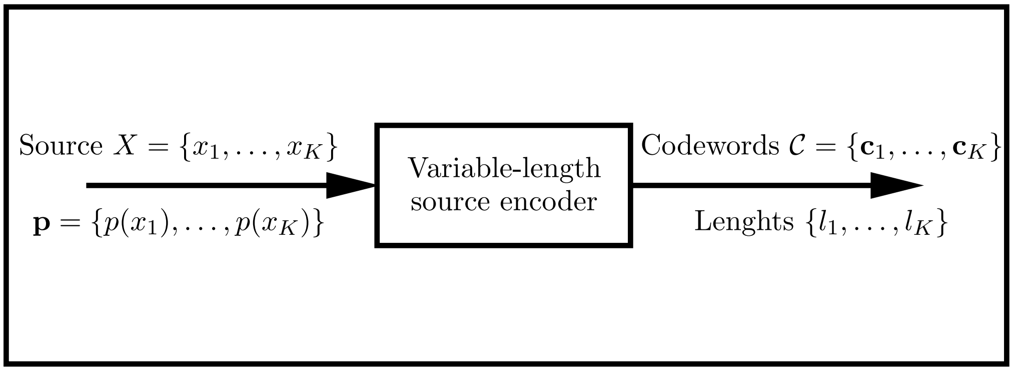

Source $X = \{x_1, \ldots, x_K\}$
$\mathbf{p} = \{p(x_1), \ldots, p(x_K)\}$
Variable-length source encoder
Codewords $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$
Lenghts $\{l_1, \ldots, l_K\}$

Figure 2.1: Variable length source encoder.

The source $X = \{x_1, \ldots, x_k, \ldots, x_K\}$, K being the source alphabet size, is characterized by its letter distribution: $\mathbf{p} = \{p(x_1), \ldots, p(x_k), \ldots, p(x_K)\}$ and its entropy $H(X)$.

The variable length source code $\mathcal{C}$ is a set of codewords $\{\mathbf{c}_1, \ldots, \mathbf{c}_k, \ldots, \mathbf{c}_K\}$ which consists each in l_k symbols taken from an output alphabet $Y = \{y_1, \ldots, y_j, \ldots, y_J\}$. In other words, the l^{th} element of the k^{th} codeword $c_{k,l} \in \{y_1, \ldots, y_J\}$ (where $1 \leq l \leq l_k$):

Source Symbol	Codeword	Codeword Length
x_1	$\mathbf{c}_1 = (c_{1,1}, c_{1,2}, \ldots, c_{1,l}, \ldots, c_{1,l_1})$	l_1
x_2	$\mathbf{c}_2 = (c_{2,1}, c_{2,2}, \ldots, c_{2,l}, \ldots, c_{2,l_2})$	l_2
$\vdots$	$\vdots$	$\vdots$
x_k	$\mathbf{c}_k = (c_{k,1}, c_{k,2}, \ldots, c_{k,l}, \ldots, c_{k,l_k})$	l_k
$\vdots$	$\vdots$	$\vdots$
x_K	$\mathbf{c}_K = (c_{K,1}, c_{K,2}, \ldots, c_{K,l}, \ldots, c_{K,l_K})$	l_K

The expected code length $L(\mathcal{C})$ of the variable length source code is determined by the source symbols distribution $\mathbf{p}$ and the length of the individual codewords.

$$L(\mathcal{C}) = \sum_{k=1}^{K} p(x_k) l_k \tag{2.27}$$

2.4.1 Uniquely decodable codes

The transmission of data from an information source (e.g. transmitter) to an information sink (e.g. receiver) is generally in the form of a continuous data stream; at the receiving end, one should be able to reconstruct without any ambiguity the source symbol sequence from the received sequence of codewords. However, some conditions must be imposed on the choice of the set of codewords, or code $\mathcal{C}$, to insure that a received sequence would uniquely determine the original transmitted information sequence generated from the source X. The ensemble (or universe) of all possible codes $\{\mathcal{C}\}$ can be subdivided into smaller sets of codes:

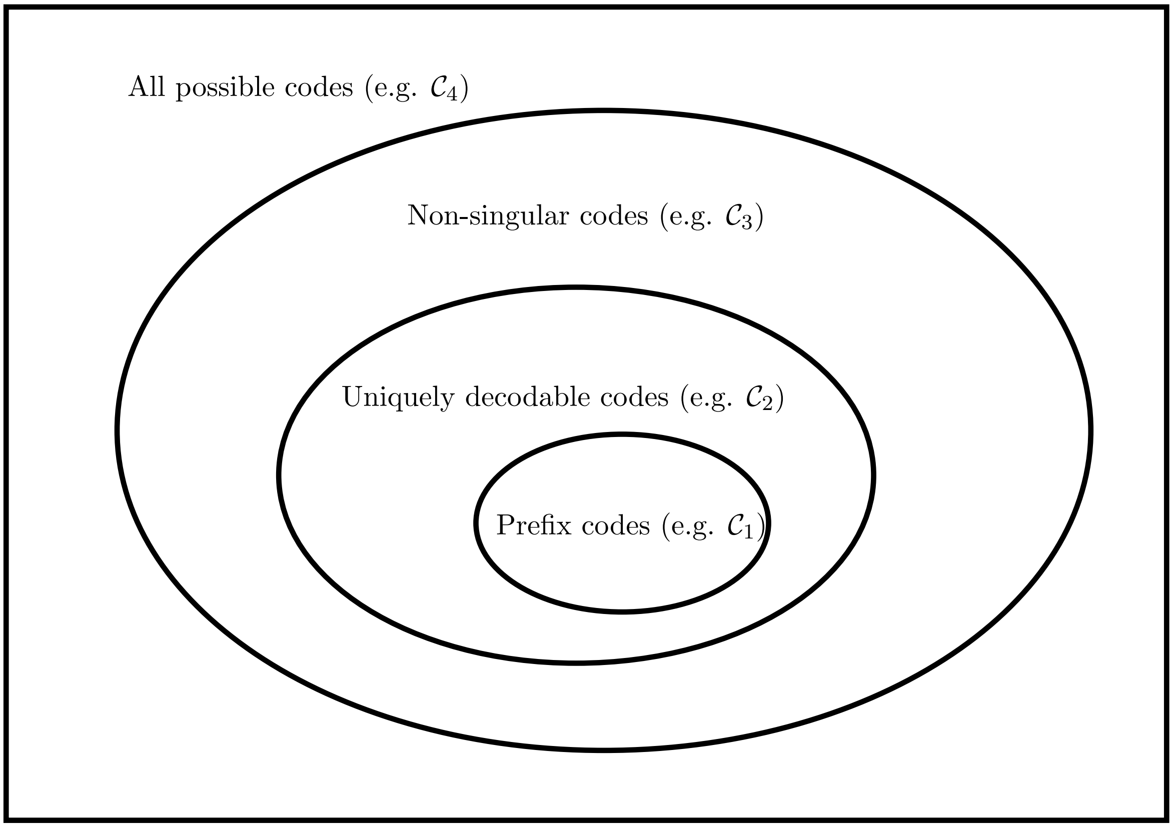

All possible codes (e.g. $\mathcal{C}_4$)
Non-singular codes (e.g. $\mathcal{C}_3$)
Uniquely decodable codes (e.g. $\mathcal{C}_2$)
Prefix codes (e.g. $\mathcal{C}_1$)

Figure 2.2: Source codes' subdivisions.

a) **Prefix code**:

A code is called a *prefix code*, or sometimes *instantaneous code*, if no codeword $\mathbf{c}_k$ is a prefix of any other codeword $\mathbf{c}_{k'}$ in the code $\mathcal{C}$. For instance, the code $\mathcal{C}_1$ is a prefix code:

$$\mathcal{C}_1 = \{0, 10, 110, 111\}$$

b) **Uniquely decodable code**:

A code is called a *uniquely decodable code* if each possible sequence of codewords can be produced only by a unique sequence of source symbols.

$$\mathcal{C}_2 = \{0, 01, 011, 0111\}$$

The code $\mathcal{C}_2$ is not a prefix code: the codeword $\mathbf{c}_1 = (0)$ is a prefix of $\mathbf{c}_2 = (01)$, $\mathbf{c}_3 = (011)$ and $\mathbf{c}_4 = (0111)$. Nevertheless, there is no ambiguity in the decoding process for such a code. The received sequence "001110100110" for instance corresponds to the source symbol sequence "$x_1, x_4, x_2, x_1, x_3, x_1$" and no other one.

c) **Non-singular code**:

The only condition for a *non-singular code* is that all codewords in such a code is different from the other codewords, i.e. $\mathbf{c}_k \neq \mathbf{c}_{k'}$ if $k \neq k'$.

$$\mathcal{C}_3 = \{0, 1, 00, 11\}$$

Here, code $\mathcal{C}_3$ is neither prefix code nor a uniquely decodable code. The received string "01000110111" can be decoded in many ways as "$x_1, x_2, x_1, x_1, x_1, x_2, x_2, x_1, x_2, x_2, x_2$" or "$x_1, x_2, x_3, x_1, x_4, x_1, x_2, x_4$", etc.

d) **All possible codes**:

Here, one considers all possible mappings from the set of K symbols $\{x_1, \ldots, x_K\}$ into K codewords $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, and this without any conditions.

$$\mathcal{C}_4 = \{0, 10, 11, 10\}$$

Both source letters x_2 and x_4 are encoded with the same string "10".

2.4.2 Kraft Inequality and Optimum Codes

2.4.2.1 Kraft Inequality for Prefix Codes

Theorem *(Kraft Inequality for Prefix Codes)*

A prefix code $\mathcal{C}$, with K codewords $\{\mathbf{c}_1, \ldots, \mathbf{c}_k, \ldots, \mathbf{c}_K\}$ of lengths $l_1, \ldots,\ l_k, \ldots,\ l_K$ and using an alphabet of size J, must satisfy the following inequality (Kraft Inequality):

$$\boxed{\sum_{k=1}^{K} J^{-l_k} \leq 1}$$

Proof:

A prefix code $\mathcal{C}$ can be represented as a tree where each branch of the tree represents a symbol from a codeword, and a codeword is represented by a path from the root to a leaf (see Figure 2.3).

For a prefix code, no codeword can be the *prefix* of another codeword. On the tree, because of this prefix condition, no branch (i.e., no codeword, or part of a codeword) extends beyond a given leaf (that is from a shorter codeword). In other words, a given *ancestor* codeword (leaf) *disables* all *descendants* codewords (branches).

Let the lengths $l_1 \leq \ldots \leq l_k \leq \ldots \leq l_K$. If the length of a first codeword $\mathbf{c}_1$ is $l_1 = 1$, then one of the J branches at level 1 in the code tree is disabled. Thus, a fraction J^{-1}, or J^{-l_1}, of the total number of branches in the tree is disabled. Now, if a second codeword $\mathbf{c}_2$ of the same length $l_2 = 1$ is used, then an additional fraction $J^{-l_2} = J^{-1}$ is disabled in the code tree.

A codeword $\mathbf{c}_k$ of length l_k once chosen will result in another J^{-l_k} of the code tree to be again disabled. Since the prefix code $\mathcal{C}$ consists in the K codewords $\{\mathbf{c}_1,\ \ldots,\ \mathbf{c}_k,\ \ldots,\ \mathbf{c}_K\}$, of respective lengths $l_1, \ldots,\ l_k, \ldots,\ l_K$, then, once all codewords are used, the sum of the fractions of the branches of the code tree will be given by:

$$\sum_{k=1}^{K} J^{-l_k} \leq 1 \qquad (2.28)$$

the *sum of fractions of the total number of branches being less or at most equal to unity.*

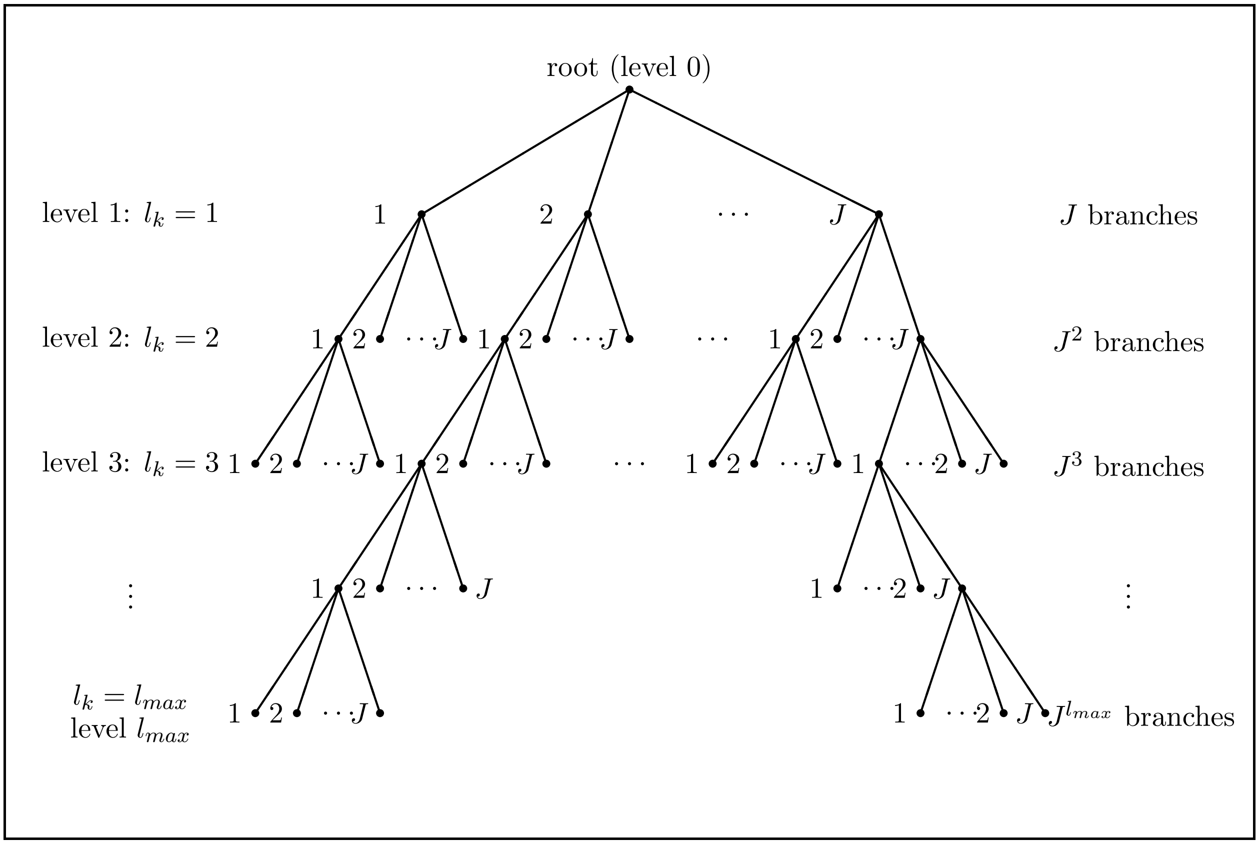


Figure 2.3: Code tree (Kraft inequality): ancestors and descendants.

QED

2.4.2.2 Kraft Inequality for Uniquely Decodable Codes

To be able to decode without ambiguity the codewords one does not necessarily need to chose a prefix code. We have seen that the *larger set of uniquely decodable codes* can be used for source compaction coding. Since the set of uniquely decodable codes contains the set of prefix codes, then it seems that one can contruct a more efficient code due to its greater *flexibility*. However, quite surprisingly, a uniquely decodable code still need to satisfy the Kraft inequality.

Theorem *(Uniquely Decodable Code):*

A uniquely decodable code $\mathcal{C}$, defined on an alphabet of size J, with K codewords having the lengths $l_1, \ldots, l_k, \ldots, l_K$, must satisfy the Kraft inequality:

$$\boxed{\sum_{k=1}^{K} J^{-l_k} \leq 1}$$

Proof: Consider a string of N concatenated codewords (i.e., sequence of codewords). Assume that the lengths of the codewords are arranged in a ascending order:

$$l_1 \leq \ldots \leq l_k \leq \ldots \leq l_K \tag{2.29}$$

Consider the following sum over each of the K codewords, for all N codewords in the sequence:

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N = \underbrace{\sum_{k_1=1}^{K} J^{-l_{k_1}}}_{\text{first codeword}} \underbrace{\sum_{k_2=1}^{K} J^{-l_{k_2}}}_{\text{second codeword}} \ldots \underbrace{\sum_{k_N=1}^{K} J^{-l_{k_N}}}_{\text{last codeword}} \tag{2.30}$$

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N = \underbrace{\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \ldots \sum_{k_N=1}^{K}}_{\text{sum over all possible sequences}} \underbrace{J^{-(l_{k_1}+l_{k_2}+\ldots+l_{k_N})}}_{\text{all strings of } N \text{ codewords}} \tag{2.31}$$

The exponent of J, ignoring the minus sign $(-)$, represents the total length of a particular sequence of codewords:

$$l_{k_1} + l_{k_2} + \ldots + l_{k_N} \tag{2.32}$$

Since the codewords are arranged according to their respective lengths, then the minimum and maximum of the exponent is given by:

$$l_{min} = \min\left[l_{k_1} + l_{k_2} + \ldots + l_{k_N}\right] \quad = \quad N\ l_1 \tag{2.33}$$

$$l_{max} = \max\left[l_{k_1} + l_{k_2} + \ldots + l_{k_N}\right] \quad = \quad N\ l_K \tag{2.34}$$

Thus the total length l of a particular sequence of codewords ranges from $l_{min} = N\ l_1$ to $l_{max} = N\ l_K$:

$$l_{min} = N\ l_1 \leq l \leq N\ l_K = l_{max} \tag{2.35}$$

Let the parameter A_l be an *enumerator* indicating the number of sequences of N codewords for which the total length is exactly l. We can then write that:

$$\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \cdots \sum_{k_N=1}^{K} J^{-(l_{k_1} + l_{k_2} + \ldots + l_{k_N})} = \sum_{l=l_{min}}^{l_{max}} A_l J^{-l} \tag{2.36}$$

Now, since code $\mathcal{C}$ is a uniquely decodable code, there is a maximum number of *distincts sequences* of length l which is equal to:

$$\max(A_l) = J^l \tag{2.37}$$

thus $A_l \leq J^l$ and:

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N \quad = \quad \sum_{l=l_{min}}^{l_{max}} A_l J^{-l} \tag{2.38}$$

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N \quad \leq \quad \sum_{l=l_{min}}^{l_{max}} J^l J^{-l} \tag{2.39}$$

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N \quad \leq \quad l_{max} - l_{min} + 1 \tag{2.40}$$

$$\left(\sum_{k=1}^{K} J^{-l_k}\right)^N \quad \leq \quad N(l_K - l_1) + 1 \tag{2.41}$$

Taking the Nth root on both sides:

$$\sum_{k=1}^{K} J^{-l_k} \quad \leq \quad \left[N(l_K - l_1) + 1\right]^{1/N} \tag{2.42}$$

Choosing the sequence length N to be arbitrary large, i.e. $N \rightarrow \infty$ or $1/N \rightarrow 0$:

$$\lim_{N\to\infty} [N(l_K - l_1) + 1]^{1/N} = 1 \tag{2.43}$$

and therefore,

$$\sum_{k=1}^{K} J^{-l_k} \leq 1$$

QED

2.4.2.3 Lower Bound on the Average Codeword Length

In this section, we consider a uniquely decodable (source) code $\mathcal{C}$ and determine a lower bound on its average, or expected, codeword length $L(\mathcal{C})$.

Let X be a memoryless source of alphabet size K having the distribution: $\mathbf{p}(x) = \{p(x_1), \ldots, p(x_k), \ldots, p(x_K)\}$. This source of information is to be represented by a variable length code $\mathcal{C}$ $=\{\mathbf{c}_1, \ldots, \mathbf{c}_k, \ldots, \mathbf{c}_K\}$, where each component $c_{k,l} \in \{0, \ldots, J-1\}$, i.e. taken from an alphabet of size J.

Theorem *(Lower Bound on the Average Codeword Length):*

The average codeword length $L(\mathcal{C})$ of a uniquely decodable code $\mathcal{C}$ is lower bounded by the source entropy $H(X)$:

$$\boxed{L(\mathcal{C}) \log_b J \geq H(X) \qquad \text{or equivalently:} \qquad L(\mathcal{C}) \geq \frac{H(X)}{\log_b J}}$$

where b is the logarithmic base used to compute the source entropy $H(X)$.

Proof:

If the theorem statement is true then:

$$\begin{aligned} L(\mathcal{C}) \log_b J - H(X) &\geq 0 \qquad \text{or} && (2.44)\\ H(X) - L(\mathcal{C}) \log_b J &\leq 0 \qquad \text{or using natural logarithms:} && (2.45)\\ (\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] &\leq 0 && (2.46) \end{aligned}$$

By definition, the entropy $H(X)$ and the average codeword lenght $L(\mathcal{C})$ are:

$$H(X) = -\sum_{k=1}^{K} p(x_k) \ln p(x_k) \qquad \text{and} \qquad L(\mathcal{C}) = \sum_{k=1}^{K} p(x_k) l_k \tag{2.47}$$

The left-hand side of the previous inequality becomes:

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] = (\log_b e)\left[-\sum_{k=1}^{K} p(x_k) \ln p(x_k) - \left(\sum_{k=1}^{K} p(x_k) l_k\right) \ln J\right] \tag{2.48}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] = (\log_b e)\left[\sum_{k=1}^{K} p(x_k)\left(-\ln p(x_k) - l_k \ln J\right)\right] \tag{2.49}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] = (\log_b e)\left[\sum_{k=1}^{K} p(x_k)\left(\ln J^{-l_k} - \ln p(x_k)\right)\right] \tag{2.50}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] = (\log_b e)\left[\sum_{k=1}^{K} p(x_k) \ln\left(\frac{J^{-l_k}}{p(x_k)}\right)\right] \tag{2.51}$$

Since the alphabet size J and the probabilities $\{p(x_k)\}$ are always positive, then the ratio $\frac{J^{-l_k}}{p(x_k)}$ is also positive. Since for $x \geq 0$, $\ln x \leq x - 1$, then

$$\ln\left[\frac{J^{-l_k}}{p(x_k)}\right] \leq \left[\frac{J^{-l_k}}{p(x_k)} - 1\right] \qquad \text{for } k = 1, \ldots, K \tag{2.52}$$

and therefore,

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] \leq (\log_b e)\left[\sum_{k=1}^{K} p(x_k)\left[\frac{J^{-l_k}}{p(x_k)} - 1\right]\right] \tag{2.53}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] \leq (\log_b e)\left[\sum_{k=1}^{K} J^{-l_k} - \sum_{k=1}^{K} p(x_k)\right] \tag{2.54}$$

Since the variable length code ($\mathcal{C}$) is uniquely decodable, it must satisfies the Kraft inequality, i.e. $\sum_{k=1}^{K} J^{-l_k} \leq 1$. On the other hand, by definition, the sum of probabilities $\sum_{k=1}^{K} p(x_k) = 1$, and thus,

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] \leq (\log_b e)\left[\underbrace{\sum_{k=1}^{K} J^{-l_k}}_{\leq 1} - \underbrace{\sum_{k=1}^{K} p(x_k)}_{=1}\right] \tag{2.55}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] \leq (\log_b e)\left[\underbrace{\sum_{k=1}^{K} J^{-l_k} - \sum_{k=1}^{K} p(x_k)}_{\leq 0}\right] \tag{2.56}$$

$$(\log_b e)\left[H(X) - L(\mathcal{C}) \ln J\right] \leq 0 \tag{2.57}$$

Therefore, for any logarithmic base b:

$$H(X) - L(\mathcal{C}) \log_b J \leq 0 \tag{2.58}$$

and the average codeword length is larger or equal to the source entropy:

$$L(\mathcal{C}) \geq \frac{H(X)}{\log_b J} \tag{2.59}$$

QED

2.4.2.4 Upper Bound on the Average Codeword Length

In the previous section, we have seen that the average codeword length $L(\mathcal{C})$ of a uniquely decodable source compaction code $\mathcal{C}$ is larger or equal to the entropy $H(X)$ of the source X. In this section, we show that it is always possible to construct a uniquely decodable with an average codeword length which is arbitrarily close to the source's entropy.

Theorem *(Upper Bound on the Average Codeword Length):*

Given a memoryless source $X = \{x_1, \ldots, x_k, \ldots, x_K\}$, it is possible to construct a uniquely decodable code $\mathcal{C}$ for which the average codeword length $L(\mathcal{C})$ is upper bounded by:

$$\boxed{L(\mathcal{C}) < \frac{H(X)}{\log_b J} + 1}$$

Proof:

For this proof, we can choose the (J-ary) Shannon code construction where each codeword has a specific length l_k such that:

$$l_k = \lceil -\log_J p(x_k) \rceil \tag{2.60}$$

or, by definition of the ceiling function[1]:

$$-\log_J p(x_k) \leq l_k = \lceil -\log_J p(x_k) \rceil < -\log_J p(x_k) + 1 \tag{2.61}$$

We must ensure that the code is uniquely decodable. Does it satisfy the Kraft inequality? From the above inequality, we know that:

$$l_k \quad \geq \quad -\log_J p(x_k) \qquad \text{or} \tag{2.62}$$

$$-l_k \quad \leq \quad \log_J p(x_k) \tag{2.63}$$

Raising J by the left and right hand sides of the inequality,

$$J^{-l_k} \quad \leq \quad J^{\log_J p(x_k)} = p(x_k) \qquad \text{for } k = 1, \ldots, K. \tag{2.64}$$

[1] $\lceil x \rceil$: smallest integer larger than or equal to the argument x.

Summing over the set of source symbols:

$$\sum_{k=1}^{K} J^{-l_k} \quad \leq \quad \sum_{k=1}^{K} p(x_k) = 1 \tag{2.65}$$

This implies that the code satisfies the Kraft inequality and can then be represented as a prefix code or a uniquely decodable code.

Now consider that for the Shannon code, we also have:

$$l_k < -\log_J p(x_k) + 1 \tag{2.66}$$

Averaging on both sides of this inequality over all source symbols, i.e. for $k = 1, \ldots, K$:

$$\begin{aligned}
\sum_{k=1}^{K} p(x_k) l_k \quad &< \quad \sum_{k=1}^{K} p(x_k) \left[-\log_J p(x_k) + 1 \right] && (2.67)\\
\sum_{k=1}^{K} p(x_k) l_k \quad &< \quad -\sum_{k=1}^{K} p(x_k) \log_J p(x_k) + \sum_{k=1}^{K} p(x_k) && (2.68)\\
L(\mathcal{C}) \quad &< \quad H(X) + 1 && (2.69)
\end{aligned}$$

where the entropy $H(X)$ is the entropy of the source X expressed using base J. Converted to an arbitry base b, one obtains:

$$L(\mathcal{C}) < \frac{H(X)}{log_b J} + 1 \tag{2.70}$$

QED

Thus, a uniquely decodable code is lower bounded and upper bounded as:

$$\boxed{\underbrace{\frac{H(X)}{\log_b J}}_{\text{unique codewords}} \quad \leq L(\mathcal{C}) < \quad \underbrace{\frac{H(X)}{\log_b J} + 1}_{\text{existence of the code}}}$$

2.4.2.5 Encoding N i.i.d. source symbols

Let $\mathbf{X} = (X_1, \ldots, X_n, \ldots, X_N)$ be a sequence of N independent and identically distributed (i.i.d.) random variables (or a N-dimensional random vector) of entropy:

$$H(\mathbf{X}) = H(X_1) + \ldots + H(X_n) + \ldots + H(X_N) = NH(X_n) = NH(X) \tag{2.71}$$

The average codeword length needed to represent the random vector $\mathbf{X}$ of alphabet size K^N (i.e. $\{x_1, \ldots, x_K, x_{K+1}, \ldots, x_{K^N}\}$). The average (concatenated) codeword length $L_N(\mathcal{C})$ is N times the average codeword length of the code use to encode a single source X_n, $L(\mathcal{C})$:

$$L_N(\mathcal{C}) = NL(\mathcal{C}) \tag{2.72}$$

The average concatenated codeword length is bounded by:

$$\begin{aligned} \frac{H(\mathbf{X})}{\log_b J} \leq \quad L_N(\mathcal{C}) \quad &< \frac{H(\mathbf{X})}{\log_b J} + 1 \qquad \text{or} & (2.73) \\ \frac{NH(X)}{\log_b J} \leq \quad NL(\mathcal{C}) \quad &< \frac{NH(X)}{\log_b J} + 1 & (2.74) \end{aligned}$$

Then, dividing both sides of the equation by N,

$$\frac{H(X)}{\log_b J} \leq \quad L(\mathcal{C}) \quad < \frac{H(X)}{\log_b J} + \frac{1}{N}$$

where the average codeword length $L(\mathcal{C})$ per source symbol can be made arbitrary close to the entropy per symbol $H(X)$ by increasing the number of symbols N being encoded, hence reducing the ratio $\frac{1}{N}$.

2.4.3 Optimum coding (Huffman code)

Consider a source $X = \{x_1, \cdots, x_K\}$ with a distribution $\mathbf{p}(x) = \{p(x_1), \cdots, p(x_K)\}$. The problem of source coding is to minimize the average codeword length $L(\mathcal{C})$ of a uniquely decodable code $\mathcal{C}$:

$$\begin{array}{ccccc} x_1 & \Rightarrow & \mathbf{c}_1 & = & (c_{1,1}, \ldots, c_{1,l_1}) \\ \vdots & & \vdots & & \vdots \\ x_k & \Rightarrow & \mathbf{c}_k & = & (c_{k,1}, \ldots, c_{k,l_k}) \\ \vdots & & \vdots & & \vdots \\ x_K & \Rightarrow & \mathbf{c}_K & = & (c_{K,1}, \ldots, c_{K,l_K}) \end{array} \tag{2.75}$$

where l_k is, as previously, the length of the codeword $\mathbf{c_k}$ used to represent the symbol x_k and M the codeword symbol alphabet size. The problem consists in minimizing the expected length of the code for a given input distribution $\mathbf{p}(\mathbf{x})$.

$$\min_{\{\mathcal{C}\}} L(\mathcal{C}) = \min_{\{\mathcal{C}\}} \sum_{k=1}^{K} p(x_k) l_k \tag{2.76}$$

Let the distribution of the source symbols be arranged in a decreasing order of probability:

$$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_k) \ldots \geq p(x_{K-1}) \geq p(x_K) \tag{2.77}$$

The source symbols are to be encoded using a prefix code (and thus a uniquely decodable code) where the length of the codewords are $l_1, \ldots, l_K$.

If for $k < j$, which means that $p(x_k) \geq p(x_j)$, the length $l_k > l_j$ (which is not wanted), then one can exchange the 2 codewords. The improvement, or reduction, ΔL in the average codeword length, due to this permutation of codewords, is equal to:

$$\begin{aligned} \Delta L &= [p(x_k) l_j + p(x_j) l_k] - [p(x_k) l_k + p(x_j) l_j] \\ &= [p(x_j) - p(x_k)]\ [l_k - l_j] \\ \Delta L &\leq 0 \end{aligned} \tag{2.78}$$

If the original code was already an optimum code then $\Delta L = 0$.

2.4.3.1 Procedure to construct a binary Huffman code

a) Arrange the source symbols in order of decreasing probabilities:

$$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_k) \ldots \geq p(x_{K-1}) \geq p(x_K)$$

For instance:

$$p(\text{“E”}) \approx 10.3\% \quad \geq \quad p(\text{“T”}) \approx 7.96\% \quad \geq \quad \ldots \quad \geq \quad p(\text{“Z”}) \approx 0.05\%$$

b) Assign a “1” (or “0”) to the last digit of the K^{th} codeword $\mathbf{c}_K$ and a “0” (or “1”) to the last digit of codeword $\mathbf{c}_{K-1}$. Then the 2 codewords $\mathbf{c}_K$ and $\mathbf{c}_{K-1}$ have the same codeword length $l_K = l_{K-1}$.

$$\begin{aligned} \mathbf{c}_{K-1} &\Rightarrow (c_{K-1,1}, \ldots, c_{K-1,l_{(K-1)}-1}, 0) \\ \mathbf{c}_K &\Rightarrow (c_{K,1}, \ldots, c_{K,l_K-1}, 1) \end{aligned}$$

c) Form a new source X' where $x'_k = x_k$ for $k = 1, 2, \ldots, K-2$, and create a new “pseudosymbol” $x'_{K-1} = x_{K-1} \cup x_K$. The resulting new distribution $\mathbf{p}'$ is then given by:

$$\begin{aligned} p(x'_k) &= p(x_k) \qquad \text{for } 1 \leq k \leq K-2 \text{ and} \\ p(x'_{K-1}) &= p(x_{K-1}) + p(x_K) \end{aligned}$$

d) Rearrange the new set of probabilities (or distribution) such as:

$$p(x'_1) \geq \ldots \geq p(x'_k) \ldots \geq p(x'_{K-1})$$

e) Repeat steps 2 to 5 until all original source symbols $\{x_k\}$ have been encoded.

Example *(Huffman code for single symbols):*

Consider a source $X = \{x_1, x_2, x_3, x_4, x_5\}$ characterized with the following distribution $\mathbf{p}(x)$: $p(x_1) = 0.35$, $p(x_2) = 0.22$, $p(x_3) = 0.18$, $p(x_4) = 0.15$ and $p(x_5) = 0.10$. The entropy $H(X)$ of this source of information is then equal to:

$$H(X) = -\sum_{k=1}^{5} p(x_k) \log_2 p(x_k) = 2.1987 \; Sh$$

This source can be encoded using a binary Huffman code. Table 2.2 and Figure 2.4 shown below, indicate the resulting codewords $\{\mathbf{c}_k = (c_{k,1}, \cdots, c_{k,l_k})\}$ along with the codeword length l_k for this particular source of information.

The average codeword length L is equal to:

$$L = \sum_{k=1}^{5} p(x_k) \; l_k = 2.25 \; bits/source\ symbol$$

providing a variable-length source code efficiency ξ of 97.7%:

$$\xi = \frac{H(X)}{L} = \frac{2.1987 \; Sh}{2.25 \; bits} = 97.7\%$$

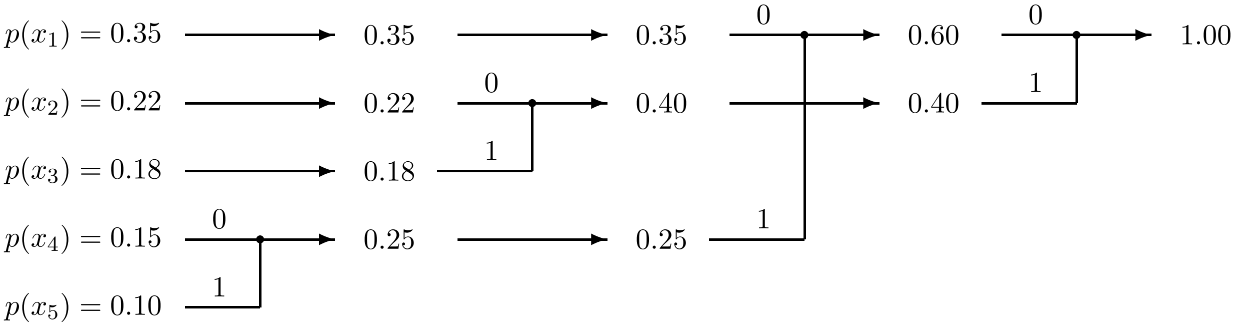


Figure 2.4: Binary Huffman code structure

Table 2.2: Huffman code for single symbols

Symbol x_k	Probability $p(x_k)$	Codeword $\mathbf{c}_k = (c_{k,1}, \cdots, c_{k,l_k})$	Length l_k
x_1	$p(x_1) = 0.35$	0 0	2
x_2	$p(x_2) = 0.22$	1 0	2
x_3	$p(x_3) = 0.18$	1 1	2
x_4	$p(x_4) = 0.15$	0 1 0	3
x_5	$p(x_5) = 0.10$	0 1 1	3

Example *(Huffman code for pairs of symbols):*

Let now construct another binary Huffman code, but this time to encode pairs of source symbols $\mathbf{x}_k$ (i.e., digrams): $\mathbf{x}_k \equiv (x_i, x_j)$. The entropy $H(\mathbf{X})$ of the source of digrams is now:

$$H(\mathbf{X}) = -\sum_{k=1}^{25} p(\mathbf{x}_k) \log_2 p(\mathbf{x}_k) = 4.3974 \; Sh/digram = 2.1987 \; Sh/source\ symbol$$

while the average codeword length L becomes (see table 2.3 on next page):

$$L = \sum_{k=1}^{25} p(\mathbf{x}_k) \; l_k = 4.4196 \; bits/digram = 2.2098 \; bits/source\ symbol$$

The Huffman code efficiency ξ has then increased to the ratio:

$$\xi = \frac{H(\mathbf{X})}{L} = \frac{4.3974\ Sh/digram}{4.4196\ bits/digram} = 99.5\%$$

Table 2.3: Huffman code for digram sourcewords $\mathbf{x}_k = (x_i, x_j)$

Sourceword $\mathbf{x}_k = (x_i, x_j)$	Probability $p(\mathbf{x}_k) = p(x_i)p(x_j)$	Codeword $c_k = (c_{k,1}, \cdots, c_{k,l_k})$	Length l_k
$\mathbf{x}_1 = (x_1, x_1)$	0.1225	1 0 0	3
$\mathbf{x}_2 = (x_1, x_2)$	0.0770	0 0 0 1	4
$\mathbf{x}_3 = (x_2, x_1)$	0.0770	0 0 1 0	4
$\mathbf{x}_4 = (x_1, x_3)$	0.0630	0 1 1 0	4
$\mathbf{x}_5 = (x_3, x_1)$	0.0630	0 1 1 1	4
$\mathbf{x}_6 = (x_1, x_4)$	0.0525	1 0 1 1	4
$\mathbf{x}_7 = (x_4, x_1)$	0.0525	1 1 0 0	4
$\mathbf{x}_8 = (x_2, x_2)$	0.0484	1 1 1 0	4
$\mathbf{x}_9 = (x_2, x_3)$	0.0396	0 0 1 1 0	5
$\mathbf{x}_{10} = (x_3, x_2)$	0.0396	0 0 0 0 1	5
$\mathbf{x}_{11} = (x_1, x_5)$	0.0350	0 0 1 1 1	5
$\mathbf{x}_{12} = (x_5, x_1)$	0.0350	0 1 0 0 0	5
$\mathbf{x}_{13} = (x_2, x_4)$	0.0330	0 1 0 0 1	5
$\mathbf{x}_{14} = (x_4, x_2)$	0.0330	0 1 0 1 0	5
$\mathbf{x}_{15} = (x_3, x_3)$	0.0324	1 0 1 0 0	5
$\mathbf{x}_{16} = (x_3, x_4)$	0.0270	1 0 1 0 1	5
$\mathbf{x}_{17} = (x_4, x_3)$	0.0270	1 1 0 1 0	5
$\mathbf{x}_{18} = (x_4, x_4)$	0.0225	1 1 1 1 0	5
$\mathbf{x}_{19} = (x_2, x_5)$	0.0220	1 1 1 1 1	5
$\mathbf{x}_{20} = (x_5, x_2)$	0.0220	0 0 0 0 0 0	6
$\mathbf{x}_{21} = (x_3, x_5)$	0.0180	0 0 0 0 0 1	6
$\mathbf{x}_{22} = (x_5, x_3)$	0.0180	0 1 0 1 1 0	6
$\mathbf{x}_{23} = (x_4, x_5)$	0.0150	0 1 0 1 1 1	6
$\mathbf{x}_{24} = (x_5, x_4)$	0.0150	1 1 0 1 1 0	6
$\mathbf{x}_{25} = (x_5, x_5)$	0.0100	1 1 0 1 1 1	6

2.4.3.2 Non-binary Huffman codes

Let $X = \{x_1, \cdots, x_K\}$ be the source of information with the distribution $\mathbf{p}(x) = \{p(x_1), \cdots, p(x_K)\}$ and let the codeword symbol alphabet size $M \neq 2$. A non-binary Huffman code $\mathcal{C}$ can be contructed as indicated below on Figure 2.5. Note that, this time, since the number of source symbols K may not be exactly equal to $c(M-1)+M$, where c is an arbitrary integer.

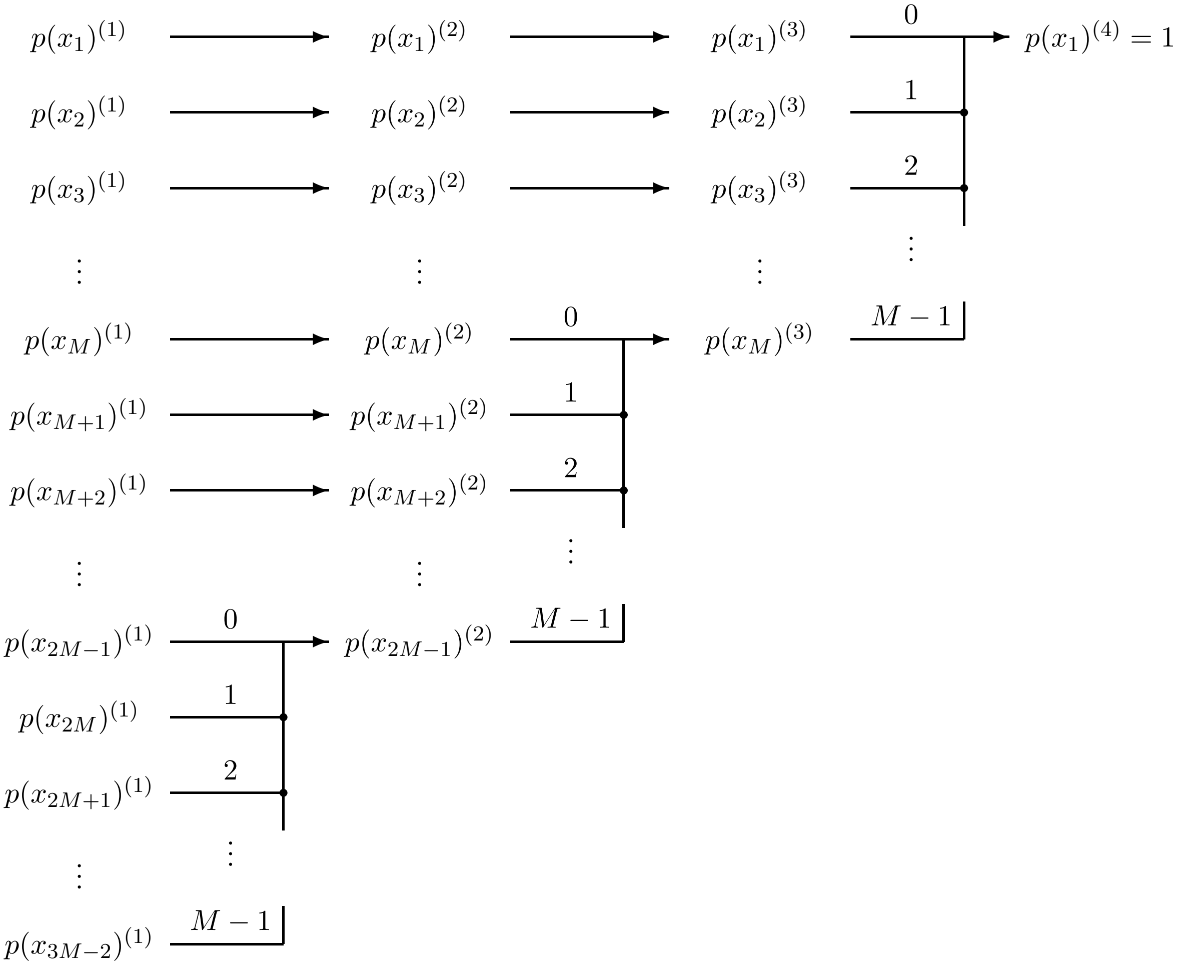


Figure 2.5: Structure of a non-binary Huffman code

2.5 Fixed length source compaction codes (Shannon source coding theorem)

Consider a sourcewords, or sequences of information, of length N, generated by a discrete memoryless source X of independent and identically distributed (i.i.d.) random variables with of entropy $H(X)$. These sourcewords are to be Consider a sourcewords, or sequences of information, of length N, generated by a discrete memoryless source X of encoded into codewords of blocklength L where each component is taken from an alphabet of size J.

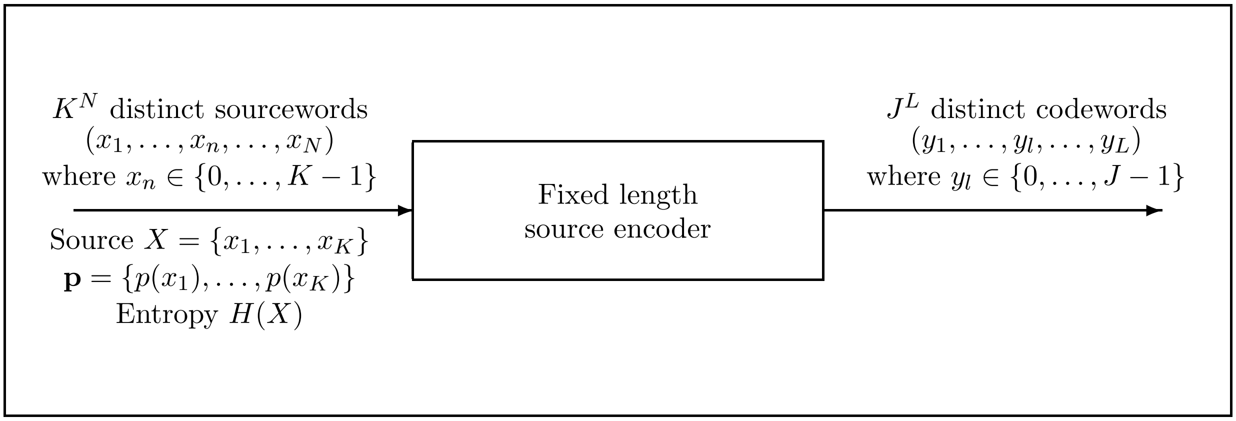

K^N distinct sourcewords
$(x_1, \ldots, x_n, \ldots, x_N)$
where $x_n \in \{0, \ldots, K-1\}$
Source $X = \{x_1, \ldots, x_K\}$
$\mathbf{p} = \{p(x_1), \ldots, p(x_K)\}$
Entropy $H(X)$
Fixed length source encoder
J^L distinct codewords
$(y_1, \ldots, y_l, \ldots, y_L)$
where $y_l \in \{0, \ldots, J-1\}$

Figure 2.6: Fixed length source compaction encoder.

Theorem *(Shannon Source Coding Theorem)*

Consider a memoryless source X of entropy $H(X)$. It is possible to construct a fixed length source compaction code that encode sourcewords of length N into codewords of length L, having an arbitrary small *block decoding failure probability* P_e, provided that:

a) $L \log_b J > NH(X)$ and

b) $N \geq N_0$ (sufficiently large)

where b is the base of the source entropy $H(X)$.

Proof:

The set of all K^N sourcewords can be partionned into the sets of typical sequences, i.e. $\mathcal{T}_X(\delta)$, and the set of non typical, or unlikely, sequences $\mathcal{T}_X^c(\delta)$. The number of typical sequences $\|\mathcal{T}_X(\delta)\|$ is bounded as:

$$(1-\epsilon) b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]} \tag{2.79}$$

The maximum number of possible codewords of length L is J^L whereas the number of typical sequences of the information source X is upper bounded by $b^{N[H(X)+\delta]}$.

If, as required by the theorem, the alphabet size J and the codeword length L are such that $L \log_b J > NH(X)$, then, raising b by both terms of the inequality:

$$\begin{aligned} L \log_b J &> NH(X) \qquad \text{implies that} & (2.80)\\ b^{L \log_b J} &> b^{NH(X)} & (2.81)\\ J^L &> b^{NH(X)} & (2.82) \end{aligned}$$

The parameter δ that defines the typical sequences, can be chosen such that both sides are equal (by allowing a sufficient number of sequences to be considered as *typical*:

$$\underbrace{J^L}_{\text{codewords}} \geq \underbrace{b^{N[H(X)+\delta]}}_{\text{typical sequences}} \tag{2.83}$$

hence providing a *unique codeword for each typical sequence*, the number of which can not be greater that $b^{N[H(X)+\delta]}$.

Therefore, the set of non encodable sequences (sourcewords) is contained in the set of atypical sequences $\mathcal{T}_X^c(\delta)$. For N *sufficiently large*, i.e. $N \geq N_0$, the probability that a source sequence is in the set of atypical sequences $\mathcal{T}_X^c(\delta)$ is smaller than ϵ. Therefore the probability of having a sourceword that is not typical, or the probability of an error decoding failure can be made arbitrary small: $P_e \leq \epsilon$.

QED

Example *(Fixed Length Source Compaction Encoder):*

Let X represent a memoryless source of information with the probabilities $p(x_1) = 0.1$ and $p(x_2) = 0.9$. Its entropy $H(X)$ is:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i) = -\left[(0.1)\log_2(0.1) + (0.9)\log_2(0.9)\right] \\ H(X) &= 0.4690 \qquad \text{(Shannons)} \end{aligned}$$

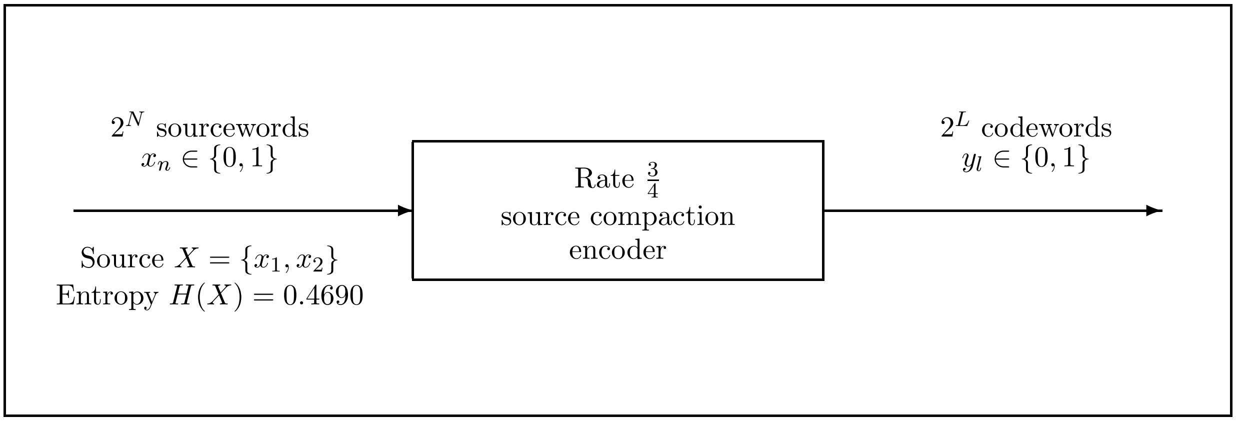


Figure 2.7: Rate $\frac{3}{4}$ fixed length source compaction encoder.

Suppose that we use the fixed length source compaction encoder depicted on Figure 2.7 to encode N-bit sourcewords into L bits binary codewords, where $N = 4$ and $L = 3$, hence resulting in a rate $\frac{3}{4}$ source encoder. We note that the condition of Shannon source coding theorem is satisfied, that is:

$$H(X) = 0.4690 \leq \frac{L}{N} = \frac{3}{4} = 0.75 \qquad \text{(with } K = J = 2\text{)}$$

There are $2^4 = 16$ possible sourcewords of length 4. However, there are only $2^3 = 8$ possible codewords of length L.

We can partition the 16 sourcewords into a set of 7 typical sequences $\mathcal{T}_X(\delta)$ which will be assigned to a unique codeword and a set of 9 non typical sequences $\mathcal{T}_X^c(\delta)$ which will be represented by a *default codeword*. The probabilities of the sourcewords are, in increasing order:

$$\begin{aligned}
p(x_1)^4 &= 0.0001 \qquad \tbinom{4}{4} = 1 \text{ sourceword}\\
p(x_1)^3 p(x_2) &= 0.0009 \qquad \tbinom{4}{3} = 4 \text{ sourcewords}\\
p(x_1)^2 p(x_2)^2 &= 0.0081 \qquad \tbinom{4}{2} = 6 \text{ sourcewords}\\
p(x_1) p(x_2)^3 &= 0.0729 \qquad \tbinom{4}{1} = 4 \text{ sourcewords}\\
p(x_2)^4 &= 0.6561 \qquad \tbinom{4}{0} = 1 \text{ sourceword}
\end{aligned}$$

The probability $1 - P_e$ of *faithful decoding*, or of transmitting a sourceword which is in the set of typical sequence is:

$$\begin{aligned}
1 - P_e &= \underbrace{\binom{4}{0}}_{1} p(x_2)^4 + \underbrace{\binom{4}{1}}_{4} p(x_1)p(x_2)^3 + \underbrace{\min\left(\binom{4}{2}, 2\right)}_{2} p(x_1)^2 p(x_2)^2\\
1 - P_e &= 0.6561 + (4 \times 0.0729) + (2 \times 0.0081) = 9.6390 \times 10^{-1}
\end{aligned}$$

The decoding error probability P_e is then equal to 3.6100×10^{-2} or 3.61%.

Now, let the sourceword blocklength be increased from $N = 4$ to $N = 8$ and the codeword blocklength increased from $L = 3$ to $L = 6$, thus keeping the code rate $R = \frac{6}{8} = 0.75$ as before. The entropy per source symbol $H(X)$ remains the same as well, that is $H(X) = 0.4690$.

There are now $2^8 = 256$ 8-bit sourcewords to be encoded into $2^6 = 64$ 6-bit codewords. A unique 6-bit codeword is assigned to each of the 63 most likely sourcewords, or typical sequences (in $\mathcal{T}_X(\delta)$), and the remaining 193 sourcewords (atypical sequences in $\mathcal{T}_X^c(\delta)$) are encoded into the default 6-bit codeword. The probabilities of the 8-bit sourcewords are:

$$\begin{aligned}
p(x_1)^8 &= 1.0000 \times 10^{-8} \qquad \tbinom{8}{8} = 1 \text{ sourceword}\\
p(x_1)^7 p(x_2) &= 9.0000 \times 10^{-8} \qquad \tbinom{8}{7} = 8 \text{ sourcewords}\\
p(x_1)^6 p(x_2)^2 &= 8.1000 \times 10^{-7} \qquad \tbinom{8}{6} = 28 \text{ sourcewords}\\
p(x_1)^5 p(x_2)^3 &= 7.2900 \times 10^{-6} \qquad \tbinom{8}{5} = 56 \text{ sourcewords}\\
p(x_1)^4 p(x_2)^4 &= 6.5610 \times 10^{-5} \qquad \tbinom{8}{4} = 70 \text{ sourcewords}\\
p(x_1)^3 p(x_2)^5 &= 5.9049 \times 10^{-4} \qquad \tbinom{8}{3} = 56 \text{ sourcewords}\\
p(x_1)^2 p(x_2)^6 &= 5.3144 \times 10^{-3} \qquad \tbinom{8}{2} = 28 \text{ sourcewords}\\
p(x_1) p(x_2)^7 &= 4.7830 \times 10^{-2} \qquad \tbinom{8}{1} = 8 \text{ sourcewords}\\
p(x_2)^8 &= 4.3047 \times 10^{-1} \qquad \tbinom{8}{0} = 1 \text{ sourceword}
\end{aligned}$$

The faithful decoding probability $1 - P_e$ is then:

$$
\begin{aligned}
1 - P_e &= \underbrace{\binom{8}{0}}_{1} p(x_2)^8 + \underbrace{\binom{8}{1}}_{8} p(x_1)p(x_2)^7 + \underbrace{\binom{8}{2}}_{28} p(x_1)^2 p(x_2)^6 \\
&\quad + \underbrace{\min\left(\binom{8}{3}, 26\right)}_{26} p(x_1)^3 p(x_2)^5 \\
1 - P_e &= 4.3047 \times 10^{-1} + (8 \times 4.7830 \times 10^{-2}) + (28 \times 5.3144 \times 10^{-3}) \\
&\quad +(26 \times 5.9049 \times 10^{-4}) \\
1 - P_e &= 9.7726 \times 10^{-1}
\end{aligned}
$$

The decoding error probability P_e is then equal to 2.2739×10^{-2} or 2.2739%. Therefore, for the same source entropy $H(X) = 0.4690$ and the same code rate $R = \frac{L}{N} = \frac{6}{8} = \frac{3}{4} = 0.75$, the decoding error probability decreased from $P_e = 3.61\%$ to $P_e = 2.2739\%$ by increasing the sourceword blocklength from $N = 4$ to $N = 8$.

Theorem *(Converse of the Source Coding Theorem)*

Let $\epsilon > 0$. Given a memoryless source X of entropy $H(X)$, a codeword alphabet size J and a codeword length L, if:

a) $L \log_b J < NH(X)$ and

b) $N \geq N_0$

then the probability of decoding failure P_e is lower bounded by:

$$P_e \quad > \quad 1 - \epsilon$$

Example *(Fixed Length Source Compaction Encoder Revisited):*

Let X be a memoryless source of information but this times with the following probabilities $p(x_1) = 0.3$ and $p(x_2) = 0.7$. The new source entropy $H(X)$ is:

$$\begin{array}{rcl} H(X) & = & -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i) = -\left[(0.3)\log_2(0.3) + (0.7)\log_2(0.7)\right] \\ H(X) & = & 0.88129 \qquad \text{(Shannons)} \end{array}$$

Suppose that the sourcewords are again encoded with the same source compaction encoder of rate $R = \frac{L}{N} = 0.75$ (see Figure 2.8) as was used in the previous example (where the source entropy was only 0.4690 bits). Therefore,

$$R = \frac{L}{N} = 0.75 \quad < \quad H(X) = 0.88129$$

and this code do not satisfy the condition of the source coding theorem.

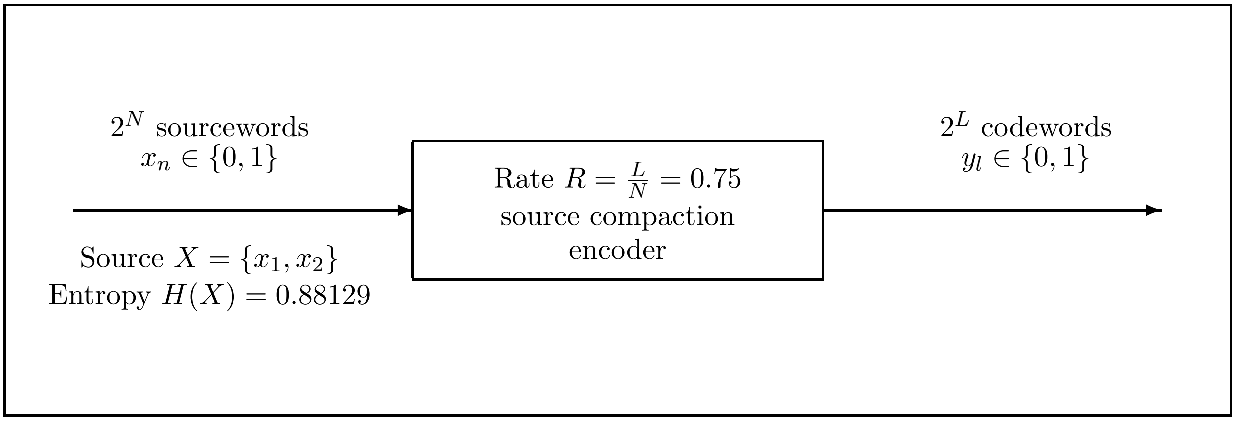


Figure 2.8: Rate $R = \frac{L}{N}$ fixed length source compaction encoder.

If we encode the N-bit sourcewords into L bits binary codewords with $N = 4$ and $L = 3$, and partition the 16 sourcewords into the set of 7 typical sequences $\mathcal{T}_X(\delta)$ and the set of 9 non typical sequences $\mathcal{T}_X^c(\delta)$, then the probabilities of the sourcewords will be:

$$\begin{array}{rcll} p(x_1)^4 & = & 2.4010 \times 10^{-1} & \binom{4}{4} = 1 \text{ sourceword} \\ p(x_1)^3 p(x_2) & = & 1.0290 \times 10^{-1} & \binom{4}{3} = 4 \text{ sourcewords} \\ p(x_1)^2 p(x_2)^2 & = & 4.4100 \times 10^{-2} & \binom{4}{2} = 6 \text{ sourcewords} \\ p(x_1) p(x_2)^3 & = & 1.8900 \times 10^{-2} & \binom{4}{1} = 4 \text{ sourcewords} \\ p(x_2)^4 & = & 8.1000 \times 10^{-3} & \binom{4}{0} = 1 \text{ sourceword} \end{array}$$

and the probability $1 - P_e$ is in that case given by:

$$\begin{aligned}
1 - P_e &= p(x_2)^4 + (4 \times p(x_1)p(x_2)^3) + (2 \times p(x_1)^2 p(x_2)^2) \\
1 - P_e &= 2.4010 \times 10^{-1} + (4 \times 1.0290 \times 10^{-1}) + (2 \times 4.4100 \times 10^{-2}) \\
1 - P_e &= 7.3990 \times 10^{-1}
\end{aligned}$$

leading to a decoding error probability P_e of 2.6010×10^{-1} or 26.01%.

Increasing the sourceword blocklength to $N = 8$ and the codeword blocklength to $L = 6$, for the same code rate $R = \frac{6}{8} = 0.75$, the 8-bit sourceword probabilities become:

$$\begin{aligned}
p(x_1)^8 &= 6.5610 \times 10^{-5} & \tbinom{8}{8} &= 1 \text{ sourceword} \\
p(x_1)^7 p(x_2) &= 1.5309 \times 10^{-4} & \tbinom{8}{7} &= 8 \text{ sourcewords} \\
p(x_1)^6 p(x_2)^2 &= 3.5721 \times 10^{-4} & \tbinom{8}{6} &= 28 \text{ sourcewords} \\
p(x_1)^5 p(x_2)^3 &= 8.3349 \times 10^{-4} & \tbinom{8}{5} &= 56 \text{ sourcewords} \\
p(x_1)^4 p(x_2)^4 &= 1.9448 \times 10^{-3} & \tbinom{8}{4} &= 70 \text{ sourcewords} \\
p(x_1)^3 p(x_2)^5 &= 4.5379 \times 10^{-3} & \tbinom{8}{3} &= 56 \text{ sourcewords} \\
p(x_1)^2 p(x_2)^6 &= 1.0588 \times 10^{-2} & \tbinom{8}{2} &= 28 \text{ sourcewords} \\
p(x_1) p(x_2)^7 &= 2.4706 \times 10^{-2} & \tbinom{8}{1} &= 8 \text{ sourcewords} \\
p(x_2)^8 &= 5.7648 \times 10^{-2} & \tbinom{8}{0} &= 1 \text{ sourceword}
\end{aligned}$$

The 256 8-bit sourcewords are encoded into 63 unique 6-bit codewords and the 193 sourcewords encoded into a default codeword. The faithful decoding probability $1 - P_e$ is then:

$$\begin{aligned}
1 - P_e &= 5.7648 \times 10^{-2} + (8 \times 2.4706 \times 10^{-2}) + (28 \times 1.0588 \times 10^{-2}) \\
&\quad +(26 \times 4.5379 \times 10^{-3}) \\
1 - P_e &= 6.6976 \times 10^{-1}
\end{aligned}$$

and the decoding error probability $P_e = 3.3024 \times 10^{-1}$. Therefore, by increasing the sourceword length from $N = 4$ to $N = 8$, the decoding error probability did *increase* from 26.01% to 33.024%!

2.6 Discrete sources with memory

So far, we have only considered sequences of independent, identically distributed (i.i.d.) random variables. We now consider information sources represented by random variables that are *dependent* from each other. The probability of a random vector, $p(\mathbf{x})$, is characterized by:

$$p(\mathbf{x}) = p(x_1, \ldots, x_n, \ldots, x_N) \neq \prod_{n=1}^{N} p(x_n) \tag{2.84}$$

We can represent a sequence of random variables $\{X_n\}_{n=1,\ldots,N}$, as a discrete-time random process. Here, we will assume that the information source can be modeled as a *stationary random process.*

Definition *(Stationary Random Process):*

A random process is said to be stationary if the joint distribution of any subset of the sequence of random variables, is invariant with respect to a time shift τ.

$$\boxed{f_{\mathbf{X}}(x_1, \ldots, x_N; t_1, \ldots, t_N) = f_{\mathbf{X}}(x_1, \ldots, x_N; t_1 + \tau, \ldots, t_N + \tau) \qquad \text{where } \tau \in \mathcal{R}}$$

For a discrete-time random process, this stationarity property can be written as:

$$\boxed{Pr\{X_1 = x_1, \ldots, X_N = x_N\} = Pr\{X_{1+l} = x_1, \ldots, X_{N+l} = x_N\}}$$

where l is a discrete-time shift.

Definition *(Markovian Random Process):*

A discrete-time random process $\{X_1, \ldots, X_n, \ldots, X_N\}$ is termed a *Markov* process if, for $1 \leq n \leq N$:

$$\boxed{Pr\{X_{n+1} = x_{n+1} | X_1 = x_1, \ldots, X_n = x_n\} = Pr\{X_{n+1} = x_{n+1} | X_n = x_n\}}$$

The *joint probability* can be rewritten from the general expression:

$$\begin{aligned} p(x_1,\ldots,x_n) &= p(x_1)\ p(x_2|x_1)\ p(x_3|x_1,x_2)\ldots p(x_n|x_1,\ldots,x_{n-1}) \\ p(x_1,\ldots,x_n) &= p(x_1)\ p(x_2|x_1)\ p(x_3|x_2)\ldots p(x_n|x_{n-1}) \qquad (2.85) \end{aligned}$$

Definition *(Time-invariant Markov Chain):*

A Markov process, or Markov chain, is said to be *time-invariant*, if the conditional probabilities $\{p(x_{n+1}|x_n)\}$ do not depend on the time index, that is, if:

$$\boxed{Pr(X_2 = x_2|X_1 = x_1) = Pr(X_{n+1} = x_2|X_n = x_1) \qquad \text{for } 1 \leq n \leq N}$$

Definition *(Entropy Rate):*

For a *source with memory*, the entropy rate $H_R(X)$ is defined as the average information content per source symbol:

$$\boxed{H_R(X) = \lim_{N\to\infty} \frac{1}{N} H(X_1,\ldots,X_n,\ldots,X_N)}$$

where the limit exists.

Example *(Entropy Rate of a Memoryless Source):*

Consider a *memoryless source* X where the random variables are, by definition, independent *but not necessarily* identically distributed. The entropy rate $H_R(X)$ is:

$$H_R(X) = \lim_{N\to\infty} \frac{1}{N}\ H(X_1,\ldots,X_n,\ldots,X_N) \qquad (2.86)$$

Using the chain rule for the entropy:

$$\begin{aligned} H_R(X) &= \lim_{N\to\infty} \frac{1}{N} \left[H(X_1) + H(X_2|X_1) + \ldots + H(X_N|X_1,X_2,\ldots,X_{N-1})\right] \\ &= \lim_{N\to\infty} \frac{1}{N} \left[H(X_1) + H(X_2) + \ldots + H(X_N)\right] \qquad (2.87) \\ H_R(X) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n) \qquad \text{(independence of variables)} \end{aligned}$$

Therefore,

$$\boxed{H_R(X) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n)}$$

Note that the limit may, or may not, exist.

a) if the random variables $\{X_n\}_{n=1,\ldots,N}$ are identically distributed, then the entropy $H(X_n)$ will be the same for all n.

$$X_n \rightsquigarrow X \Rightarrow \mathbf{p}(x) = \{p(x_k)\} \qquad \text{(unique distribution)}$$

The entropy rate $H_R(X)$ for the independent, identically distributed random variables case is:

$$\begin{aligned} H_R(X) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X) \\ &= \lim_{N\to\infty} \frac{N}{N} H(X) \\ H_R(X) &= H(X) \end{aligned}$$

The entropy rate $H_R(X)$ of a memoryless source of i.i.d. random variables is simply the entropy $H(X)$ of the random variable X.

b) if the random variables $\{X_n\}_{n=1,\ldots,N}$ are not identically distributed, the limit may not exist. For instance, consider the following binary distribution $\mathbf{p}(x) = \{p(x_{1,n}), p(x_{2,n})\}$:

$$p(x_{1,n}) = \begin{cases} 0.5 & \text{for} \quad 2k' < \log_b \log_b n \leq 2k' + 1 \\ 0.0 & \text{for} \quad 2k' + 1 < \log_b \log_b n \leq 2k' + 2 \end{cases}$$

where k' is a positive integer.

i) for $2 < n \leq 4$: $p(x_1) = 0.5$, $p(x_2) = 0.5$ and $H(X_3) = H(X_4) = 1\ Sh$;

ii) for $4 < n \leq 16$: $p(x_1) = 0.0$, $p(x_2) = 1.0$ and $H(X_5) = \ldots = H(X_{16}) = 0\ Sh$;

iii) for $16 < n \leq 256$: $p(x_1) = 0.5$, $p(x_2) = 0.5$ and $H(X_{17}) = \ldots = H(X_{256}) = 1\ Sh$;

iv) for $256 < n \leq 65,536$: $p(x_1) = 0.0$, $p(x_2) = 1.0$ and $H(X_{257}) = \ldots = H(X_{65,536}) = 0\ Sh$; and so on.

For this specific distribution, we observe that the *running average* of the entropy $H(X_n)$, as n increases, oscillates from $H(X_n) = 0\ Sh$ to $H(X_n) = 1\ Sh$. Then

$$H_R(X) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n)$$

does not converge (see Figure 2.9). Then the entropy rate $H_R(X)$ is not defined on this particular distribution.

$n =$	0	$\log_2 n =$	$-\infty$	$\log_2 \log_2 n =$	$\emptyset$
$n =$	1	$\log_2 n =$	$0.000\ 000$	$\log_2 \log_2 n =$	$-\infty$
$n =$	2	$\log_2 n =$	$1.000\ 000$	$\log_2 \log_2 n =$	$0.000\ 000$
$n =$	3	$\log_2 n =$	$1.584\ 963$	$\log_2 \log_2 n =$	$0.664\ 449$
$n =$	4	$\log_2 n =$	$2.000\ 000$	$\log_2 \log_2 n =$	$1.000\ 000$
$n =$	5	$\log_2 n =$	$2.321\ 928$	$\log_2 \log_2 n =$	$1.215\ 323$
$n =$	6	$\log_2 n =$	$2.584\ 963$	$\log_2 \log_2 n =$	$1.370\ 143$
$n =$	7	$\log_2 n =$	$2.807\ 355$	$\log_2 \log_2 n =$	$1.489\ 211$
$n =$	8	$\log_2 n =$	$3.000\ 000$	$\log_2 \log_2 n =$	$1.584\ 963$
$\vdots$		$\vdots$		$\vdots$	
$n =$	15	$\log_2 n =$	$3.906\ 891$	$\log_2 \log_2 n =$	$1.966\ 021$
$n =$	16	$\log_2 n =$	$4.000\ 000$	$\log_2 \log_2 n =$	$2.000\ 000$
$n =$	17	$\log_2 n =$	$4.087\ 463$	$\log_2 \log_2 n =$	$2.031\ 206$
$\vdots$		$\vdots$		$\vdots$	
$n =$	32	$\log_2 n =$	$5.000\ 000$	$\log_2 \log_2 n =$	$2.321\ 928$
$\vdots$		$\vdots$		$\vdots$	
$n =$	64	$\log_2 n =$	$6.000\ 000$	$\log_2 \log_2 n =$	$2.584\ 963$
$\vdots$		$\vdots$		$\vdots$	
$n =$	128	$\log_2 n =$	$7.000\ 000$	$\log_2 \log_2 n =$	$2.807\ 355$
$\vdots$		$\vdots$		$\vdots$	
$n =$	255	$\log_2 n =$	$7.994\ 353$	$\log_2 \log_2 n =$	$2.998\ 981$
$n =$	256	$\log_2 n =$	$8.000\ 000$	$\log_2 \log_2 n =$	$3.000\ 000$
$n =$	257	$\log_2 n =$	$8.005\ 625$	$\log_2 \log_2 n =$	$3.001\ 014$
$\vdots$		$\vdots$		$\vdots$	
$n =$	$65,535$	$\log_2 n =$	$15.999\ 978$	$\log_2 \log_2 n =$	$3.999\ 998$
$n =$	$65,536$	$\log_2 n =$	$16.000\ 000$	$\log_2 \log_2 n =$	$4.000\ 000$
$n =$	$65,537$	$\log_2 n =$	$16.000\ 022$	$\log_2 \log_2 n =$	$4.000\ 002$
$\vdots$		$\vdots$		$\vdots$	

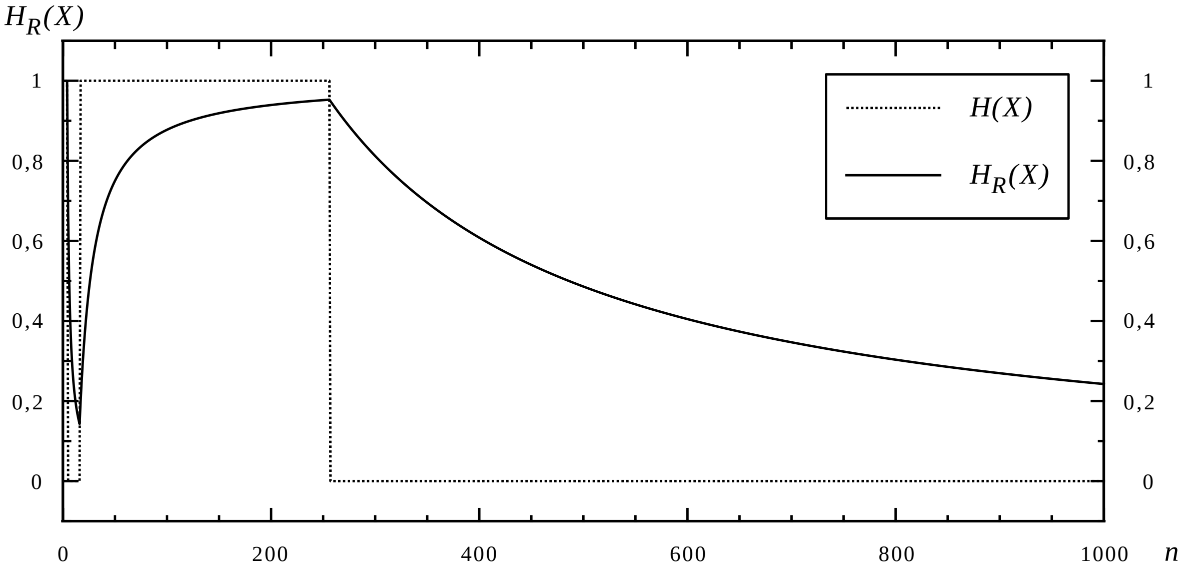


Figure 2.9: Entropy rate $H_R(X)$ and entropy $H(X)$ of source X.

2.7 Properties of a stationary source:

Consider a stationary source of information X.

Property I:

$$\boxed{H(X_N|X_1,\ldots,X_{N-1}) \leq H(X_{N-1}|X_1,\ldots,X_{N-2})}$$

Proof:

$$H(X_{N-1}|X_1,\ldots,X_{N-2}) = H(X_N|X_2,\ldots,X_{N-1})$$

by stationarity (time-shift $l = 1$). But

$$H(X_N|X_2,\ldots,X_{N-1}) \geq H(X_N|X_1,X_2,\ldots,X_{N-1})$$

On the right hand side of the inequality, the uncertainty about X_N is reduced, or at most equal, by the observation of X_1. Therefore, as expected:

$$H(X_N|X_1,\ldots,X_{N-1}) \leq H(X_{N-1}|X_1,\ldots,X_{N-2})$$

QED

Property II:

$$\boxed{H(X_N|X_1,\ldots,X_{N-1}) \leq \frac{1}{N}\, H(X_1,\ldots,X_N)}$$

Proof:

Consider the entropy of the random vector $\mathbf{X} = (X_1,\ldots,X_N)$:

$$\begin{aligned} H(X_1,\ldots,X_N) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_N|X_1,\ldots,X_{N-1}) \\ H(X_1,\ldots,X_N) &= \sum_{n=1}^{N} H(X_n|X_1,\ldots,X_{n-1}) \end{aligned}$$

but we know, from property I, that:

$$H(X_n|X_1,\ldots,X_{n-1}) \leq H(X_{n-1}|X_1,\ldots,X_{n-2}) \qquad \text{for } 1 \leq n \leq N$$

or

$$H(X_N|X_1,\ldots,X_{N-1}) \leq H(X_n|X_1,\ldots,X_{n-1})$$

for $1 \leq n \leq N$. Summing each side over N:

$$\begin{aligned} \sum_{n=1}^{N} H(X_N|X_1,\ldots,X_{N-1}) &\leq \sum_{n=1}^{N} H(X_n|X_1,\ldots,X_{n-1}) \\ N\ H(X_N|X_1,\ldots,X_{N-1}) &\leq H(X_1,\ldots,X_N) \qquad \text{or} \\ H(X_N|X_1,\ldots,X_{N-1}) &\leq \frac{1}{N}\ H(X_1,\ldots,X_N) \end{aligned}$$

QED

Property III:

$$\boxed{\frac{1}{N}\ H(X_1,\ldots,X_N) \leq \left(\frac{1}{N-1}\right)\ H(X_1,\ldots,X_{N-1})}$$

Proof: The entropy $H(\mathbf{X})$ can be expressed as the following sum of entropy and equivocation:

$$H(X_1,\ldots,X_N) = H(X_1,\ldots,X_{N-1}) + H(X_N|X_1,\ldots,X_{N-1})$$

Using property II, we obtain the inequality:

$$\begin{aligned} H(X_1,\ldots,X_N) &\leq H(X_1,\ldots,X_{N-1}) + \frac{1}{N}\ H(X_1,\ldots,X_N) \\ \left(\frac{N-1}{N}\right)\ H(X_1,\ldots,X_N) &\leq H(X_1,\ldots,X_{N-1}) \qquad \text{or} \\ H(X_1,\ldots,X_N) &\leq \left(\frac{N}{N-1}\right)\ H(X_1,\ldots,X_{N-1}) \qquad \text{or} \end{aligned}$$

Dividing both sides of the inequality by N:

$$\frac{1}{N}\ H(X_1,\ldots,X_N) \leq \left(\frac{1}{N-1}\right)\ H(X_1,\ldots,X_{N-1})$$

QED

Property IV:

$$\boxed{\lim_{N\to\infty} \frac{1}{N}\ H(X_1,\ldots,X_N) = \lim_{N\to\infty} H(X_N|X_1,\ldots,X_{N-1})}$$

Proof: We prove this result in two steps:

a) Consider property II:

$$H(X_N|X_1,\ldots,X_{N-1}) \leq \frac{1}{N}H(X_1,\ldots,X_N)$$

Since both sides of the inequality are positive and decreasing functions of the blocklength N, the limit as N goes to infinity exists. Taking the limit as $N \to \infty$, the inequality holds:

$$\begin{aligned}
\lim_{N\to\infty} H(X_N|X_1,\ldots,X_{N-1}) &\leq \lim_{N\to\infty}\frac{1}{N}H(X_1,\ldots,X_N) \\
\lim_{N\to\infty}\frac{1}{N}H(X_1,\ldots,X_N) &\geq \lim_{N\to\infty} H(X_N|X_1,\ldots,X_{N-1})
\end{aligned}$$

b) Consider the $(N+l)$ terms in the entropy per symbol, where l is a positive integer indicating a discrete time-shift.

$$\begin{aligned}
\left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N+l}) &= \left(\frac{1}{N+l}\right)[H(X_1,\ldots,X_{N-1}) + H(X_N|X_1,\ldots,X_{N-1}) \\
&\quad + \ldots + H(X_{N+l}|X_1,\ldots,X_{N+l-1})] \\
\left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N+l}) &= \left(\frac{1}{N+l}\right)\left[H(X_1,\ldots,X_{N-1}) + \sum_{n=N}^{N+l} H(X_n|X_1,\ldots,X_{n-1})\right]
\end{aligned}$$

But, since the process is a stationary process, for $N \leq n \leq N+l$, we must have, by property I:

$$H(X_N|X_1,\ldots,X_{N-1}) \geq H(X_n|X_1,\ldots,X_{n-1})$$

Summing both sides from $n = N$ to $n = N+l$:

$$\begin{aligned}
\sum_{n=N}^{N+l} H(X_N|X_1,\ldots,X_{N-1}) &\geq \sum_{n=N}^{N+l} H(X_n|X_1,\ldots,X_{n-1}) \\
(l+1)\ H(X_N|X_1,\ldots,X_{N-1}) &\geq \sum_{n=N}^{N+l} H(X_n|X_1,\ldots,X_{n-1})
\end{aligned}$$

Then, the previous inequality can be rewritten as:

$$\begin{aligned}
\left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N+l}) &\leq \left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N-1}) \\
&\quad + \left(\frac{l+1}{N+l}\right)H(X_N|X_1,\ldots,X_{N-1})
\end{aligned}$$

For $l \to \infty$:

$$\begin{aligned}
\lim_{l\to\infty}\left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N+l}) &\leq \lim_{l\to\infty}\left[\left(\frac{1}{N+l}\right)H(X_1,\ldots,X_{N-1})\right. \\
&\quad \left. + \left(\frac{l+1}{N+l}\right)H(X_N|X_1,\ldots,X_{N-1})\right]
\end{aligned}$$

As $l \to \infty$, $\left(\frac{1}{N+l}\right) \to 0$ and $\left(\frac{l+1}{N+l}\right) \to 1$, while both $H(X_1, \ldots, X_{N-1})$ and $H(X_N|X_1, \ldots, X_{N-1})$ terms are finite. Thus,

$$\lim_{l\to\infty} \left(\frac{1}{N+l}\right) H(X_1, \ldots, X_{N+l}) \leq H(X_N|X_1, \ldots, X_{N-1})$$

The above inequality holds true for any value of N. Now, taking the limit as $N \to \infty$ on both sides, yields to:

$$\begin{aligned} \underbrace{\lim_{N\to\infty} \lim_{l\to\infty} \left(\frac{1}{N+l}\right)}_{N+l\to\infty} H(X_1, \ldots, X_{N+l}) &\leq \lim_{N\to\infty} H(X_N|X_1, \ldots, X_{N-1}) \\ \lim_{N\to\infty} \frac{1}{N} H(X_1, \ldots, X_N) &\leq \lim_{N\to\infty} H(X_N|X_1, \ldots, X_{N-1}) \end{aligned}$$

Then, having also that opposite inequality (from the first part of the proof):

$$\lim_{N\to\infty} \frac{1}{N} H(X_1, \ldots, X_N) \geq \lim_{N\to\infty} H(X_N|X_1, \ldots, X_{N-1})$$

This implies that:

$$\lim_{N\to\infty} \frac{1}{N} H(X_1, \ldots, X_N) = \lim_{N\to\infty} H(X_N|X_1, \ldots, X_{N-1})$$

which is the entropy rate $H_R(X)$ for a stationary source.

QED

Example *(Markovian Source):*

Consider a stationary binary source of information having memory, which can be represented as a time-invariant Markovian source, as shown on Figure 2.10, where the transition probabilities $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$.

The transition probability matrix $\mathbf{P}$ between the two states (here each state represents a binary symbol from the source, i.e., x_1 and x_2) of the Markov chain is given by:

$$\mathbf{P} = \begin{bmatrix} p(x_1^{(n+1)}|x_1^{(n)}) & p(x_2^{(n+1)}|x_1^{(n)}) \\ p(x_1^{(n+1)}|x_2^{(n)}) & p(x_2^{(n+1)}|x_2^{(n)}) \end{bmatrix} = \begin{bmatrix} (1-\alpha) & \alpha \\ \beta & (1-\beta) \end{bmatrix}$$

Since the distribution is assumed stationary, then the distribution at discrete time $(n+1)$ is equal to the distribution at time n, i.e,

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)}\ \mathbf{P} = \mathbf{p}^{(n)} = \mathbf{p}$$

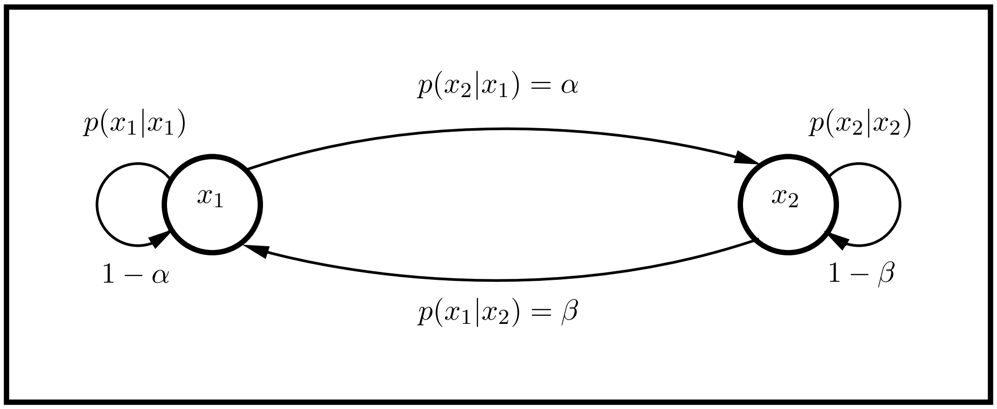


Figure 2.10: Two-state Markovian source X_n.

Then, the stationary distribution $\mathbf{p}$ has the following property:

$$\begin{aligned}
\mathbf{p} &= [p(x_1), p(x_2)] \\
&= [p(x_1), p(x_2)] \begin{bmatrix} p(x_1|x_1) & p(x_2|x_1) \\ p(x_1|x_2) & p(x_2|x_2) \end{bmatrix} \\
\mathbf{p} &= \{[p(x_1)\ p(x_1|x_1) + p(x_2)\ p(x_1|x_2)], [p(x_1)\ p(x_2|x_1) + p(x_2)\ p(x_2|x_2)]\}
\end{aligned}$$

and therefore,

$$\begin{aligned}
p(x_1) &= p(x_1)\ (1-\alpha) + p(x_2)\ \beta \qquad \text{and} \\
p(x_2) &= p(x_1)\ \alpha + p(x_2)\ (1-\beta)
\end{aligned}$$

From the first equation, since $p(x_2) = 1 - p(x_1)$, the probability $p(x_1)$ can be expressed as a function of the transition probabilities:

$$\begin{aligned}
p(x_1) &= p(x_1)\ (1-\alpha) + p(x_2)\ \beta \\
p(x_1) &= p(x_1)\ (1-\alpha) + [1 - p(x_1)]\ \beta \\
p(x_1) &= p(x_1)\ (1-\alpha-\beta) + \beta \\
p(x_1)\ (\alpha+\beta) &= \beta \\
p(x_1) &= \frac{\beta}{\alpha+\beta}
\end{aligned}$$

Similarly, $p(x_2)$ is a function of the transition probabilities.

$$\begin{aligned}
p(x_2) &= p(x_1)\ \alpha + p(x_2)\ (1-\beta) \\
p(x_2) &= [1 - p(x_2)]\ \ \alpha + p(x_2)\ (1-\beta) \\
p(x_2) &= p(x_2)\ (-\alpha + 1 - \beta) + \alpha \\
p(x_2)\ (\alpha+\beta) &= \alpha \\
p(x_2) &= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

The stationary distribution $\mathbf{p}$ is then given by the following expressions: (which is not a function of the time index n):

$$\boxed{p(x_1) = \frac{\beta}{\alpha+\beta} \qquad \text{and} \qquad p(x_2) = \frac{\alpha}{\alpha+\beta}}$$

The entropy $H(X_n)$ of this Markovian source X_n (at discrete time n) is:

$$\begin{aligned}
H(X_n) &= -\sum_{k=1}^{2} p(x_k) \log_b p(x_k) \\
H(X_n) &= -\left[\left(\frac{\beta}{\alpha+\beta}\right) \log_b \left(\frac{\beta}{\alpha+\beta}\right) + \left(\frac{\alpha}{\alpha+\beta}\right) \log_b \left(\frac{\alpha}{\alpha+\beta}\right)\right]
\end{aligned}$$

or equivalently:

$$\boxed{H(X_n) = H\left[\left(\frac{\beta}{\alpha+\beta}\right), \left(\frac{\alpha}{\alpha+\beta}\right)\right] = H(X)}$$

Note that the entropy $H(X_n) = H(X)$ is not a function of the time index n.

How does the joint entropy $H(X_1, \ldots, X_n, \ldots)$ grow as a function of time n? The answer is provided by the entropy rate function $H_R(X)$:

$$H_R(X) = \lim_{N\to\infty} \frac{1}{N}\ H(X_1, \ldots, X_N)$$

Since this process is assumed *stationary*, we can write that:

$$H_R(X) = \lim_{N\to\infty} H(X_N | X_1, \ldots, X_{N-1})$$

Furthermore, since this particular process is also a stationary *Markovian* process:

$$H_R(X) = \lim_{N\to\infty} H(X_N|X_{N-1})$$

Finally, the stationary process is *time-invariant*:

$$\begin{aligned} H_R(X) &= \lim_{N\to\infty} H(X_N|X_{N-1}) \\ &= \lim_{N\to\infty} H(X_2|X_1) \\ H_R(X) &= H(X_2|X_1) \end{aligned}$$

For this specific case, the entropy rate $H_R(X)$ is equal to the equivocation of X_2 given X_1:

$$\begin{aligned} H_R(X) &= H(X_2|X_1) \\ H_R(X) &= -\sum_{k=1}^{2}\sum_{j=1}^{2} p(x_k^{(1)})\ p(x_j^{(2)}|x_k^{(1)}) \log_b p(x_j^{(2)}|x_k^{(1)}) \\ H_R(X) &= -\sum_{k=1}^{2} p(x_k^{(1)}) \sum_{j=1}^{2} p(x_j^{(2)}|x_k^{(1)}) \log_b p(x_j^{(2)}|x_k^{(1)}) \\ H_R(X) &= -\left\{\left(\frac{\beta}{\alpha+\beta}\right)[(1-\alpha)\log_b(1-\alpha)+\alpha\log_b\alpha] \right. \\ &\quad \left. +\left(\frac{\alpha}{\alpha+\beta}\right)[\beta\log_b\beta+(1-\beta)\log_b(1-\beta)]\right\} \end{aligned}$$

The entropy rate $H_R(X)$ for the stationary Markovian source is:

$$\boxed{H_R(X) = \underbrace{\left(\frac{\beta}{\alpha+\beta}\right)}_{p(x_1)} H(\alpha) + \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{p(x_2)} H(\beta)}$$

2.8 Universal Source Coding

For a source of information X of known distribution $\mathbf{p} = \{p(x_k)\}_{k=1,\ldots,K}$, it is possible to design source compaction code $\mathcal{C}$ for which the average codeword length $L(\mathcal{C})$ is close to the source entropy $H(X)$. Huffman codes, for instance, are variable-length prefix codes which make use of the source statistics to assign short codewords to those source symbols which occurs often whereas unlikely symbols are encoded with longer codewords.

Unfortunately, how can one transmit the symbols generated by a source of information of unknown statistics, i.e., for which we don't know *a priori* the relative frequency of each symbol? A solution for this problem consists in a coding scheme, called *universal source coding*, which compacts the information from X without the knowledge of the source statistics.

2.8.1 Lempel-Ziv Code

We describe here a simple source compaction coding scheme, the Lempel-Ziv algorithm, which is a universal coding algorithm. The Lempel-Ziv algorithm is used often to compact data files for which the input distribution $\mathbf{p}$ is unknown.

Example *(Lempel-Ziv coding):*

Let X be a source of information for which we do not know the distribution $\mathbf{p}$. Suppose that we want to *source encode* the following sequence S generated by the source X:

$$S = 001000101110000011011010111101\ldots$$

Since the sequence is binary, often the two subsequences $S_1 = 0$ and $S_2 = 1$ are already stored. We perform the Lempel-Ziv encoding process by searching the original sequence S for those new subsequences which are the shortest and identify them as such:

$$\begin{aligned}
S &= \underbrace{00}_{S_3=00}\ 1000101110000011011010111101\ldots \\
S &= 00\ \underbrace{10}_{S_4=10}\ 00101110000011011010111101\ldots \\
S &= 0010\ \underbrace{001}_{S_5=001}\ 01110000011011010111101\ldots \\
S &= 0010001\ \underbrace{01}_{S_6=01}\ 110000011011010111101\ldots \\
S &= 001000101\ \underbrace{11}_{S_7=11}\ 0000011011010111101\ldots \\
S &= 00100010111\ \underbrace{000}_{S_8=000}\ 0011011010111101\ldots \\
S &= 00100010111000\ \underbrace{0011}_{S_9=0011}\ 011010111101\ldots \\
S &= 001000101110000011\ \underbrace{011}_{S_{10}=011}\ 010111101\ldots \\
S &= 001000101110000011011\ \underbrace{010}_{S_{11}=010}\ 111101\ldots
\end{aligned}$$

$$\begin{aligned} S &= 001000101110000011011010\ \underbrace{111}_{S_{12}=111}\ 101\ldots \\ S &= 001000101110000011011010111\ \underbrace{101}_{S_{13}=101}\ \ldots \end{aligned}$$

We then proceed to complete the Lempel-Ziv encoding process by arranging the subsequences in order of occurrence, or position, in the sequence S as shown on Table 2.4.

Table 2.4: Example of a Lempel-Ziv code.

position	subsequence		numerical	binary
	S_n		representation	codeword
1	S_1	0		
2	S_2	1		
3	S_3	00	1 1	001 0
4	S_4	10	2 1	010 0
5	S_5	001	3 2	011 1
6	S_6	01	1 2	001 1
7	S_7	11	2 2	010 1
8	S_8	000	3 1	011 0
9	S_9	0011	5 2	101 1
10	S_{10}	011	6 2	110 1
11	S_{11}	010	6 1	110 0
12	S_{12}	111	7 2	111 1
13	S_{13}	101	4 2	100 1

The numerical representation is obtained by concatenating the previous subsequences to make longer ones. For instance, the subsequence $S_3 = 00$ is the concatenation of subsequence $S_1 = 0$ with itself, whereas the subsequence $S_9 = 0011$ is obtained from $S_5 = 001$ and $S_2 = 1$. The first part of the subsequence is called *root sequence* or *pointer* while the last part is termed *innovation symbol.*

The numerical representation of the subsequence is then binary encoded as shown on the last column of Table 2.4. Note that there are only two different innovation symbols, namely 1 and 2, which are binary encoded as 0 and 1. For the binary representation of the pointer, the standard binary representation is used, e.g., the pointer 6 is encoded as 110.

Note that the Lempel-Ziv code is a *fixed-length code*, unlike the Huffman code which is a variable length code. In practice, the blocklength of a Lempel-Ziv code 12 bits long which corresponds to $2^{12} = 4096$ different entries.

The decoding process should allow for the *unique decoding* of the coded sequence into the original source sequence S. Here the Lempel-Ziv encoded stream will be:

$$S_C = 0010\ 0100\ 0111\ 0011\ 0101\ 0110\ 1011\ 1101\ 1100\ 1111\ 1001$$

The source decoder uses the pointer to determine the root subsequence and appends the innovation symbol. For instance, the last codeword $c^{(11)} = 1001$ as the pointer $100 = 4$ which represents $S_4 = 10$, and appens to it the innovation symbol $S_2 = 1$, leading to the source subsequence $S_{13} = 101$.

Note: The actual compaction ratio obtained for standard English text files is about 55%.

2.9 Problems

Problem 2.1: A source produces a sequence $\overline{X} = \{X_1, \ldots, X_N\}$ of statistically independent binary digits with the probabilities $p(x_1) = 0.995$ and $p(x_2) = 0.005$. These digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer 1's.

a) If the codewords are all of the same length, find the minimum length required to provide the specified set of codewords.

b) Find the probability of getting a source sequence for which no codeword has been provided.

c) Use the *Weak Law of Large Numbers* to bound the probability of getting a sequence for which no codeword has been provided and compare with part (b).

Problem 2.2: An information source X produces statistically independent binary digits with the following probabilities: $p(x_1) = 3/4$ and $p(x_2) = 1/4$. Consider sequences of N binary digits, where the probability of unlikely sequences $\mathcal{T}_X^c(\delta)$ is bounded as:

$$Pr\left\{\left|-\frac{1}{N}\sum_{n=1}^{N}\log_2 p(x_n) - H(X)\right| \geq \delta\right\} \leq \epsilon \tag{2.88}$$

a) Using the *Weak Law of Large Numbers*, determine the minimum sequence length N_0 such that for $N \geq N_0$ the inequality holds when $\delta = 5 \times 10^{-2}$ and $\epsilon = 10^{-1}$.

b) Repeat for $\delta = 10^{-3}$ and $\epsilon = 10^{-6}$.

c) For these two cases, find the lower and upper bounds for the number of typical sequences $\left\|\mathcal{T}_X(\delta)\right\|$.

Problem 2.3: For each of the following discrete memoryless sources, construct a binary and a ternary Huffman code and find the corresponding average codeword length L in each case.

a) A source X with a six-letter alphabet having these probabilities: $p(x_1) = .33$, $p(x_2) = .23$, $p(x_3) = .12$, $p(x_4) = .12$, $p(x_5) = .10$ and $p(x_6) = .10$.

b) A source X with a seven-letter alphabet having these probabilities: $p(x_1) = .35$, $p(x_2) = .20$, $p(x_3) = .15$, $p(x_4) = .15$, $p(x_5) = .10$, $p(x_6) = .03$ and $p(x_7) = .02$.

c) For the code in (b), construct two different binary Huffman codes with the same (minimum) average codeword length L but different variances. Which code is preferable in practice and why?

Problem 2.4: A source of information X generates binary sourcewords of length $n = 4$ with a binomial distribution:

$$p(x_k) = \binom{n}{k} p^k q^{n-k}, \qquad \text{for } 0 \leq k \leq n.$$

where $p(x_k)$ represents the probability of having a sourceword with k ones (1's) and $n-k$ zeroes (0's).

a) Determine the source entropy (per 4-tuples) $H(X)$ in *Sh* if $p = 0.1$ and $q = 0.9$.

b) Contruct a binary Huffman code $\mathcal{C}$ for that source. What is the code efficiency ξ?

c) Now suppose that the probabilities are changed to: $p = 0.35$ and $q = 0.65$. What is the entropy $H(X)$? What is the efficiency of the Huffman code?

Problem 2.5: *(Computer-oriented problem)*

An information source X with a five-letter alphabet $\{x_1, \ldots, x_5\}$ has the probabilities: $p(x_1) = .04$, $p(x_2) = .09$, $p(x_3) = .12$, $p(x_4) = .29$ and $p(x_5) = .46$.

a) Construct a binary Huffman code for this source and compare the average codeword length L with the source entropy $H(X)$.

b) Consider now a new source X' consisting of pairs (or digrams) of the five original letters:

$$X' \equiv \{(x_1, x_1), (x_1, x_2), \cdots, (x_5, x_5)\}$$

Construct a binary Huffman code for this new source X' and compare its efficiency with the single symbol Huffman code of a). Assume independent random variables, i.e. $p(x_i, x_j) = p(x_i)p(x_j)$, $\forall i, j$.

c) Repeat b) for trigrams, i.e.:

$$X'' \equiv \{(x_1, x_1, x_1), (x_1, x_1, x_2), \cdots, (x_5, x_5, x_5)\}$$

Once again, assume independence: $p(x_i, x_j, x_k) = p(x_i)p(x_j)p(x_k)$, $\forall i, j$ and k.

Problem 2.6: A binary Markov source generates two symbols, x_1 and x_2. The transition probabilities between the Markov chain states are:

$$\begin{aligned} p\left(x_1^{(2)}|x_1^{(1)}\right) &= p\left(x_2^{(2)}|x_2^{(1)}\right) = \rho \qquad \text{and} \\ p\left(x_2^{(2)}|x_1^{(1)}\right) &= p\left(x_1^{(2)}|x_2^{(1)}\right) = 1 - \rho \end{aligned}$$

a) Compute the source entropy $H(X)$ (per source symbol).

b) Let $\rho = 0.25$. Construct a binary Huffman code $\mathcal{C}$ which encodes sourcewords of blocklength $N = 3$ (i.e., blocks of 3 source symbols).

c) What is the average codeword length $L(\mathcal{C})$ (per source symbol) of this Huffman code?

d) Compute the code efficiency $\eta_{\mathcal{C}}$.

Chapter 3

Channel Coding for Noisy Channels

3.1 Convex sets and convex functions

3.1.1 Convex sets

Definition *(Convex set):*

A set of points $\mathcal{S}$ is *convex* if, for any pair of points $p_1 \in \mathcal{S}$ and $p_2 \in \mathcal{S}$, any point p on the *straight line* connecting p_1 and p_2 will be also in the set $\mathcal{S}$.

In other words, if the point $p_1 \in \mathcal{S}$ and the point $p_2 \in \mathcal{S}$ then for any point $p = \lambda p_1 + (1 - \lambda) p_2$, where $\lambda \in [0, 1]$, will be contained in the same set $\mathcal{S}$.

Let $\mathbf{p}_1 = (x_1, \ldots, x_N)$ and $\mathbf{p}_2 = (y_1, \ldots, y_N)$ be two points in an N-dimensional space. Any point $\mathbf{p}$ on the line connecting p_1 and p_2 in that N-dimensional space can be expressed as:

$$\mathbf{p} = \lambda \mathbf{p}_1 + (1 - \lambda) \mathbf{p}_2 \qquad \text{where } \lambda \in [0, 1] \tag{3.1}$$

That is, for $\lambda = 1$, $\mathbf{p} = \mathbf{p}_1$, whereas for $\lambda = 0$, $\mathbf{p} = \mathbf{p}_1$, and for $0 < \lambda < 1$, the point $\mathbf{p}$ is located on the line connecting p_1 and p_2 on the N-dimensional space. Figure 3.1 illustrates a convex set and a non-convex.

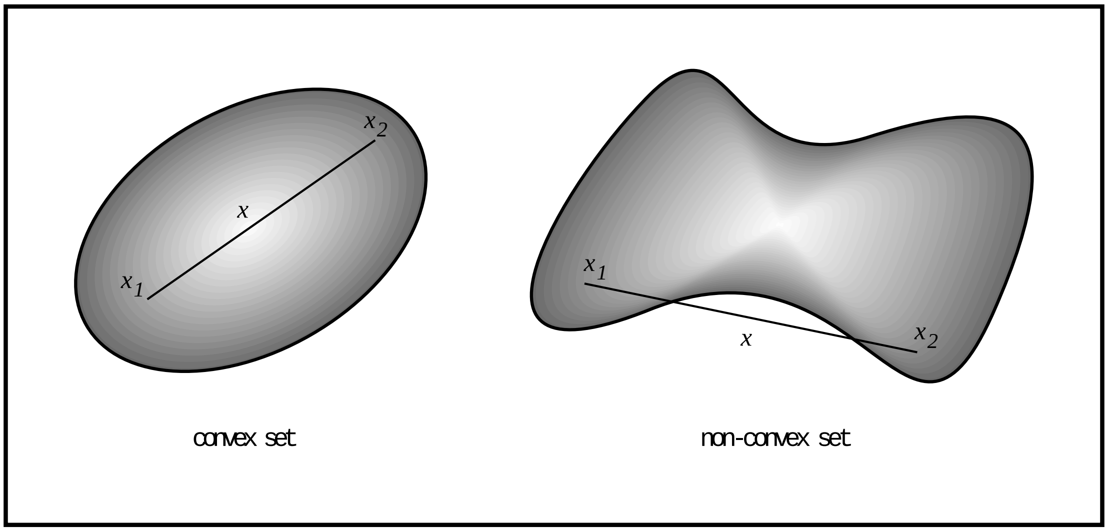


Figure 3.1: Convex and non-convex sets.

Example *(Convex set):*

Let $\mathbf{p}_1 = (x_1, x_2, x_3)$ and $\mathbf{p}_2 = (y_1, y_2, y_3)$ be two 3-dimensional probability distributions, i.e., two points in a 3-dimensional space. Let the set $\mathcal{S}_{\mathbf{p}} = \{\mathbf{p}\}$ be the set of all possible 3-dimensional probability distributions. Therefore, for any valid 3-dimensional probability distribution $\mathbf{p}$, we must have the two following conditions:

$$p(x_k) \geq 0 \qquad \text{for } k = 1, 2, 3$$

as well as

$$\sum_{k=1}^{3} p(x_k) = 1$$

We can view the distribution $\mathbf{p}$ as a 3-dimensional vector, as depicted on figure 3.2.

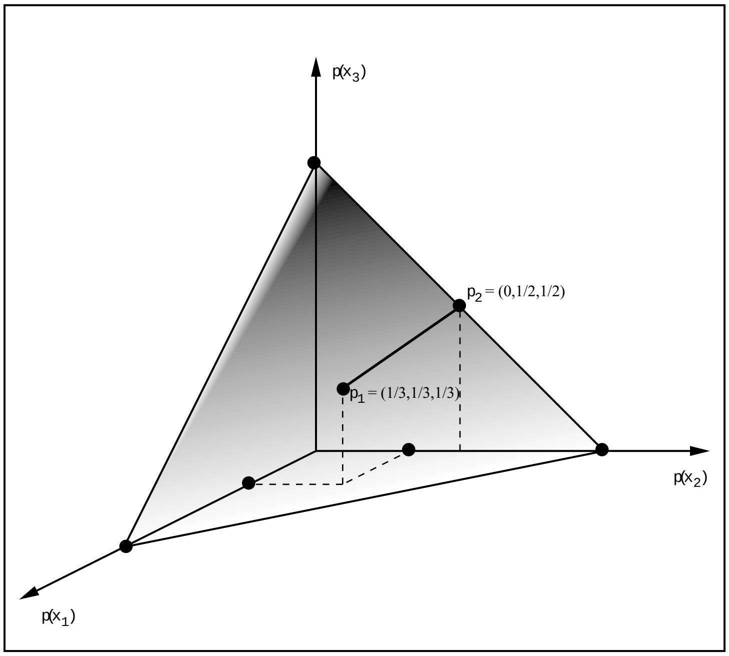


Figure 3.2: Example of a 3-dimensional probability distribution.

If we connect $\mathbf{p}_1 = (x_1, x_2, x_3)$ and $\mathbf{p}_2 = (y_1, y_2, y_3)$ with a straight line $\mathbf{p}$ (i.e. the *set of distributions* between $\mathbf{p}_1$ and $\mathbf{p}_2$), then:

$$\begin{aligned} \mathbf{p} &= \lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2 \\ p(x_k) &= \lambda p_1(x_k) + (1 - \lambda) p_2(x_k) \qquad \text{for } k = 1, 2, 3 \end{aligned}$$

For each input symbol x_k, since $p_1(x_k)$, $p_2(x_k)$, λ, and $(1-\lambda)$ are all positive, we must have that

$$p(x_k) = \lambda p_1(x_k) + (1-\lambda) p_2(x_k) \geq 0 \qquad \forall k$$

We note also that

$$\begin{aligned}
\sum_{k=1}^{3} p(x_k) &= \sum_{k=1}^{3} \left[\lambda p_1(x_k) + (1-\lambda) p_2(x_k)\right] \\
\sum_{k=1}^{3} p(x_k) &= \lambda \sum_{k=1}^{3} p_1(x_k) + (1-\lambda) \sum_{k=1}^{3} p_2(x_k) \\
\sum_{k=1}^{3} p(x_k) &= \lambda + (1-\lambda) = 1
\end{aligned}$$

Therefore, any point $\mathbf{p} = \lambda \mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$ between two distributions $\mathbf{p}_1$ and $\mathbf{p}_2$ is also a valid distribution, and this for any choice of pairs of distributions. Thus, the set $\mathcal{S}_{\mathbf{p}}$ of all possible N-dimensional distributions (in this example $N = 3$) forms a convex set.

3.1.2 Convex functions

Definition *(Convex function):*

A real function $f(x)$, defined on a convex set $\mathcal{S}$ (e.g., input symbol distributions), is *concave* (*convex down*, *convex "cap"* or *convex* $\cap$) if, for any point x on the straight line between the pair of points x_1 and x_2, i.e., $x = \lambda x_1 + (1-\lambda)x_2$ ($\lambda \in [0,1]$), in the convex set $\mathcal{S}$:

$$\boxed{f(x) \geq \lambda f(x_1) + (1-\lambda) f(x_2)}$$

otherwise, if:

$$\boxed{f(x) \leq \lambda f(x_1) + (1-\lambda) f(x_2)}$$

then the function is said to be simply *convex* (*convex up*, *convex "cup"* or *convex* $\cup$).

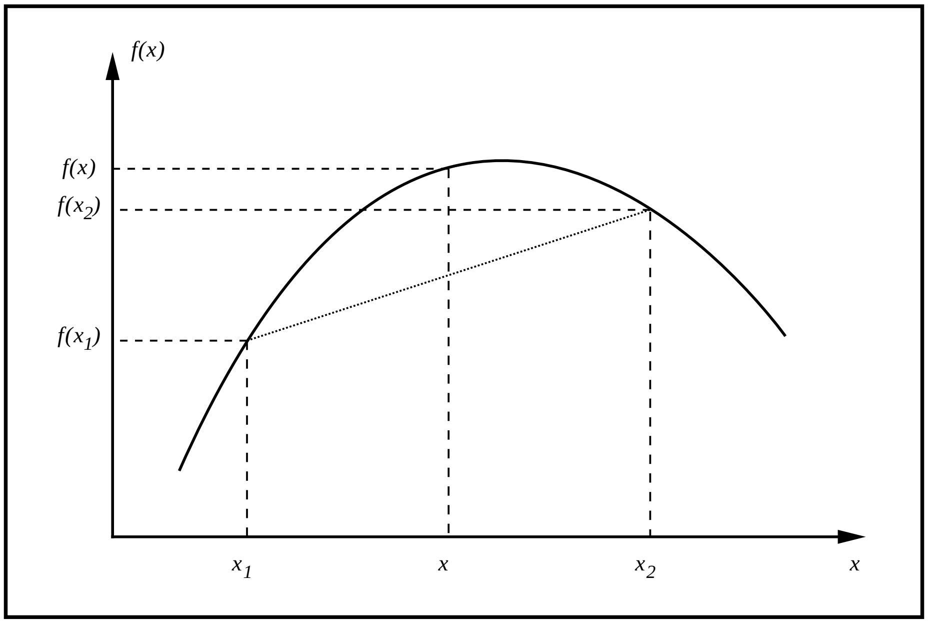


Figure 3.3: Convex $\cap$ (*convex down* or *convex "cap"*) function.

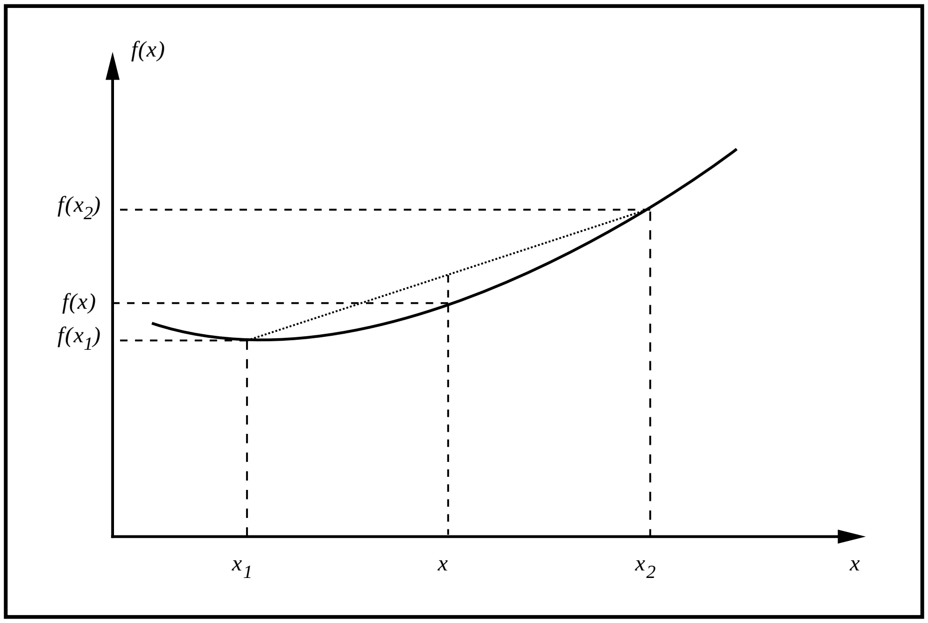


Figure 3.4: Convex $\cup$ (*convex up* or *convex "cup"*) function.

3.1.3 Convexity ($\cap$) of mutual information over input distributions

Theorem *(Convexity of the mutual information function):*

The (average) mutual information $I(X;Y)$ is a *concave* (or *convex "cap"*, or *convex* $\cap$) function over the *convex set* $\mathcal{S}_{\mathbf{p}}$ of all possible input distributions $\{\mathbf{p}\}$.

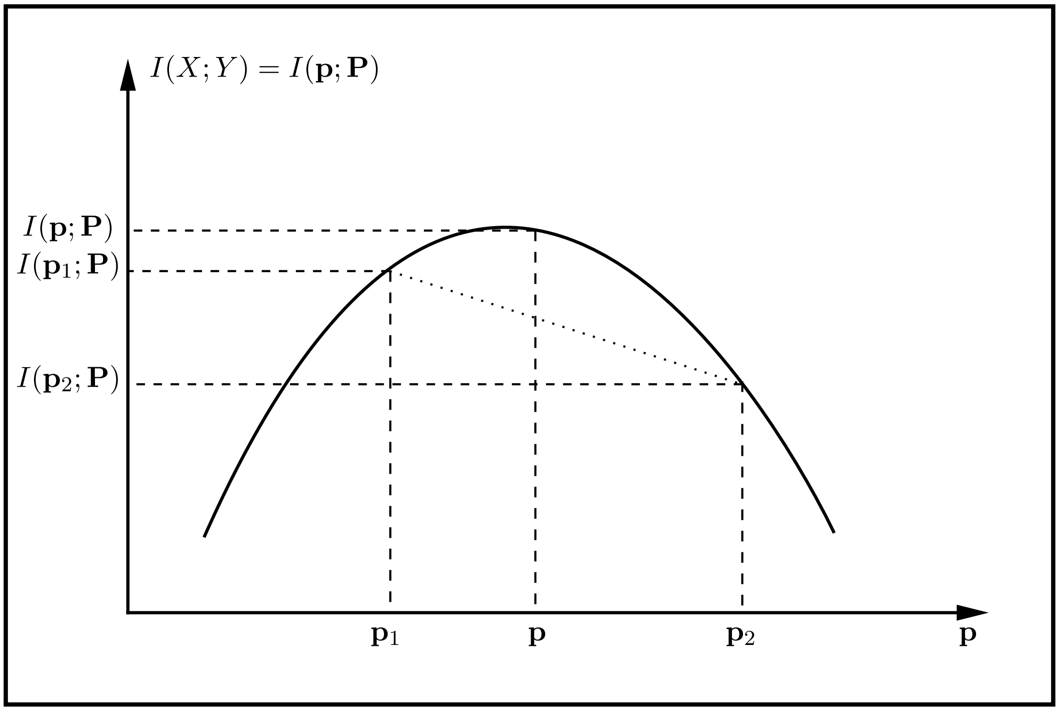


Figure 3.5: Convexity ($\cap$) of mutual information function over the set of input symbol distributions $\mathcal{S}_{\mathbf{p}}$.

Proof:

The (average) mutual information function $I(X;Y)$ is a function of both the input symbol distribution $\mathbf{p} = \{p(x_k)\}_{k=1,\ldots,K}$ and the channel transition probabilities' matrix $\mathbf{P} = \{p(y_j|x_k)\}_{\substack{k=1,\ldots,K \\ j=1,\ldots,J}}$.

$$\begin{aligned} I(X;Y) &= \sum_{k=1}^{K} \sum_{j=1}^{J} p(x_k) p(y_j|x_k) \log_b \left[\frac{p(y_j|x_k)}{\sum_{l=1}^{K} p(x_l) p(y_j|x_l)} \right] \qquad (3.2) \\ &= f\left[p(x_k), p(y_j|x_k)\right] \\ I(X;Y) &= f\left(\mathbf{p}, \mathbf{P}\right) \end{aligned}$$

For channel coding, we want to evaluate the *maximum transfer of information* on a given channel (i.e., for a given transition probability matrix $\mathbf{P}$) over all possible input distributions $\mathcal{S}_{\mathbf{p}} = \{\mathbf{p}\}$. Consider two different input distributions:

$$\mathbf{p}_1 = \{p_1(x_k)\}_{k=1,\ldots,K} \qquad \text{and} \qquad \mathbf{p}_2 = \{p_2(x_k)\}_{k=1,\ldots,K}$$

The distribution $\mathbf{p}$, *between* distributions $\mathbf{p}_1$ and $\mathbf{p}_2$ in the convex set $\mathcal{S}_{\mathbf{p}}$, can be expressed as:

$$\begin{aligned} \mathbf{p} &= \lambda\ \mathbf{p}_1 + (1-\lambda)\ \mathbf{p}_2 \qquad \text{or} \\ p(x_k) &= \lambda\ p_1(x_k) + (1-\lambda)\ p_2(x_k) \qquad \text{for } k = 1,\ldots,K \end{aligned} \tag{3.3}$$

The corresponding output symbol distribution $\{p(y_j)\}$ is given by:

$$\begin{aligned} p(y_j) &= \sum_{k=1}^{K} p(x_k)\ p(y_j|x_k) \qquad \text{for } j = 1,\ldots,J \\ &= \sum_{k=1}^{K} [\lambda\ p_1(x_k) + (1-\lambda)\ p_2(x_k)]\ \ p(y_j|x_k) \\ &= \lambda \sum_{k=1}^{K} p_1(x_k)\ p(y_j|x_k) + (1-\lambda) \sum_{k=1}^{K} p_2(x_k)\ p(y_j|x_k) \\ p(y_j) &= \lambda p_1(y_j) + (1-\lambda) p_2(y_j) \end{aligned} \tag{3.4}$$

which is also a *convex set*. We want to show that the mutual information is a *concave* (i.e., convex $\cap$) function of the input distribution $\mathbf{p}$, that is:

$$\begin{aligned} f(x) &\geq \lambda\ f(x_1) + (1-\lambda)\ f(x_2) \quad \text{for } x = \lambda\ x_1 + (1-\lambda)\ x_2 \\ I(X;Y) &\geq \lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2) \quad \text{over the convex set } \mathcal{S}_{\mathbf{p}} \end{aligned} \tag{3.5}$$

Consider the difference between the right-hand side and left-hand side terms in the above Equation (3.5). If the statement is true about the convexity of $I(X;Y)$ then the difference is negative (or at most equal to zero).

$$\begin{aligned} \lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2) - I(X;Y) = \lambda \sum_{k=1}^{K}\sum_{j=1}^{J} p_1(x_k) p(y_j|x_k) \log_b \frac{p(y_j|x_k)}{p_1(y_j)} + \\ (1-\lambda) \sum_{k=1}^{K}\sum_{j=1}^{J} p_2(x_k) p(y_j|x_k) \log_b \frac{p(y_j|x_k)}{p_2(y_j)} - \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k) p(y_j|x_k) \log_b \frac{p(y_j|x_k)}{p(y_j)} \\ = \lambda \sum_{k=1}^{K}\sum_{j=1}^{J} p_1(x_k) p(y_j|x_k) \log_b \frac{p(y_j|x_k)}{p_1(y_j)} + (1-\lambda) \sum_{k=1}^{K}\sum_{j=1}^{J} p_2(x_k) p(y_j|x_k) \log_b \frac{p(y_j|x_k)}{p_2(y_j)} - \end{aligned} \tag{3.6}$$

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{j=1}^{J}[\lambda\ p_1(x_k)+(1-\lambda)\ p_2(x_k)]\,p(y_j|x_k)\log_b\frac{p(y_j|x_k)}{p(y_j)} && (3.7)\\
=\ &\lambda\sum_{k=1}^{K}\sum_{j=1}^{J}p_1(x_k)p(y_j|x_k)\left[\log_b\frac{p(y_j|x_k)}{p_1(y_j)}-\log_b\frac{p(y_j|x_k)}{p(y_j)}\right]+ && (3.8)\\
&(1-\lambda)\sum_{k=1}^{K}\sum_{j=1}^{J}p_2(x_k)p(y_j|x_k)\left[\log_b\frac{p(y_j|x_k)}{p_2(y_j)}-\log_b\frac{p(y_j|x_k)}{p(y_j)}\right]
\end{aligned}$$

$$\begin{aligned}
&\lambda\ I(X_1;Y_1)+(1-\lambda)\ I(X_2;Y_2)-I(X;Y)= && (3.9)\\
&\quad\lambda\sum_{k=1}^{K}\sum_{j=1}^{J}p_1(x_k)p(y_j|x_k)\log_b\frac{p(y_j)}{p_1(y_j)}+(1-\lambda)\sum_{k=1}^{K}\sum_{j=1}^{J}p_2(x_k)p(y_j|x_k)\log_b\frac{p(y_j)}{p_2(y_j)}
\end{aligned}$$

but since the ratios $\frac{p(y_j)}{p_1(y_j)}$ and $\frac{p(y_j)}{p_2(y_j)}$ are strictly positive, and that $\log_b(x)=\log_b(e)\ \ln(x)$, then Equation (3.9) can be rewritten as:

$$\begin{aligned}
&\lambda\ I(X_1;Y_1)+(1-\lambda)\ I(X_2;Y_2)-I(X;Y)= && (3.10)\\
&\quad(\log_b e)\left[\lambda\sum_{k=1}^{K}\sum_{j=1}^{J}p_1(x_k)p(y_j|x_k)\ln\frac{p(y_j)}{p_1(y_j)}+(1-\lambda)\sum_{k=1}^{K}\sum_{j=1}^{J}p_2(x_k)p(y_j|x_k)\ln\frac{p(y_j)}{p_2(y_j)}\right]
\end{aligned}$$

Changing the summation order;

$$\begin{aligned}
&\lambda\ I(X_1;Y_1)+(1-\lambda)\ I(X_2;Y_2)-I(X;Y)= && (3.11)\\
&\quad(\log_b e)\left[\lambda\sum_{j=1}^{J}\sum_{k=1}^{K}p_1(x_k)p(y_j|x_k)\ln\frac{p(y_j)}{p_1(y_j)}+(1-\lambda)\sum_{j=1}^{J}\sum_{k=1}^{K}p_2(x_k)p(y_j|x_k)\ln\frac{p(y_j)}{p_2(y_j)}\right]
\end{aligned}$$

$$\begin{aligned}
&\lambda\ I(X_1;Y_1)+(1-\lambda)\ I(X_2;Y_2)-I(X;Y)= && (3.12)\\
&\quad(\log_b e)\left[\lambda\sum_{j=1}^{J}p_1(y_j)\ln\frac{p(y_j)}{p_1(y_j)}+(1-\lambda)\sum_{j=1}^{J}p_2(y_j)\ln\frac{p(y_j)}{p_2(y_j)}\right]
\end{aligned}$$

Therefore, since $\ln(x)\leq(x-1)$ for $x>0$ then:

$$\begin{aligned}
&\lambda\ I(X_1;Y_1)+(1-\lambda)\ I(X_2;Y_2)-I(X;Y)\leq && (3.13)\\
&\quad(\log_b e)\left\{\lambda\sum_{j=1}^{J}p_1(y_j)\left[\frac{p(y_j)}{p_1(y_j)}-1\right]+(1-\lambda)\sum_{j=1}^{J}p_2(y_j)\left[\frac{p(y_j)}{p_2(y_j)}-1\right]\right\}
\end{aligned}$$

$$\begin{aligned}\lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2) - I(X;Y) \leq& \\ (\log_b e)\left\{\lambda\left[\sum_{j=1}^{J} p(y_j) - \sum_{j=1}^{J} p_1(y_j)\right] + (1-\lambda)\left[\sum_{j=1}^{J} p(y_j) - \sum_{j=1}^{J} p_2(y_j)\right]\right\}&\end{aligned} \tag{3.14}$$

$$\lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2) - I(X;Y) \leq (\log_b e)\left\{\lambda\,[1-1] + (1-\lambda)\,[1-1]\right\} \tag{3.15}$$

This implies that $\lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2) - I(X;Y) \leq 0$, that is:

$$\boxed{I(X;Y) \geq \lambda\ I(X_1;Y_1) + (1-\lambda)\ I(X_2;Y_2)}$$

for $p(x_k) = \lambda\ p_1(x_k) + (1-\lambda)\ p_2(x_k)$, for $k = 1, \ldots .K$. The mutual information $I(X;Y)$ does have a maximum over the set $\mathcal{S}_{\mathbf{p}}$ of all possible input distributions $\{\mathbf{p}\}$.

QED

3.1.4 Convexity ($\cup$) of mutual information over channel transition probability matrices

---

Theorem *(Convexity of the mutual information function):*

The (average) mutual information $I(X;Y)$ is a *convex* (or *convex "cup"*, or *convex* $\cup$) function over the *convex set* $\mathcal{S}_{\mathbf{P}}$ of all possible transition probability matrices $\{\mathbf{P}\}$.

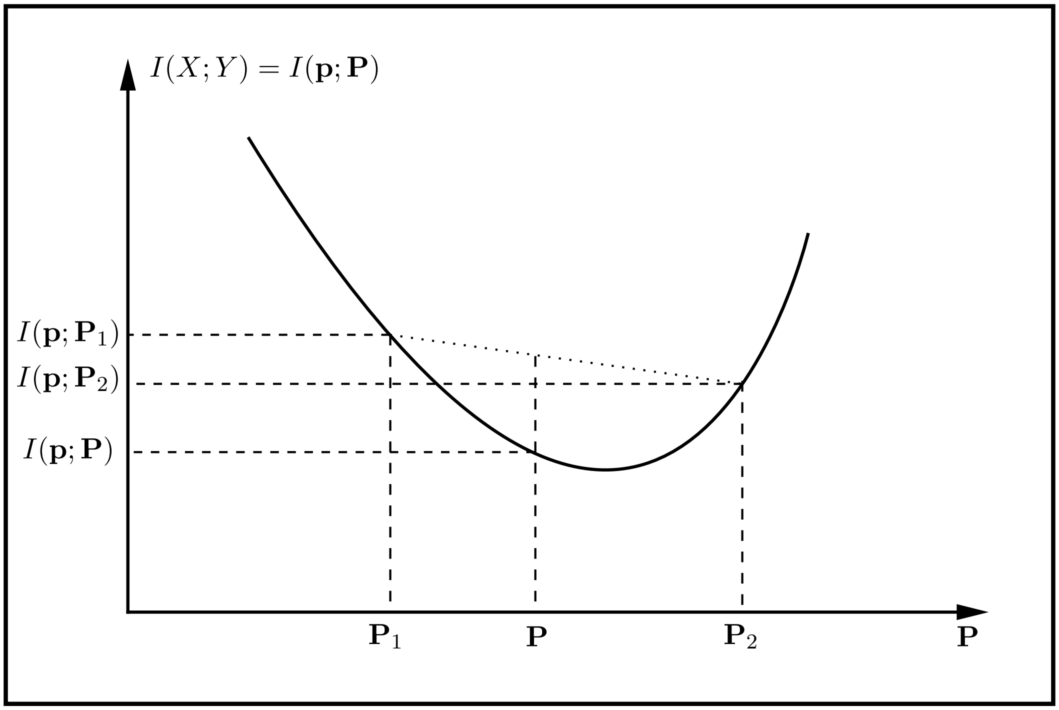

$I(X;Y) = I(\mathbf{p};\mathbf{P})$
$I(\mathbf{p};\mathbf{P}_1)$
$I(\mathbf{p};\mathbf{P}_2)$
$I(\mathbf{p};\mathbf{P})$
$\mathbf{P}_1$
$\mathbf{P}$
$\mathbf{P}_2$
$\mathbf{P}$

Figure 3.6: Convexity ($\cup$) of mutual information function over the set of transition probability matrices $\mathcal{S}_{\mathbf{P}}$.

The proof is similar to the proof of the convexity ($\cap$) of the mutual information over the input distributions.

3.2 Capacity of memoryless channel

3.2.1 Capacity of symmetric channels

A discrete memoryless channel is said to be *symmetric* if the set of output symbols $\{y_j\}_{j=1,\ldots,J}$ can be partitionned into subsets such that for each subset of the matrix of transition probabilities, each column is a permutation of each other and each row is also a permutation of each other row.

For instance, the binary symmetric channel has a transition probability matrix $\mathbf{P_1}$:

$$\mathbf{P_1} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

where each column and row are permutations of others.

However, if the crossover probabilities of another binary transition probability matrix $\mathbf{P_2}$ are of different values:

$$\mathbf{P_2} = \begin{pmatrix} 1-\epsilon_1 & \epsilon_1 \\ \epsilon_2 & 1-\epsilon_2 \end{pmatrix}$$

then the binary channel is no longer symmetrical.

Consider now a binary input and ternary output channel characterized by transition probability matrix $\mathbf{P_3}$:

$$\mathbf{P_3} = \begin{pmatrix} 1-\epsilon_1-\epsilon_2 & \epsilon_2 \\ \epsilon_1 & \epsilon_1 \\ \epsilon_2 & 1-\epsilon_1-\epsilon_2 \end{pmatrix}$$

In this 2×3 matrix, the second column is a permutation of the first one but the row permutation condition is not respected. However, if we partition the set of outputs $\{y_1, y_2, y_3\}$ into the two subsets $\{y_1, y_3\}$ and $\{y_2\}$ we obtain the following submatrices:

$$\mathbf{P'_3} = \begin{pmatrix} 1-\epsilon_1-\epsilon_2 & \epsilon_2 \\ \epsilon_2 & 1-\epsilon_1-\epsilon_2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{P''_3} = \begin{pmatrix} \epsilon_1 & \epsilon_1 \end{pmatrix}$$

For each submatrix, each row and each column is a permutation of another. This results in a channel which is said to be a *weakly symmetric channel.*

If we modify matrix $\mathbf{P_3}$ into $\mathbf{P_4}$ by exchanging the crossover probabilites as:

$$\mathbf{P_4} = \begin{pmatrix} 1-\epsilon_1-\epsilon_2 & \epsilon_1 \\ \epsilon_1 & \epsilon_2 \\ \epsilon_2 & 1-\epsilon_1-\epsilon_2 \end{pmatrix}$$

then the channel is no longer symmetric since it is impossible to partition the set of outputs such has to obey the row and column permutation condition.

Theorem *(Capacity of a symmetric channel):*

For a discrete symmetric channel, the channel capacity C is achieved with an equiprobable input distribution, i.e., $p(x_k) = \frac{1}{K}, \forall k$, and is given by:

$$\boxed{C = \left[\sum_{j=1}^{J} p(y_j|x_k) \log_b p(y_j|x_k)\right] + \log_b J}$$

In other words, the capacity of the channel is given by the maximum of the mutual information over all possible input distributions for a fixed set of channel transition probabilities. For symmetric channels this maximum is obtained with an equiprobable input distribution.

Proof:

$$
\begin{aligned}
C &= \max_{\mathcal{S}_\mathbf{p}} I(X;Y) && (3.16)\\
C &= \max_{\mathcal{S}_\mathbf{p}} \left[\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k, y_j) \log_b \frac{p(y_j|x_k)}{p(y_j)}\right] && (3.17)\\
C &= \max_{\mathcal{S}_\mathbf{p}} \left[\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(y_j|x_k) \log_b p(y_j|x_k) - \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(y_j|x_k) \log_b p(y_j)\right] && (3.18)\\
C &= \max_{\mathcal{S}_\mathbf{p}} \left[\sum_{k=1}^{K} p(x_k) \sum_{j=1}^{J} p(y_j|x_k) \log_b p(y_j|x_k) - \sum_{j=1}^{J} \log_b p(y_j) \sum_{k=1}^{K} p(x_k, y_j)\right] && (3.19)\\
C &= \max_{\mathcal{S}_\mathbf{p}} \left[\sum_{k=1}^{K} p(x_k) \sum_{j=1}^{J} p(y_j|x_k) \log_b p(y_j|x_k) - \sum_{j=1}^{J} p(y_j) \log_b p(y_j)\right] && (3.20)
\end{aligned}
$$

Since the channel is symmetric and thus each row and column is a permutation of the other in the transition probability matrix, the sum $\sum_{j=1}^{J} p(y_j|x_k) \log_b p(y_j|x_k)$ in the first term is independent of the input k and thus does not affect the maximization of the mutual information.

The second term, $-\sum_{j=1}^{J} p(y_j) \log_b p(y_j)$ is simply the entropy $H(Y)$ of the output Y:

$$\max_{\mathcal{S}_\mathbf{p}} \left[-\sum_{j=1}^{J} p(y_j) \log_b p(y_j)\right] = \max_{\mathcal{S}_\mathbf{p}} H(Y) \qquad (3.21)$$

and $H(Y)$ is maximized when the outputs are equiprobable: $p(y_j) = \frac{1}{J}$ for $j = 1, \ldots, J$.

$$H(Y) = -\sum_{j=1}^{J} \frac{1}{J} \log_b \frac{1}{J} = \log_b J \tag{3.22}$$

which is obtained when the inputs are also equiprobable: $p(x_k) = \frac{1}{K}$ (since the channel is symmetric):

$$p(y_j) = \sum_{k=1}^{K} p(x_k, y_j) = \sum_{k=1}^{K} p(x_k) p(y_j|x_k) = \frac{1}{J} \tag{3.23}$$

and then the capacity is:

$$C = \left[\sum_{j=1}^{J} p(y_j|x_k) \log_b p(y_j|x_k) \right] + \log_b J \tag{3.24}$$

QED

3.2.2 Blahut-Arimoto algorithm (capacity of asymmetric channels)

Step 1: Choose an arbitrary input distribution:

$$\mathbf{p}^{(0)} = \{p^{(0)}(x_k)\}_{k=1,\ldots,K}$$

A good choice for the initial distribution $\mathbf{p}^{(0)}$ is the equiprobable distribution.

Step 2: Compute the following terms:

a) coefficients c_k:

$$c_k = \exp\left[\sum_{j=1}^{J} p(y_j|x_k) \ln\left(\frac{p(y_j|x_k)}{\sum_{l=1}^{K} p(x_l)\ p(y_j|x_l)}\right)\right] \qquad \text{for } k = 1, \ldots, K$$

b) lower bound on $I(X;Y)$:

$$I_L = \ln \sum_{k=1}^{K} p(x_k)\ c_k$$

c) upper bound on $I(X;Y)$:

$$I_U = \ln\left(\max_{k=1,\ldots,K} c_k\right)$$

Step 3: Test if the difference between I_U and I_L is smaller than a fixed tolerance ϵ:

$$I_U - I_L \overset{?}{\leq} \epsilon$$

yes: If the answer is **yes** then

$$\boxed{C = I_L}$$

and stop the program.

no: Otherwise, if the answer is **no**, then change the input distribution:

$$p^{(n+1)}(x_k) = \frac{p^{(n)}(x_k)\ c_k}{\sum_{l=1}^{K} p(x_l)^{(n)}\ c_l}$$

and go back to **step 2**.

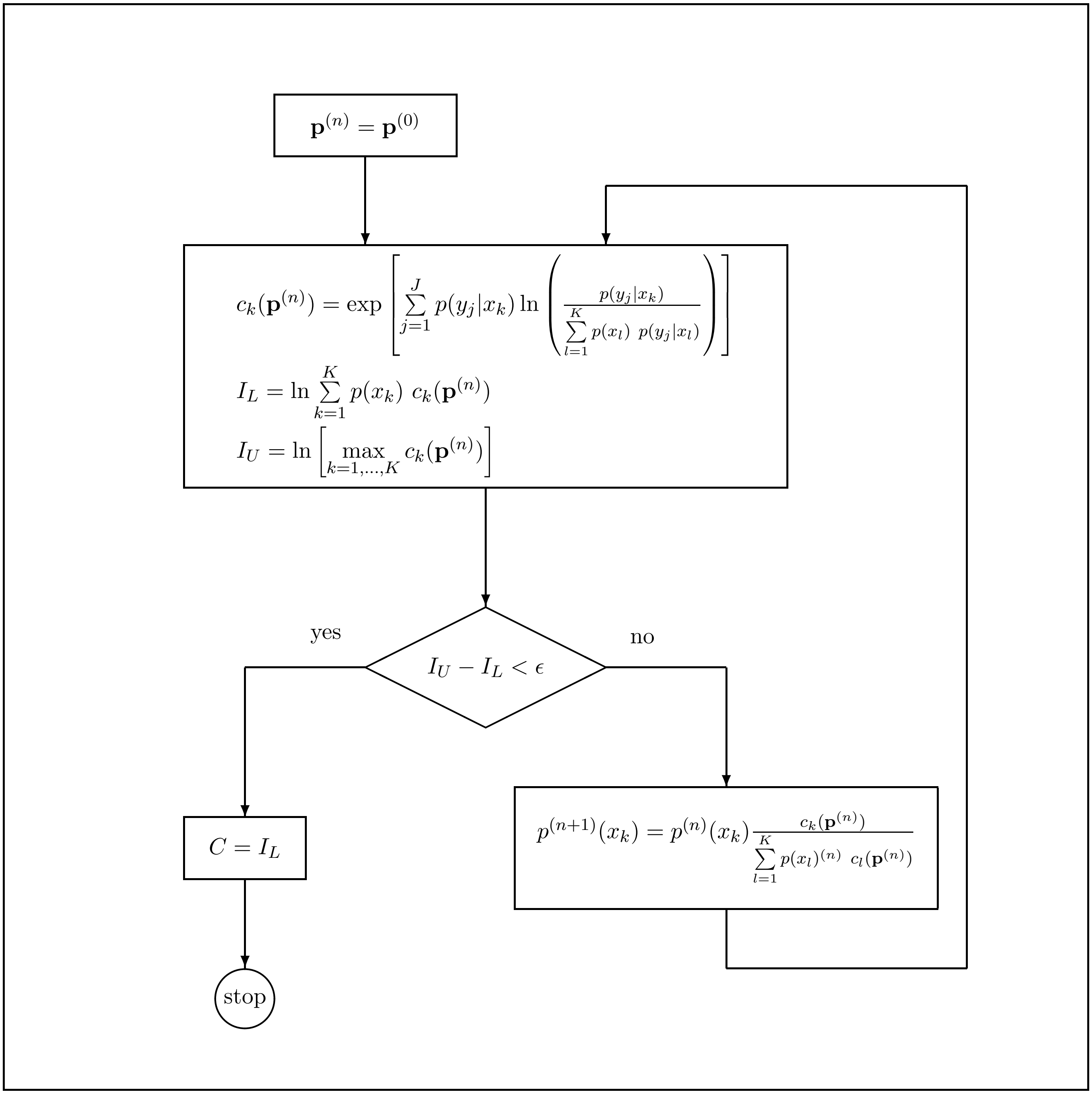

$\mathbf{p}^{(n)} = \mathbf{p}^{(0)}$
$c_k(\mathbf{p}^{(n)}) = \exp\left[\sum_{j=1}^{J} p(y_j|x_k) \ln\left(\frac{p(y_j|x_k)}{\sum_{l=1}^{K} p(x_l)\ p(y_j|x_l)}\right)\right]$
$I_L = \ln \sum_{k=1}^{K} p(x_k)\ c_k(\mathbf{p}^{(n)})$
$I_U = \ln\left[\max_{k=1,\ldots,K} c_k(\mathbf{p}^{(n)})\right]$
yes
$I_U - I_L < \epsilon$
no
$C = I_L$
$p^{(n+1)}(x_k) = p^{(n)}(x_k)\frac{c_k(\mathbf{p}^{(n)})}{\sum_{l=1}^{K} p(x_l)^{(n)}\ c_l(\mathbf{p}^{(n)})}$
stop

Figure 3.7: Blahut-Arimoto algorithm for computing the capacity of asymmetric channels (from *"Principles and Practice of Information Theory"* by Richard E. Blahut).

Example *(Capacity of a Non Symmetric Channel (Blahut-Arimoto Algorithm)):*

For this example, we want to compute the capacity C of a channel which has four inputs and four inputs. In order to verify that the program functions properly, we begin with a symmetric channel with the following transition probability matrix:

$$\mathbf{P}_1 = \begin{pmatrix} 0.4000 & 0.3000 & 0.2000 & 0.1000 \\ 0.1000 & 0.4000 & 0.3000 & 0.2000 \\ 0.3000 & 0.2000 & 0.1000 & 0.4000 \\ 0.2000 & 0.1000 & 0.4000 & 0.3000 \end{pmatrix}$$

We know that for the input distribution $\mathbf{p}^*$ that maximizes the mutual information is the equiprobable distribution, that is $\mathbf{p}^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. We set the threshold value $\epsilon = 10^{-6}$ to stop the iterative algorithm. If we begin the Blahut-Arimoto algorithm with $\mathbf{p}^*$, then we must obtain the channel capacity C at the first iteration, i.e. without updating the input distribution since it is the optimum one already:

n	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	I_U	I_L	ϵ
1	0.2500	0.2500	0.2500	0.2500	0.1064	0.1064	0.0000

And the channel capacity is $C = 0.1064$ logons or 0.1536 shannons (dividing by $\ln(2)$).

Now, let's compute the channel capacity of the same symmetric channel $\mathbf{P}$ (using the Blahut-Arimoto algorithm) but starting this time with a non equiprobable input distribution: $\mathbf{p}_1 = (0.1, 0.6, 0.2, 0.1)$.

At the beginning, the algorithm shows different values of I_U and I_L for that distribution. After a few iterations, the algorithm converges rapidly towards the ideal distribution $\mathbf{p}^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and the capacity is obtained: $C = 0.1064$ logons (or $C = 0.1536$ shannons).

n	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	I_U	I_L	ϵ
1	0.1000	0.6000	0.2000	0.1000	0.1953	0.0847	0.1106
2	0.1073	0.5663	0.2155	0.1126	0.1834	0.0885	0.0949
3	0.1141	0.5348	0.2287	0.1249	0.1735	0.0916	0.0819
4	0.1204	0.5061	0.2394	0.1369	0.1650	0.0942	0.0708
5	0.1264	0.4802	0.2480	0.1484	0.1576	0.0963	0.0613
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	0.1524	0.3867	0.2668	0.1963	0.1326	0.1024	0.0302
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	0.1912	0.3037	0.2604	0.2448	0.1219	0.1057	0.0162
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
40	0.2320	0.2622	0.2476	0.2578	0.1110	0.1064	0.0046
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
80	0.2482	0.2515	0.2486	0.2516	0.1068	0.1064	0.0004
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
202	0.2500	0.2500	0.2500	0.2500	0.1064	0.1064	0.0000

Now consider the following non symmetric channel's transition probability matrix $\mathbf{P}_2$:

$$\mathbf{P}_2 = \begin{pmatrix} 0.1000 & 0.2500 & 0.2000 & 0.1000 \\ 0.1000 & 0.2500 & 0.6000 & 0.2000 \\ 0.8000 & 0.2500 & 0.1000 & 0.2000 \\ 0.1000 & 0.2500 & 0.1000 & 0.5000 \end{pmatrix}$$

We initialize the input distribution $\mathbf{p}_2 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (a good choice is the equiprobable distribution even if the channel is not symmetric).

The algorithm provides the following results:

n	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	I_U	I_L	ϵ
1	0.2500	0.2500	0.2500	0.2500	0.4498	0.2336	0.2162
2	0.3103	0.1861	0.2545	0.2228	0.4098	0.2504	0.1595
3	0.3640	0.1428	0.2653	0.2036	0.3592	0.2594	0.0998
4	0.4022	0.1118	0.2804	0.1899	0.3289	0.2647	0.0642
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
8	0.4522	0.0450	0.3389	0.1639	0.2988	0.2763	0.0225
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
16	0.4629	0.0076	0.3732	0.1565	0.2848	0.2830	0.0018
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
32	0.4641	0.0002	0.3769	0.1588	0.2846	0.2844	0.0003
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
64	0.4640	0.0000	0.3768	0.1592	0.2844	0.2844	0.0000
65	0.4640	0.0000	0.3768	0.1592	0.2844	0.2844	0.0000

The capacity is $C = 0.2844$ logons (or $C = 0.4103$ shannons) and is obtained with the optimum input distribution for this assymetric channel: $\mathbf{p}^* = (0.4640, 0.0000, 0.3768, 0.1592)$. Note that the second symbol x_2 should not be used if we want to reach the channel capacity!

3.3 Capacity of channels with memory

As we have seen, the capacity per symbol of discrete channels having memory is given by the limit, as the blocklength N goes to infinity, of the maximum of the mutual information over the set $\mathcal{S}_{\mathbf{p}(X_1,\ldots,X_N)}$ of all possible input vectors, or *sourcewords*:

$$\boxed{C = \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{p}(X_1,\ldots,X_N)}} \frac{1}{N} I(X_1,\ldots,X_n,\ldots,X_N;Y_1,\ldots,Y_n,\ldots,Y_N)} \tag{3.25}$$

The mutual information between the input and output vectors, $\mathbf{X}$ and $\mathbf{Y}$ is the difference between the entropy of the input vector $H(\mathbf{X})$ and its equivocation $H(\mathbf{X}|\mathbf{Y})$ given the output vector.

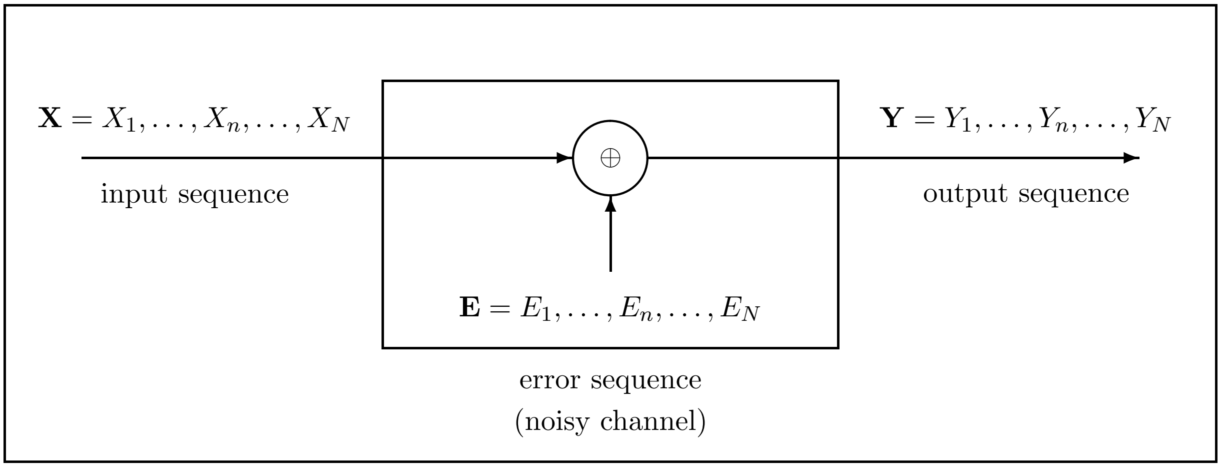

$\mathbf{X} = X_1,\ldots,X_n,\ldots,X_N$
input sequence
$\oplus$
$\mathbf{Y} = Y_1,\ldots,Y_n,\ldots,Y_N$
output sequence
$\mathbf{E} = E_1,\ldots,E_n,\ldots,E_N$
error sequence
(noisy channel)

Figure 3.8: Binary channel with memory.

Consider, as a noisy channel over which we want to send information data, a binary channel with memory (see Figure 3.8). The vector $\mathbf{E} = (E_1,\ldots,E_n,\ldots,E_N)$ is an *error sequence* which indicates if the channel is in error or not at discrete time n.

The mutual information can be expressed as a function

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \\ I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{X}) - H(\mathbf{X}|(\mathbf{X}\oplus\mathbf{E})) \\ I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{X}) - H(\mathbf{X}|\mathbf{E}) \end{aligned} \tag{3.26}$$

But since the uncertainty (i.e., equivocation) about the input vector $(X_1,\ldots,X_N)$ at the receiving end, depends solely on the error sequence $(E_1,\ldots,E_N)$ then the mutual information $I(X_1,\ldots,X_N;Y_1,\ldots,Y_N)$ is equal to the uncertainty about $(X_1,\ldots,X_N)$ less the uncertainty about the error sequence $(E_1, \dots, E_N)$ itself.

$$I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{E}) \tag{3.27}$$

From Equation (3.25), the channel capacity C becomes:

$$C = \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{p}(\mathbf{x})}} [H(\mathbf{X}) - H(\mathbf{E})] \tag{3.28}$$

The entropy $H(E_1, \dots, E_N)$ is often called *channel entropy.*

Example *(Capacity of a binary symmetric channel):*

This is a channel without memory. Nevertheless, its capacity can be determined using the general expression (i.e., Equation (3.28)). It is also a symmetric channel: therefore, the input distribution $\mathbf{p}^*$ which leads to the channel capacity C is an equiprobable source distribution: $p(x_1) = p(x_2) = \frac{1}{2}$.

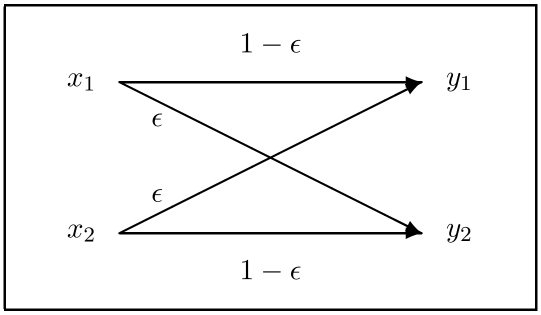


Figure 3.9: Binary symmetric channel.

The entropy (per symbol) of the source is then $H(X) = 1$ *Sh*. The channel entropy $H(E)$ is:

$$\begin{aligned} H(E) &= -\sum_{j=1}^{2} p(y_j|x_k)\log_b p(y_j|x_k) \qquad \text{for } k = 1,2 \\ H(E) &= -[(1-\epsilon)\log_2(1-\epsilon) + \epsilon\log_2\epsilon] \end{aligned}$$

The channel capacity C is:

$$\begin{aligned} C &= \max_{\mathcal{S}_{\mathbf{p}}} I(X;Y) \\ C &= H(X) - H(E) \\ C &= 1 + [(1-\epsilon)\log_2(1-\epsilon) + \epsilon\log_2\epsilon] \end{aligned}$$

For instance, if the crossover probability $\epsilon = 0.025$, then the channel capacity $C = 0.830\ Sh$. Note that the corresponding *channel bit error rate* (BER) is equal to ϵ, i.e., $BER = 0.025$:

$$\begin{aligned} BER &= \sum_{k=1}^{K} p(x_k) \sum_{\substack{j=1 \\ j\neq k}}^{J} p(y_j|x_k) \\ BER &= p(x_1)\ p(y_2|x_1) + p(x_2)\ p(y_1|x_2) \\ BER &= \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon \end{aligned}$$

Example *(Capacity of a binary symmetric channel with memory):*

Consider now a binary symmetric channel with memory for which the crossover probability $p(y_j|x_k)_{j\neq k} = \epsilon$ at every discrete time instant n but for which the occurrence of errors are not independent (for instance, E_n may not be independent of E_{n-1}):

$$E_n = \begin{cases} 1 & \text{with probability } p^{(n)}(y_j|x_k)_{j\neq k} = \epsilon \\ 0 & \text{with probability } p^{(n)}(y_j|x_k)_{j=k} = 1-\epsilon \end{cases}$$

The BER is still equal to ϵ, but the channel capacity C is affected by the memory of the channel. Since the channel is symmetric, we know that the capacity is achieved with the equiprobable distribution $\mathbf{p}^*$, that is: $p(x_1) = p(x_2) = \frac{1}{2}$. If the error generation process in the noisy channel is independent of the input, which is a fair assumption, then we can assume that the identically distributed input random variables $\{X_n\}_{n=1,\ldots,N}$ are also independent.

$$\begin{aligned} BER &= \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{k=1}^{K} p^{(n)}(x_k) \sum_{\substack{j=1 \\ j\neq k}}^{J} p^{(n)}(y_j|x_k)\right] \\ BER &= \frac{1}{N}\sum_{n=1}^{N}\left[p^{(n)}(x_1)\ p^{(n)}(y_2|x_1) + p^{(n)}(x_2)\ p^{(n)}(y_1|x_2)\right] \\ BER &= \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{2}\epsilon + \frac{1}{2}\epsilon\right] \\ BER &= \frac{1}{2}\epsilon + \frac{1}{2}\epsilon \\ BER &= \epsilon \end{aligned}$$

The entropy per symbol of the source is $H(X) = 1\ Sh$. The channel capacity C is:

$$C = \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{p}(X_1,\ldots,X_N)}} \left[I(X_1,\ldots,X_n,\ldots,X_N;Y_1,\ldots,Y_n,\ldots,Y_N)\right]$$

The capacity per symbol is:

$$
\begin{aligned}
C &= \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[\frac{1}{N} I(X_1,\ldots,X_n,\ldots,X_N;Y_1,\ldots,Y_n,\ldots,Y_N)\right] \\
C &= \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[\frac{1}{N} H(X_1,\ldots,X_n,\ldots,X_N) - \frac{1}{N} H(X_1,\ldots,X_n,\ldots,X_N|Y_1,\ldots,Y_n,\ldots,Y_N)\right] \\
C &= \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[\frac{1}{N} \sum_{n=1}^{N} H(X_n) - \frac{1}{N} H(E_1,\ldots,E_n,\ldots,E_N|Y_1,\ldots,Y_n,\ldots,Y_N)\right] \\
C &= \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} H(E_1,\ldots,E_n,\ldots,E_N|Y_1,\ldots,Y_n,\ldots,Y_N)\right]
\end{aligned}
$$

since the variables $X_1,\ldots,X_n,\ldots,X_N$ are i.i.d. Considering that the equivocation of the error sequence $H(E_1,\ldots,E_n,\ldots,E_N|Y_1,\ldots,Y_n,\ldots,Y_N)$, after the observation of the received symbols $Y_1,\ldots,Y_n,\ldots,Y_N$, can be at most equal to the entropy itself $H(E_1,\ldots,E_n,\ldots,E_N)$, then:

$$
\begin{aligned}
C &= \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} H(E_1,\ldots,E_n,\ldots,E_N|Y_1,\ldots,Y_n,\ldots,Y_N)\right] \\
C &\geq \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} H(E_1,\ldots,E_n,\ldots,E_N)\right] \\
C &\geq \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} \sum_{n=1}^{N} H(E_n)\right]
\end{aligned}
$$

because for the error generation process with memory, $H(E_1,\ldots,E_n,\ldots,E_N) = \sum_{n=1}^{N} H(E_n|E_1,\ldots,E_{n-1}) \leq \sum_{n=1}^{N} H(E_n)$. However, the E_n are identically distributed:

$$
\begin{aligned}
C &\geq \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} \sum_{n=1}^{N} H(E_n)\right] \\
C &\geq \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} \left[H(X) - \frac{1}{N} \sum_{n=1}^{N} H(E)\right] \\
C &\geq \lim_{N\to\infty} \max_{\mathcal{S}_{\mathbf{PX}}} [H(X) - H(E)] \\
C &\geq H(X) - H(E) \\
C &\geq Cmemoryless
\end{aligned}
$$

Even if the bit error rate is the same in both memory and memoryless channels cases, the channel capacity is greater or equal than the capacity of the memoryless channel. In fact, for a given BER, the memory of the channel can be exploited, e.g, the correlation between the successive noise samples, to increase the effective throughput of information (e.g., channel equalization, error correcting codes).

3.4 Jointly typical pairs of sequences

Definition *(Jointly typical pair of sequences):*

Given a memoryless pair of random variables (X, Y) with a joint probability distribution $\{p(\mathbf{x}, \mathbf{y})\}$ and a joint entropy $H(XY)$, the set of jointly typical pairs of sequences $\mathcal{T}_{XY}(\delta)$ of block-length N are the pairs $(\mathbf{x}, \mathbf{y})$ in the set:

$$\boxed{\mathcal{T}_{XY}(\delta) \equiv \left\{(\mathbf{x}, \mathbf{y}) : \left| -\frac{1}{N} \log_b p(\mathbf{x}, \mathbf{y}) - H(XY) \right| < \delta \right\}} \tag{3.29}$$

provided that $\mathbf{x}$ and $\mathbf{y}$ are respective elements of the typical sequences sets $\mathcal{T}_X(\delta)$ and $\mathcal{T}_Y(\delta)$:

$$\begin{aligned}
\mathcal{T}_X(\delta) &= \left\{\mathbf{x} \text{ such that: } \left| -\frac{1}{N} \log_b p(\mathbf{x}) - H(X) \right| < \delta \right\} \\
\mathcal{T}_Y(\delta) &= \left\{\mathbf{y} \text{ such that: } \left| -\frac{1}{N} \log_b p(\mathbf{y}) - H(Y) \right| < \delta \right\}
\end{aligned}$$

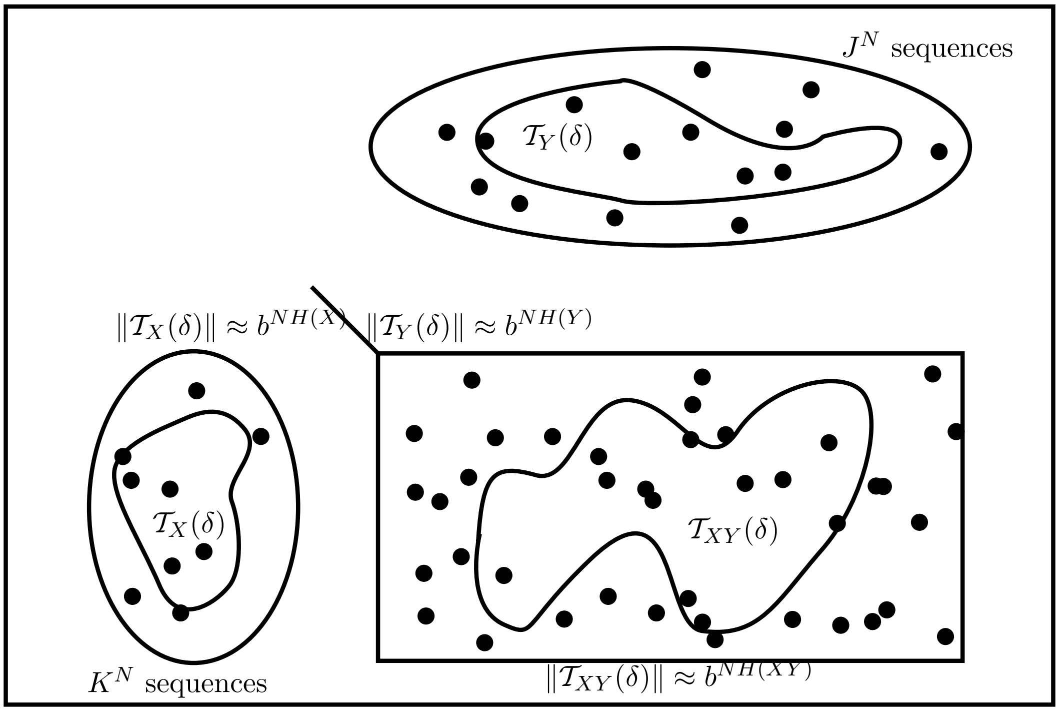

J^N sequences
$\mathcal{T}_Y(\delta)$
$\|\mathcal{T}_X(\delta)\| \approx b^{NH(X)}$
$\|\mathcal{T}_Y(\delta)\| \approx b^{NH(Y)}$
$\mathcal{T}_X(\delta)$
$\mathcal{T}_{XY}(\delta)$
K^N sequences
$\|\mathcal{T}_{XY}(\delta)\| \approx b^{NH(XY)}$

Figure 3.10: Relationship between the sets of jointly typical pairs of sequences $\mathcal{T}_{XY}(\delta)$, and the sets of typical sequences $\mathcal{T}_X(\delta)$ and $\mathcal{T}_Y(\delta)$.

Theorem *(Shannon-McMillan theorem for jointly typical pairs of sequences):*

Given a dependent pair of memoryless sources of joint entropy $H(XY)$, for a blocklength N sufficiently large (i.e. $N \geq N_0$), the set of pairs of vectors $\{(\mathbf{x}, \mathbf{y})\}$ can be partitioned into a set of jointly typical pairs of sequences $\mathcal{T}_{XY}(\delta)$ and a set of jointly atypical pairs of sequences $\mathcal{T}^c_{XY}(\delta)$ for which:

a) The probability of atypical sequences is upperbounded as:

$$\boxed{Pr\left[(\mathbf{x}, \mathbf{y}) \in \mathcal{T}^c_{XY}(\delta)\right] < \epsilon} \tag{3.30}$$

b) If the pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$ then:

$$\boxed{b^{-N[H(XY)+\delta]} < p(\mathbf{x}, \mathbf{y}) < b^{-N[H(XY)-\delta]}} \tag{3.31}$$

c) The number of elements in the set of jointly typical pairs of sequences $\mathcal{T}_{XY}(\delta)$ is upperbounded by:

$$\boxed{\|\mathcal{T}_{XY}(\delta)\| < b^{N[H(XY)+\delta]}} \tag{3.32}$$

d) If $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$ and $\mathbf{x}$ is fixed, then the conditional probability $p(\mathbf{y}|\mathbf{x})$ is bounded as:

$$\boxed{b^{-N[H(Y|X)+2\delta]} < p(\mathbf{y}|\mathbf{x}) < b^{-N[H(Y|X)-2\delta]}} \tag{3.33}$$

Proof:

The proof of the first three properties of jointly typical pairs of sequences is similar to that for typical sequences. The proof of the fourth property of jointly typical pairs of sequences follows.

The conditional probability $p(\mathbf{y}|\mathbf{x})$ can be expressed as:

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} \tag{3.34}$$

and the conditional entropy $H(Y|X)$, or equivocation of Y given X, is the difference between the joint entropy $H(XY)$ and the entropy $H(X)$:

$$H(Y|X) = H(XY) - H(X) \tag{3.35}$$

Considering that, if the pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$, and that, by definition, $\mathbf{x} \in \mathcal{T}_X(\delta)$ and $\mathbf{y} \in \mathcal{T}_Y(\delta)$, then;

$$\begin{aligned} \left|-\frac{1}{N}\log_b p(\mathbf{x}, \mathbf{y}) - H(XY)\right| &< \delta \\ \left|-\frac{1}{N}\log_b p(\mathbf{x}) - H(X)\right| &< \delta \qquad \text{and} \\ \left|-\frac{1}{N}\log_b p(\mathbf{y}) - H(Y)\right| &< \delta \end{aligned} \tag{3.36}$$

Considering the difference between $\frac{1}{N}\log_b p(\mathbf{y}|\mathbf{x})$ and the equivocation $H(Y|X)$, then:

$$\begin{aligned} \left|-\frac{1}{N}\log_b p(\mathbf{y}|\mathbf{x}) - H(Y|X)\right| &= \left|-\frac{1}{N}\log_b \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} - [H(XY) - H(X)]\right| \\ &= \left|-\frac{1}{N}\log_b p(\mathbf{x}, \mathbf{y}) + \frac{1}{N}\log_b p(\mathbf{x}) - H(XY) + H(X)\right| \\ &= \left|-\frac{1}{N}\log_b p(\mathbf{x}, \mathbf{y}) - H(XY) + \frac{1}{N}\log_b p(\mathbf{x}) + H(X)\right| \\ \left|-\frac{1}{N}\log_b p(\mathbf{y}|\mathbf{x}) - H(Y|X)\right| &< \delta + \delta \end{aligned} \tag{3.37}$$

and therefore:

$$\left| -\frac{1}{N} \log_b p(\mathbf{y}|\mathbf{x}) - H(Y|X) \right| < 2\delta \tag{3.38}$$

QED

3.5 Channel coding theorem

3.5.1 Shannon's channel (second) coding theorem

Figure 3.11 illustrates a communication link. A source of information $\mathbf{W}$ generates messages as symbols, or sourcewords, from a set of M possible messages.

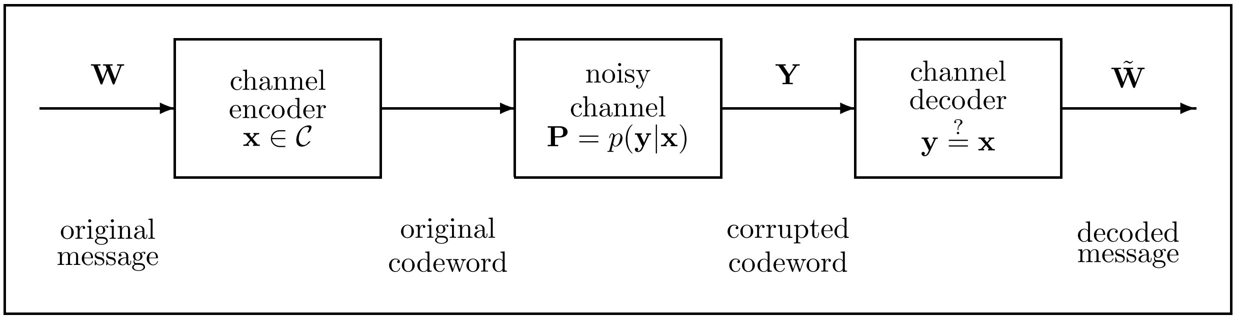

$\mathbf{W}$
channel encoder $\mathbf{x} \in \mathcal{C}$
noisy channel $\mathbf{P} = p(\mathbf{y}|\mathbf{x})$
$\mathbf{Y}$
channel decoder $\mathbf{y} \overset{?}{=} \mathbf{x}$
$\tilde{\mathbf{W}}$
original message
original codeword
corrupted codeword
decoded message

Figure 3.11: Noisy communication channel with channel coding.

We assume that each message w is chosen with equal probability:

$$p(w) = \frac{1}{M} = b^{-NR} \tag{3.39}$$

where R is the code rate and N is the codewords' blocklength. The information rate into a channel is $R = \frac{1}{N}H(\mathbf{X})$ where $H(\mathbf{X})$ is the entropy of the set of M *sourcewords* (or input vectors) of lenght N. Then, if we consider an equiprobable source $R = \frac{1}{N}\log_b M$.

A channel encoder maps each message W_m, $m = 1, \ldots, M$, as a unique codeword $\mathbf{c}_m$ of blocklength N. We assume also that the codewords are b-ary ($c_{m,n} \in \{0, 1, \ldots, b-1\} \forall m, n$). A *code* $\mathcal{C}$ consists in M codewords:

$$\mathcal{C} = \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,n} & \cdots & c_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,n} & \cdots & c_{M,N} \end{bmatrix} \tag{3.40}$$

There are $M = b^{NR}$ different messages and therefore the code $\mathcal{C}$ consists in b^{NR} codewords. The number of possible *codes* $\mathcal{S}_{\mathcal{C}} \equiv \{\mathcal{C}\}$ is equal to the number of possible matrices: $b^{N \times M} = b^{\left(N \times 2^{NR}\right)}$.

For instance, consider a binary code for which the blocklength $N = 20$. If the code rate $R = \frac{1}{2}$, then there are:

$$\begin{array}{rcccc} 2^{\left(N\times 2^{NR}\right)} & = & 2^{\left(20\times 2^{10}\right)} & = & 2^{(20\times 1024)} \\ 2^{\left(N\times 2^{NR}\right)} & = & 2^{20480} & = & 10^{6165} \end{array}$$

possible codes, which is quite large. For larger values of N, the number of codes becomes enormous.

Theorem *(Shannon's channel coding theorem):*

Let C be the information transfer capacity of a memoryless channel defined by its transition probabilities matrix $\mathbf{P} = \{p(\mathbf{y}|\mathbf{x})\}$. If the code rate $R < C$, then there *exists* a channel code $\mathcal{C}$ of size M and blocklength N, such that the probability of decoding error P_e is *upperbounded* by an arbitrarily small number ϵ;

$$\boxed{P_e \leq \epsilon}$$

provided that the blocklength N is sufficiently large (i.e., $N \geq N_0$).

Proof:

The proof of Shannon's channel coding theorem derived below is based on the *random* selection of a set of codes $\mathcal{S}_{\mathcal{C}} = \{\mathcal{C}\}$ and the average probability of decoding errors over this set of codes $\mathcal{S}_{\mathcal{C}}$. A decoding rule using jointly typical pairs of sequences is considered. This isn't an optimal decoding rule but it is probably the simplest rule to prove the channel coding theorem (also know as Shannon's second coding theorem).

The codes are chosen with the following *random coding* rule: the b^{NR} codewords are selected accordingly to a fix distribution $p(x)$. In other words, each of the $N \times b^{NR}$ elements of the code matrix $\mathcal{C}$ is chosen independently of each other with the same probability $p(x)$ (which maximizes the mutual information):

$$\mathcal{C} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \\ \vdots \\ \mathbf{x}_M \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} & \cdots & x_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,n} & \cdots & x_{M,N} \end{bmatrix} \tag{3.41}$$

The probability $p(\mathcal{C})$ of selecting a particular code $\mathcal{C}$ is:

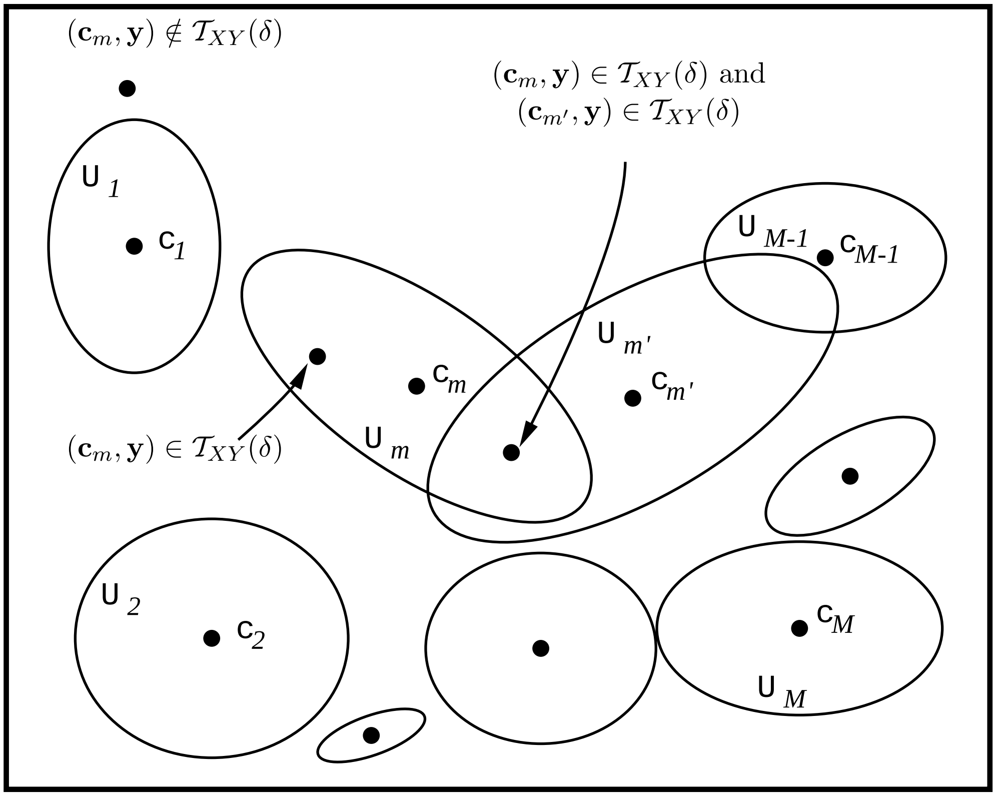
$$\begin{aligned} p(\mathcal{C}) &= \prod_{m=1}^{M} p(\mathbf{x}_m) \\ p(\mathcal{C}) &= \prod_{m=1}^{M} \prod_{n=1}^{N} p(x_{m,n}) \end{aligned} \tag{3.42}$$

Note that some codes will be *bad* codes. The mutual information $I(\mathbf{X}, \mathbf{Y})$ between the channel input and output is a function of the codewords' elements distribution $\mathbf{p} = \{p(\mathbf{x})\}$ as well as the transition probability matrix $\mathbf{P} = \{p(\mathbf{y}|\mathbf{x})\}$ of the noisy channel. The channel transition probabilities are considered also i.i.d. (independent and identically distributed).

The decoding rule is based on the definition of jointly typical sequences. A received (and possibly corrupted) codeword $\mathbf{y}$ is *mapped* into a valid codeword $\mathbf{x}_m$, or $\mathbf{c}_m$, if the pair of sequences $(\mathbf{y}, \mathbf{c}_m)$ are jointly typical, i.e., $(\mathbf{y}, \mathbf{c}_m) \in \mathcal{T}_{XY}(\delta)$. The decoded message is then w_m.

As shown on Figure 3.12 two types of decoding errors may occur:

- $(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta) \qquad \text{for } m = 1, \ldots, M$
- $(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta) \qquad \text{for } m' \neq m$



$(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)$
$(\mathbf{c}_m, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$ and
$(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$
U 1
c1
U M-1
cM-1
U m'
cm'
cm
$(\mathbf{c}_m, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)$
U m
U 2
c2
cM
U M

set $\{\mathbf{y}\}$ of all possible received sequences

Figure 3.12: Decoding decision regions for jointly typical pairs of sequences $\mathcal{T}_{XY}(\delta)$.

As mentionned previously, this decoding rule is not an optimal one, but it provides a *relatively* simple derivation of the channel coding theorem.

The probability of a decoding error $P_{e|m}$, given that message w_m was transmitted is given by the probability of the *union* of error events:

$$P_{e|m} = Pr\left\{[(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)] \bigcup \bigcup_{\substack{m'=1\\m'\neq m}}^{M} [(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)]\right\} \tag{3.43}$$

$$P_{e|m} \leq Pr\left[(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)\right] + \sum_{\substack{m'=1\\m'\neq m}}^{M} Pr\left[(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.44}$$

by the union bound property: $Pr\left[\bigcup_{m=1}^{M} \mathcal{E}_i\right] \leq \sum_{m=1}^{M} Pr\left[\mathcal{E}_i\right]$.

From the Shannon-McMillan theorem for jointly typical pairs of sequences $(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)$ (for N sufficiently large):

$$Pr\left[(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)\right] \leq \epsilon_1 \tag{3.45}$$

Then, for a given transmitted codeword $\mathbf{c}_m$, the error probability is bounded as:

$$P_{e|m} \leq \epsilon_1 + \sum_{\substack{m'=1\\m'\neq m}}^{M} Pr\left[(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.46}$$

The second term on the right-hand side is not necessarily small. However, we will see that, on the average, this term is small.

Define an *indicator function* $\phi(\mathbf{x}, \mathbf{y})$ such that:

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & \text{if } (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta) \\ 0 & \text{if } (\mathbf{x}, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta) \end{cases} \tag{3.47}$$

Then, for a given transmitted codeword $\mathbf{c}_m$,

$$Pr\left[(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] = \sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\; p(\mathbf{y}|\mathbf{c}_m) \tag{3.48}$$

where $p(\mathbf{y}|\mathbf{c}_m)$ represents the probability of receiving the vector $\mathbf{y}$ given that the m^{th} codeword $\mathbf{c}_m$ was transmitted, and the sum $\sum_{\mathbf{y}}$ is over all received sequences $\{\mathbf{y}\}$.

Then, for all $m' \neq m$:

$$P_{e|m} \leq \epsilon_1 + \sum_{\substack{m'=1\\m'\neq m}}^{M} \sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\; p(\mathbf{y}|\mathbf{c}_m) \tag{3.49}$$

We now use the *random coding* scheme where each element of a code $\mathcal{C}$ is chosen according to a fix distribution $p(x)$. Instead of determining the error probability of a given code $\mathcal{C}$, we will determine the *expected* error decoding probability *over all possible codes*, i.e., $\mathcal{S}_{\mathcal{C}}$ with distribution $p(x)$.

The expected error decoding probability, given that codeword $\mathbf{c}_m$ was transmitted, over the ensemble of randomly chosen codes $\mathcal{S}_{\mathcal{C}}$ is:

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\epsilon_1 + \sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\ p(\mathbf{y}|\mathbf{c}_m)\right] \tag{3.50}$$

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\epsilon_1\right] + \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\ p(\mathbf{y}|\mathbf{c}_m)\right] \tag{3.51}$$

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \epsilon_1 + \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\ p(\mathbf{y}|\mathbf{c}_m)\right] \tag{3.52}$$

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \epsilon_1 + \sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\ p(\mathbf{y}|\mathbf{c}_m)\right] \tag{3.53}$$

But since the term $\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\sum_{\mathbf{y}} \phi(\mathbf{c}_{m'}, \mathbf{y})\ p(\mathbf{y}|\mathbf{c}_m)\right]$ is an expectation over the ensemble of codes $\mathcal{S}_{\mathcal{C}}$ (randomly chosen), it can be rewritten as $Pr\left[(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right]$. The expected probability of error becomes:

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \epsilon_1 + \sum_{\substack{m'=1 \\ m'\neq m}}^{M} Pr\left[(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.54}$$

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \epsilon_1 + (M-1)\ Pr\left[(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.55}$$

since the random codeword $\mathbf{x}$ is not a function of the received codeword index m'.

Now, let's consider the expected probability of decoding errors P_e over the set $\{\mathbf{c}_m\}$ of codewords (i.e., the code $\mathcal{C}$).

$$P_e = \sum_{m=1}^{M} p(\mathbf{c}_m)\ P_{e|m} \tag{3.56}$$

The expected probability of error $\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e]$ over the set of codes $\mathcal{S}_{\mathcal{C}}$ is:

$$\begin{aligned}
\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] &= \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[\sum_{m=1}^{M} p(\mathbf{c}_m)\; P_{e|m}\right] && (3.57)\\
\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] &= \sum_{m=1}^{M} \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[p(\mathbf{c}_m)\; P_{e|m}\right] && (3.58)\\
\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] &= \sum_{m=1}^{M} p(\mathbf{c}_m) \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] && (3.59)\\
\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] &= \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \left[P_{e|m}\right] && (3.60)
\end{aligned}$$

Therefore we can write the expected error decoding probability over $\mathcal{S}_{\mathcal{C}}$ as:

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \leq \epsilon_1 + (M-1)\; Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.61}$$

A decoding error will occur if the pair of transmitted and received codewords, $(\mathbf{x},\mathbf{y})$, are *jointly typical*, which implies that:

$$\begin{aligned}
\mathbf{x} \in \mathcal{T}_X(\delta) &\;\rightarrow\; b^{-N[H(X)+\delta]} \leq p(\mathbf{x}) \leq b^{-N[H(X)-\delta]}\\
\mathbf{y} \in \mathcal{T}_Y(\delta) &\;\rightarrow\; b^{-N[H(Y)+\delta]} \leq p(\mathbf{y}) \leq b^{-N[H(Y)-\delta]}\\
(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta) &\;\rightarrow\; \|\mathcal{T}_{XY}(\delta)\| \leq b^{N[H(XY)+\delta]}
\end{aligned}$$

Then:

$$\begin{aligned}
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &= \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)} p(\mathbf{x},\mathbf{y}) && (3.62)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &= \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)} p(\mathbf{x})p(\mathbf{y}) && (3.63)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &= \|\mathcal{T}_{XY}(\delta)\| p(\mathbf{x})p(\mathbf{y}) && (3.64)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &\leq \|\mathcal{T}_{XY}(\delta)\|\; b^{-N[H(X)-\delta]}\; b^{-N[H(Y)-\delta]} && (3.65)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &\leq b^{N[H(XY)+\delta]}\; b^{-N[H(X)-\delta]}\; b^{-N[H(Y)-\delta]} && (3.66)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &\leq b^{-N[H(X)+H(Y)-H(XY)-3\delta]} && (3.67)\\
Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] &\leq b^{-N[I(X;Y)-3\delta]} && (3.68)
\end{aligned}$$

The expected error decoding probability over $\mathcal{S}_{\mathcal{C}}$ becomes:

$$\begin{aligned}
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] &\leq \epsilon_1 + (M-1)\ Pr\left[(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] && (3.69)\\
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] &\leq \epsilon_1 + (M-1)\ b^{-N[I(X;Y)-3\delta]} && (3.70)\\
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] &\leq \epsilon_1 + M\ b^{-N[I(X;Y)-3\delta]} && (3.71)\\
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] &\leq \epsilon_1 + b^{NR}\ b^{-N[I(X;Y)-3\delta]} && (3.72)\\
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] &\leq \epsilon_1 + b^{-N[I(X;Y)-R-3\delta]} && (3.73)
\end{aligned}$$

The second term in the above equation can be made arbitrarily small provided that the code rate R is smaller than $I(X;Y) - 3\delta$ (making the exponent negative), and provided that the blocklength N is sufficiently large. If the fixed input distribution is chosen such as to maximize the mutual information $I(X;Y)$ then the probability of error can be made arbitrarily small provided that $R < C - 3\delta$. For $p(\mathbf{x}) = p^*(\mathbf{x})$:

$$\boxed{\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e] \leq \epsilon_1 + \epsilon_2}$$

Finally, if the average probability of error $\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\,[P_e]$ can be made smaller than $\epsilon_1 + \epsilon_2$, then there must exists a code $\mathcal{C}^*$ for which the error probability P_e is at least as good as the average:

$$\boxed{P_e \leq \epsilon}$$

QED

3.5.2 Converse of the Channel Coding Theorem

Shannon's coding theorem states that *there exists a channel code $\mathcal{C}$ of size M and blocklength N, such that the probability of decoding error P_e is arbitrarily small provided that the rate $R < C$, and that N is sufficiently large.* What happens now if we try to transmit information at a rate R above the channel capacity C? The converse of the channel coding theorem stipulates that:

Theorem *(Converse of the channel coding theorem):*

Let a memoryless channel with capacity C be used to transmit codewords of blocklength N and input information R. Then the error decoding probability P_e satisfies the following inequality:

$$\boxed{P_e(N,R) \geq 1 - \frac{C}{R} - \frac{1}{NR}}$$

If the rate $R > C$, then the error decoding probability P_e is bounded away from zero.

Proof:

The Fano's inequality provides a lower bound on the probability of error $P_e(N,R)$ in terms of the conditional entropy $H(\mathbf{X}|\mathbf{Y})$:

$$\boxed{H(\mathbf{X}|\mathbf{Y}) \leq 1 + NR\ P_e(N,R)}$$

The decoding error probability $P_e(N,R)$ is given by

$$P_e(N,R) = Pr\left(\tilde{\mathbf{W}} \neq \mathbf{W}\right) \tag{3.74}$$

Define a binary error variable E with entropy $H(E) \leq 1\ Sh$ such that:

$$E = \begin{cases} 1 & \text{if } \tilde{\mathbf{W}} \neq \mathbf{W} \text{ (an error)} \\ 0 & \text{if } \tilde{\mathbf{W}} = \mathbf{W} \text{ (no error)} \end{cases} \tag{3.75}$$

Consider the following equivocation: $H(E,\mathbf{W}|\mathbf{Y})$. It can be expanded as:

$$\begin{aligned} H(E,\mathbf{W}|\mathbf{Y}) &= H(E|\mathbf{Y}) + H(\mathbf{W}|E,\mathbf{Y}) \\ H(E,\mathbf{W}|\mathbf{Y}) &= H(\mathbf{W}|\mathbf{Y}) + H(E|\mathbf{W},\mathbf{Y}) \\ H(E|\mathbf{Y}) + H(\mathbf{W}|E,\mathbf{Y}) &= H(\mathbf{W}|\mathbf{Y}) + H(E|\mathbf{W},\mathbf{Y}) \end{aligned} \tag{3.76}$$

However, the term $H(E|\mathbf{W},\mathbf{Y}) = 0$ since, given both $\mathbf{W},\mathbf{Y}$ and thus $\tilde{\mathbf{W}}$, there is no uncertainty about E (i.e. we know for sure from the observation of both $\mathbf{W}$ and $\mathbf{Y}$ if there is an error or not).

Furthermore, $H(E|\mathbf{Y}) \leq H(E) \leq 1$.

$$\begin{aligned} H(E|\mathbf{Y}) + H(\mathbf{W}|E,\mathbf{Y}) &= H(\mathbf{W}|\mathbf{Y}) \\ 1 + H(\mathbf{W}|E,\mathbf{Y}) &\geq H(\mathbf{W}|\mathbf{Y}) \end{aligned} \tag{3.77}$$

Finally, consider the term $H(\mathbf{W}|E,\mathbf{Y})$:

$$\begin{aligned} H(\mathbf{W}|E,\mathbf{Y}) &= Pr(E=0)H(\mathbf{W}|E=0,\mathbf{Y}) + Pr(E=1)H(\mathbf{W}|E=1,\mathbf{Y}) \\ H(\mathbf{W}|E,\mathbf{Y}) &\leq Pr(E=0)0 + Pr(E=1)\log_2(M-1) \\ H(\mathbf{W}|E,\mathbf{Y}) &\leq Pr(E=1)\log_2(M) \\ H(\mathbf{W}|E,\mathbf{Y}) &\leq Pr(E=1)\log_2(2^{NR}) \\ H(\mathbf{W}|E,\mathbf{Y}) &\leq Pr(E=1)NR \\ H(\mathbf{W}|E,\mathbf{Y}) &\leq P_e(N,R)\ NR \end{aligned} \tag{3.78}$$

Then,

$$\begin{aligned} 1 + H(\mathbf{W}|E,\mathbf{Y}) &\geq H(\mathbf{W}|\mathbf{Y}) \\ 1 + P_e(N,R)\ NR &\geq H(\mathbf{W}|\mathbf{Y}) \end{aligned} \tag{3.79}$$

Again, we consider a source of information $\mathbf{W}$ which generates messages as symbols, or sourcewords, from a set of M possible messages with equal probability, i.e. $p(w) = \frac{1}{M} = 2^{-NR}$. The source entropy $H(\mathbf{W})$ is simply equal to $-\log_2 2^{-NR} = NR$. Using the relationships between entropies, equivocations and mutual information, can rewrite NR as:

$$\begin{aligned} NR &= H(\mathbf{W}) \\ NR &= H(\mathbf{W}|\mathbf{Y}) + I(\mathbf{W};\mathbf{Y}) \\ NR &= H(\mathbf{X}|\mathbf{Y}) + I(\mathbf{X};\mathbf{Y}) \end{aligned} \tag{3.80}$$

Using the Fano's inequality, i.e. $H(\mathbf{X}|\mathbf{Y}) \leq 1 + NR\ P_e(N,R)$, then:

$$\begin{aligned} NR &= H(\mathbf{X}|\mathbf{Y}) + I(\mathbf{X};\mathbf{Y}) \\ NR &\leq 1 + NR\ P_e(N,R) + I(\mathbf{X};\mathbf{Y}) \end{aligned} \tag{3.81}$$

The mutual information $I(\mathbf{X};\mathbf{Y})$ between the input vectors (original codewords of blocklength N) and output vectors (codewords corrupted by the memoryless channel) can be expressed as:

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - \sum_{n=1}^{N} H(Y_n|X_n) \\ I(\mathbf{X};\mathbf{Y}) &\leq \sum_{n=1}^{N} H(Y_n) - \sum_{n=1}^{N} H(Y_n|X_n) \end{aligned}$$

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &\leq \sum_{n=1}^{N} I(X_n;Y_n) \\ I(\mathbf{X};\mathbf{Y}) &\leq \sum_{n=1}^{N} C \\ I(\mathbf{X};\mathbf{Y}) &\leq NC \end{aligned} \tag{3.82}$$

using the facts that: *(i)* the joint entropy of vector $H(\mathbf{Y})$, which is a sum of entropies and equivocations, is less or equal to the sum of individual entropies $\sum_{n=1}^{N} H(Y_n)$; and *(ii)* the capacity C is the maximum of the mutual information $I(X_n;Y_n)$. Therefore, the source entropy $H(\mathbf{W}) = NR$ satisfies the inequality

$$\begin{aligned} NR &\leq 1 + NR\; P_e(N,R) + I(\mathbf{X};\mathbf{Y}) \\ NR &\leq 1 + NR\; P_e(N,R) + NC \end{aligned} \tag{3.83}$$

and finally, dividing by NR, we have that

$$1 \leq \frac{1}{NR} + P_e(N,R) + \frac{C}{R} \tag{3.84}$$

$$\boxed{P_e(N,R) \geq 1 - \frac{C}{R} - \frac{1}{NR}}$$

QED

Therefore, for $R > C$, we cannot achieve an arbitrarily low error decoding probability P_e. This results is known as the *weak converse* of the channel coding theorem. There is also a *strong converse* of the channel coding theorem, based on the probability of decoded symbol error, which states that, at rates R above channel capacity C, the error decoding probability P_e tends towards one.

3.6 Channel reliability function

3.6.1 Random coding exponent

The conditional error decoding probability $P_{e|m}$, given that the m^{th} message is generated by the source of information and encoded with codeword $\mathbf{c}_m$, is equal to (using as the M decoding regions $\{\mathcal{U}_m\}_{m=1}^{M}$ the jointly typical pairs of sequences):

$$P_{e|m} \leq Pr\left[(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)\right] + \sum_{\substack{m'=1 \\ m' \neq m}}^{M} Pr\left[(\mathbf{c}_{m'}, \mathbf{y}) \in \mathcal{T}_{XY}(\delta)\right] \tag{3.85}$$

The decoding error probability P_e of a code $\mathcal{C}$ is in fact a function $P_e(N, R)$ of the choosen code blocklength N, the code rate R, and (the channel capacity C being determined by the channel itself):

$$P_e(N, R) \leq \underbrace{Pr\left[(\mathbf{c}_m, \mathbf{y}) \notin \mathcal{T}_{XY}(\delta)\right]}_{\epsilon_1} + \underbrace{b^{-N[C-R-3\delta]}}_{\epsilon_2} = \epsilon \tag{3.86}$$

where the first term ϵ_1 is smaller than δ for a sufficiently large blocklength N, while the second term ϵ_2 decreases exponentially with N, provided that the code rate $R < C - 3\delta$. The error decoding probability $P_e(N, R)$, can be expressed as [Bla87]:

$$\boxed{P_e(N, R) \leq b^{-NE_r(R)}} \tag{3.87}$$

where the function $E_r(R)$ is called the *random coding exponent.*

The random coding exponent is defined as [Bla87]:

$$E_r(R) \equiv \max_{s \in [0,1]} \max_{\mathcal{S}_{\mathbf{p}}} \left[-sR - \log_b \sum_{j=1}^{J} \left(\sum_{k=1}^{K} p(x_k) p(y_j|x_k)^{\frac{1}{(1+s)}} \right)^{1+s} \right] \tag{3.88}$$

Since the random coding exponent can be written in terms of a double maximum over the input symbols distribution $\mathcal{S}_{\mathbf{p}}$ and a parameter $s \in [0, 1]$, the decoding error probability $P_e(N, R)$ can be written as a double minimum over the same sets:

$$P_e(N, R) \leq \min_{s \in [0,1]} \min_{\mathcal{S}_{\mathbf{p}}} \left[b^{sNR} \sum_{\mathbf{y}} \left(\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\frac{1}{(1+s)}} \right)^{1+s} \right] \tag{3.89}$$

Figure 3.13 illustrates the typical shape of the random coding exponent $E_r(R)$ as a function of the code rate R. The larger is the random coding exponent $E_r(R)$, the smaller will be the error decoding probability $P_e(N, R)$ for a given blocklength. As shown on this figure, the maximum of the random coding exponent $E_r(R)$, over the ensemble of all possible input distributions $\mathcal{S}_{\mathbf{p}}$ and parameter $s \in [0, 1]$, is achieved with $s = 1$ for low code rates. The largest code rate R at which the random coding exponent $E_r(R)$ decreases with a slope $s = -1$ is called the *critical rate* R_{crit}. The *cut-off rate* R_0 corresponds to the code rate R, at which the *tangent* of the random coding exponent $E_r(R)$ intersects with the rate axis. The cut-off rate R_0 is often considered as the code rate limit beyond which it is *very expensive* to communicate reliably [Bla87] over a noisy communication channel.

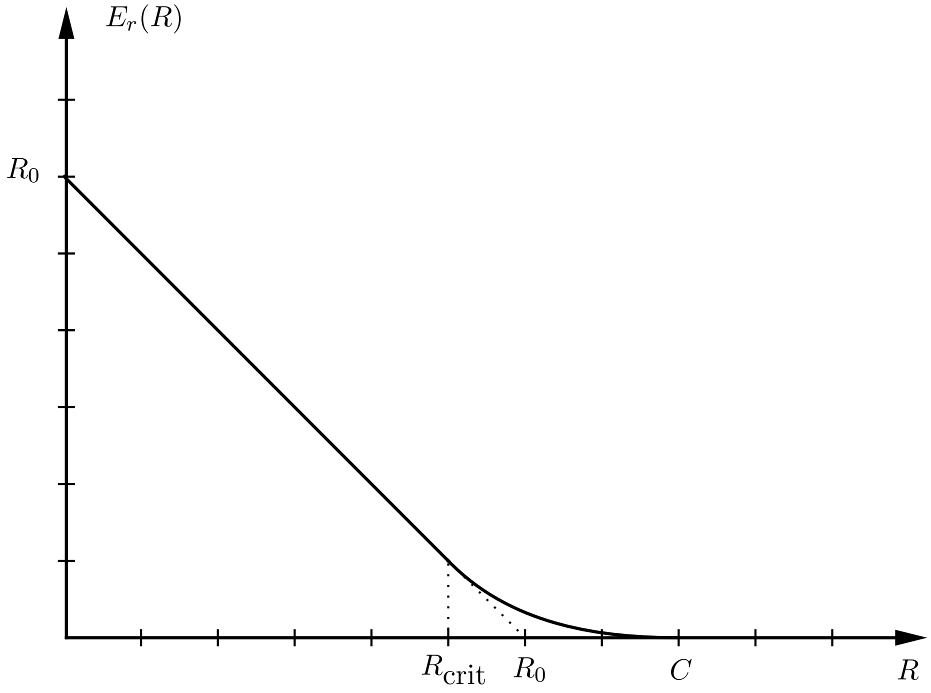


Figure 3.13: Random coding exponent $E_r(R)$ for block codes for BSC with $\epsilon = 10^{-2}$.

3.6.2 Error bounds and channel reliability function

The random coding exponent $E_r(R)$ provides an upperbound on the error decoding probability $P_e(N, R)$. This bound is obtained from a *random coding* argument. The question is: "*How tight is this bound?*". Other bounds on the probability of error have been derived based on different arguments. Some of them, such as the random coding bound, give an upperbound on the probability of error (i.e. *there exists a code for which P_e is smaller than ϵ_{upper}*), while others are *lower bounds* on P_e (i.e. *it is impossible to create codes such that $P_e < \epsilon_{lower}$*).

The *channel reliability function* $E^*(R)$ takes into account the upperbounds, as well as the lowerbounds, on the decoding error probability $P_e(\mathcal{C})$ [Bla87] to define an area where should lie the actual exponent of the probability of decoding error $P_e(N, R)$.

3.6.3 Error bounds on $P_e(N, R)$

a) The *random coding bound* $E_r(R)$ is a *lower bound* on the channel reliability function $E^*(R)$ and thus an *upper bound* on the error decoding probability $P_e(N, R)$. As we have already seen, this bound is obtained by *random codes selection* coding and a *maximum likelyhood* decoding rule (or also a decoding rule based on *jointly typical pairs of sequences*)

$$\boxed{P_e(N, R) \leq b^{-NE_r(R)}} \tag{3.90}$$

b) The *expurgated bound* $E_x(R)$ also provides a *lower bound* on $E^*(R)$ (and an *upper bound* on $P_e(N, R)$. Its derivation is based on a *random selection* of codes (i.e., $\mathcal{S}_{\mathcal{C}}$) and a *maximum likelyhood* decoding rule (as for the random coding bound). For small code rates R, many of the bad codes obtained by random selection can be *improved* before the expectation of the probability of error $\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C}) \, [P_e]$ is computed, leading to a lower decoding error probability.

$$\boxed{P_e(N, R) \leq b^{-NE_x(R)}} \tag{3.91}$$

c) The *space-partitioning bound* $E_p(R)$, however, is an *upper bound* on $E^*(R)$ and thus a *lower bound* on the decoding error probability $P_e(N, R)$. For this bound, the *space* $\mathcal{S}_{\mathbf{y}}$ of received (and corrupted) codewords is *partitioned* into a set of M decoding regions: $\mathcal{U}_1$, …, $\mathcal{U}_m$, …, $\mathcal{U}_M$. A message w_m is encoded with the codeword $\mathbf{c}_m$ before transmission in the noisy channel. A received vector $\mathbf{y}$ will be correctly decoded as message w_m if $\mathbf{y} \in \mathcal{U}_m$, but incorrectly decoded if $\mathbf{y} \notin \mathcal{U}_m$. Since the received vector $\mathbf{y}$ can be anywhere in the *space* $\mathcal{S}_{\mathbf{y}}$, this leads to a minimum probability of error $P_e(N, R)$ [Bla87]: it is impossible to find a code $\mathcal{C}$ with a lower probability of error.

$$\boxed{P_e(N, R) \geq b^{-NE_p(R)}} \tag{3.92}$$

d) The *sphere-packing bound* $E_s(R)$ is another *upper bound* on $E^*(R)$ (i.e. a *lower bound* on $P_e(N, R)$). Here the codewords of a code $\mathcal{C}$ are represented as points on the set of all possible received vectors $\mathcal{S}_{\mathbf{y}}$ in an N-dimensional space. The decoding regions: $\mathcal{U}_1$, …, $\mathcal{U}_m$, …, and $\mathcal{U}_M$; are represented as N-dimensional spheres. A decoding error occurs when the received codeword $\mathbf{y} \notin \mathcal{U}_m$, assuming that the correct codeword is $\mathbf{c}_m$, or when there is an overlapping of spheres: $\mathbf{y} \in \mathcal{U}_m$ and $\mathbf{y} \in \mathcal{U}'_m$ $(m \neq m')$. The problem is how much decoding spheres can be packed in the N-dimensional space with minimum (or no) overlapping?

$$\boxed{P_e(N,R) \geq b^{-NE_s(R)}} \tag{3.93}$$

e) Finally, the *straight-line bound* $E_l(R)$ is an *upper bound* on the reliability function $E^*(R)$ and therefore a *lower bound* on $P_e(N,R)$. It is based on the conjecture (not proven yet) that the channel reliability function is a *convex* $\cup$ function over the rate R of the code. If this argument is true, then any point on a straight line between any two points on the *upperbounds* of $E^*(R)$ will be also an upperbound of $E^*(R)$.

$$\boxed{P_e(N,R) \geq b^{-NE_l(R)}} \tag{3.94}$$

3.6.4 Channel reliability function $E^*(R)$

The channel reliability function $E^*(R)$ is established from these bounds as shown below on Figure 3.14.

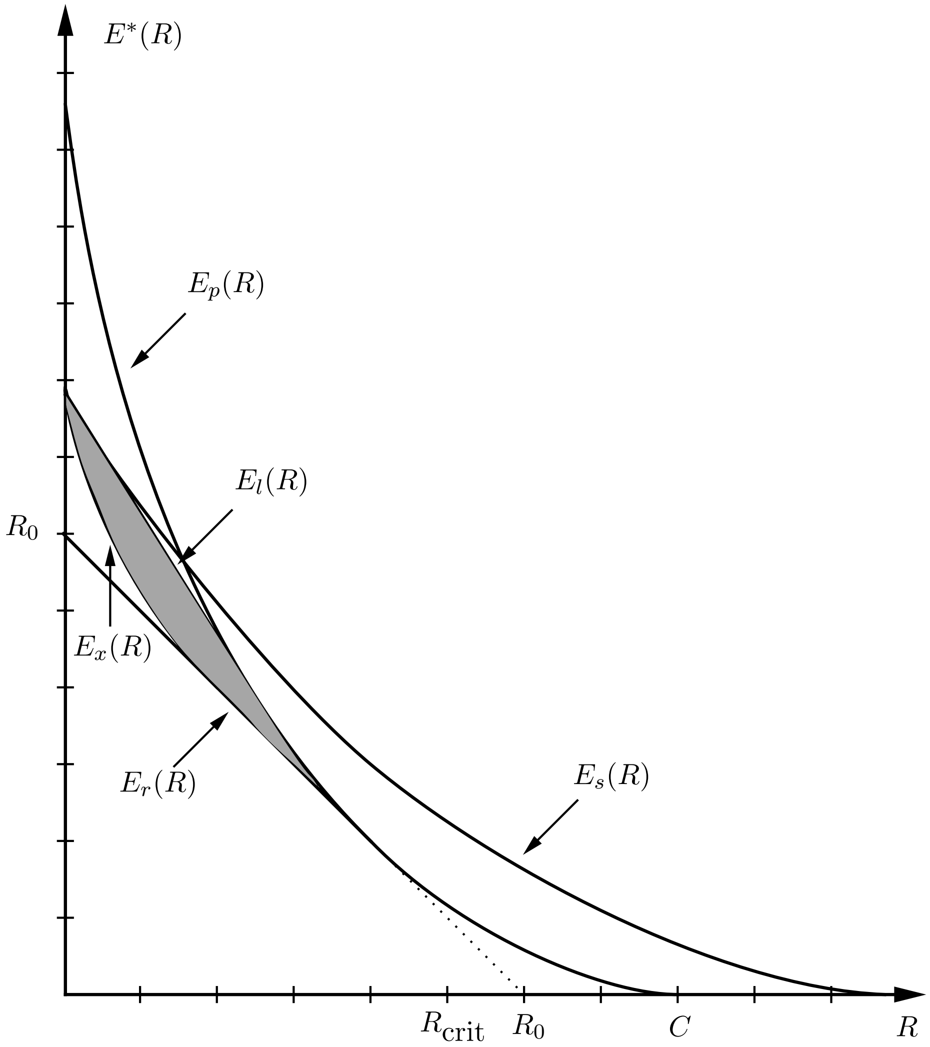


Figure 3.14: Channel reliability function $E^*(R)$.

3.7 Channel coding theorem revisited

3.7.1 Shannon's channel coding theorem (random coding exponent)

In the proof of Shannon's channel coding theorem, we used a random coding argument along with a decision rule based on *jointly typical pairs of sequences.* This was useful since we knew the properties of such sequences, and it provided a bounded expression for the probability of decoding errors P_e: an exponent of base b, function of the difference between the rate R of error control code and the channel capacity C, as well as the blocklength N of the codewords (we implicitly assumed a fixed length code).

In this section, we determine the probability of decoding errors P_e, using once again the random coding argument. However, this time, we will use a *maximum likehood* decoding rule.

Theorem *(Shannon's channel coding theorem (random coding exponent)):*

Let C be the information transfer capacity of a memoryless channel defined by its transition probabilities matrix $\mathbf{P}$. If the code rate $R < C$, then there *exists* a channel code $\mathcal{C}$ of size M and blocklength N, such that the probability of decoding error P_e is *upperbounded* as:

$$\boxed{P_e(N,R) \leq \min_{s\in[0,1]} \min_{\mathcal{S}_{\mathbf{p}}} \left[b^{sNR} \sum_{\mathbf{y}} \left(\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\frac{1}{(1+s)}} \right)^{1+s} \right]}$$

provided that the blocklength N is sufficiently large (i.e., $N \geq N_0$).

Proof:

The proof is based as before on the random selection of a set of codes $\mathcal{S}_{\mathcal{C}} = \{\mathcal{C}\}$ and the expected error probability P_e over that set. Once again we assume that the source of information $\mathbf{W}$ is an equiprobable source:

$$p(w_m) = \frac{1}{M} = b^{-NR} \qquad \forall m \tag{3.95}$$

where R is the code rate, N the blocklength. The encoder assigns the unique codeword $\mathbf{c}_m$ to the message w_m.

$$\mathcal{C} = \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,n} & \cdots & c_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,n} & \cdots & c_{M,N} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} & \cdots & x_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,n} & \cdots & x_{M,N} \end{bmatrix}$$

The b^{NR} codewords of the code $\mathcal{C}$ are chosen such as the distribution of the codewords elements $p(x)$ maximize the mutual information $I(X;Y)$ and hence lead to the channel capacity C. At the channel output, a maximum likehood decoding decoder assigns to the received (and probably corrupted) codeword $\mathbf{y}$, the *original* codeword $\mathbf{c}_m$ according to the maximum likelyhood rule:

$$\mathcal{U}_m = \{\mathbf{y} : p(\mathbf{y}|\mathbf{c}_m) > p(\mathbf{y}|\mathbf{c}_{m'}) \qquad \text{for all } m \neq m'\}$$

The error probability $P_{e|m}$, given that codeword $\mathbf{c}_m$ is transmitted, is given by:

$$P_{e|m} = \sum_{\mathbf{y} \notin \mathcal{U}_m} p(\mathbf{y}|\mathbf{c}_m) \tag{3.96}$$

Let $\phi_m(\mathbf{y})$ be an indicator function defined as:

$$\phi_m(\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{y} \in \mathcal{U}_m \\ 0 & \text{if } \mathbf{y} \notin \mathcal{U}_m \end{cases}$$

This indicator function in bounded as:

$$1 - \phi_m(\mathbf{y}) \leq \left\{ \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \left[\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)} \right]^{\left(\frac{1}{1+s}\right)} \right\}^s \qquad \text{for } s \in [0,1] \tag{3.97}$$

The expression $1-\phi_m(\mathbf{y})$ can take only two values, i.e., $1-\phi_m(\mathbf{y}) = 0$ or $1-\phi_m(\mathbf{y}) = 1$. For $1-\phi_m(\mathbf{y}) = 0$, the inequality holds true since the right-hand side of the inequality is positive; it is a sum a positive terms raised at a positive power:

$$\left\{ \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \left[\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)} \right]^{\left(\frac{1}{1+s}\right)} \right\}^s \geq 0 \tag{3.98}$$

Now, if $1-\phi_m(\mathbf{y}) = 1$, i.e., $\phi_m(\mathbf{y}) = 0$, then this indicates that there is at least a codeword $\mathbf{c}_{m'}$ such that

$$p(\mathbf{y}|\mathbf{c}_{m'}) \geq p(\mathbf{y}|\mathbf{c}_m)$$

(thus leading to a decoding error). Then, for this codeword $\mathbf{c}_{m'}$:

$$\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)} \geq 1 \tag{3.99}$$

Raising to a positive power $\left(\frac{1}{1+s}\right)$ and adding $M-2$ non negative terms maintains the inequality:

$$\sum_{\substack{m'=1\\m'\neq m}}^{M} \left[\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)}\right]^{\left(\frac{1}{1+s}\right)} \geq 1 \tag{3.100}$$

and raising once more to a positive exponent s still preserves the inequality:

$$\left\{\sum_{\substack{m'=1\\m'\neq m}}^{M} \left[\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)}\right]^{\left(\frac{1}{1+s}\right)}\right\}^{s} \geq 1 \tag{3.101}$$

The error probability can be expressed as a function of the indicator function and then upperbounded as follows:

$$\begin{aligned}
P_{e|m} &= \sum_{\mathbf{y}\notin\mathcal{U}_m} p(\mathbf{y}|\mathbf{c}_m) && (3.102)\\
P_{e|m} &= \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)[1-\phi_m(\mathbf{y})] && (3.103)\\
P_{e|m} &\leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)\left\{\sum_{\substack{m'=1\\m'\neq m}}^{M} \left[\frac{p(\mathbf{y}|\mathbf{c}_{m'})}{p(\mathbf{y}|\mathbf{c}_m)}\right]^{\left(\frac{1}{1+s}\right)}\right\}^{s} && (3.104)\\
P_{e|m} &\leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)\left\{p(\mathbf{y}|\mathbf{c}_m)^{\left(-\frac{1}{1+s}\right)} \sum_{\substack{m'=1\\m'\neq m}}^{M} [p(\mathbf{y}|\mathbf{c}_{m'})]^{\left(\frac{1}{1+s}\right)}\right\}^{s} && (3.105)\\
P_{e|m} &\leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)p(\mathbf{y}|\mathbf{c}_m)^{\left(-\frac{s}{1+s}\right)}\left\{\sum_{\substack{m'=1\\m'\neq m}}^{M} [p(\mathbf{y}|\mathbf{c}_{m'})]^{\left(\frac{1}{1+s}\right)}\right\}^{s} && (3.106)\\
P_{e|m} &\leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1+s-s}{1+s}\right)}\left\{\sum_{\substack{m'=1\\m'\neq m}}^{M} [p(\mathbf{y}|\mathbf{c}_{m'})]^{\left(\frac{1}{1+s}\right)}\right\}^{s} && (3.107)\\
P_{e|m} &\leq \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)}\left\{\sum_{\substack{m'=1\\m'\neq m}}^{M} [p(\mathbf{y}|\mathbf{c}_{m'})]^{\left(\frac{1}{1+s}\right)}\right\}^{s} && (3.108)
\end{aligned}$$

Considering the random selection scheme where each element of a code $\mathcal{C}$ is chosen according to the distribution $p(x)$, we will determine the average error decoding probability over the set $\mathcal{S}_{\mathcal{C}}$ of all possible codes (with distribution $p(x)$).

$$\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left\{ \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)} \left[\sum_{\substack{m'=1\\m'\neq m}}^{M} p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)} \right]^s \right\} \tag{3.109}$$

The expectation of a sum being equal to the sum of expectations,

$$\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathbf{y}} \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left\{ p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)} \left[\sum_{\substack{m'=1\\m'\neq m}}^{M} p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)} \right]^s \right\} \tag{3.110}$$

Note that the first term on the right-hand side is a function of the selected codeword $\mathbf{c}_m$ while the second term depends only on the codeword $\mathbf{c}_{m'}$ which are selected randomly from each other. Therefore the error expectation $\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right]$ can be written as:

$$\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathbf{y}} \left\{ \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)}\right] \times \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[\sum_{\substack{m'=1\\m'\neq m}}^{M} p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)} \right]^s \right\} \tag{3.111}$$

Using the Jensen's inequality[1]:

$$\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathbf{y}} \left\{ \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)}\right] \times \left[\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \sum_{\substack{m'=1\\m'\neq m}}^{M} p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)} \right]^s \right\} \tag{3.112}$$

$$\sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathbf{y}} \left\{ \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{c}_m)^{\left(\frac{1}{1+s}\right)}\right] \times \left\{ \sum_{\substack{m'=1\\m'\neq m}}^{M} \sum_{\mathcal{S_C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)}\right] \right\}^s \right\} \tag{3.113}$$

Since the information source $\mathbf{W}$ generates the messages with equal probability, then the codewords occur with equal probability:

$$p(\mathbf{c}_m) = \frac{1}{M} = b^{-NR}$$

and the second term $\left[p(\mathbf{y}|\mathbf{c}_{m'})^{\left(\frac{1}{1+s}\right)}\right]$ do not depend on the actual message, or codeword, transmitted.

[1] The Jensen's inequality states that $E\left[x^s\right] \leq \left[E(x)\right]^s$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq \sum_{\mathbf{y}} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right] \times \left\{ \sum_{\substack{m'=1 \\ m'\neq m}}^{M} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right] \right\}^{s} \quad (3.114)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq (M-1)^s \sum_{\mathbf{y}} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right] \times \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{s} \quad (3.115)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq (M-1)^s \sum_{\mathbf{y}} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{(s+1)} \quad (3.116)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq M^s \sum_{\mathbf{y}} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{(s+1)} \quad (3.117)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \leq b^{NRs} \sum_{\mathbf{y}} \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{(s+1)} \quad (3.118)$$

Rewritting the right-hand side in terms of the probability $p(\mathbf{x})$ of selecting the random codeword $\mathbf{x}$ over the set of codes $\mathcal{S}_\mathcal{C}$ and considering the expected probability of error for any transmitted codeword.

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_e\right] = \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[\sum_{m=1}^{M} p(\mathbf{c}_m)\; P_{e|m}\right] \quad (3.119)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_e\right] = \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_{e|m}\right] \quad (3.120)$$

$$\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \left[P_e\right] \leq b^{NRs} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{(s+1)} \quad (3.121)$$

$$P_e \leq b^{NRs} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)}\right]^{(s+1)} \quad (3.122)$$

because there must exist one code $\mathcal{C}^*$ for which P_e is, at most, as small as the the expected probability of decoding error over the set of codes $\mathcal{S}_\mathcal{C}$. This holds true for value of parameter $s \in [0,1]$. Therefore, the error probability P_e can be written as a *double minimum* over the input distribution (to maximize the mutual information and hence reach the channel capacity C) and the positive exponent parameter s:

$$\boxed{P_e(N,R) \leq \min_{s\in[0,1]} \min_{\mathcal{S}_\mathbf{p}} \left[b^{sNR} \sum_{\mathbf{y}} \left(\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\frac{1}{(1+s)}} \right)^{1+s} \right]} \quad (3.123)$$

The error probability $P_e(N,R)$ can also be expressed in terms of the individual codeword elements. Since we consider a memoryless channel and the random generation of codewords, we have that:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} p(y_j|x_k) \qquad \text{and} \tag{3.124}$$

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_k) \tag{3.125}$$

then:

$$P_e \leq b^{NRs} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \tag{3.126}$$

$$P_e \leq b^{NRs} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} \prod_{n=1}^{N} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \tag{3.127}$$

$$P_e \leq b^{NRs} \sum_{y_1=1}^{J} \sum_{y_2=1}^{J} \cdots \sum_{y_N=1}^{J} \left[\sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \cdots \sum_{x_N=1}^{K} \prod_{n=1}^{N} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \tag{3.128}$$

The sums of products can be replaced by a product of sums using the rule:

$$\sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \cdots \sum_{x_N=1}^{K} \left[\prod_{n=1}^{N} A(x_n) \right] = \prod_{n=1}^{N} \left[\sum_{x_N=1}^{K} A(x_n) \right] \tag{3.129}$$

and therefore,

$$P_e \leq b^{NRs} \sum_{y_1=1}^{J} \sum_{y_2=1}^{J} \cdots \sum_{y_N=1}^{J} \left[\sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \cdots \sum_{x_N=1}^{K} \prod_{n=1}^{N} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \tag{3.130}$$

$$P_e \leq b^{NRs} \prod_{n=1}^{N} \left[\sum_{y_N=1}^{J} \left[\sum_{x_N=1}^{K} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \right] \tag{3.131}$$

$$P_e \leq b^{NRs} \left[\sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \right]^{N} \tag{3.132}$$

$$\tag{3.133}$$

Maximizing over the range s and the input distributions, the random coding error bound can expressed as:

$$\boxed{P_e(N,R) \leq \min_{s\in[0,1]} \min_{\mathcal{S}_{\mathbf{p}}} \left[b^{NRs} \left[\sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(x_k) p(y_j|x_k)^{\left(\frac{1}{1+s}\right)} \right]^{(s+1)} \right]^{N} \right]} \tag{3.134}$$

The random coding exponent $E_r(R) \approx -\frac{1}{N}\log_b P_e$ can be expressed as *double maximum* over the same sets $\mathcal{S}_{\mathbf{p}}$ and $s \in [0,1]$:

$$\boxed{E_r(R) \equiv \max_{s\in[0,1]} \max_{\mathcal{S}_{\mathbf{p}}} \left[-sR - \log_b \sum_{j=1}^{J} \left(\sum_{k=1}^{K} p(x_k) p(y_j|x_k)^{\frac{1}{(1+s)}} \right)^{1+s} \right]} \tag{3.135}$$

QED

3.8 Problems

Problem 3.1: The average mutual information $I(X;Y)$ is a convex $\cap$ function (i.e. "concave" function) over the convex set of input symbols distributions $\{\bar{p}\} = \{\{p(x_k)\}\}$. However, $I(X;Y)$ is a convex $\cup$ function (or "convex" function) over the convex set of channel transition probabilities matrices $\{\mathbf{P}\} = \{\{p(y_j|x_k)\}\}$.

a) Show that the set of channel transition probabilities matrices $\{\mathbf{P}\} = \{\{p(y_j|x_k)\}\}$ forms a convex set.

b) Show that over this convex set $\{\mathbf{P}\} = \{\{p(y_j|x_k)\}\}$, the average mutual information is a convex $\cup$ function.

Problem 3.2: The channel transition probability matrix $\mathbf{P}$ of a ternary communication channel is given by:

$$\mathbf{P} = \begin{pmatrix} (1-2\epsilon) & \epsilon & \epsilon \\ \epsilon & (1-2\epsilon) & \epsilon \\ \epsilon & \epsilon & (1\ -2\epsilon) \end{pmatrix}$$

a) What is capacity C of the ternary channel as a function of the crossover probability ϵ?

b) If this ternary channel is cascaded with another identical ternary channel what will be the resulting channel capacity $C_{cascade}$?

c) Draw on the same figure both capacities (i.e. C and $C_{cascade}$) for $\epsilon \in [0, 0.5]$.

d) How does the cascading of these ternary channels affect the rate of transfer of information (i.e. $I(X;Y)$)?

Problem 3.3: Consider all binary sequences of blocklength $N = 7$. We wish to choose a code $\mathcal{C}$ of size M:

$$\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_m, \cdots, \mathbf{c}_M\}$$

To correct all single errors during transmission through a binary channel, the minimum distance d_{min} between any pairs of codewords $(\mathbf{c}_m, \mathbf{c}_{m'})$ should be ≥ 3.

a) How many binary sequences have a distance less than or equal to 1 from a given codeword $\mathbf{c}_m$?

b) What is the maximum possible value of the code size M?

c) Assuming that the maximum value of M can be achieved, what is the rate R of generation of information that can be transmitted over the channel?

Problem 3.4: Consider a binary noisy channel for which the noise affects the transmitted bits in blocks of 15 bits. For this specific channel, a block is either transmitted without error, with one error or with 2 errors out of the 15 bits. Each combination of no error, single error or double errors occur with equal probability. Let $\mathbf{X}$ represent the 15-bits source sequences whereas $\mathbf{Y}$ represent the 15-bits sequences received from the noisy channel.

a) Indicate the number of error patterns and their probability of occurrence.

b) Determine the equivocation $H(\mathbf{Y}|\mathbf{X})$.

c) What is the maximum value of the mutual information between the input and output sequences $I(\mathbf{X};\mathbf{Y})$? Under what conditions is achieved the maximum of $I(\mathbf{X};\mathbf{Y})$?

d) Now what is the capacity C of the original binary channel (i.e. considering a transmission bit per bit but under the same channel conditions)?

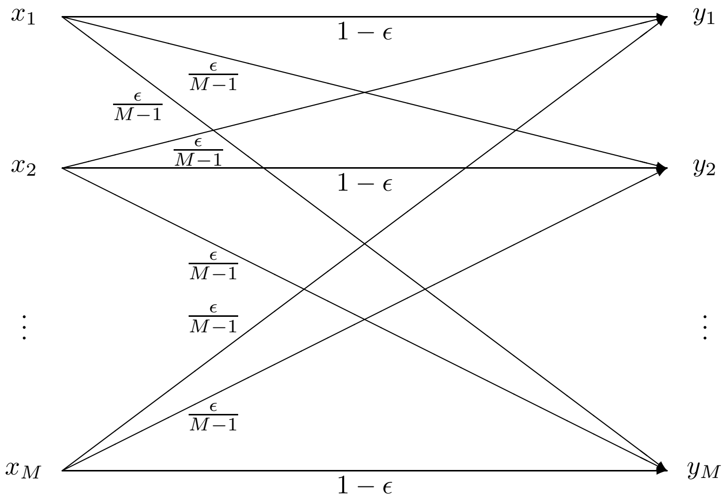
Problem 3.5: Consider n identical binary symmetric channels with crossover probability $0 < \epsilon < \frac{1}{2}$.

a) Find the channel capacity C_2 when two of these BSC channels are cascaded.

b) Write an expression for the capacity C_n of n cascaded channels.

c) Suppose that the number of BSC channels $n \to \infty$. What is the resulting capacity C_∞? Justify your answer and explain the result.

Problem 3.6: Consider the following M-ary memoryless channel.



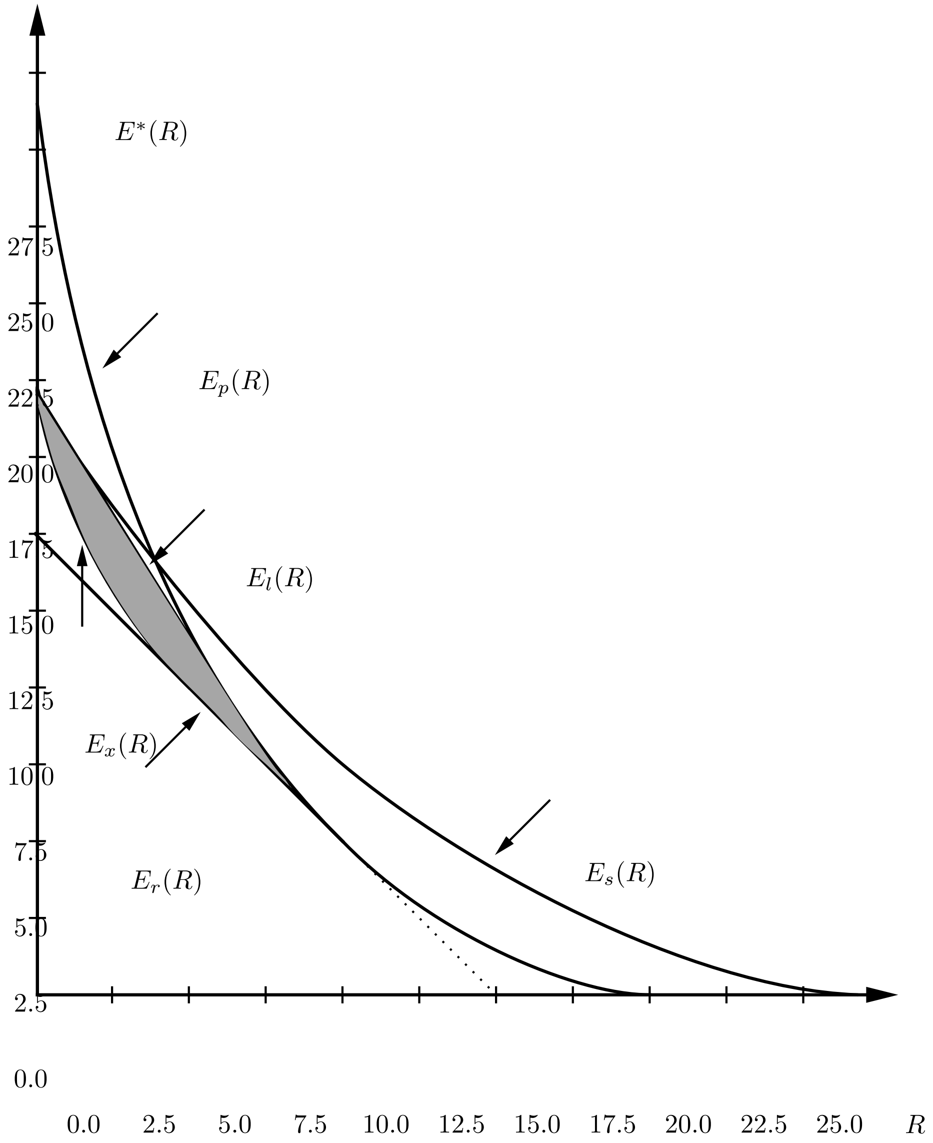
a) Find the expression for the channel capacity C as a function of the parameter ϵ.

b) Draw the channel capacity function C for $0 \leq \epsilon \leq 1$.

c) If we put two of these M-ary memoryless channels in cascade what will be the expression of the resulting *composite channel* C_2?

d) Sketch the composite channel capacity C_2 over the same range of parameter ϵ.

Problem 3.7: The channel reliability function of a given transmission channel is given below. We want to use channel coding in order to insure reliable communication over this noisy channel.

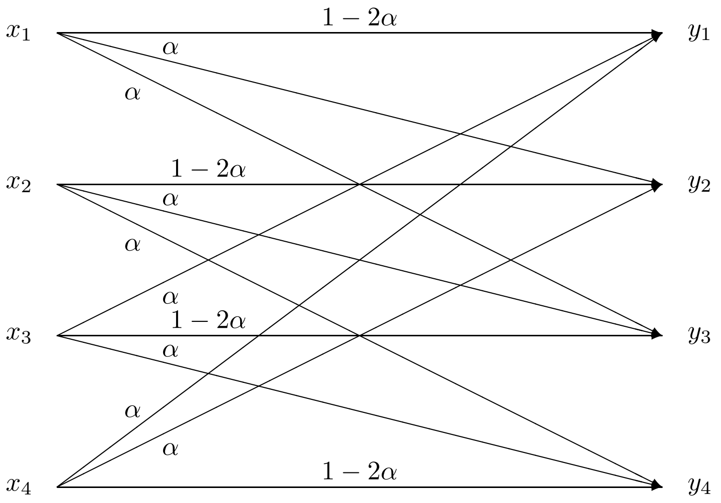


a) What is the maximum code rate $R = \frac{k}{n}$ for this channel. Give the theoretical as well as practical limits.

b) How would you assess the probability of error P_e for a code with a rate $R = 10$ and a code length $n = 255$?

c) What can be said about the probability of error for a code rate $R = 2.5$ and a code length $n = 63$?

Problem 3.8: Consider the following quaternary memoryless channel.



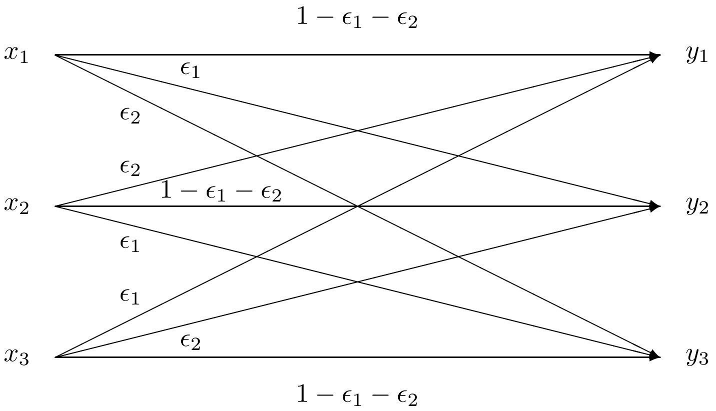
a) What is the transition probability matrix **P**?

b) Find the expression for the channel capacity C as a function of the parameter α.

c) Compute the values of the channel capacity function C for $\alpha = 0$, 0.05, 0.10, … , 0.40, 0.45, 0.50, and draw this channel capacity function C for that range of α values.

Problem 3.9: For each of the following channels, give the transition probability matrix **P**, determine the expression for the channel capacity (in *Sh* or bits) and the corresponding input symbols' distribution.

a) Ternary memoryless channel:



b) Pentagonal channel, where for $0 \leq k \leq 4$ and $0 \leq j \leq 4$:

$$p(y_j|x_k) = \begin{cases} \frac{1}{2} & \text{if } j = k \pm 1 \bmod 5 \\ 0 & \text{if } j \neq k \pm 1 \bmod 5 \end{cases}$$

Problem 3.10: *Computer-oriented problem*

Write a program which computes the capacity C of asymmetric channels. Use the Blahut-Arimoto algorithm described in class. You have to provide the program's listing. (Note: for further references you may wish to read section 5.4 of "Principles and Practice of Information Theory" by Richard E. Blahut and/or section 13.8 of "Elements of Information Theory" by Thomas M. Cover and Joy A. Thomas. Compute the channel capacity of the following channels:

a)

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.1 & 0.7 \end{pmatrix}$$

b)

$$\mathbf{P} = \begin{pmatrix} 0.4 & 0.2 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.2 & 0.4 & 0.0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.4 & 0.0 & 0.0 & 0.4 & 0.2 \\ 0.2 & 0.4 & 0.0 & 0.0 & 0.4 \end{pmatrix}$$

Chapter 4

Rate Distortion Theory

4.1 Rate distortion function

Definition *(Rate distortion function):*

The *rate distortion function* $R(D)$ is defined, for a given distortion *criterion* (or *fidelity criterion*) D, as the minimum of the mutual information $I(X;\hat{X})$ over the set of D-admissible transition matrices:

$$\boxed{R(D) \equiv \min_{\mathcal{P}_D} I(X;\hat{X})}$$

where $\mathcal{P}_D$ is the set of all D-admissible (distortion) transition probability matrices, defined as:

$$\boxed{\mathcal{P}_D \equiv \left\{ \mathbf{P} : \sum_{k=1}^{K} \sum_{j=1}^{J} p(x_k) p(\hat{x}_j|x_k) d(x_k, \hat{x}_j) \leq D \right\}}$$

Theorem *(Convexity of the rate distortion function):*

The rate distortion function $R(D)$ is a *convex cup*, or convex $\cup$, function over the convex set $\mathcal{D} \equiv \{D\}$, decreasing in the distortion interval: $[d_{min}, d_{max}]$.

Proof:

Consider the set $\mathcal{P}_D$ of all D-admissible transition probability matrices (which are function of the fidelity criterion D).

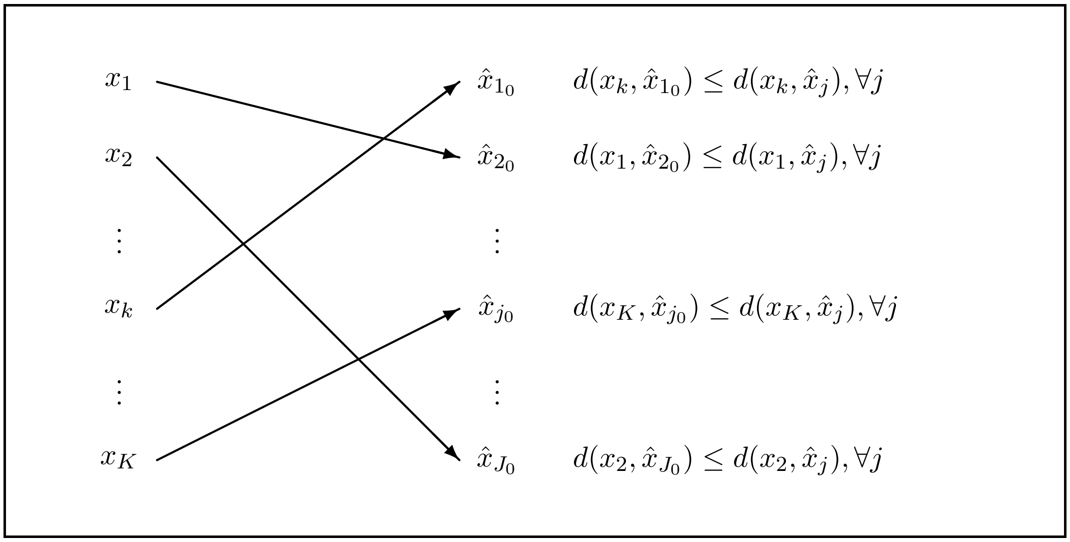
$$\mathcal{P}_D = \left\{ \mathbf{P} = \{p(\hat{x}_j|x_k)\} : \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j)}_{\text{average distortion}} \leq \underbrace{D}_{\substack{\text{distortion}\\ \text{criterion}}} \right\} \tag{4.1}$$

a) If the distortion criterion D is negative, then the set $\mathcal{P}_D$ is empty, which implies that the rate distortion function $R(D)$ is not defined (this is due to the fact that the average distortion is defined as being always positive).

b) If D is positive but smaller than d_{min}, then the set $\mathcal{P}_D$ is still empty, since it is not possible to find a matrix $\mathbf{P}$ giving an average distortion $\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k,\hat{x}_j)d(x_k,\hat{x}_j)$ smaller than D:

$$d_{min} = \sum_{k=1}^{K} p(x_k)d(x_k,\hat{x}_{j_0}) \tag{4.2}$$

where $d(x_k,\hat{x}_{j_0}) \leq d(x_k,\hat{x}_j)$, for $j = 1,\ldots,J$.


x_1
x_2
$\vdots$
x_k
$\vdots$
x_K
$\hat{x}_{1_0}$
$\hat{x}_{2_0}$
$\vdots$
$\hat{x}_{j_0}$
$\vdots$
$\hat{x}_{J_0}$
$d(x_k,\hat{x}_{1_0}) \leq d(x_k,\hat{x}_j), \forall j$
$d(x_1,\hat{x}_{2_0}) \leq d(x_1,\hat{x}_j), \forall j$
$d(x_K,\hat{x}_{j_0}) \leq d(x_K,\hat{x}_j), \forall j$
$d(x_2,\hat{x}_{J_0}) \leq d(x_2,\hat{x}_j), \forall j$

c) For $d_{min} \leq D \leq D' \leq d_{max}$, the set $\mathcal{P}_D \in \mathcal{P}_{D'}$ because the set $\mathcal{P}_{D'}$ *admits* transition probability matrices for which the average distortion can attain D' while the set $\mathcal{P}_D$ includes only those matrices with average distortion smaller or equal to D.

As the allowable distortion D increases, the set of D-admissible transition probability matrices $\mathcal{P}_D$, satisfying the condition $E[d(X,\hat{X})] \leq D$, increases. This provides more matrices over which the average mutual information $I(X;\hat{X})$ can be minimized. Therefore:

$$R(D') \quad = \quad \min_{\mathcal{P}_{D'}} I(X;\hat{X}) \tag{4.3}$$

$$\begin{aligned}
R(D') &= \min_{\mathcal{P}_{D'}} \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)\log_b\left[\frac{p(\hat{x}_j|x_k)}{\sum_{l=1}^{K} p(x_l)p(\hat{x}_j|x_l)}\right] && (4.4)\\
R(D') &\leq \min_{\mathcal{P}_D\in\mathcal{P}_{D'}} \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)\log_b\left[\frac{p(\hat{x}_j|x_k)}{\sum_{l=1}^{K} p(x_l)p(\hat{x}_j|x_l)}\right] && (4.5)\\
R(D') &\leq \min_{\mathcal{P}_D\in\mathcal{P}_{D'}} I(X;\hat{X}) && (4.6)
\end{aligned}$$

d) The maximum distortion d_{max} is defined as the *minimum distortion* D for which the mutual information is null between the input X and its reproduction $\hat{X}$; i.e., $I(X;\hat{X}) = 0$. Therefore, at this rate, there is no transfer of information and $\hat{X}$ is independent of X. The transition probabilities $p(\hat{x}_j|x_k)$ are then equal to the marginal probabilities $p(\hat{x}_j)$.

The expected (or average) distortion $E[d(X,\hat{X})]$ is:

$$\begin{aligned}
E[d(X,\hat{X})] &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j) && (4.7)\\
E[d(X,\hat{X})] &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j)d(x_k,\hat{x}_j) && (4.8)\\
E[d(X,\hat{X})] &= \sum_{j=1}^{J} p(\hat{x}_j)\sum_{k=1}^{K} p(x_k)d(x_k,\hat{x}_j) && (4.9)
\end{aligned}$$

The maximum distortion d_{max} is the smallest distortion D with zero information transfer [1]:

$$d_{max} = \min_{j=1,\ldots,J} \sum_{k=1}^{K} p(x_k)d(x_k,\hat{x}_j) \qquad (4.10)$$

that is, choosing the reproducing letter to minimize the distortion D.

e) We now proove that the rate distortion function $R(D)$ is a convex $\cup$ function of the distortion D. This is done in two steps: first, the distortion D is shown to be a convex set, and secondly, we show that $R(D)$ is indeed a convex $\cup$ function of D.

i) Let's define two transition probability matrices $\mathbf{P}'$ and $\mathbf{P}''$ such that:

$$\begin{aligned}
\mathbf{P}' &= \{p'(\hat{x}_j|x_k)\} \quad\Rightarrow\quad D' \text{ and } R(D')\\
\mathbf{P}'' &= \{p''(\hat{x}_j|x_k)\} \quad\Rightarrow\quad D'' \text{ and } R(D'')
\end{aligned}$$

that is:

$$\begin{aligned}
\mathbf{P}'\in\mathcal{P}_{D'} &= \left\{\mathbf{P}' : \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p'(\hat{x}_j|x_k)d(x_k,\hat{x}_j) \leq D'\right\}\\
\mathbf{P}''\in\mathcal{P}_{D''} &= \left\{\mathbf{P}'' : \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p''(\hat{x}_j|x_k)d(x_k,\hat{x}_j) \leq D''\right\}
\end{aligned}$$

[1] Note that there can be greater rates at which $I(X;\hat{X}) = 0$, but here we are interested only in the smallest rate.

As seen before, the set of transition probability matrices $\{\mathbf{P}\}$ is a convex set, thus we can choose another matrix $\mathbf{P}$ as:

$$p(\hat{x}_j|x_k) = \lambda p'(\hat{x}_j|x_k) + (1-\lambda)p''(\hat{x}_j|x_k) \tag{4.11}$$

by definition of a convex set, where $\lambda \in [0,1]$. If we take D equal to the expected distortion $E[d(X,\hat{X})]$:

$$\begin{aligned}
d(X,\hat{X}) &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j) && (4.12)\\
d(X,\hat{X}) &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)\left[\lambda p'(\hat{x}_j|x_k) + (1-\lambda)p''(\hat{x}_j|x_k)\right]d(x_k,\hat{x}_j) && (4.13)\\
d(X,\hat{X}) &= \lambda\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p'(\hat{x}_j|x_k)d(x_k,\hat{x}_j)\\
&+ (1-\lambda)\sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p''(\hat{x}_j|x_k)d(x_k,\hat{x}_j) && (4.14)\\
d(X,\hat{X}) &= \lambda d'(X,\hat{X}) + (1-\lambda)d''(X,\hat{X}) && (4.15)
\end{aligned}$$

For the pseudo-channels for which $d(X,\hat{X}) = D$ then:

$$D = \lambda D' + (1-\lambda)D'' \tag{4.16}$$

ii) We consider now the convexity of the rate distortion function $R(D)$ itself. The mutual information $I(X;\hat{X})$ is a convex $\cup$ function over the convex set of transition probabilities' matrices.

$$R(D) = \min_{\mathcal{P}_D} I(X;\hat{X})$$

Let $\mathbf{P}' = \{p'(\hat{x}_j|x_k)\}$ be a D'-admissible *pseudo-channel* (or reproduction channel) with mutual information $I'(X;\hat{X}) = R(D')$, and $\mathbf{P}'' = \{p''(\hat{x}_j|x_k)\}$ a D''-admissible channel with $I''(X;\hat{X}) = R(D'')$. Construct another transition probability matrix $\mathbf{P}$ as follows:

$$\mathbf{P} = \lambda\mathbf{P}' + (1-\lambda)\mathbf{P}''$$

or equivalently:

$$p(\hat{x}_j|x_k) = \lambda p'(\hat{x}_j|x_k) + (1-\lambda)p''(\hat{x}_j|x_k)$$

This new reproduction channel $\mathbf{P}$ is then $D = \lambda D' + (1-\lambda)D''$ admissible. Therefore the rate distortion function is:

$$R(D) = \min_{\mathcal{P}_D} I(X;\hat{X}) \leq I(X;\hat{X}) \tag{4.17}$$

But the mutual information $I(X;\hat{X})$ is a convex $\cup$, function over the convex set of transition matrices $\{\mathbf{P}\}$, that is:

$$I(X;\hat{X}) \leq \lambda I'(X;\hat{X}) + (1-\lambda)I''(X;\hat{X})$$

and therefore;

$$\begin{aligned} R(D) &= R(\lambda D' + (1-\lambda)D'') && (4.18)\\ R(D) &\leq I(X;\hat{X}) && (4.19)\\ R(D) &\leq \lambda I'(X;\hat{X}) + (1-\lambda)I''(X;\hat{X}) && (4.20)\\ R(D) &\leq \lambda R(D') + (1-\lambda)R(D'') && (4.21) \end{aligned}$$

The rate distortion function $R(D)$ is a convex $\cup$ function, and a decreasing function over the convex set of distortion criterion D, from d_{min} to d_{max}.

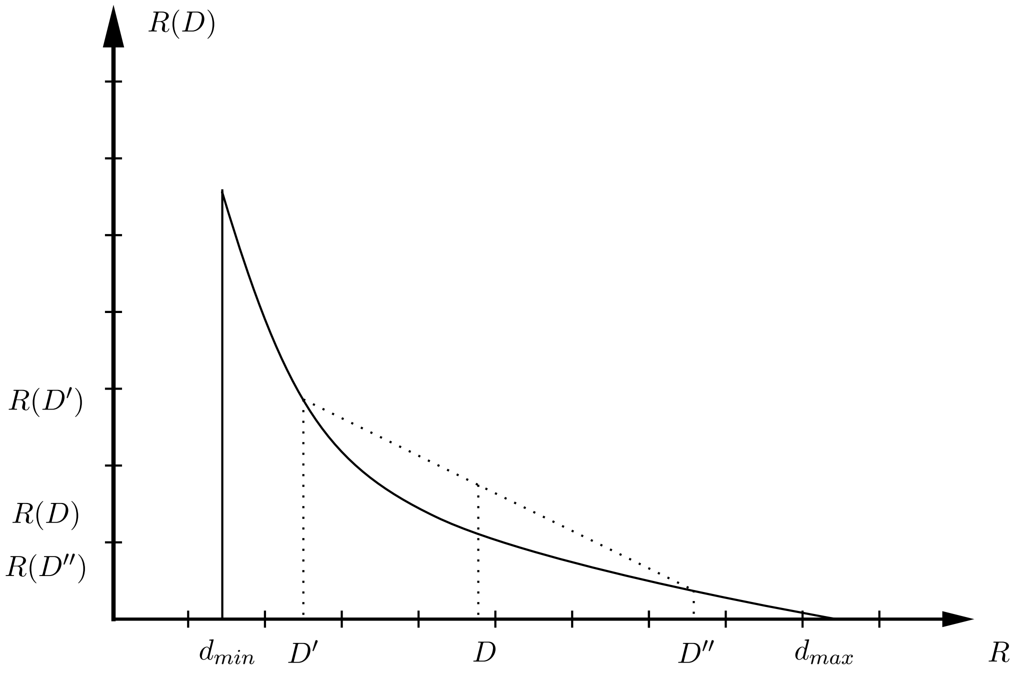


Figure 4.1: Convexity of the rate distortion function $R(D)$.

QED

Example *(Rate distortion function):*

In general, the computation of the rate distortion function requires an iterative algorithm such as the Blahut-Arimoto algorithm. For this example, however, because of the many symmetries, it is possible to find a closed-form expression for the rate-distortion function.

Consider a ternary source of information X characterized with the input symbol distribution: $p(x_1) = p(x_2) = p(x_3) = \frac{1}{3}$, that is X is an equiprobable source. Then the source entropy $H(X)$, expressed in shannons, is given by:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{3} p(x_i) \log_2 p(x_i) \\ H(X) &= -3 \left[\left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right) \right] \\ H(X) &= 1.584 \qquad \text{(Shannons)} \end{aligned}$$

This source of information is to be compressed by a source compression code to reduce the rate needed to transmit it over a communication channel. At the receiving end, a source compression decoder reproduces, with some amount of distortion the information. As shown on Figure 4.2, the reproduction $\hat{X}$ is binary instead of ternary to reduce its maximum entropy $H(\hat{X})$ to 1 shannon instead of 1.584, hence reducing the information rate $R(D)$.

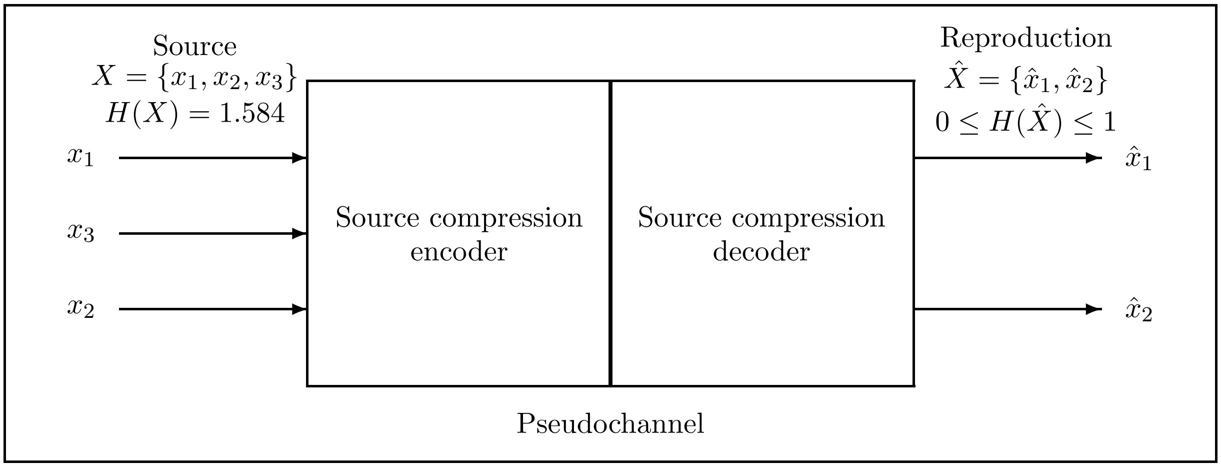

Source
$X = \{x_1, x_2, x_3\}$
$H(X) = 1.584$
x_1
x_3
x_2
Source compression encoder
Source compression decoder
Reproduction
$\hat{X} = \{\hat{x}_1, \hat{x}_2\}$
$0 \leq H(\hat{X}) \leq 1$
$\hat{x}_1$
$\hat{x}_2$
Pseudochannel

Figure 4.2: Source compression encoder and decoder, or pseudochannel, for the computation of the rate distortion function $R(D)$.

As depicted on Figure 4.3, the distortion measures associated with the mapping of the source symbols into the reproduction symbols are given in the distortion matrix **D**:

$$\mathbf{D} = \left[\begin{array}{ccc} d(x_1,\hat{x}_1) & d(x_2,\hat{x}_1) & d(x_3,\hat{x}_1) \\ d(x_1,\hat{x}_2) & d(x_2,\hat{x}_2) & d(x_3,\hat{x}_2) \end{array} \right] = \left[\begin{array}{ccc} 1 & 2 & 1 \\ 2 & 1 & 1 \end{array} \right]$$

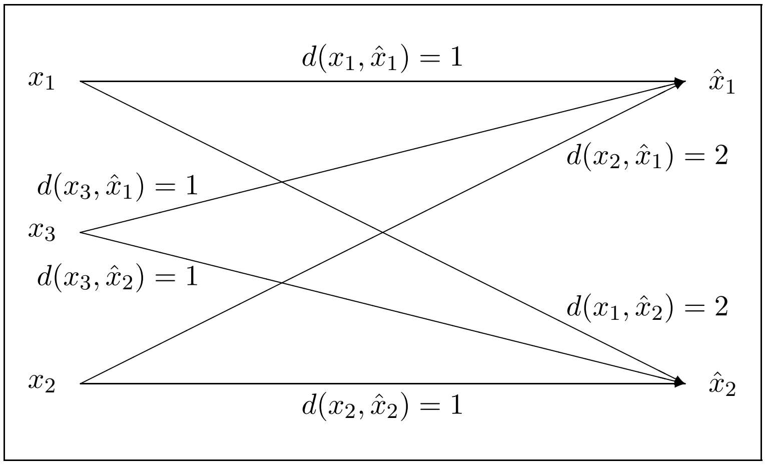


Figure 4.3: Distortion measures associated with the distortion matric **D**.

The minimum distortion d_{min} is given by:

$$\begin{aligned}
d_{min} &= \sum_{k=1}^{3} p(x_k) d(x_k,\hat{x}_{j_0}) \\
d_{min} &= p(x_1)d(x_1,\hat{x}_1) + p(x_2)d(x_2,\hat{x}_2) + p(x_3) \underbrace{d(x_3,\hat{x}_1)}_{\text{or } d(x_3,\hat{x}_2)} \\
d_{min} &= \left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times 1\right) \\
d_{min} &= 1
\end{aligned}$$

The maximum distortion d_{max} is equal to:

$$\begin{aligned}
d_{max} &= \min_{j=1,2} \sum_{k=1}^{3} p(x_k) d(x_k,\hat{x}_j) \\
d_{max} &= \min_{j=1,2} \left\{ \left[p(x_1)d(x_1,\hat{x}_1) + p(x_2)d(x_2,\hat{x}_1) + p(x_3)d(x_3,\hat{x}_1) \right], \right. \\
&\quad \left. \left[p(x_1)d(x_1,\hat{x}_2) + p(x_2)d(x_2,\hat{x}_2) + p(x_3)d(x_3,\hat{x}_2) \right] \right\} \\
d_{max} &= \min_{j=1,2} \left\{ \left[\left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times 2\right) + \left(\frac{1}{3}\times 1\right) \right], \left[\left(\frac{1}{3}\times 2\right) + \left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times 1\right) \right] \right\} \\
d_{max} &= \min_{j=1,2} \left(\frac{4}{3}, \frac{4}{3}\right) \\
d_{max} &= \frac{4}{3}
\end{aligned}$$

Therefore, the rate distortion function $R(D)$ is defined only in the distortion criterion range, that is:

$$d_{min} = 1 \leq D \leq d_{max} = \frac{4}{3}$$

We now can determine the rate distortion function as:

$$\begin{aligned} R(D) &= \min_{\mathcal{P}_D} I(X;\hat{X}) \\ R(D) &= \min_{\mathcal{P}_D} \sum_{k=1}^{3} \sum_{j=1}^{2} p(x_k) p(\hat{x}_j|x_k) \log_b \left[\frac{p(\hat{x}_j|x_k)}{\sum_{l=1}^{3} p(x_l) p(\hat{x}_j|x_l)} \right] \end{aligned}$$

where the set of admissible channels, $\mathcal{P}_D$, for a given allowable amount of distortion D is given by:

$$\mathcal{P}_D \equiv \left\{ \mathbf{P} : \sum_{k=1}^{3} \sum_{j=1}^{2} p(x_k) p(\hat{x}_j|x_k) d(x_k, \hat{x}_j) \leq D \right\}$$

We need to indicate the transition probabilities associated with the (cascaded source compression encoder and decoder) pseudochannel. Figure 4.4 shows the transition probabilities:

$$\mathbf{P} = \left[\begin{array}{ccc} p(\hat{x}_1|x_1) & p(\hat{x}_1|x_2) & p(\hat{x}_1|x_3) \\ p(\hat{x}_2|x_1) & p(\hat{x}_2|x_2) & p(\hat{x}_2|x_3) \end{array} \right] = \left[\begin{array}{ccc} 1-\alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1-\alpha_2 & 1-\alpha_3 \end{array} \right]$$

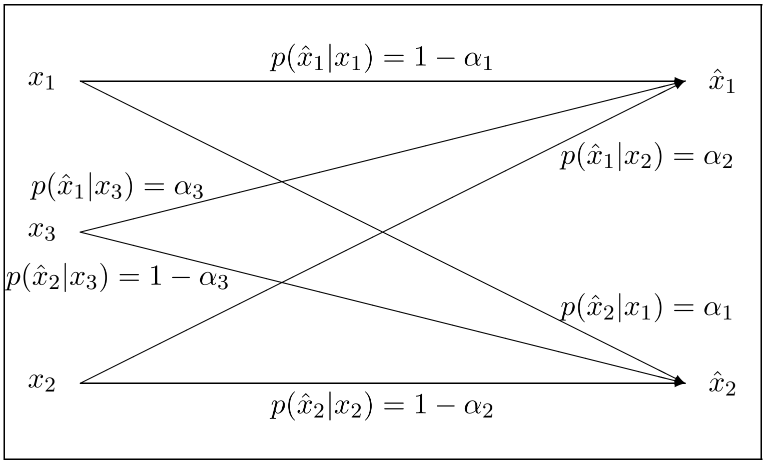

x_1
$p(\hat{x}_1|x_1) = 1-\alpha_1$
$\hat{x}_1$
$p(\hat{x}_1|x_2) = \alpha_2$
$p(\hat{x}_1|x_3) = \alpha_3$
x_3
$p(\hat{x}_2|x_3) = 1-\alpha_3$
$p(\hat{x}_2|x_1) = \alpha_1$
x_2
$\hat{x}_2$
$p(\hat{x}_2|x_2) = 1-\alpha_2$

Figure 4.4: Transition probability matrix $\mathbf{P}$ for the source compression encoder and decoder (pseudochannel).

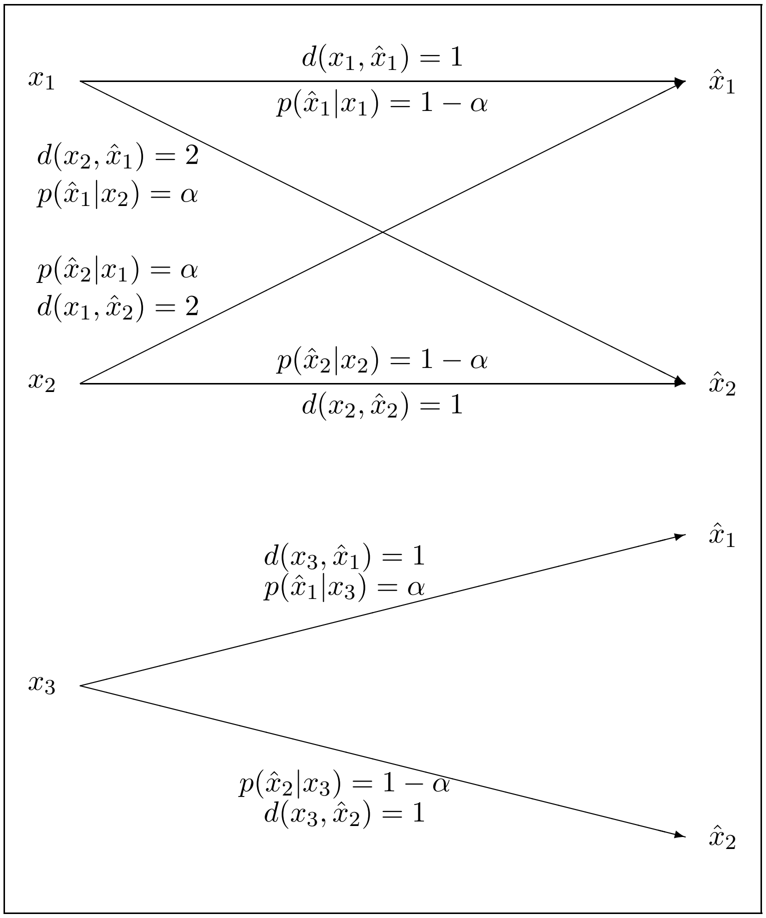
This means that if we want to find the set of D-admissible channels, we need to search the set of transition probability matrices over $0 \leq \alpha_1 \leq 1$, $0 \leq \alpha_2 \leq 1$, and $0 \leq \alpha_3 \leq 1$ and determine the range over which $d(X, \hat{X}) = \sum_{k=1}^{3} \sum_{j=1}^{2} p(x_k) p(\hat{x}_j|x_k) d(x_k, \hat{x}_j) \leq D$.

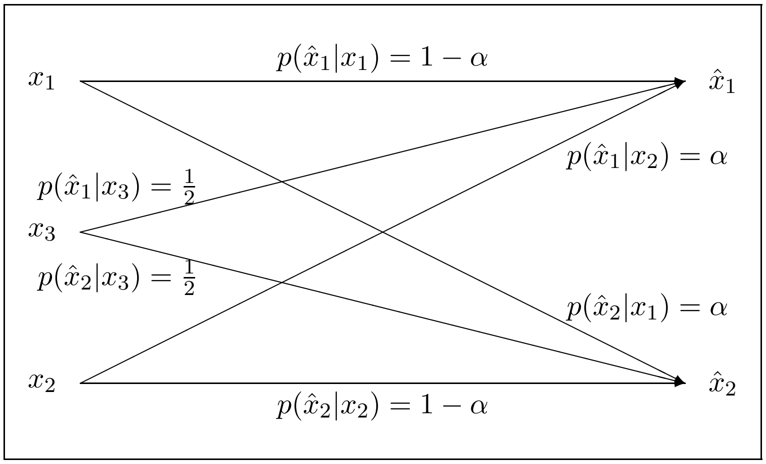
$$\begin{aligned}
d(X,\hat{X}) &= \sum_{k=1}^{3}\sum_{j=1}^{2} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j) \\
d(X,\hat{X}) &= p(x_1)\left[p(\hat{x}_1|x_1)d(x_1,\hat{x}_1)+p(\hat{x}_2|x_1)d(x_1,\hat{x}_2)\right]+ \\
&\quad p(x_2)\left[p(\hat{x}_1|x_2)d(x_2,\hat{x}_1)+p(\hat{x}_2|x_2)d(x_2,\hat{x}_2)\right]+ \\
&\quad p(x_3)\left[p(\hat{x}_1|x_3)d(x_3,\hat{x}_1)+p(\hat{x}_2|x_3)d(x_3,\hat{x}_2)\right] \\
d(X,\hat{X}) &= \frac{1}{3}\left[(1-\alpha_1)\times 1+\alpha_1\times 2\right]+\frac{1}{3}\left[\alpha_2\times 2+(1-\alpha_2)\times 1\right]+\frac{1}{3}\left[\alpha_3\times 1+(1-\alpha_3)\times 1\right]
\end{aligned}$$

Fortunately here, because of the symmetry in the distortion matrix $\mathbf{D}$ and the equiprobable source of information, we have that for x_1: if $\alpha_1 = 1$ then x_1 contributes for $\frac{1}{3}$ to the average distortion $d(X, \hat{X})$ while if $\alpha_1 = 0$ then x_1 adds $\frac{2}{3}$ to the average distortion $d(X, \hat{X})$. Similarly, for input symbol x_2, the distortion contribution are respectively: $\frac{1}{3}$ for $\alpha_2 = 0$ and $\frac{2}{3}$ for $\alpha_2 = 1$. Then both α_1 and α_2 in the expression of the transition probabilities affect equally the average distortion $d(X, \hat{X})$.

Now consider the distortion caused by the mapping of source symbol x_3 into either reproduction symbol, i.e. $\hat{x}_1$ or $\hat{x}_2$. The contribution of x_3 in the average distortion $d(X, \hat{X})$ is independent of the value of α_3. However, since we want to obtain the rate distortion function by minimizing the mutual information, we have to set α_3 to $\frac{1}{2}$ such that the contribution in the expression of the mutual information will be equal to zero:

$$\begin{aligned}
R(D) &= \min_{\mathcal{P}_D} I(X;\hat{X}) \\
R(D) &= \min_{\mathcal{P}_D}\left\{\sum_{k=1}^{3}\sum_{j=1}^{2} p(x_k)p(\hat{x}_j|x_k)\log_2\left[\frac{p(\hat{x}_j|x_k)}{\sum_{l=1}^{3}p(x_l)p(\hat{x}_j|x_l)}\right]\right\} \\
R(D) &= \min_{\mathcal{P}_D}\left\{p(x_1)\left[p(\hat{x}_1|x_1)\log_2\frac{p(\hat{x}_1|x_1)}{p(\hat{x}_1)}+p(\hat{x}_2|x_1)\log_2\frac{p(\hat{x}_2|x_1)}{p(\hat{x}_2)}\right]\right. \\
&\quad +p(x_2)\left[p(\hat{x}_1|x_2)\log_2\frac{p(\hat{x}_1|x_2)}{p(\hat{x}_1)}+p(\hat{x}_2|x_2)\log_2\frac{p(\hat{x}_2|x_2)}{p(\hat{x}_2)}\right] \\
&\quad \left.+\,p(x_3)\left[p(\hat{x}_1|x_3)\log_2\frac{p(\hat{x}_1|x_3)}{p(\hat{x}_1)}+p(\hat{x}_2|x_3)\log_2\frac{p(\hat{x}_2|x_3)}{p(\hat{x}_2)}\right]\right\} \\
R(D) &= \min_{\mathcal{P}_D}\left\{\frac{1}{3}\left[(1-\alpha_1)\log_2\frac{(1-\alpha_1)}{p(\hat{x}_1)}+\alpha_1\log_2\frac{\alpha_1}{p(\hat{x}_2)}\right]\right. \\
&\quad +\frac{1}{3}\left[\alpha_2\log_2\frac{\alpha_2}{p(\hat{x}_1)}+(1-\alpha_2)\log_2\frac{(1-\alpha_2)}{p(\hat{x}_2)}\right] \\
&\quad \left.+\,\frac{1}{3}\left[(1-\alpha_3)\log_2\frac{(1-\alpha_3)}{p(\hat{x}_1)}+\alpha_3\log_2\frac{\alpha_3}{p(\hat{x}_2)}\right]\right\}
\end{aligned}$$


x_1
$d(x_1,\hat{x}_1)=1$
$p(\hat{x}_1|x_1)=1-\alpha$
$\hat{x}_1$
$d(x_2,\hat{x}_1)=2$
$p(\hat{x}_1|x_2)=\alpha$
$p(\hat{x}_2|x_1)=\alpha$
$d(x_1,\hat{x}_2)=2$
x_2
$p(\hat{x}_2|x_2)=1-\alpha$
$d(x_2,\hat{x}_2)=1$
$\hat{x}_2$
$\hat{x}_1$
$d(x_3,\hat{x}_1)=1$
$p(\hat{x}_1|x_3)=\alpha$
x_3
$p(\hat{x}_2|x_3)=1-\alpha$
$d(x_3,\hat{x}_2)=1$
$\hat{x}_2$


x_1
$p(\hat{x}_1|x_1)=1-\alpha$
$\hat{x}_1$
$p(\hat{x}_1|x_2)=\alpha$
$p(\hat{x}_1|x_3)=\frac{1}{2}$
x_3
$p(\hat{x}_2|x_3)=\frac{1}{2}$
$p(\hat{x}_2|x_1)=\alpha$
x_2
$\hat{x}_2$
$p(\hat{x}_2|x_2)=1-\alpha$

Minimizing the last term of the mutual information $I(X;\hat{X})$, that is over the output symbol x_3 does not affect (i.e. increase or decrease) the average distortion $d(X,\hat{X})$. However this is not the case for input symbols x_1 and x_2. If we choose $p(\hat{x}_1|x_3) = (1-\alpha_3) = \alpha_3 = \frac{1}{2}$ then:

$$\log_2 \frac{p(\hat{x}_1|x_3)}{p(\hat{x}_1)} = \log_2 \frac{\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)} = 0$$

By symmetry, since x_1 and x_2 contribute to $d(X,\hat{X})$ and $I(X;\hat{X})$ in a similar manner, we can choose $\alpha_1 = \alpha_2 = \alpha$ and $\alpha_3 = \frac{1}{2}$ because $p(x_1) = p(x_2) = p(x_3) = \frac{1}{3}$, and then

$$p(\hat{x}_1) = \frac{1}{3}(1-\alpha) + \frac{1}{3}\alpha + \frac{1}{3}\frac{1}{2} = \frac{1}{2} = p(\hat{x}_2)$$

The average distortion between X and $\hat{X}$ can now be expressed as:

$$\begin{aligned}
d(X,\hat{X}) &= \sum_{k=1}^{3}\sum_{j=1}^{2} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j) \\
d(X,\hat{X}) &= \frac{1}{3}\left[(1-\alpha)\times 1 + \alpha \times 2\right] + \frac{1}{3}\left[\alpha \times 2 + (1-\alpha)\times 1\right] + \frac{1}{3}\left[\left(\frac{1}{2}\right)\times 1 + \left(\frac{1}{2}\right)\times 1\right] \\
d(X,\hat{X}) &= \frac{1}{3} - \frac{1}{3}\alpha + \frac{2}{3}\alpha + \frac{2}{3}\alpha + \frac{1}{3}\alpha - \frac{1}{3}\alpha + \frac{1}{6} + \frac{1}{6} \\
d(X,\hat{X}) &= 1 + \frac{2}{3}\alpha
\end{aligned}$$

The distortion range being $d_{min} = 1 \le D \le \frac{4}{3} = d_{max}$ and taking $d(X,\hat{X}) = D$ it follows that $0 \le \alpha \le \frac{1}{2}$, that is:

$$D = 1 + \frac{2}{3}\alpha \qquad \text{or equivalently} \qquad \alpha = \frac{3}{2}(D-1)$$

The rate $R(D)$ is also a function of the unique parameter α:

$$\begin{aligned}
R(D) &= \min_{\mathcal{P}_D} I(X;\hat{X}) \\
R(D) &= \min_{\mathcal{P}_D} \left\{ \frac{1}{3}\left[(1-\alpha_1)\log_2\frac{(1-\alpha_1)}{p(\hat{x}_1)} + \alpha_1\log_2\frac{\alpha_1}{p(\hat{x}_2)}\right] + \frac{1}{3}\left[\alpha_2\log_2\frac{\alpha_2}{p(\hat{x}_1)} + (1-\alpha_2)\log_2\frac{(1-\alpha_2)}{p(\hat{x}_2)}\right] \right. \\
&\quad \left. + \frac{1}{3}\left[(1-\alpha_3)\log_2\frac{(1-\alpha_3)}{p(\hat{x}_1)} + \alpha_3\log_2\frac{\alpha_3}{p(\hat{x}_2)}\right]\right\} \\
R(D) &= \min_{\mathcal{P}_D} \left\{ \frac{1}{3}\left[(1-\alpha)\log_2\frac{(1-\alpha)}{\left(\frac{1}{2}\right)} + \alpha\log_2\frac{\alpha}{\left(\frac{1}{2}\right)}\right] + \frac{1}{3}\left[\alpha\log_2\frac{\alpha}{\left(\frac{1}{2}\right)} + (1-\alpha)\log_2\frac{(1-\alpha)}{\left(\frac{1}{2}\right)}\right] \right. \\
&\quad \left. + \frac{1}{3}\left[\left(\frac{1}{2}\right)\log_2\frac{\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)} + \left(\frac{1}{2}\right)\log_2\frac{\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)}\right]\right\} \\
R(D) &= \min_{\mathcal{P}_D} \left\{ \frac{1}{3}\left[(1-\alpha)\log_2\frac{(1-\alpha)}{\left(\frac{1}{2}\right)} + \alpha\log_2\frac{\alpha}{\left(\frac{1}{2}\right)}\right] + \frac{1}{3}\left[\alpha\log_2\frac{\alpha}{\left(\frac{1}{2}\right)} + (1-\alpha)\log_2\frac{(1-\alpha)}{\left(\frac{1}{2}\right)}\right]\right\}
\end{aligned}$$

This can be simplified as:

$$\begin{aligned}
R(D) &= \min_{\mathcal{P}_D} I(X;\hat{X}) \\
R(D) &= \min_{\mathcal{P}_D} \frac{2}{3}(1-\alpha)\log_2 2(1-\alpha) + \frac{2}{3}\alpha \log_2 2\alpha \\
R(D) &= \min_{\mathcal{P}_D} \frac{2}{3}\left[\log_2 2 + (1-\alpha)\log_2(1-\alpha) + \alpha \log_2 \alpha\right] \\
R(D) &= \min_{\mathcal{P}_D} \frac{2}{3}\left[1 + (1-\alpha)\log_2(1-\alpha) + \alpha \log_2 \alpha\right] \\
R(D) &= \min_{\mathcal{P}_D} \frac{2}{3}\left[1 - H(\alpha)\right] \qquad \text{or} \\
R(D) &= \min_{\mathcal{P}_D} \frac{2}{3}\left[1 - H\left(\frac{3}{2}(D-1)\right)\right]
\end{aligned}$$

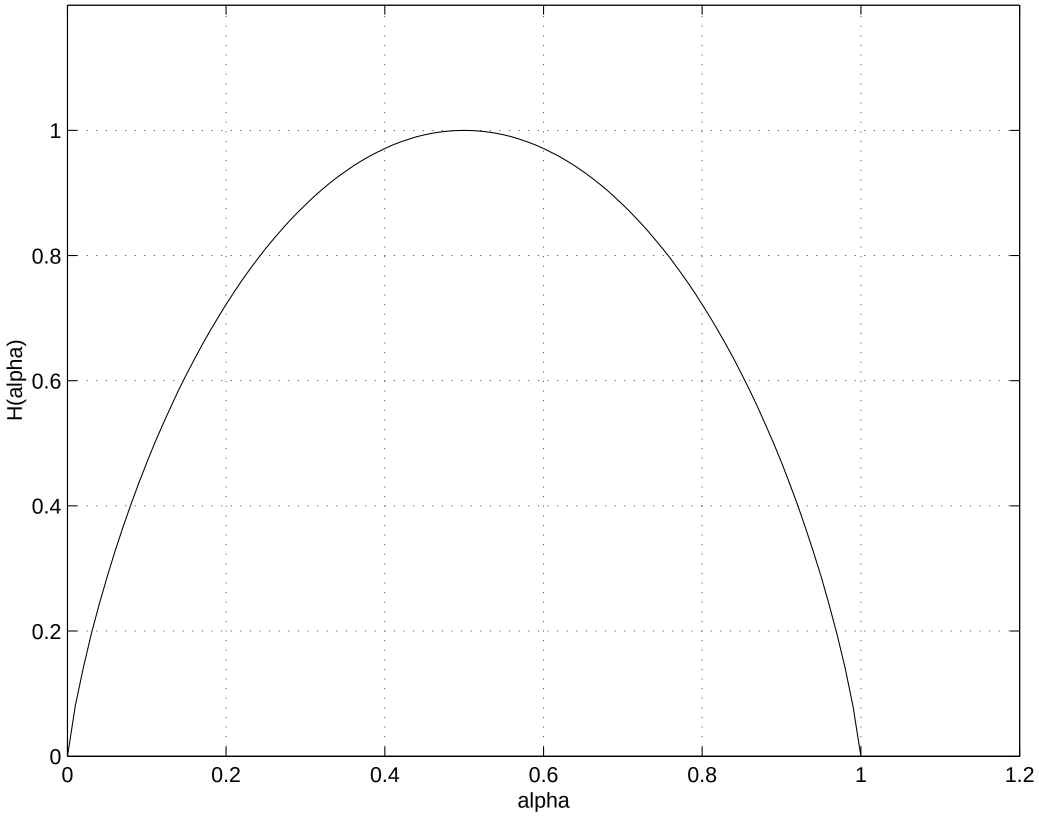


Figure 4.5: Entropy function $H(\alpha)$ of a binary source.

Then the rate distortion function $R(D)$ is a function of the parameter α and can be expressed with the parametric equations:

$$\boxed{R(D) = \frac{2}{3}\left[1 - H(\alpha)\right] \qquad \text{and} \qquad D = 1 + \frac{2}{3}\alpha \qquad \text{for } 0 \leq \alpha \leq \tfrac{1}{2}.}$$

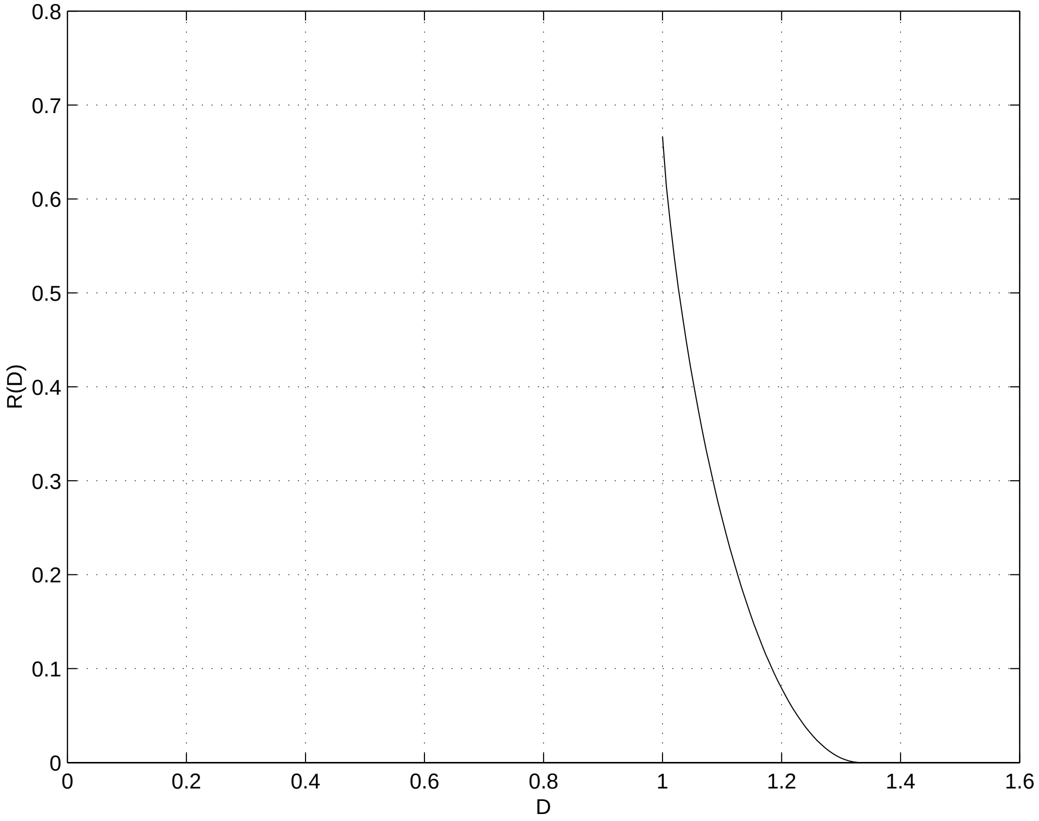


Figure 4.6: Rate distortion function $R(D)$ with $R(\alpha)$ and $D(\alpha)$, $0 \leq \alpha \leq \frac{1}{2}$.

4.2 Computation of the rate distortion function

The following algorithm (see [2]) computes the rate distortion function $R(D)$ iteratively. Also, see Berger[3] for an in-depth treatment of the rate distortion function.

Step 1: Initialization:

- **1.1:** Set the rate-distortion function $R(D)$ slope s, $-\infty < s < 0$.
- **1.2:** Set the initial reproduction set distribution $\mathbf{p}^{(0)}$, where $\mathbf{p}^{(0)} = \{p^{(0)}(\hat{x}_j)\}$. Choose $p^{(0)}(\hat{x}_j) \neq 0$ for $j = 1, \ldots, J$ and, obviously $\sum_{j=1}^{J} p^0(\hat{x}_j) = 1$.
- **1.3:** Compute the matrix $\mathbf{A} = \{A_{k,j}\}$ from the distortion matrix:

$$A_{k,j} = b^{sd(x_k,\hat{x}_j)} \qquad \text{for } k \in [1, \ldots, K] \text{ and } j \in [1, \ldots, J]$$

Step 2: Iterative algorithm:

- **2.1:** Compute the J coefficients $\{c_j\}$, for $1 \leq j \leq J$, as:

$$c_j = \sum_{k=1}^{K} \frac{p(x_k) A_{k,j}}{\sum_{l=1}^{J} p^{(r)}(\hat{x}_l) A_{k,l}}$$

- **2.2:** Update the reproduction set distribution $\mathbf{p}^{(r+1)}$:

$$p^{(r+1)}(\hat{x}_j) = c_j p^{(r)}(\hat{x}_j) \qquad \text{for } j \in [1, \ldots, J]$$

- **2.3:** Compute the "lower value" T_L as:

$$T_L = \sum_{j=1}^{J} p^{(r+1)}(\hat{x}_j) \log_b c_j$$

- **2.4:** Compute the "upper value" T_U as:

$$T_U = \max_{j=1}^{J} \log_b c_j$$

Step 3: Test if the difference between T_U and T_L is smaller than a fixed tolerance ϵ:

$$\text{If} \qquad T_U - T_L \geq \epsilon \qquad \text{then go back to } \textbf{Step 2}\text{, else continue}$$

[2] Blahut, R. E., "Principles and Practice of Information Theory", Addison-Wesley, Reading, Massachusetts, 1987.

[3] Berger, T., "Rate Distortion Theory", Prentice-Hall, Englewood Cliffs, N.J., 1971.

Step 4: Distortion and rate computation:

4.1: Compute the transition probabilities of the composite source compression encoder (or reproduction) channel $\mathbf{P}^{(r+1)} = \{p^{(r+1)}(\hat{x}_j|x_k)\}$, for $k \in [1, \ldots, K]$ and $j \in [1, \ldots, J]$:

$$p^{(r+1)}(\hat{x}_j|x_k) = \frac{p^{(r+1)}(\hat{x}_j)A_{k,j}}{\sum_{l=1}^{J} p^{(r+1)}(\hat{x}_l)A_{k,l}}$$

4.2: Compute the distortion D:

$$D = \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p^{(r+1)}(\hat{x}_j|x_k)d(x_k, \hat{x}_j)$$

4.3: Compute the rate distortion function $R(D)$:

$$R(D) = sD - \sum_{k=1}^{K} p(x_k)\log_b \sum_{j=1}^{J} p^{(r)}(\hat{x}_j)A_{k,j} - \sum_{j=1}^{J} p^{(r+1)}(\hat{x}_j)\log_b c_j$$

Step 5: Program termination:

5.1: Change the value of s.

5.2: Go back to **Step 1** or halt the program.

A flowchart of the above iterative algorithm [Bla87] is is shown on Figure 4.7.

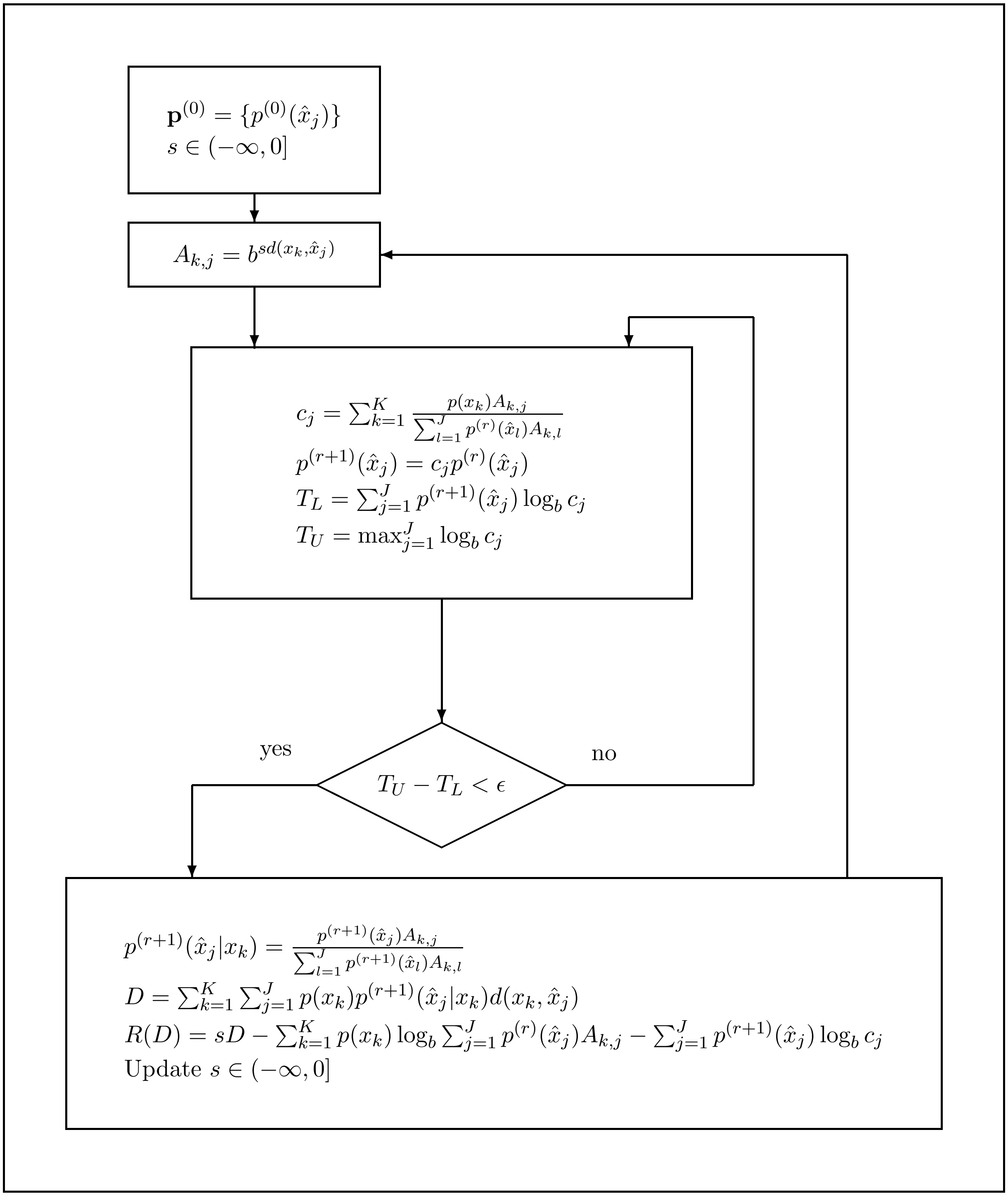

$\mathbf{p}^{(0)} = \{p^{(0)}(\hat{x}_j)\}$
$s \in (-\infty, 0]$
$A_{k,j} = b^{sd(x_k,\hat{x}_j)}$
$c_j = \sum_{k=1}^{K} \frac{p(x_k)A_{k,j}}{\sum_{l=1}^{J} p^{(r)}(\hat{x}_l)A_{k,l}}$
$p^{(r+1)}(\hat{x}_j) = c_j p^{(r)}(\hat{x}_j)$
$T_L = \sum_{j=1}^{J} p^{(r+1)}(\hat{x}_j)\log_b c_j$
$T_U = \max_{j=1}^{J} \log_b c_j$
yes
$T_U - T_L < \epsilon$
no
$p^{(r+1)}(\hat{x}_j|x_k) = \frac{p^{(r+1)}(\hat{x}_j)A_{k,j}}{\sum_{l=1}^{J} p^{(r+1)}(\hat{x}_l)A_{k,l}}$
$D = \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p^{(r+1)}(\hat{x}_j|x_k)d(x_k,\hat{x}_j)$
$R(D) = sD - \sum_{k=1}^{K} p(x_k)\log_b \sum_{j=1}^{J} p^{(r)}(\hat{x}_j)A_{k,j} - \sum_{j=1}^{J} p^{(r+1)}(\hat{x}_j)\log_b c_j$
Update $s \in (-\infty, 0]$

Figure 4.7: Iterative algorithm for computing the rate distortion function $R(D)$ (from by *"Principles and Practice of Information Theory"* Richard E. Blahut).

4.3 Rate distortion theorem

Definition *(Distortion jointly typical pair of sequences):*

Given a memoryless pair of random variables $(X;\hat{X})$ with a joint probability distribution $\{p(\mathbf{x},\hat{\mathbf{x}})\}$ and a joint entropy $H(\mathbf{X},\hat{\mathbf{X}})$, the set of ϵ-distortion jointly typical pairs of sequences $\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)$ of blocklength N are the pairs $(\mathbf{x},\hat{\mathbf{x}})$ defined as:

$$\begin{aligned}
\mathcal{T}_{\mathbf{X}}(\epsilon) &\equiv \left\{\mathbf{x} \text{ such that: } \left|-\frac{1}{N}\log_b p(\mathbf{x}) - H(\mathbf{X})\right| < \epsilon\right\} \\
\mathcal{T}_{\hat{\mathbf{X}}}(\epsilon) &\equiv \left\{\hat{\mathbf{x}} \text{ such that: } \left|-\frac{1}{N}\log_b p(\hat{\mathbf{x}}) - H(\hat{\mathbf{X}})\right| < \epsilon\right\} \\
\mathcal{T}_{\mathbf{X}\hat{\mathbf{X}}}(\epsilon) &\equiv \left\{(\mathbf{x},\hat{\mathbf{x}}) \text{ such that: } \left|-\frac{1}{N}\log_b p(\mathbf{x},\hat{\mathbf{x}}) - H(\mathbf{X},\hat{\mathbf{X}})\right| < \epsilon\right\} \\
\mathcal{T}_{\mathbf{X}\hat{\mathbf{X}}d}(\epsilon) &\equiv \left\{(\mathbf{x},\hat{\mathbf{x}}) \text{ such that: } \left|-d(\mathbf{x},\hat{\mathbf{x}}) - E\left[d(\mathbf{X},\hat{\mathbf{X}})\right]\right| < \epsilon\right\}
\end{aligned}$$

4.3.1 Shannon's rate distortion (third) theorem

Theorem *(Rate distortion theorem):*

For an independent identically distributed (i.i.d.) source $\mathbf{X}$, with distribution $\mathbf{p} = \{p(\mathbf{x})\}$ and bounded distortion measure $d(\mathbf{x},\hat{\mathbf{x}})$, it is possible to find a source compression code $\mathcal{C}$ of rate R such that the average distortion per symbol is less than a distortion criterion $D+\epsilon_1$, provided that the code rate R is larger than the rate distortion function $R(D) = \min_{\mathcal{P}_D} I(X;\hat{X})$ and the blocklength N is sufficiently large ($N \geq N_0$):

$$\boxed{R > R(D) + \epsilon_2}$$

where $\mathcal{P}_D = \left\{\mathbf{P} : \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_k)p(\hat{x}_j|x_k)d(x_k,\hat{x}_j) \leq D\right\}$.

Proof:

The expected distortion $E\left[d(\mathbf{X},\hat{\mathbf{X}})\right]$ over all codes and all transmitted vectors $\{\mathbf{X}\}$ of blocklength N is:

$$\begin{aligned}
E\left[d(\mathbf{X},\hat{\mathbf{X}})\right] &= \sum_{(\mathbf{x},\hat{\mathbf{x}})} p(\mathbf{x},\hat{\mathbf{x}})d(\mathbf{x},\hat{\mathbf{x}}) && (4.22)\\
E\left[d(\mathbf{X},\hat{\mathbf{X}})\right] &= \underbrace{\sum_{(\mathbf{x},\hat{\mathbf{x}})\in\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)} p(\mathbf{x},\hat{\mathbf{x}})}_{\leq 1}\underbrace{d(\mathbf{x},\hat{\mathbf{x}})}_{\leq D+\epsilon} && (4.23)\\
&+ \sum_{(\mathbf{x},\hat{\mathbf{x}})\notin\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)} p(\mathbf{x},\hat{\mathbf{x}})\underbrace{d(\mathbf{x},\hat{\mathbf{x}})}_{\leq d_{max}}\\
E\left[d(\mathbf{X},\hat{\mathbf{X}})\right] &\leq D+\epsilon+P_e d_{max} && (4.24)
\end{aligned}$$

P_e is the probability that there does not exist a sequence $\hat{\mathbf{x}}$ which is ϵ-distortion typical with any of the possible input sequences $\{\mathbf{x}\}$ of length N (random coding argument).

Let the indicator function $\phi(\mathbf{x},\hat{\mathbf{x}})$ be defined as:

$$\phi(\mathbf{x},\hat{\mathbf{x}}) = \begin{cases} 1 & \text{if } (\mathbf{x},\hat{\mathbf{x}})\in\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)\\ 0 & \text{if } (\mathbf{x},\hat{\mathbf{x}})\notin\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)\end{cases}$$

The probability P_e can be expressed as the sum over all codebooks, or codes $\{\mathcal{C}\}$, of the probability of a non-ϵ-distortion typical sequence in a given codebook $\mathcal{C}$:

$$P_e = \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \sum_{\mathbf{x}\notin J(\mathcal{C})} p(\mathbf{x}) \tag{4.25}$$

where $J(\mathcal{C})$ is the set of source sequences that have, at least, one codeword $\hat{\mathbf{x}}$ which is ϵ-distortion typical with $\mathbf{x}$. Consider a single randomly chosen codeword $\mathbf{x}$:

$$Pr\left[(\mathbf{x},\hat{\mathbf{x}})\notin\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)\right] = 1 - \underbrace{\sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}})\phi(\mathbf{x},\hat{\mathbf{x}})}_{Pr\left[(\mathbf{x},\hat{\mathbf{x}})\in\mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)\right]} \tag{4.26}$$

where $\hat{\mathbf{x}}$ is the reproduction sequence.

$$\begin{aligned}
P_e &= \sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C}) \sum_{\mathbf{x}\notin J(\mathcal{C})} p(\mathbf{x}) && (4.27)\\
P_e &= \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathcal{S}_\mathcal{C}:\mathbf{x}\notin J(\mathcal{C})} Pr(\mathcal{C}) && (4.28)
\end{aligned}$$

In the first equation, P_e is the probability of occurrence of all sequences not represented by a codeword with a fidelity criterion D, averaged over the set of all codes $\mathcal{S}_\mathcal{C}$. Th second equation indicates the probability of choosing a code $\mathcal{C}$ that does not *well* represent the input sequence $\mathbf{x}$, averaged over all input sequences $\{\mathbf{x}\}$. The error probability P_e can thus be written as:

$$\begin{aligned} P_e &= \sum_{\mathbf{x}} p(\mathbf{x}) \left[1 - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}) \phi(\mathbf{x}, \hat{\mathbf{x}})\right]^{M} && (4.29) \\ P_e &= \sum_{\mathbf{x}} p(\mathbf{x}) \left[1 - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}) \phi(\mathbf{x}, \hat{\mathbf{x}})\right]^{2^{NR}} && (4.30) \end{aligned}$$

since there are $M = 2^{NR}$ independently chosen codewords $\{\hat{\mathbf{x}}\}$.

Consider the probability $p(\hat{\mathbf{x}})$. If $(\mathbf{x}, \hat{\mathbf{x}}) \in \mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)$ then, by definition of ϵ-distortion typical pair of sequences:

$$\begin{aligned} 2^{-N[H(X)+\epsilon]} \leq &\quad p(\mathbf{x}) \quad \leq 2^{-N[H(X)-\epsilon]} && (4.31) \\ 2^{-N\left[H(\hat{X})+\epsilon\right]} \leq &\quad p(\hat{\mathbf{x}}) \quad \leq 2^{-N\left[H(\hat{X})-\epsilon\right]} && (4.32) \\ 2^{-N\left[H(X,\hat{X})+\epsilon\right]} \leq &\quad p(\mathbf{x}, \hat{\mathbf{x}}) \quad \leq 2^{-N\left[H(X,\hat{X})-\epsilon\right]} && (4.33) \end{aligned}$$

The conditional probability $p\left(\hat{\mathbf{x}}|\mathbf{x}\right)$ can be expressed as:

$$\begin{aligned} p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &= \frac{p(\mathbf{x}, \hat{\mathbf{x}})}{p(\mathbf{x})} && (4.34) \\ p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &= \frac{p(\hat{\mathbf{x}})}{p(\hat{\mathbf{x}})} \frac{p(\mathbf{x}, \hat{\mathbf{x}})}{p(\mathbf{x})} && (4.35) \\ p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &\leq p(\hat{\mathbf{x}}) \frac{2^{-N\left[H(X,\hat{X})-\epsilon\right]}}{2^{-N\left[H(\hat{X})+\epsilon\right]}\ 2^{-N[H(X)+\epsilon]}} && (4.36) \\ p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &\leq p(\hat{\mathbf{x}})\ 2^{-N\left[H(X,\hat{X})-\epsilon-H(\hat{X})-\epsilon-H(X)-\epsilon\right]} && (4.37) \\ p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &\leq p(\hat{\mathbf{x}})\ 2^{-N\left[-I(X;\hat{X})-3\epsilon\right]} && (4.38) \\ p\left(\hat{\mathbf{x}}|\mathbf{x}\right) &\leq p(\hat{\mathbf{x}})\ 2^{N\left[I(X;\hat{X})+3\epsilon\right]} && (4.39) \end{aligned}$$

This implies that:

$$\begin{aligned} p(\hat{\mathbf{x}}) \geq p\left(\hat{\mathbf{x}}|\mathbf{x}\right)\ 2^{-N\left[I(X;\hat{X})+3\epsilon\right]} &\qquad \text{and therefore} && (4.40) \\ \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}})\phi(\mathbf{x}, \hat{\mathbf{x}}) \geq \sum_{\hat{\mathbf{x}}} p\left(\hat{\mathbf{x}}|\mathbf{x}\right)\ 2^{-N\left[I(X;\hat{X})+3\epsilon\right]}\phi(\mathbf{x}, \hat{\mathbf{x}}) & && (4.41) \end{aligned}$$

Then P_e becomes:

$$P_e \leq \sum_{\mathbf{x}} p(\mathbf{x}) \left[1 - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, 2^{-N\left[I(\mathbf{X};\hat{\mathbf{X}})+3\epsilon\right]} \phi(\mathbf{x}, \hat{\mathbf{x}})\right]^{2^{NR}} \tag{4.42}$$

Let α be defined as $2^{-N\left[I(\mathbf{X};\hat{\mathbf{X}})+3\epsilon\right]}$ and β be $\sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, \phi(\mathbf{x}, \hat{\mathbf{x}})$. It can be shown that:

$$(1-\alpha\beta)^M \leq 1 - \beta + e^{-\alpha M}$$

Note that the product: $0 \leq \alpha\beta \leq 1$, since $0 \leq \alpha = 2^{-N\left[I(\mathbf{X};\hat{\mathbf{X}})+3\epsilon\right]} \leq 1$ and $0 \leq \beta = \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) \leq 1$.

$$\begin{aligned} (1-\alpha\beta)^M &= e^{\ln(1-\alpha\beta)^M} \\ (1-\alpha\beta)^M &= e^{M\ln(1-\alpha\beta)} \end{aligned}$$

but we know that, for $x > 0$, $\ln x \leq (x-1)$, which implies, for the range of interest of the product $\alpha\beta$, that:

$$\ln(1-\alpha\beta) \leq (1-\alpha\beta) - 1 = -\alpha\beta$$

$$\begin{aligned} (1-\alpha\beta)^M &= e^{M\ln(1-\alpha\beta)} \\ (1-\alpha\beta)^M &\leq e^{M(-\alpha\beta)} \\ (1-\alpha\beta)^M &\leq e^{-M\alpha\beta} \end{aligned}$$

Furthermore, as shown on Figure 4.8,

$$(1-\alpha\beta)^M \leq e^{-M\alpha\beta} \leq 1 - \beta + e^{-\alpha M} \qquad \text{for } 0 \leq \beta \leq 1$$

For $\beta = 0$ and $\beta = 1$:

$$\begin{aligned} e^{-M\alpha\beta} &= e^{-M\alpha 0} = 1 \leq 1 + e^{-\alpha M} && (\text{for } \beta = 0) \\ e^{-M\alpha\beta} &= e^{-M\alpha} \leq 1 - 1 + e^{-\alpha M} = e^{-\alpha M} && (\text{for } \beta = 1) \end{aligned}$$

Therefore:

$$\begin{aligned} P_e &\leq \sum_{\mathbf{x}} p(\mathbf{x}) \left[1 - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, 2^{-N\left[I(\mathbf{X};\hat{\mathbf{X}})+3\epsilon\right]} \phi(\mathbf{x}, \hat{\mathbf{x}})\right]^{2^{NR}} && (4.43) \\ P_e &\leq \sum_{\mathbf{x}} p(\mathbf{x}) \left\{1 - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) + e^{-\left[2^{-N\left[I(\mathbf{X};\hat{\mathbf{X}})+3\epsilon\right]} 2^{NR}\right]}\right\} && (4.44) \\ P_e &\leq 1 + e^{-2^{N\left[R - I(\mathbf{X};\hat{\mathbf{X}})-3\epsilon\right]}} - \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\mathbf{x})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) && (4.45) \end{aligned}$$

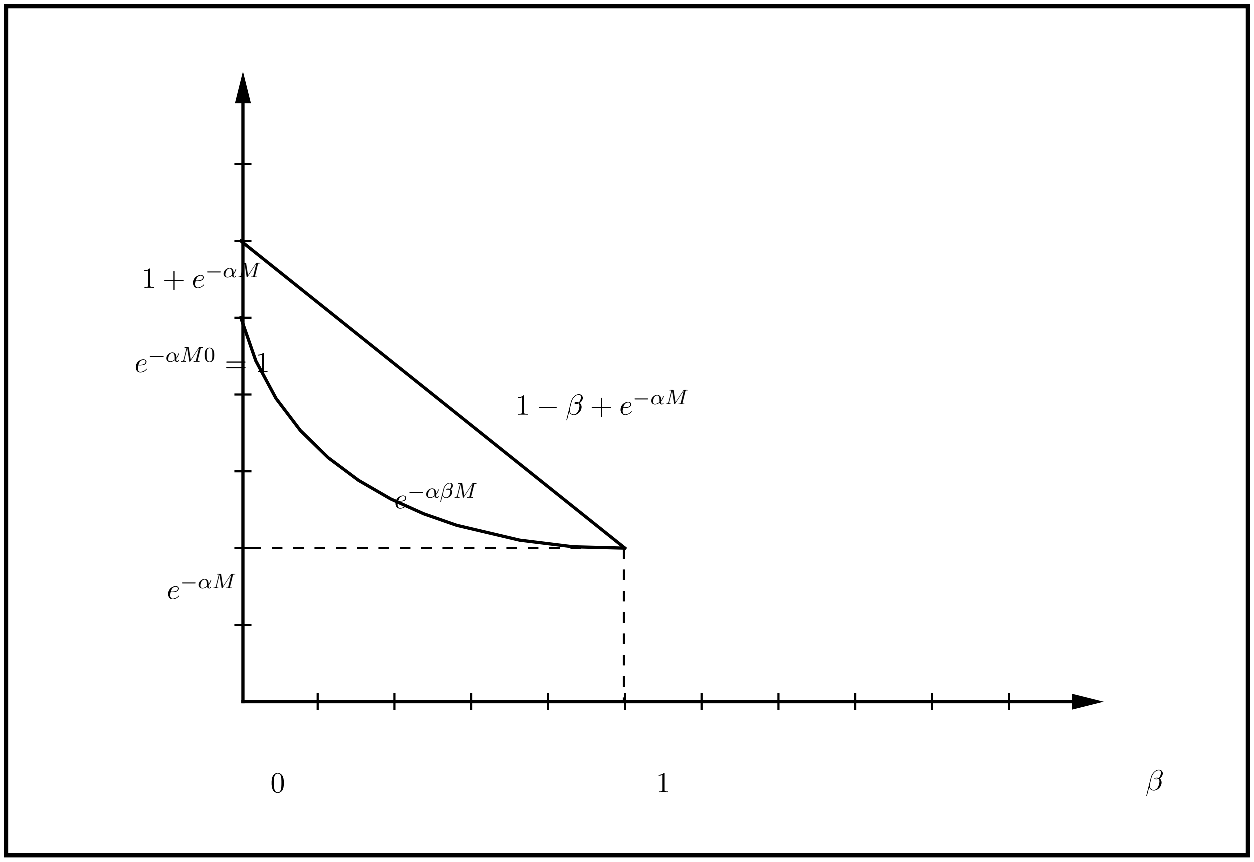


Figure 4.8: Illustration of the inequality $(1-\alpha\beta)^M \leq 1 - \beta + e^{-\alpha M}$.

Since the joint probability $p(\mathbf{x}, \hat{\mathbf{x}}) = p(\mathbf{x})\, p(\hat{\mathbf{x}}|\mathbf{x})$;

$$P_e \leq 1 - \sum_{\mathbf{x}} \sum_{\hat{\mathbf{x}}} p(\mathbf{x}, \hat{\mathbf{x}})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) + e^{-2^{N\left[R - I(\mathbf{X};\hat{\mathbf{X}}) - 3\epsilon\right]}} \tag{4.46}$$

where

$$\begin{aligned} 1 - \sum_{\mathbf{x}} \sum_{\hat{\mathbf{x}}} p(\mathbf{x}, \hat{\mathbf{x}})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) &= Pr\left[\left(\mathbf{X}, \hat{\mathbf{X}}\right) \notin \mathcal{T}_{\mathbf{X},\hat{\mathbf{X}}d}(\epsilon)\right] \\ 1 - \sum_{\mathbf{x}} \sum_{\hat{\mathbf{x}}} p(\mathbf{x}, \hat{\mathbf{x}})\, \phi(\mathbf{x}, \hat{\mathbf{x}}) &< \epsilon_1 \end{aligned}$$

by definition of a ϵ-typical pair of sequences. Then,

$$P_e \leq \epsilon_1 + \underbrace{e^{-2^{N\left[R - I(\mathbf{X};\hat{\mathbf{X}}) - 3\epsilon\right]}}}_{\epsilon_2} = \epsilon' \tag{4.47}$$

The term ϵ_2 goes to 0, as N increases to infinity, if the exponent of e, i.e. $-2^{N\left[R - I(\mathbf{X};\hat{\mathbf{X}}) - 3\epsilon\right]}$

tends towards $-\infty$, or equivalently, if the exponent of e^{-2}, i.e. $R - I(\mathbf{X};\hat{\mathbf{X}}) - 3\epsilon > 0$. In other words, if the rate R is greater than $I(\mathbf{X};\hat{\mathbf{X}}) + 3\epsilon$ (where ϵ can be made arbitrarily small). The above is true for any D-admissible channel.

Choosing the channel $\mathbf{P} = \{p\,(\hat{\mathbf{x}}|\mathbf{x})\}$ such as to minimize the mutual information $I(\mathbf{X};\hat{\mathbf{X}})$, leads to the inequality:

$$R > R(D) + 3\epsilon$$

Then, there exists at least one code $\mathcal{C}^*$ such that:

a) the average distortion $d(\mathbf{X},\hat{\mathbf{X}})$ is upperbounded as:

$$d(\mathbf{X},\hat{\mathbf{X}}) < D + \epsilon$$

b) the code rate $R(D)$ is lowerbounded by:

$$R > R(D) + \epsilon'$$

QED

4.3.2 Information transmission theorem

Shannon's channel coding theorem states that the transmission of information over a noisy channel can be as *reliable* as one desires, as long as the error control code rate, i.e. R_1, is smaller than the channel's capacity C. On the other hand, the rate distortion theorem states that it is possible to find a source compression code for which the average distortion (fidelity criterion) is arbitrarily close to a predetermined fidelity criterion D, provided that the rate R_2 of the source compression code is greater than the value of the rate distortion function $R(D)$ at the expected distortion level.

Figure 4.9 illustrates a communication link with source compression coding as well as channel coding. For both source compression and channel coding theorems, the codeword blocklength N should be sufficiently large ($N \geq N_0$).

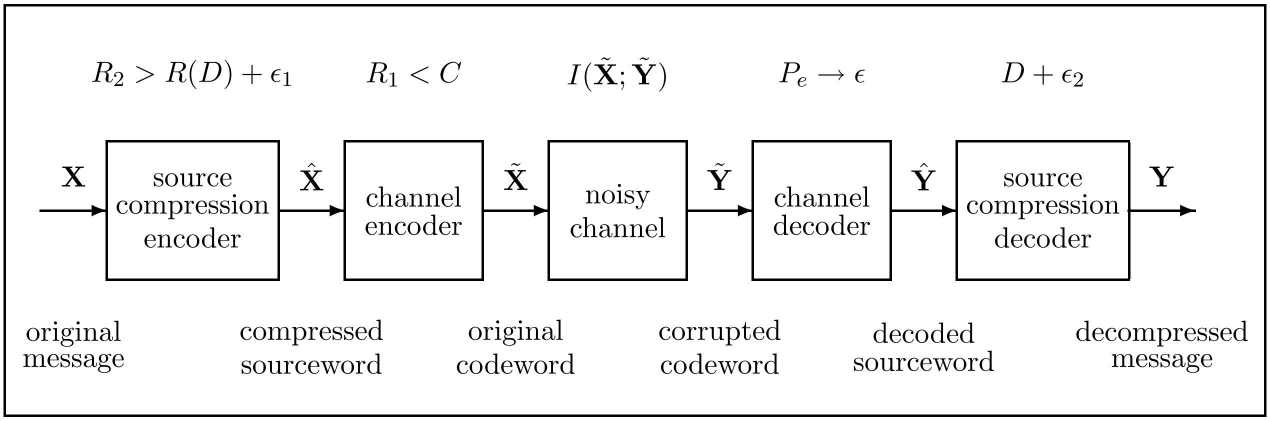


Figure 4.9: Illustration of the information transmission theorem.

The information transmission theorem combines these two theorems:

Theorem *(Information transmission theorem):*

The output sequence of a discrete memoryless source, obtained by source compression with a rate distortion function $R(D)$, can be reproduced with at most D average distortion at the output of any discrete memoryless channel having a capacity C, provided that:

$$\boxed{R(D) < C}$$

if the blocklength N is sufficiently large (i.e., $N \geq N_0$).

4.4 Problems

Problem 4.1: Consider a binary source with input distribution $\mathbf{p} = \{p(x_1) = \alpha,\ p(x_2) = 1 - \alpha\}$, where $\alpha \in [0, 0.5]$. The distortion matrix $\mathbf{d}$ is given by (i.e. error probability distortion matrix):

$$\mathbf{d} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Determine the expression of the rate distortion function $R(D)$ as a function of α and D. Draw the $R(D)$ as a function of D for $\alpha = 0.1$, 0.2, 0.3, 0.4 and 0.5.

Problem 4.2: A binary equiprobable memoryless source X generates 4.8 *kbits/s*. A source compression encoder, with a ternary reproduction alphabet $\hat{X}$, is used to compress the data prior transmission. The distortion matrix $\mathbf{d}$ is given by (note that an infinite distortion $d(x_k, \hat{x}_j) = \infty$ indicates that there is no transition from x_k to $\hat{x}_j$):

$$\mathbf{d} = \begin{pmatrix} 0 & \infty & 1 \\ \infty & 0 & 1 \end{pmatrix}$$

a) Express the probability transition matrix $\mathbf{P}$ as a function of the distortion D (the average per-symbol distortion $d(X, \hat{X})$ is used as the distortion criterion D).

b) Compute and draw the rate-distortion function $R(D)$.

c) Determine the minimum code rate R at which the information can be transmitted if the distortion criterion is to be kept at $D \leq 20\%$? What is the corresponding information transfer rate, expressed in *kbits/s*?

d) Find a simple source compression encoding scheme that achieves any desired rate R at the distortion level D determined from $R = R(D)$.

Problem 4.3: A memoryless source X generates bits with the input symbol distribution: $p(x_1) = \frac{3}{4}$ and $p(x_2) = \frac{1}{4}$. A source compression encoder, with a binary reproduction alphabet $\hat{X}$, is used to compress the data prior transmission. The distortion matrix $\mathbf{d}$ is given by:

$$\mathbf{d} = \begin{bmatrix} d(x_1, \hat{x}_1) & d(x_2, \hat{x}_1) \\ d(x_1, \hat{x}_2) & d(x_2, \hat{x}_2) \end{bmatrix} = \begin{bmatrix} 0 & \infty \\ 4 & 0 \end{bmatrix}$$

where an infinite distortion, i.e., $d(x_k, \hat{x}_j) = \infty$, indicates that there is no transition from x_k to $\hat{x}_j$. The transition probability matrix $\mathbf{P}$ is given by:

$$\mathbf{P} = \begin{bmatrix} p(\hat{x}_1|x_1) & p(\hat{x}_1|x_2) \\ p(\hat{x}_2|x_1) & p(\hat{x}_2|x_2) \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ (1-\alpha) & 1 \end{bmatrix}$$

a) Find the source entropy $H(X)$.

b) Give the reproduction symbol probabilities as a function of α.

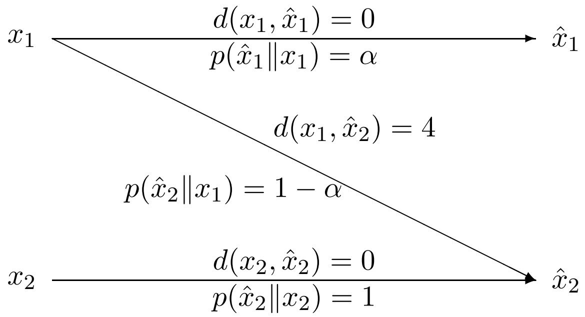
c) Find the minimum d_{min} and the maximum distortion d_{max}.

d) Give the expression for the average per-symbol distortion $d(X, \hat{X})$ as a function of α.

e) Give the expression for the pseudochannel mutual information $I(X; \hat{X})$ as a function of α.

f) Compute the rate-distortion function $R(D)$.

g) Draw the rate-distortion function $R(D)$ from $(d_{min} - 2) \leq D \leq (d_{max} + 2)$.


x_1
$d(x_1, \hat{x}_1) = 0$
$p(\hat{x}_1 \| x_1) = \alpha$
$\hat{x}_1$
$d(x_1, \hat{x}_2) = 4$
$p(\hat{x}_2 \| x_1) = 1 - \alpha$
x_2
$d(x_2, \hat{x}_2) = 0$
$p(\hat{x}_2 \| x_2) = 1$
$\hat{x}_2$

Chapter 5

Multiterminal Networks and Information Theory

In this chapter, we study two fundamental multiterminal networks: the multiple access channel and the broadcast channel. The capacity region of these two types of networks will be derived.

5.1 Multiple Access Networks

5.1.1 Capacity Region of a Two-source Multiple Access Channel

In this section, we retrict ourselves to the simple case of a multiple access network having only two independent sources of information. We will use this simple case to define and derive the capacity region of a multiple access channel. Later, we will consider the more general and realistic case of a m-user network.

Theorem *(Multiple Access Channel Capacity):*

The capacity region C of a memoryless multiple access channel is the closure of the convex hull of the set of all rates R_1 and R_2 for which:

$$\begin{aligned} R_1 &\leq I(X_1; Y|X_2), \\ R_2 &\leq I(X_2; Y|X_1), \text{and} \\ R_1 + R_2 &\leq I(X_1, X_2; Y) \end{aligned}$$

for some product distribution $\{(p_1(x_{1,k}), p_2(x_{2,j})\}$ on the input pair (X_1, X_2).

Proof:

The proof of this theorem is very similar to the proof of of Shannon's channel coding theorem on the achievability of the capacity of a single channel. We will use again the random coding argument and the expectation of probability of error over an ensemble $\mathcal{S}_{\mathcal{C}} = \{\mathcal{C}\}$ of codes. Furthermore, a decoding strategy based on jointly typical pairs of sequences will be considered for deriving a lower bound on decoding error probabilities.

A simple multiple access communication network consisting of two information sources W_1 and W_2 is depicted in Figure 5.1.

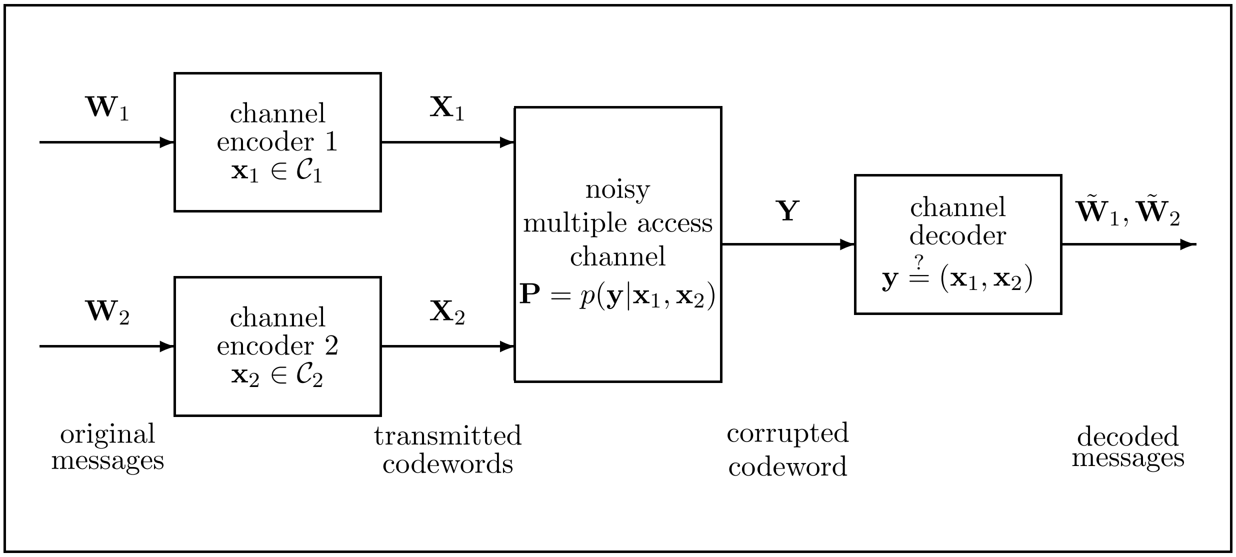

$\mathbf{W}_1$
channel encoder 1 $\mathbf{x}_1 \in \mathcal{C}_1$
$\mathbf{X}_1$
$\mathbf{W}_2$
channel encoder 2 $\mathbf{x}_2 \in \mathcal{C}_2$
$\mathbf{X}_2$
noisy multiple access channel $\mathbf{P} = p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)$
$\mathbf{Y}$
channel decoder $\mathbf{y} \overset{?}{=} (\mathbf{x}_1, \mathbf{x}_2)$
$\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2$
original messages
transmitted codewords
corrupted codeword
decoded messages

Figure 5.1: Simple two-source multiple access communication network.

We will assume that the messages from both sources are equiprobable to support the random coding argument.

$$p(w_1) = \frac{1}{M_1} = 2^{-NR_1} \qquad \text{and} \qquad p(w_2) = \frac{1}{M_2} = 2^{-NR_2} \tag{5.1}$$

where R_1 and R_2 are the code rates whereas N is the codewords' blocklength for the two distinct information sources.

We consider here a block code $\mathcal{C}$ for the two-source multiple access channel as a composite code consisting of two component codes, $\mathcal{C}_1$ and $\mathcal{C}_2$. The first component code $\mathcal{C}_1$ maps each message $W^1_{m_1}$, $m_1 = 1, \ldots, M_1$, from the first source as a unique codeword $\mathbf{c}^1_m$ of blocklength N. The codewords are assumed binary ($c^1_{m,n} \in \{0, 1\}$). Similarly, the second component code, $\mathcal{C}_2$, encodes each message from the second source, $W^2_{m_2}$, as a unique codeword $\mathbf{c}^2_m$ of blocklength N where $m_2 = 1, \ldots, M_2$:

$$
\mathcal{C}_1 = \begin{bmatrix} \mathbf{c}_1^1 \\ \vdots \\ \mathbf{c}_{m_1}^1 \\ \vdots \\ \mathbf{c}_{M_1}^1 \end{bmatrix} = \begin{bmatrix} c_{1,1}^1 & \cdots & c_{1,n_1}^1 & \cdots & c_{1,N}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m_1,1}^1 & \cdots & c_{m_1,n_1}^1 & \cdots & c_{m_1,N}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M_1,1}^1 & \cdots & c_{M_1,n_1}^1 & \cdots & c_{M_1,N}^1 \end{bmatrix}
$$

and

$$
\mathcal{C}_2 = \begin{bmatrix} \mathbf{c}_1^2 \\ \vdots \\ \mathbf{c}_{m_2}^2 \\ \vdots \\ \mathbf{c}_{M_2}^2 \end{bmatrix} = \begin{bmatrix} c_{1,1}^2 & \cdots & c_{1,n_2}^2 & \cdots & c_{1,N}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m_2,1}^2 & \cdots & c_{m_2,n_2}^2 & \cdots & c_{m_2,N}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M_2,1}^2 & \cdots & c_{M_2,n_2}^2 & \cdots & c_{M_2,N}^2 \end{bmatrix} \qquad (5.2)
$$

As mentionned previously, the decoding rule is based on the definition of jointly typical sequences. Again, this decoding rule is not optimal but its use simplify the derivation of the capacity of a multiple access network. The composite codeword $\mathbf{y}$ received from the multiple access channel is mapped into two valid (component code) codewords $(\mathbf{c}_{m_1}, \mathbf{c}_{m_2})$, if the triplet of sequences $(\mathbf{c}_{m_1}, \mathbf{c}_{m_2}, \mathbf{y})$ are jointly typical, (i.e., $(\mathbf{c}_{m_1}, \mathbf{c}_{m_2}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)$ for a given arbitrarily small offset δ. The decoded messages from the two information sources are then (w_{m_1}, w_{m_2}). There are four different types of decoding errors that may occur. These are:

- $(\mathbf{c}_{m_1}, \mathbf{c}_{m_2}, \mathbf{y}) \notin \mathcal{T}_{X_1 X_2 Y}(\delta)$ for $m_1 = 1, \ldots, M_1$ and $m_2 = 1, \ldots, M_2$
- $(\mathbf{c}_{m_1'}, \mathbf{c}_{m_2}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)$ for $m_1' \neq m_1$
- $(\mathbf{c}_{m_1}, \mathbf{c}_{m_2'}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)$ for $m_2' \neq m_2$
- $(\mathbf{c}_{m_1'}, \mathbf{c}_{m_2'}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)$ for $m_1' \neq m_1$ and $m_2' \neq m_2$

The probability of a decoding error $P_{e|m_1,m_2}$ given messages w_{m_1} and w_{m_2} is determined by the union of error events.

$$
\begin{aligned}
P_{e|m_1,m_2} \quad = \quad Pr\Bigg\{ & [(\mathbf{c}_{m_1}, \mathbf{c}_{m_2}, \mathbf{y}) \notin \mathcal{T}_{X_1 X_2 Y}(\delta)] \bigcup \bigcup_{\substack{m_1'=1 \\ m_1' \neq m_1}}^{M_1} \left[(\mathbf{c}_{m_1'}, \mathbf{c}_{m_2}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)\right] \qquad (5.3) \\
& \bigcup_{\substack{m_2'=1 \\ m_2' \neq m_2}}^{M_2} \left[(\mathbf{c}_{m_1}, \mathbf{c}_{m_2'}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)\right] \bigcup_{\substack{m_1'=1 \\ m_1' \neq m_1}}^{M_1} \bigcup_{\substack{m_2'=1 \\ m_2' \neq m_2}}^{M_2} \left[(\mathbf{c}_{m_1'}, \mathbf{c}_{m_2'}, \mathbf{y}) \in \mathcal{T}_{X_1 X_2 Y}(\delta)\right] \Bigg\}
\end{aligned}
$$

Using the union bound the error decoding probability can be rewritten as the following inequality:

$$\begin{aligned}
P_{e|m_1,m_2} \quad \leq \quad & Pr\left[(\mathbf{c}_{m_1},\mathbf{c}_{m_2},\mathbf{y}) \notin \mathcal{T}_{X_1X_2Y}(\delta)\right] + \sum_{\substack{m_1'=1\\ m_1'\neq m_1}}^{M_1} Pr\left[(\mathbf{c}_{m_1'},\mathbf{c}_{m_2},\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] \qquad (5.4)\\
& + \sum_{\substack{m_2'=1\\ m_2'\neq m_2}}^{M_2} Pr\left[(\mathbf{c}_{m_1},\mathbf{c}_{m_2'},\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] + \sum_{\substack{m_1'=1\\ m_1'\neq m_1}}^{M_1}\sum_{\substack{m_2'=1\\ m_2'\neq m_2}}^{M_2} Pr\left[(\mathbf{c}_{m_1'},\mathbf{c}_{m_2'},\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right]
\end{aligned}$$

This last expression can be written as a sum of four possible terms: ϵ_1, ϵ_2, ϵ_3, and ϵ_4. The first term, ϵ_1, can be made arbitraily small by the definition of jointly typical pairs of sequences and using a sufficiently large blocklength N. Unfortunately, the three other terms are not necessarily arbitrary small. We will show next that using random coding, the expection of each of these three terms can be made arbitrary small on the ensemble average of the codes provided that the blocklength N is large enough and that the individual rates obey thes conditions: $R_1 \leq C_1$, $R_2 \leq C_2$, and $R_1 + R_2 \leq C_1 + C_2$.

As for the case of single channels, we use the random coding construction scheme for the ensemble of codes and determine, on this ensemble of multiaccess codes $\mathcal{S}_\mathcal{C}$, the average of the error probability. The expected error decoding probability over the ensemble of randomly chosen codes is:

$$\begin{aligned}
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\left[P_e\right] \quad \leq \quad & \epsilon_1 + \sum_{\substack{m_1'=1\\ m_1'\neq m_1}}^{M_1} Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] + \sum_{\substack{m_2'=1\\ m_2'\neq m_2}}^{M_2} Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right]\\
& + \sum_{\substack{m_1'=1\\ m_1'\neq m_1}}^{M_1}\sum_{\substack{m_2'=1\\ m_2'\neq m_2}}^{M_2} Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] \qquad (5.5)
\end{aligned}$$

and since the multiaccess codewords $\mathbf{x}_1, \mathbf{x}_2$ are not a function of the received codeword indices m_1', m_2', the above expression reduces to:

$$\begin{aligned}
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\left[P_e\right] \quad \leq \quad & \epsilon_1 + (M_1-1)Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] + (M_2-1)Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right]\\
& + (M_1-1)(M_2-1)Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] \qquad (5.6)\\
\sum_{\mathcal{S}_\mathcal{C}} Pr(\mathcal{C})\left[P_e\right] \quad \leq \quad & \epsilon_1 + 2^{NR_1}Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] + 2^{NR_2}Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right]\\
& + 2^{NR_1}2^{NR_2}Pr\left[(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \in \mathcal{T}_{X_1X_2Y}(\delta)\right] \qquad (5.7)
\end{aligned}$$

$$\begin{aligned}\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \;\; &\leq \;\; \epsilon_1 + 2^{NR_1} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} p(\mathbf{x}_1)p(\mathbf{x}_2,\mathbf{y}) + 2^{NR_2} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} p(\mathbf{x}_2)p(\mathbf{x}_1,\mathbf{y}) \\ &\quad +2^{N(R_1+R_2)} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} p(\mathbf{x}_1,\mathbf{x}_2)p(\mathbf{y}) \qquad (5.8)\end{aligned}$$

From the definition and properties of jointly typical pairs of sequences we know that these probabilities are bounded as:

$$\begin{array}{ccc} p(\mathbf{x}_1) \leq 2^{-N[H(X_1)-\delta]}; & p(\mathbf{x}_2) \leq 2^{-N[H(X_2)-\delta]}; & p(\mathbf{x}_1,\mathbf{x}_2) \leq 2^{-N[H(X_1X_2)-\delta]}; \\ p(\mathbf{y}) \leq 2^{-N[H(Y)-\delta]}; & p(\mathbf{x}_1,\mathbf{y}) \leq 2^{-N[H(X_1Y)-\delta]}; & p(\mathbf{x}_2,\mathbf{y}) \leq 2^{-N[H(X_2Y)-\delta]} \end{array}$$

Therefore, the expected probability of decoding error in the multiple access case is:

$$\begin{aligned}\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \;\; &\leq \;\; \epsilon_1 + 2^{NR_1} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} 2^{-N[H(X_1)-\delta]}2^{-N[H(X_2Y)-\delta]} \\ &\quad +2^{NR_2} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} 2^{-N[H(X_2)-\delta]}2^{-N[H(X_1Y)-\delta]} \\ &\quad +2^{N(R_1+R_2)} \sum_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in\mathcal{T}_{X_1X_2Y}(\delta)} 2^{-N[H(X_1X_2)-\delta]}2^{-N[H(Y)-\delta]} \qquad (5.9)\end{aligned}$$

The maximum number of jointly typical pairs of sequences $\|\mathcal{T}_{X_1X_2Y}(\delta)\| \leq 2^{N[H(X_1X_2Y)+\delta]}$ and therefore:

$$\begin{aligned}\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \;\; &\leq \;\; \epsilon_1 + 2^{NR_1}2^{N[H(X_1X_2Y)+\delta]}2^{-N[H(X_1)-\delta]}2^{-N[H(X_2Y)-\delta]} \\ &\quad +2^{NR_2}2^{N[H(X_1X_2Y)+\delta]}2^{-N[H(X_2)-\delta]}2^{-N[H(X_1Y)-\delta]} \\ &\quad +2^{N(R_1+R_2)}2^{N[H(X_1X_2Y)+\delta]}2^{-N[H(X_1X_2)-\delta]}2^{-N[H(Y)-\delta]} \qquad (5.10)\end{aligned}$$

Using the relationships between joint entropies and equivocations, $H(X_1,X_2,Y) = H(X_2) + H(X_1|X_2) + H(Y|X_1,X_2)$ and $H(X_2,Y) = H(X_2) + H(Y|X_2)$ and considering that the messages from the two information sources are independent (i.e. $H(X_1|X_2) = H(X_1)$), we rewrite the previous equation as:

$$\begin{aligned}\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \;\; &\leq \;\; \epsilon_1 + 2^{NR_1}2^{N[H(X_2)+H(X_1|X_2)+H(Y|X_1,X_2)+\delta]}2^{-N[H(X_1)-\delta]}2^{-N[H(X_2)+H(Y|X_2)-\delta]} \\ &\quad +2^{NR_2}2^{N[H(X_2)+H(X_1|X_2)+H(Y|X_1,X_2)+\delta]}2^{-N[H(X_2)-\delta]}2^{-N[H(X_1)+H(Y|X_1)-\delta]} \\ &\quad +2^{N(R_1+R_2)}2^{N[H(X_2)+H(X_1|X_2)+H(Y|X_1,X_2)+\delta]}2^{-N[H(X_2)+H(X_1|X_2)-\delta]}2^{-N[H(Y)-\delta]} \qquad (5.11) \\ \sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \;\; &\leq \;\; \epsilon_1 + 2^{NR_1}2^{N[H(X_2)+H(X_1)+H(Y|X_1,X_2)+\delta-H(X_1)+\delta-H(X_2)-H(Y|X_2)+\delta]}\end{aligned}$$

$$
\begin{aligned}
& +2^{NR_2}2^{N[H(X_2)+H(X_1)+H(Y|X_1,X_2)+\delta-H(X_2)+\delta-H(X_1)-H(Y|X_1)+\delta]} \\
& +2^{N(R_1+R_2)}2^{N[H(X_2)+H(X_1)+H(Y|X_1,X_2)+\delta-H(X_2)-H(X_1)+\delta-H(Y)+\delta]} \qquad (5.12) \\
\sum_{\mathcal{S_C}} Pr(\mathcal{C})\,[P_e] \quad \leq \quad & \epsilon_1 + 2^{NR_1}2^{N[H(Y|X_1,X_2)-H(Y|X_2)+3\delta]} + 2^{NR_2}2^{N[H(Y|X_1,X_2)-H(Y|X_1)+3\delta]} \\
& +2^{N(R_1+R_2)}2^{N[H(Y|X_1,X_2)-H(Y)+3\delta]} \qquad (5.13)
\end{aligned}
$$

Noting that the difference in the above equivocations are (average) mutual information terms:

$$
\begin{aligned}
H(Y|X_1,X_2) - H(Y|X_2) &= -I(X_1;Y|X_2), \\
H(Y|X_1,X_2) - H(Y|X_1) &= -I(X_2;Y|X_1) \qquad \text{and} \\
H(Y|X_1,X_2) - H(Y) &= -I(X_1,X_2;Y)
\end{aligned}
$$

the expected error decoding probability over the ensemble of codes is:

$$
\begin{aligned}
\sum_{\mathcal{S_C}} Pr(\mathcal{C})\,[P_e] \quad \leq \quad & \epsilon_1 + 2^{NR_1}2^{N[-I(X_1;Y|X_2)+3\delta]} + 2^{NR_2}2^{N[-I(X_2;Y|X_1)+3\delta]} \\
& +2^{N(R_1+R_2)}2^{N[-I(X_1,X_2;Y)+3\delta]} \qquad (5.14) \\
\sum_{\mathcal{S_C}} Pr(\mathcal{C})\,[P_e] \quad \leq \quad & \epsilon_1 + 2^{-N[I(X_1;Y|X_2)-R_1-3\delta]} + 2^{-N[I(X_2;Y|X_1)-R_2-3\delta]} \\
& +2^{-N[I(X_1,X_2;Y)-(R_1+R_2)-3\delta]} \qquad (5.15) \\
\sum_{\mathcal{S_C}} Pr(\mathcal{C})\,[P_e] \quad \leq \quad & \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 \qquad (5.16)
\end{aligned}
$$

As the multiple access code blocklength N increases, the term ϵ_2 will decreases only if the rate R_1 of the first component code is smaller than $I(X_1;Y|X_2) - 3\delta$, where δ is an arbitrary small positive number. Similarly, as for the second and third sterm to vanish toward zero as N increases, the following conditions must be met: $R_2 \leq I(X_2;Y|X_1) - 3\delta$ and $(R_1 + R_2) \leq I(X_1,X_2;Y) - 3\delta$. If these three conditions are met than the average error decoding probability for the multiple access code can be as low as we wish, provided that N is sufficiently large.

Therefore, there must exist at least a multiple access code $\mathcal{C}^*$ or which the error performance is at least as good as the average over the ensemble of codes. Consequently, because the error probability $P_e \leq \epsilon = \sum_{i=1}^{4} \epsilon_i$ is arbitrarily small, then every pair of rates $(R_1 + R_2)$ which obey these 3 conditions will be in the capacity region of the multiple access channel.

$$
\boxed{P_e \leq \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 = \epsilon}
$$

QED

Example *(Capacity of a Binary Erasure Multiple Access Channel):*

In this example, the multiple access channel consists of two binary sources, X_1 and X_2, an a ternary information sink, or destination, Y. The ternary output is given as the sum of the two inputs:

$$Y = X_1 + X_2$$

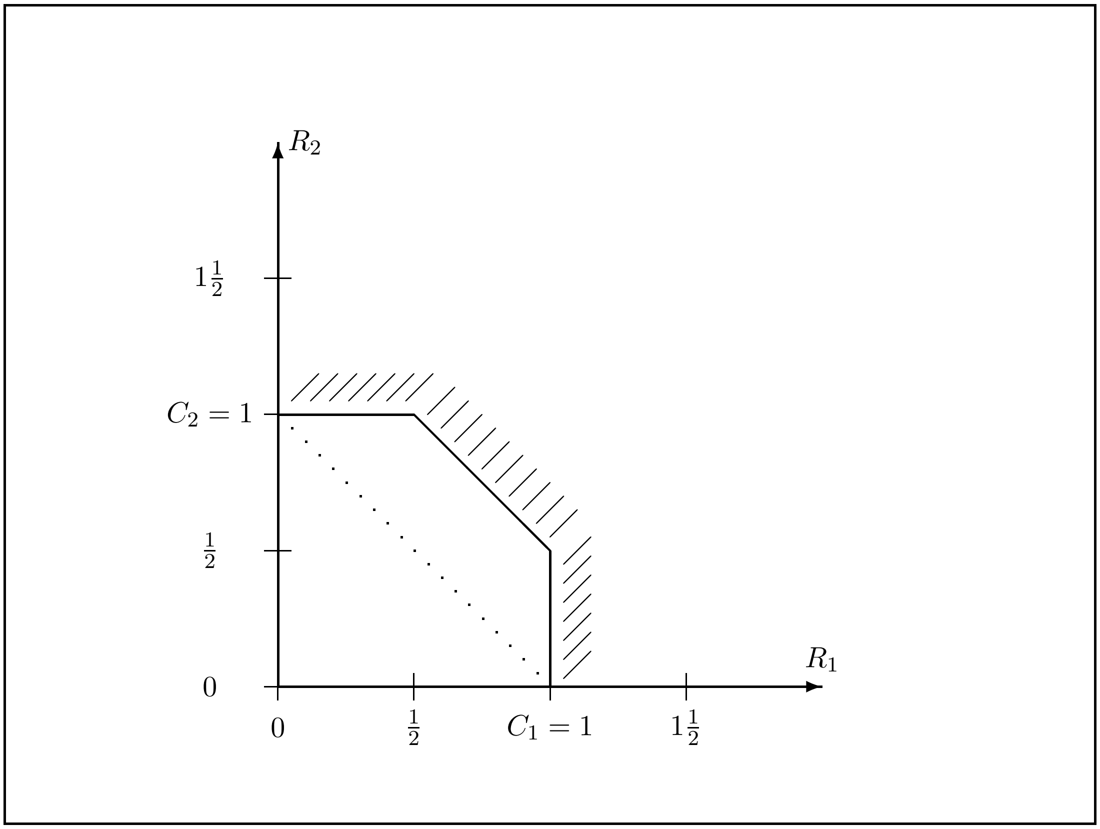


Figure 5.2: Capacity region of the binary erasure multiple access channel.

When the output $Y = 0$ or $Y = 2$, there is no ambiguity about the transmitted symbols from both input sources: if $Y = 0$ this implies that both X_1 and X_2 are equal to zero whereas $Y = 2 \leftarrow X_1 = X_2 = 1$. However, the output $Y = 1$ can be obtained either with $X_1 = 0$ and $X_2 = 1$ or with $X_1 = 1$ and $X_2 = 0$, leading to an ambiguity about the sources.

What is the capacity C of such a binary erasure multiple access channel? If we set deliberately $X_2 = 0$ (or equivalently $X_2 = 1$), then $Y = X_1$ (or $Y = X_1 + 1$), there is no longer any ambiguity and the capacity of the multiple access channel simply becomes the capacity of the channel between source X_1 and sink Y, i.e. $C = C_1$. The same applies to the second channel, from X_2 to Y, if we set $X_1 = 0$ (or $X_1 = 1$). By time-sharing the two channels, that is by allowing the first source X_1

to transmit at a fraction λ ($0 \leq \lambda \leq 1$) of the time while X_2 is set to 0 and then having source X_2 active for the remaining $(1 - \lambda)$ of the time (i.e. while $X_1 = 0$), the capacity of the multiaccess channel C consists of the straight line defined by the set of points joining the two extreme points as shown by the dotted line on Figure 5.2.

However, the capacity C is larger. Suppose that either source, for instance X_1, is already transmitting at a rate R_1. From the point of view of the second channel, meaning X_2 to Y, the source X_1 appears like a equiprobable binary noisy sequence superimposed over X_2. The channel between X_2 and Y is similar to a standard binary erasure channel (BEC) where the probability of obtaining an erasure symbol is equal to $\frac{1}{2}$ whatever the symbol from source X_2 is. We have seen previously that the capacity of a binary erasure channel is given by $C_2 = 1 - \rho$ where ρ is the transition probability from either binary input symbol to the output ternary erasure symbol. Here $\rho = \frac{1}{2}$ since source X_1 is assumed equiprobable. This means that with this multiple access scheme the first source is already exchanging information with the common receiver at a rate $R_1 = C_1 = 1$ and that an additionnal amount of information between source X_2 and Y of rate $\frac{1}{2}$. As shown by the solid line on Figure 5.2, the capacity region C of this binary erasure multiple access channel is the closure of the convex hull of the achievable rates (R_1, R_2) where $R_1 \leq 1$, $R_2 \leq 1$, and $R_1 + R_2 \leq 1\frac{1}{2}$.

5.1.2 Generalization of the Multiple Access Channel Capacity Region

Figure 5.3 illustrates a multiple access communication network with m users. Here there are m different users who transmit independent information over the same multiple access channel. The capacity region for the m-user multiple access communication network is again the convex hull of all achievable rates.

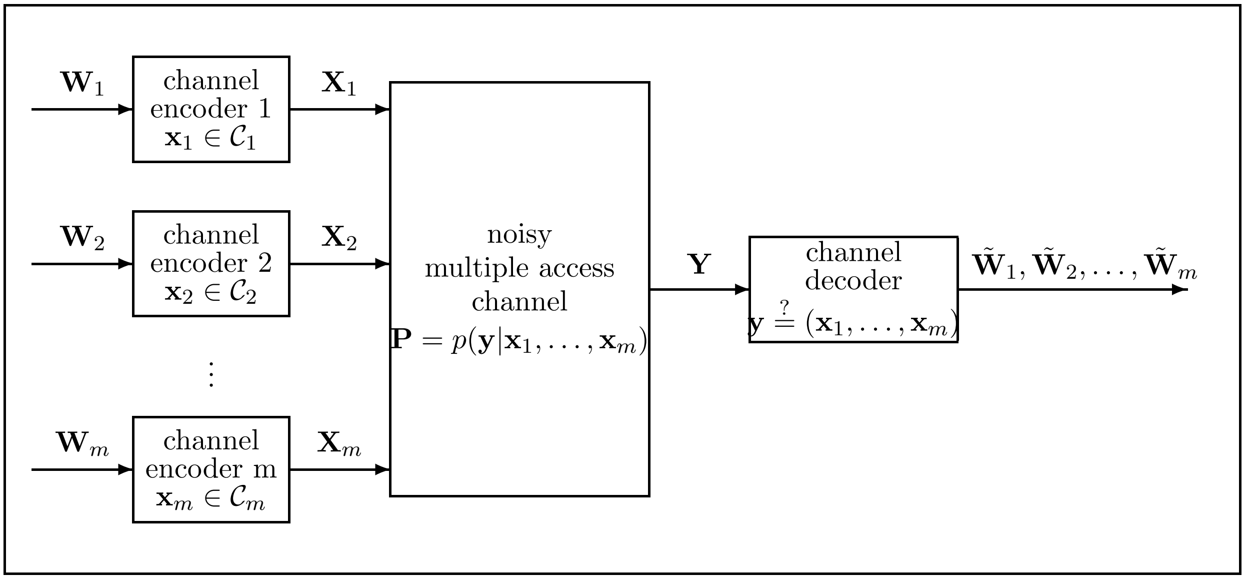

$\mathbf{W}_1$
channel encoder 1 $\mathbf{x}_1 \in \mathcal{C}_1$
$\mathbf{X}_1$
$\mathbf{W}_2$
channel encoder 2 $\mathbf{x}_2 \in \mathcal{C}_2$
$\mathbf{X}_2$
$\vdots$
$\mathbf{W}_m$
channel encoder m $\mathbf{x}_m \in \mathcal{C}_m$
$\mathbf{X}_m$
noisy multiple access channel $\mathbf{P} = p(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_m)$
$\mathbf{Y}$
channel decoder $\mathbf{y} \overset{?}{=} (\mathbf{x}_1, \ldots, \mathbf{x}_m)$
$\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2, \ldots, \tilde{\mathbf{W}}_m$

Figure 5.3: m-user multiple access communication network.

Consider the following partitionning of the set $\{1, 2, \ldots, m\}$ of all individual users: $\mathcal{S} \subseteq \{1, 2, \ldots, m\}$ (the symbol $\subseteq$ represents the set inclusion, i.e. $\mathcal{S}$ can be any subset of the set of all users $\{1, 2, \ldots, m\}$) and its complementary set $\mathcal{S}^c$. If $R(\mathcal{S}) = \sum_{i \in \mathcal{S}} R_i$ and the random source $X(\mathcal{S}) = \{X_i\}$ such that $i \in \mathcal{S}$, then one can determine the capacity region of such a m-user multiple link.

Theorem *(Capacity Region for the General Multiple Access Channel):*

The capacity region C of a m-user multiple access channel is the closure of the convex hull of the set of all rate vectors for which:

$$\boxed{R(\mathcal{S}) \leq I(X(\mathcal{S}); Y | X(\mathcal{S}^c)), \qquad \text{for all sets } \mathcal{S} \subseteq \{1, 2, \ldots, m\}}$$

for some product distribution $\{(p_1(x_{1,k_1}), p_2(x_{2,k_2}, \ldots, p_m(x_{m,k_m})\}$ on the input vector $(X_1, X_2, \ldots, X_m)$.

Proof:

The proof of this generalization theorem follows the same lines of the previous proof for the two-source case where now the number of terms in the expression of the probability of error is now $2^m - 1$ instead of 3 (excluding the first term ϵ_1).

5.2 Broadcast Networks

A broadcast communication network with m-users is depicted on Figure 5.4. A single broadcast transmitter sends information to m different users over the same broadcast channel. Here we assume that a part of the broadcast message is intended to all users, or receivers, whereas some other parts of the message are different and specific to each individual receivers.

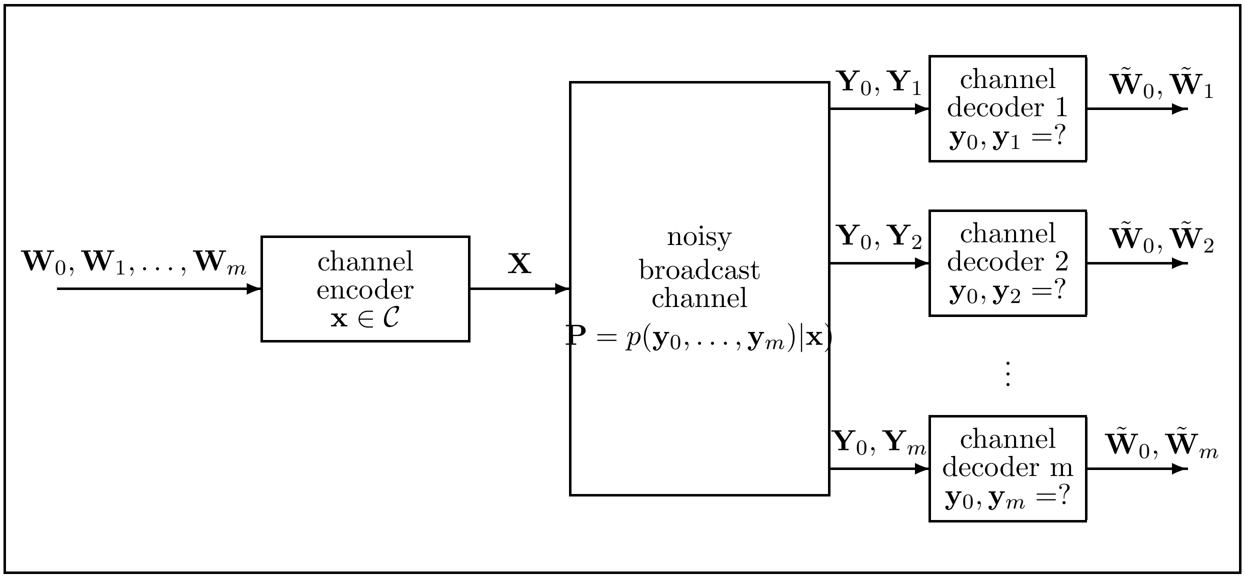

$\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_m$
channel encoder $\mathbf{x} \in \mathcal{C}$
$\mathbf{X}$
noisy broadcast channel $\mathbf{P} = p(\mathbf{y}_0, \ldots, \mathbf{y}_m)|\mathbf{x})$
$\mathbf{Y}_0, \mathbf{Y}_1$
channel decoder 1 $\mathbf{y}_0, \mathbf{y}_1 =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_1$
$\mathbf{Y}_0, \mathbf{Y}_2$
channel decoder 2 $\mathbf{y}_0, \mathbf{y}_2 =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_2$
$\vdots$
$\mathbf{Y}_0, \mathbf{Y}_m$
channel decoder m $\mathbf{y}_0, \mathbf{y}_m =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_m$

Figure 5.4: Broadcast communication network with m users.

5.2.1 Capacity Region of Broadcast Networks

The capacity region for the m-user broadcast communication link is once again the convex hull of a set of achievable rates. For the broadcast channel case, those rates are $R_0, R_1, \ldots, R_m$, where R_0 corresponds to the rate of the part of the message which is common to all m users and $R_1, \ldots, R_m$ represent those rates of the other message parts intended only for each receiver specifically.

Assume for a moment that the message sent by the broadcast transmitter is the same for all m receivers and that all transmitter to receiver links are the same (i.e. they can be represented with the same transition probability matrix). Then the problem will be equivalent to the situation where there is only one receiver and the capacity region of that degenerated case of broadcast channel would be simply the capacity of a point-to-point channel as we have considered in chapter 3.

At the other extreme, assume now that there is no common part in the broadcasted message; that is, each user receives a different message. Then the message parts can be exchanged between the single broadcast transmitter and each receiver in the network by sharing the channel. One obvious method is the time-sharing multiplexing scheme where the message part intended for a given user occupies a unique timeslot in a data stream.

Here we are interested in a more general broadcast network situation where part of the broadcast message is common to all users whereas there are also distinct messages, or parts, intended for each receiver. In this later case, a larger capacity region C can be obtained by exploiting the common and specific parts of the broadcast message.

Consider here, for simplicity, a broadcast network which consists of only two receivers. The code $\mathcal{C}$ of such a broadcast channel is defined as a block code of blocklength N and consisting of a total of $M = 2^{NR_0}2^{NR_1}2^{NR_2}$ codewords.

$$\mathcal{C} = \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_{2^{N(R_0+R_1+R_2)}} \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,n} & \cdots & c_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,n} & \cdots & c_{M,N} \end{bmatrix}$$

There are $M = 2^{N(M_0,M_1,M_2)}$ binary codewords of blocklength N. Both decoders must be able to recover the common message w_0. Furthermore, decoder 1 must decode its intended specific message w_1 whereas decoder 2 must decode w_2. In other words, the task of decoder 1 is to recover the information pair (w_0, w_1) and decoder 2 must determine the message pair (w_0, w_2). For this purpose, we can rewrite the broadcast network code by differentiating the index of the common and specific parts of the broadcasted message.

$$\mathcal{C} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,n} & \cdots & c_{m,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,n} & \cdots & c_{M,N} \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{(m_0,m_1,m_2),1} & \cdots & c_{(m_0,m_1,m_2),n} & \cdots & c_{(m_0,m_1,m_2),N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{(M_0,M_1,M_2),1} & \cdots & c_{(M_0,M_1,M_2),n} & \cdots & c_{(M_0,M_1,M_2),N} \end{bmatrix}$$

There will be a decoding error if either one of the messages ($\tilde{w}_0$, $\tilde{w}_1$,or $\tilde{w}_2$) decoded by the receivers is in error. The capacity region of broadcast channels is known only for some special cases: we will study the degraded broadcast network which is one of those known capacity region cases.

5.2.2 Capacity Region of Degraded Broadcast Networks

A degraded broadcast network is a broadcast channel which has the particular characteristic that the broadcast channel can be represented as m cascaded noisy channels as illustrated on Figure 5.5. With that channel configuration, we see that the first receiver is better than the second, which is a cascade of these first two channels, the third receiver gets a worse signal, and so on.

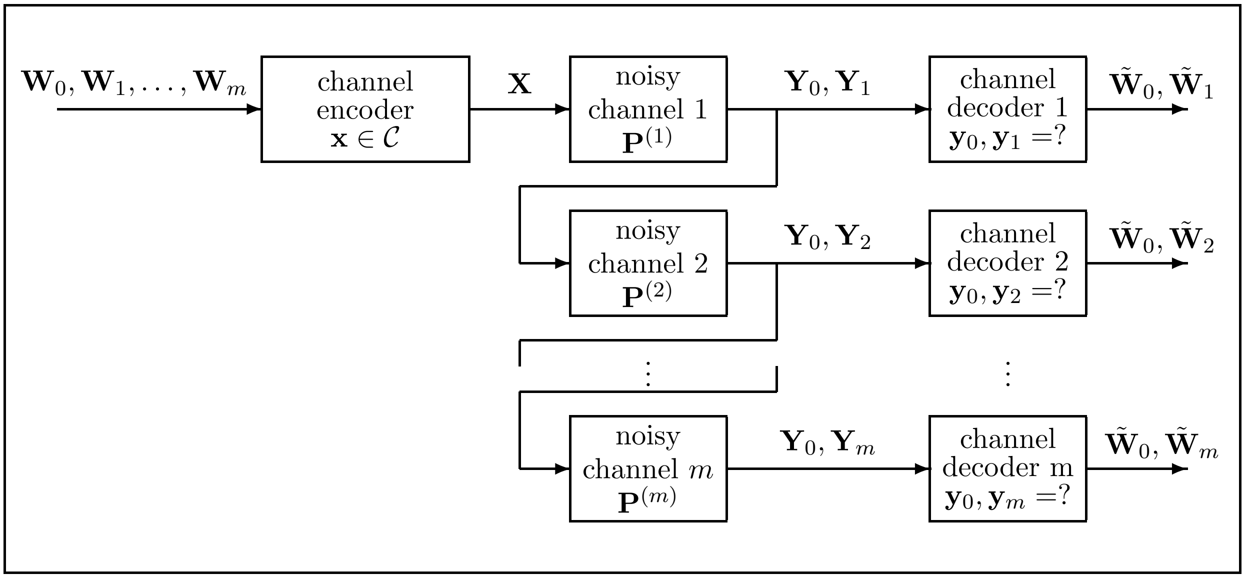

$\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_m$
channel encoder $\mathbf{x} \in \mathcal{C}$
$\mathbf{X}$
noisy channel 1 $\mathbf{P}^{(1)}$
$\mathbf{Y}_0, \mathbf{Y}_1$
channel decoder 1 $\mathbf{y}_0, \mathbf{y}_1 =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_1$
noisy channel 2 $\mathbf{P}^{(2)}$
$\mathbf{Y}_0, \mathbf{Y}_2$
channel decoder 2 $\mathbf{y}_0, \mathbf{y}_2 =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_2$
noisy channel m $\mathbf{P}^{(m)}$
$\mathbf{Y}_0, \mathbf{Y}_m$
channel decoder m $\mathbf{y}_0, \mathbf{y}_m =?$
$\tilde{\mathbf{W}}_0, \tilde{\mathbf{W}}_m$

Figure 5.5: Degraded broadcast network with m users.

Theorem *(Degraded Broadcast Channel Capacity):*

The capacity region C of a memoryless degraded broadcast channel is the closure of the convex hull of the set of all rates R_1 and R_2 for which:

$$\begin{aligned} R_1 &\leq I(X;Y_1|U), \\ R_2 &\leq I(U;Y_2) \end{aligned}$$

for some joint distribution $\{(p(u)p(x|u)p(y_1,y_2|u)\}$ on the input X, an auxiliary random variable U, and the output pair (Y_1,Y_2).

Proof:

The proof presented here involves once more the random coding argument as well as a decoding rule based on the definition of jointly typical pairs of sequences. Its derivation is similar to that used in section 3.5 for point-to-point channels and section 5.1 for multiple access channels. We briefly outline here the differences between the previous derivations and the present one for broadcast channel.

There is a broadcast channel decoding error the message intended for user 1 and/or user 2 is in error. We will use the concept of *protocodes* to show this theorem. In the degraded broadcast network configuration we have two channels: the first one between input X and output Y_1 is in fact a single point-to-point channel with high reliability. The other link between input X and output Y_2 consists of two cascaded channels and hence its performance is worse than the first channel. We recall that for cascaded channels the mutual information between the input and the output of the second channel $I(X;Y_2) \leq I(X;Y_1)$. All the codewords from the source X will be grouped in *clouds* of codewords where the cloud center is called the protocodeword of that subgroup of codewords. For the high quality broadcast link ($X \Rightarrow Y_1$), decoder 1 will attempt to decode the exact codeword transmitted by determining the transmitted protocodeword, or cloud, as well as the actual transmitted codeword. However, for the second low-quality channel ($X \Rightarrow Y_2$), decoder 2 will only try to determine the protocodeword without attempting to find the actual transmitted codeword, hence its low reliability.

An interesting example of such a degraded broadcast network is the High Definition Television (HDTV) for which the transmitted broadcast information will be high quality high images. Those users having a high quality HDTV receiver will be able to receive the intended high quality images whereas the other users with bad receivers will still be able to receive the broadcast images but with the limitation of the regular TV receivers. The common message will be the protocodes whereas the actual codewords will provide the additionnal image quality that can only be exploited by the high-end HDTV receivers. For this arrangement to work properly, a broadcast channel code should be employed.

For user 1, i.e. the single and better channel, a decoding error will occur if the triplet $(\mathbf{c}_m^\star, \mathbf{c}_{m,l}, \mathbf{y}_1)$ representing the transmitted protocodeword $\mathbf{c}_m^\star$, the transmitted codeword $\mathbf{c}_{m,l}$ and the received codeword $\mathbf{y}_1$ are not jointly typical. This first term in the error probability can be made arbitrarily small, provided that the blocklength N is sufficiently large. However, there will also be a decoding error if another triplet $(\mathbf{c}_{m'}^\star, \mathbf{c}_{m',l'}, \mathbf{y}_1)$, where either $m' \neq m$ and $l' \neq l$, happens to be jointly typical with the received codeword $\mathbf{y}_1$. In other words, the received vector $\mathbf{y}_1$ is correctly decoded only if we can determine both the correct protocodeword $\mathbf{c}_m^\star$ and the actual transmitted codeword $\mathbf{c}_{m,l}$:

$$\begin{aligned}
P_{e,\mathbf{y}_1} &\leq Pr\left[(\mathbf{c}_m^\star, \mathbf{c}_{m,l}, \mathbf{y}_1) \notin \mathcal{T}_{UXY_1}(\delta)\right] + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} \sum_{\substack{l'=1 \\ l' \neq l}}^{M_1} Pr\left[(\mathbf{c}_{m'}^\star, \mathbf{c}_{m',l'}, \mathbf{y}_1) \in \mathcal{T}_{UXY_1}(\delta)\right] \\
P_{e,\mathbf{y}_1} &\leq \epsilon_1 + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} \sum_{\substack{l'=1 \\ l' \neq l}}^{M_1} Pr\left[(\mathbf{c}_{m'}^\star, \mathbf{c}_{m',l'}, \mathbf{y}_1) \in \mathcal{T}_{UXY_1}(\delta)\right] \qquad (5.17)
\end{aligned}$$

For the second user (i.e. using the cascaded and thus the worse channel), an error in the decoding process will occur only if the decoder can not recognize the transmitted protocodeword. Again two error situations may happen: the pair $(\mathbf{c}_m^\star, \mathbf{y}_2) \notin \mathcal{T}_{UY_2}(\delta)$ (the protocodeword $\mathbf{c}_m^\star$ and the received vector $\mathbf{y}_2$ are not jointly typical), and $(\mathbf{c}_{m'}^\star, \mathbf{y}_2) \in \mathcal{T}_{UY_2}(\delta)$ (the received vector $\mathbf{y}_2$ is jointly typical with another, and thus erroneous, protocodeword $\mathbf{c}_{m'}^\star$). The first term can be made vanishlingly small by the properties of jointly typical sequences whereas the second term may not converge towards zero.

$$\begin{aligned}
P_{e,\mathbf{y}_2} &\leq Pr\left[(\mathbf{c}^{\star}_{m}, \mathbf{y}_2) \notin \mathcal{T}_{UY_2}(\delta)\right] + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} Pr\left[(\mathbf{c}^{\star}_{m'}, \mathbf{y}_2) \in \mathcal{T}_{UY_2}(\delta)\right] \\
P_{e,\mathbf{y}_2} &\leq \epsilon_2 + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} Pr\left[(\mathbf{c}^{\star}_{m'}, \mathbf{y}_2) \in \mathcal{T}_{UY_2}(\delta)\right] \qquad (5.18)
\end{aligned}$$

The error decoding probability P_e for the two-user degraded broadcast channel is then given by the union of the above two error events and, by the union bound, can be restated as:

$$\begin{aligned}
P_e &\leq P_{e,\mathbf{y}_1} + P_{e,\mathbf{y}_2} \qquad (5.19) \\
&\leq \epsilon_1 + \epsilon_2 + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} \sum_{\substack{l'=1 \\ l' \neq l}}^{M_1} Pr\left[(\mathbf{c}^{\star}_{m'}, \mathbf{c}_{m',l'}, \mathbf{y}_1) \in \mathcal{T}_{UXY_1}(\delta)\right] + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_2} Pr\left[(\mathbf{c}^{\star}_{m'}, \mathbf{y}_2) \in \mathcal{T}_{UY_2}(\delta)\right]
\end{aligned}$$

The expected probability of error over the ensemble of codes (random coding argument) is given by [Bla87, CT91]:

$$\sum_{\mathcal{S}_{\mathcal{C}}} Pr(\mathcal{C})\,[P_e] \leq \epsilon_1 + \epsilon_2 + 2^{-N[I(X;Y_1|U) - R_1 - \epsilon']} + 2^{-N[I(U;Y_2) - R_2 - \epsilon'']} \qquad (5.20)$$

The right-hand term in the last equation can be made arbitrarily small provided that:

$$\begin{aligned}
R_1 &\leq I(X;Y_1|U), \\
R_2 &\leq I(U;Y_2)
\end{aligned}$$

and provided that the blocklength N of the broadcast code is sufficiently large. Then there must exist a code for which the error decoding probability is at least as small as the average of the error decoding probability of the ensemble of codes.

$$\boxed{P_e \leq \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 = \epsilon}$$

QED

5.3 Problems

Problem 5.1: Find and draw the capacity regions for these following multiple access channels.

a) A multiple access channel which consists in two independent binary symmetric channels of capacities $C_1 = 1\ Sh$ and $C_2 = 1\ Sh$.

b) A additive *modulo-2* multiple access channel where $X_1 \in \{0, 1\}$, $X_2 \in \{0, 1\}$, and $Y = X_1 \oplus X_2$ (the symbol $\oplus$ represent the modulo-2 addition operation).

c) A *multiplicative* multiple access channel where $X_1 \in \{-1, 1\}$, $X_2 \in \{-1, 1\}$, and $Y = X_1 \times X_2$.

d) How do the capacity regions of the above three multiple access channels compare with the capacity region of the *binary erasure multiple access channel* studied in class.

Bibliography

[Bla84] R.E. Blahut. *Theory and Practice of Error Control Codes.* Addison-Wesley, Reading, Massachusetts, 1984.

[Bla87] R.E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Reading, Massachusetts, 1987.

[CCR90] T.H. Cormen, Leiserson C.E., and R.L. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, Massachusetts, 1990.

[CF94] J.-Y. Chouinard and G. Ferland. Cryptographic Degradation of DES in Block and Stream Cipher Modes in a Digital Mobile Communication Link. In *Workshop on Selected Areas in Cryptography (SAC'94)*, pages 159–169, Kingston, Canada, May 1994.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New-York, 1991.

[FC92] G. Ferland and J.-Y. Chouinard. Error Rate Performance Analysis of Stream and Block Ciphers in a Digital Mobile communication Channel. In *Third Annual Conference on Vehicular Navigation and Information Systems (VNIS 92)*, pages 426–433, Oslo, Norway, September 1992.

[FC94] G. Ferland and J.-Y. Chouinard. *Performance of BCH codes with DES encryption in a Digital Mobile Channel*, volume 793 of *Lecture Notes in Computer Science*, pages 153–172. Springer-Verlag, Berlin, 1994. *Information Theory and Applications: Third Canadian Workshop, Rockland, Ontario, Canada* (edited by A. Gulliver and N. Secord).

[For70] G.D. Forney. Convolutional Codes I: Algebraic Structure. *IEEE Transactions on Information Theory*, IT-16(6):720–738, November 1970.

[Fri67] B.D. Fritchman. A Binary Channel Characterization Using Partitioned Markov Chains. *IEEE Transactions on Information Theory*, IT-13(2):221–227, April 1967.

[Gal68] R. G. Gallagher. *Information Theory and Reliable Communications.* John Wiley and Sons, New-York, 1968.

[Gil60] E.N. Gilbert. Capacity of a Burst-Noise Channel. *Bell System Technical Journal*, 76(5):1253–1265, September 1960.

[Knu73a] D.E. Knuth. *The Art of Computer Programming: Fundamental Algorithms (volume 1).* Addison-Wesley, Reading, Massachusetts, second edition, 1973.

[Knu73b] D.E. Knuth. *The Art of Computer Programming: Sorting and Searching (volume 3).* Addison-Wesley, Reading, Massachusetts, 1973.

[Knu81] D.E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms (volume 2).* Addison-Wesley, Reading, Massachusetts, second edition, 1981.

[KS78] L.N. Kanal and A.R.K. Sastry. Models for Channels with Memory and Their Applications to Error Control. *Proceedings of the IEEE*, 66(7):724–744, July 1978.

[LC83] S. Lin and D.J. Costello. *Error Control Coding: Fundamentals and Applications.* Prentice-Hall, Englewood Cliffs, New-Jersey, 1983.

[LV93] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications.* Texts and Monographs in Computer Science. Springer-Verlag, New-York, 1993.

[Man87] M. Mansuripur. *Introduction to Information Theory.* Prentice-Hall, Englewood Cliffs, New-Jersey, 1987.

[oSAiC89] IEEE Journal on Selected Areas in Communications. Secure Communications. *IEEE Journal on Selected Areas in Communications*, SAC-7(4), May 1989.

[Osw86] J. Oswald. *Théorie de l'information ou analyse diacritique des systèmes.* Collection CNET-ENST. Masson, Paris, 1986.

[otI88] Proceedings of the IEEE. Special issue on Cryptography. *Proceedings of the IEEE*, 76(5), May 1988.

[Rez94] F.M. Reza. *An Introduction to Information Theory.* Dover Publications, New-York, 1994. (work first published by the McGraw-Hill Book Company, New-York, in 1961).

[Rom93] S. Roman. *Coding and Information Theory.* Graduate Texts in Mathematics. Springer-Verlag, New-York, 1993.

[Sha49] C.E. Shannon. Communication Theory of Secrecy Systems. *Bell System Technical Journal*, 28:656–715, October 1949.

[Sim92] G.J. Simmons. *Contemporary Cryptography: The Science of Information Integrity.* IEEE Press, New-York, 1992.

[Skl88] B. Sklar. *Digital Communications: Fundamentals and Applications.* Prentice-Hall, Englewood Cliffs, New-Jersey, 1988.

[SLCA91] A. Semmar, M. Lecours, J.-Y. Chouinard, and J. Ahern. Characterization of Error Sequences in UHF Digital Mobile Radio Channels. *IEEE Transactions on Vehicular Technology*, VT-40(4):769–775, November 1991.

[Sle73] D. Slepian. *Key Papers in the Development of Information Theory.* IEEE Press, New-York, 1973.

[SW93] N.J.A. Sloane and A.D. Wyner. *Claude Elwood Shannon: Collected Papers.* IEEE Press, New-York, 1993.

[Tor92] D.J. Torrieri. *Principles of Secure Communication Systems.* Artech House, Norwood, Massachusetts, second edition, 1992.