Theoretical physics of information: Lecture notes

Paolo Perinotti

November 21, 2022

Contents

Is information physical? Is physics informational?

Is information physical? We may be led to think so, based on the fact that every piece of information we receive about the outside world is carried by some physical system. If we start analysing the properties of information from this perspective we realise that the fundamental information processing tasks are ruled by fundamental physical laws, thus discovering fascinating results. An outstanding example is *Landauer's principle* stating that erasure of one bit of information always requires dissipation of an amount of energy equal to $KT \ln 2$, T being the temperature of the memory storing the bit.

However, if we think more deeply, we realise the way in which we become aware of physical processes is by gathering information contained in events: Physical systems and their evolution laws are a clever architecture of our mind for organising and explaining events, but they are a secondary notion, while primitive facts are events. The perception of their very reality is only a powerful belief that stands out of doubt only by virtue of its astonishing predictive power. Thus, we could reverse the starting question and ask: Is physics informational? In present-day physics, the number of scientists that started exploring the program summarised by the motto "it from bit"—coined by John A. Wheeler—is increasingly large, and this line of research is counting more and more successful results.

The aim of the course is not to answer this fundamental question, but to provide some operational and quantitative substance to the notion of information, which is one step in the long journey that takes to explore the relation between information and physics.

Messages, which are pieces of information, are encoded in some alphabet. However, as physicists we cannot accept this as the definition of a quantity, unless we provide an operational definition, which also requires a way to quantify it. The course will illustrate the basic operational contexts in which information can be quantified, along with the precise mathematical measures accomplishing this task.

The course is structured in three parts:

1. Classical information theory, where information is supposed to be carried by classical systems.

2. Quantum information theory, where information is supposed to be carried by quantum systems. This part in turn is split into two parts:

a) Encoding of classical information over quantum carriers.

b) Encoding of quantum information.

The main tasks that allow us to define precise notions of information are *compression* and *transmission over noisy channels*. While in the classical case, under rather general assumptions, both problems have been exhaustively analysed in the early work by Shannon [2], in the quantum case only compression has been thoroughly understood, while transmission is the really challenging problem, the one where the unintuitive feature of *entanglement* plays a crucial role, which is not yet completely tamed.

Part I

Classical information theory

Chapter 1

Lecture 1: Introduction

1.1 Basic ideas of classical information theory

1.1.1 Data compression

Our first question regards the amount of resources needed to *store* or *transmit* a given piece of information. This question is very vague at this stage, because we need to define what are the resources and what are the requirements of our storage. In the first place, let us set the constraint that information must be completely recoverable, without losses or errors. The question thus regards the minimal support that grants such a perfect storage or transmission. We imagine here to face the task of storage, and the resource will be memory cells.

Let us use a language L written in an alphabet Σ with four symbols, say $\Sigma = \{a, b, c, d\}$. Suppose we want to store or transmit a file written in the language L. For example, let us consider the string

$$bacadaba$$

Storing this message costs 8 memory cells, each with 4 levels. Suppose now that we use a different language L' whose alphabet X has only two symbols $X = \{0, 1\}$. If we choose the encoding

$$a \mapsto 00, \quad b \mapsto 01, \quad c \mapsto 10, \quad d \mapsto 11, \tag{1.1}$$

we can write the same string as

$$0100100011000100$$

which is longer: For the same message we now need 16 cells, each with 2 levels.

What alphabet do we choose, then? It is clear that the answer requires to establish one basic fact: whether the cost of a cell with $m \times n$ levels is higher or smaller than m times the cost of a cell with n levels. The basic assumption of information theory is that the two costs are *identical*. In this way, the issue of the choice of alphabet is solved by assuming that counting resources for storage is equivalent to counting the sheer number of levels used to store a given message. In other words, the choice of alphabet is not relevant for the purpose of counting storage resources. We can then choose one alphabet once for all, and it will be the most elementary one: $X = \{0, 1\}$. The cell for such an

alphabet is a *bit*, allocating one character which can be either 0 or 1. More formally, a bit is a variable that takes values in the alphabet X. Before analysing any given source using a different alphabet, we will then find a way to encode messages into bit strings.

Let us now come back to our original file, written in the alphabet $\Sigma = \{a, b, c, d\}$. Let us suppose that such symbols have probability distribution

$$p(a) = \frac{1}{2}, \quad p(b) = \frac{1}{4}, \quad p(c) = \frac{1}{8}, \quad p(d) = \frac{1}{8}.$$

This is the case, in particular, for our message *bacadaba*. Can we exploit our prior information about probabilities of characters to produce a better encoding? For example, we know that the character a will occur more frequently than the others, thus we choose to encode it in a short string: $a \mapsto 0$. If we now encode the second most frequent character b into 1, we will end up with an un-decodable storage: Every encoded word will be a sequence of 0 and 1, and could then be mistaken for a word made only of a's and b's. We then choose to encode $b \mapsto 10$. What about the last two characters? If we do not want the encoding of a and b to appear as a prefix of the encodings for c and d, we need to consider strings starting with 11: $c \mapsto 110$ and $d \mapsto 111$. Summarizing,

$$a \mapsto 0, \quad b \mapsto 10, \quad c \mapsto 110, \quad d \mapsto 111.$$

In this way, with respect to the previous encoding of (1.1), we have one character with a shorter encoding, one with the same length, and two with a longer encoding. However, in order to make a reasonable comparison we need to consider the *average* amount of bits per character of the two encodings. In this case, then, we have

$$l = \sum_{x=a,b,c,d} p(x)l(x) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4} \text{ bits/character},$$

which is clearly smaller than 2 bits/character provided by the encoding of equation (1.1). For example, the encoding of our string *bacadaba* is given by the following 14-bits string

$$10011001110100$$

In this mapping, every character is represented by a string of bits in such a way that no character is mapped to a string with a prefix that represents another character. This feature makes our encoding perfectly decodable. For example, the string

$$110011010110000$$

corresponds to *cacbcaaa*. We thus achieved a compression rate of 7/4 bits/character as opposed to the 2 bits/character of the encoding in equation (1.1), exploiting our prior knowledge summarized by a non-uniform *prior* distribution.

The above example leads us to the first big question in information theory:

> **First Shannon question:** *What is the smallest physical resource needed to store/transmit a given file without loss?* In other words, *What is the best compression rate that we can achieve in transmitting a given file faithfully?*

The first Shannon theorem—the so-called *source-coding theorem* will provide an answer to this question, along with a first operational way to define information.

1.1.2 Error-correction

In conjunction with the problem of information compression, the most fundamental problem in information theory is that of error correction. Dealing with compression, we assumed that our storage device is *noiseless*, i.e. an ideal storage device from which we can retrieve every bit of information untouched. Alternatively, we could have imagined that information was sent through an ideal channel, and tried to minimise the number of uses of the channel.

Suppose now that our physical resource that stores or transmits information is noisy, e.g. a disk with some bad blocks, or a transmission line that often outputs a flipped bit. In most practical cases, a more reliable resource is unaffordable, and communication can be improved only using a system solution, adding an encoder before the channel and a decoder after the channel. This kind of solution suitably exploits data redundancy and error correction techniques in the encoding.

Then, the following question naturally arises

> **Second Shannon question:** *What is the best rate for storing or transmitting a given file reliably over a noisy resource?*

The answer to this question will be given by Shannon's second theorem—the so-called noisy-channel-coding theorem. To understand how the problem of error correction is foundational for information theory, consider a fictitious scenario in which somebody is claiming to have the ability of transmitting information telepathically, but with some probability of error p. Clearly, if it turns out that $p = 1/2$, then we immediately understand that they are actually transmitting no information at all, since $p = 1/2$ is just the probability of pure-guessing. On the other hand, from the second Shannon theorem we will see that any value of p greater than $1/2$ (actually, even any value *smaller* than $1/2$!) can be used to transmit information reliably—i.e. with arbitrarily small error—upon reducing the information transmission rate. An intuitive understanding of this fact can be achieved considering a redundant encoding, e.g. substituting 00000 for 0 and 11111 for 1. In this way, it is more likely that the majority of the sent bits will be unaffected rather than the contrary. Thus, a simple *majority-voting* decoding would provide a correct transmission. Clearly, the larger is the transmitted string, the smaller is the error probability. Thus, by taking arbitrarily large encoding it is possible to make the error probability arbitrarily small. The price to pay for this strategy, however, is that the number of physical bits used to transmit a single logical bit becomes arbitrarily large. In other words, the *transmission rate*, defined as the ratio between the number of transmitted logical bits and the number of transmitted physical bits, scales down to zero if the error probability has to become arbitrarily small.

What is surprising about the second Shannon theorem, however, is the crucial result that we do not need to make the transmission rate vanishingly small: there exists a non null transmission rate at which we can transmit reliably asymptotically for large transmitted files, provided that we exploit an encoding over the entire file. What is wrong about our very naive strategy is thus the attempt to encode every single bit independently. Therefore, telepathy would have been possible if the bit error probability were just slightly different from $1/2$!

The amount of information that can be actually transmitted along a channel is connected to both the possibility of compressing and of correcting it, and the optimal encoding will address both issues jointly.

Let us start introducing the simplest example of a noisy channel, which is the one we will use to illustrate the basic ideas behind error correcting codes.

Binary symmetric channel. A *binary symmetric channel* is a channel that accepts a bit b_{in} as an input and provides a bit b_{out} as an output, with probability of correct transmission given by $1-f$, and probability of flipping the bit f, where $0 < f < 1$.

The behaviour of the binary symmetric channel is summarised by the following conditional probabilities

$$P(b_{\text{out}} = 0|b_{\text{in}} = 0) = P(b_{\text{out}} = 1|b_{\text{in}} = 1) = 1 - f,$$
$$P(b_{\text{out}} = 1|b_{\text{in}} = 0) = P(b_{\text{out}} = 0|b_{\text{in}} = 1) = f.$$

Notice that it is not restrictive to consider $0 < f \leq 1/2$, because in the opposite case we can flip the output bit $b'_{\text{out}} = b_{\text{out}} \oplus 1$, thus reducing to the case under consideration. Let us now define some notation ($\oplus$ denotes bit-wise sum modulo 2)

- s is the *source message*, a string of M bits;
- t is the *transmitted string*, a string of N bits;
- n is the *noise*, a string of N bits represening the errors introduced by the channel;
- $r = t \oplus n$ is the *received string*, a string of N bits;

Majority-voting error correction

The purpose of the encoding is to reduce the effect of the noise, at the expense of a reduced transmission rate. Roughly speaking, this means that we encode information redundantly. The most intuitive strategy is to encode every single bit in our source message s into three copies of the same bit. It will be then very unlikely that an error affects more than one bit in the three-bits block. The decoding strategy thus consists in decoding each block to the most frequent value in the block. As an example, consider the following instance of transmission

s	0	0	1	0	1	0	1
t	000	000	111	000	111	000	111
n	010	000	001	000	100	000	000
r	010	000	110	000	011	000	111

Consider now a transmitted three-bits block $\mathbf{t}_j = t_{3j}t_{3j+1}t_{3j+2}$, and analogously a received three-bits block $\mathbf{r}_j = r_{3j}r_{3j+1}r_{3j+2}$. Denote by $P(\mathbf{r}_j|s_j)$ the conditional probability that the block $\mathbf{r}_j$ is received, given that the signal bit is s_j. Then, the

posterior probability that s_j was encoded given that $\mathbf{r}_j$ has been received is given by Bayes' rule

$$P(s_j|\mathbf{r}_j) = \frac{P(\mathbf{r}_j|s_j)P(s_j)}{P(\mathbf{r}_j)}, \quad P(\mathbf{r}_j|s_j) = P(\mathbf{r}_j|\mathbf{t}_j) = \prod_{k=0}^{2} P(r_{3j+k}|t_{3j+k}), \tag{1.2}$$

where by definition $\mathbf{t}_j = s_j s_j s_j$, namely $t_{3j+k} = s_j$ for $k = 0, 1, 2$. We remind that the noisy channel is completely specified by the conditional probability $P(r|t)$, whereas $P(s_j)$ is the *prior* probability of signal s_j, and $P(\mathbf{r}_j)$ is the probability of receiving $\mathbf{r}_j$, namely the marginal probability $P(\mathbf{r}_j) = \sum_{s_j=0}^{1} P(\mathbf{r}_j|s_j)P(s_j)$. All the functions involved in the expression of equation (1.2) are thus known.

Upon receiving $\mathbf{r}_j$ we can decode the value $\bar{s}_j$ maximising the posterior probability of equation (1.2), that for equal prior probabilities $P(s_j = 0) = P(s_j = 1) = 1/2$, is equivalent to the s_j maximising the likelihood $P(\mathbf{r}_j|s_j)$, given by

$$P(\mathbf{r}_j|\mathbf{t}_j) = \prod_{k=0}^{2} P(r_{3j+k}|t_{3j+k}).$$

We remind that for the binary symmetric channel we have

$$P(r_i|t_i) = \begin{cases} 1-f & r_i = t_i, \\ f & r_i \neq t_i. \end{cases}$$

The ratio between the likelihood for $s_j = 0$ and $s_j = 1$ is thus

$$\frac{P(s_j = 1|\mathbf{r}_j)}{P(s_j = 0|\mathbf{r}_j)} = \frac{P(\mathbf{r}_j|s_j = 1)}{P(\mathbf{r}_j|s_j = 0)} = \prod_{k=0}^{2} \frac{P(r_{3j+k}|1)}{P(r_{3j+k}|0)}.$$

Defining $\gamma := (1-f)/f$, each factor in the product is equal to $\gamma > 1$ for $r_{3j+k} = 1$ and to $\gamma^{-1} < 1$ for $r_{3j+k} = 0$. Therefore, the value $\bar{s}_j$ that maximises the likelihood corresponds to the one obtained by majority-voting: every vote for 1, namely $r_{3j+k} = 1$, contributes with a factor γ in the likelihood ratio, while every vote for 0 , namely $r_{3j+k} = 0$, contributes with a factor γ^{-1}. Clearly this strategy will produce the wrong correction if two errors occur within the same block. However, assuming that errors are statistically independent, and occurring with small probability, the probability of double or triple flip within each block amounts to $3f^2(1-f) + f^3$, and is much smaller than the probability of no flip or single flip $(1-f)^3 + 3f(1-f)^2$. A practical example is the following instance of transmission

s	0	0	1	0	1	0	1
t	000	000	111	000	111	000	111
n	010	000	101	000	100	000	000
r	010	000	010	000	011	000	111
$\bar{s}$	0	0	0	0	1	0	1
successful correction	*				*		
wrong correction			*				

Using the majority-vote error correction the error probability is improved from $\sim f$ to $\sim f^2$, being now dominated by the two-flip errors. We want to emphasise that the majority-vote code has been just considered for didactic reasons, since it is very simple, but it is also very inefficient: it uses too many bits, and still has a high probability of error. Indeed, performances of different error-correction codes can be much better. For example, for a model of disk drive making errors as a binary symmetric channel with $f = 0.1$, in order to achieve a bit probability error 10^{-15} using a majority-vote code one would need *sixty* disks to store an amount of information equivalent to one disk. On the other hand, according to the second Shannon theorem we will see that only *two* disks are needed. Indeed, the astonishing fact asserted by the second Shannon theorem is that above some threshold of error probability on a single bit, it is always possible to encode information with a vanishingly small probability of error by keeping the rate of transmission *non null*. Unfortunately, a naïve majority-vote encoding would provide a vanishingly small rate versus error probability.

The $(7, 4)$ Hamming block code

In order to improve the encoding, one can exploit redundancy more conveniently by encoding an entire block of data instead of encoding every bit independently, as in the majority-voting strategy. This feature is implemented in block codes. A (N, K) block code is a rule for converting a sequence of source bits s of length K into a transmitted bit string t of length N. Redundancy is introduced taking $N > K$. In a linear block code the additional $N - K$ encoded bits are obtained as a linear function of the original K bits, and are called *parity-check* bits. An example of block code is the $(7, 4)$ Hamming code, in which one transmits $N = 7$ bits for each block of $K = 4$ source bits. The rule for the encoding is to set $t_j = s_j$ for $j = 1, 2, 3, 4$, whereas the parity-check bits $t_5 t_6 t_7$ are set so that the parity within each circle in Fig. 1.1 is even.

An instance of encoding is given in Fig. 1.2.

In order to treat the block code appropriately, let us introduce some notation here.

Definition 1.1 (Inner product)**.** Let x, y be strings in $\mathbb{F}_2^M$. Then the *inner product* of x and y is

$$x^T y = x_1 y_1 \oplus x_2 y_2 \oplus \cdots \oplus x_M y_M,$$

where $\oplus$ denotes sum modulo 2.

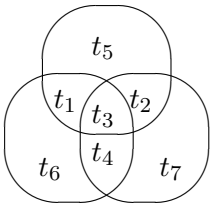


Figure 1.1 Representation of the $(7, 4)$ Hamming code.

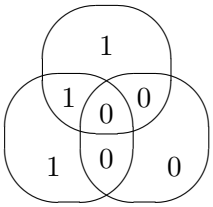


Figure 1.2 Example of $(7, 4)$ Hamming encoding.

Considering that a matrix $A \in \mathbb{F}_2^M \times \mathbb{F}_2^L$ can be written as

$$A = \begin{pmatrix} {\alpha_r^{(1)}}^T \\ {\alpha_r^{(2)}}^T \\ \vdots \\ {\alpha_r^{(M)}}^T \end{pmatrix} = (\alpha_c^{(1)} \ \alpha_c^{(2)} \ \dots \ \alpha_c^{(L)}),$$

we can then define the matrix product as follows

Definition 1.2 (Matrix product)**.** Let A be a matrix in $M_{M\times L}(\mathbb{F}_2)$ and $B \in M_{L\times J}(\mathbb{F}_2)$. Then the *matrix product* of A and B is the matrix C in $M_{M\times J}(\mathbb{F}_2)$ with elements $C_{i,j} = {\alpha_r^{(i)}}^T \beta_c^{(j)}$

The fact that the Hamming code is linear allows one to write it compactly in form of

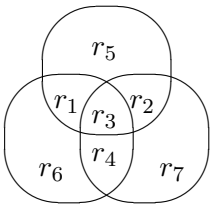
matrix multiplication as

$$t = Gs$$

where G is the *generator matrix* of the code, given by

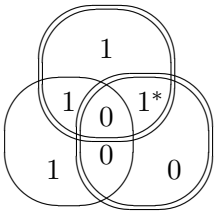
$$G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

The received vector r is represented as follows

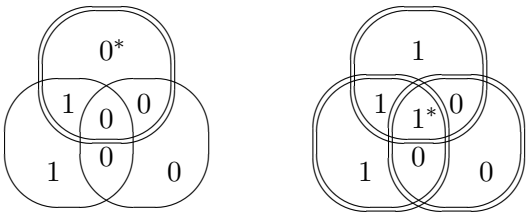


where $r = t \oplus n$. What makes it difficult to derive the decoding procedure is the fact that each bit—including the parity-check ones—may be flipped by the noise, thus we do not know where the error occurred. However, if the parity in one circle is violated, we know that one error must have occurred within that circle. We can then define three bits z_1, z_2 and z_3 corresponding to the parities of the three circles, and since their pattern heralds the occurrence of errors, the string z is called *syndrome*. One can decode the string by figuring out what are the bits affected by noise, which are supposed to be those in the intersection of circles with syndrome 1 and complements of circles with syndrome 0.

For example, in the following instance the syndrome is $z = 101$, with flipped bit r_2



where the syndrome 1 circles have been represented with a double line. Other examples are the following ones, with syndrome $z = 100$ and $z = 111$, respectively.



Also the decoding can be described in matrix form, as is the case for the encoding. In the presence of noise the received vector r is given by

$$r = Gs \oplus n$$

The matrix G can be conveniently rewritten in block form as follows

$$G = \begin{pmatrix} I_4 \\ P \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

I_4 denoting the four-dimensional identity matrix. The computation of the syndrome vector z is linear, and can be obtained as $z = Hr$, where H is called *parity-check* matrix. In our case H is given by

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} P & I_3 \end{pmatrix}. \tag{1.3}$$

Notice that H is designed to satisfy $HG = 0$, and thus, since codewords are obtained as $t = Gs$, the matrix H satisfies $Ht = 0$ for every codeword t. Therefore, for the syndrome one has

$$z = H(Gs \oplus n) = Hn$$

Indeed, there may be many noise strings n leading to the same syndrome z, and the syndrome-decoding problem is to find the most probable noise vector n satisfying the above equation, so that we can reconstruct the message through a decoding function $t = t(r)$. A decoding algorithm that solves this problem is called a maximum-likelihood decoder. For example, for small probability of flips, we have sizeable probabilities only for a single flip, and such a flip in the j-th position is described by a noise vector e_j. Therefore, one concludes that the error occurred at the j-th place for j satisfying the equation $Hr = He_j$ (single-flip syndrome), and the decoded source block would be $t = r \oplus e_j$.

The question is now: how to build a general block code? Let us analyse the question thoroughly.

Definition 1.3 (Linear block code)**.** Let $G : \mathbb{F}_2^K \to \mathbb{F}_2^N$ be a linear *encoding* function and $H : \mathbb{F}_2^N \to \mathbb{F}_2^{N-K}$ a linear *decoding* function. A (N, K) *linear block code* is a couple (G, H). Given such a code, we say that $y \in \mathbb{F}_2^N$ is a *codeword* if $y \in G(\mathbb{F}_2^K)$.

In the analysis of block codes a useful notion is that of Hamming distance. The *weight* of a bit string $x \in \mathbb{F}_2^K$ is the function $w : Z_2^K \to \mathbb{N}$ defined as $w(x) := \sum_{i=1}^K x_i$.

Definition 1.4 (Hamming distance)**.** Given $x, y \in \mathbb{F}_2^N$, their *Hamming distance* is given by the function $d : \mathbb{F}_2^N \times \mathbb{F}_2^N \to \mathbb{N}$ defined as $d(x, y) := w(x \oplus y)$.

The Hamming distance between two strings is then the number of sites at which the bits of the two strings differ. For general linear codes, the optimal decoding consists in associating the codeword t to the received string r such that $d(t, r)$ is minimised. Such a decoding is called *minimum distance decoding*. A code is characterised by its minimum distance, defined as the minimum distance between two codewords: $d_m(G, H) := \min_{x,y \in G(\mathbb{F}_2^K)} d(x, y)$. This quantity determines the performance of a code as follows.

Theorem 1.5. *Let (G, H) be a code with minimum distance $d_m(G, H) = 2j + 1$. Then the two following statements hold*

1. *(G, H) can can* correct *up to j errors in a block*

2. *(G, H) can* detect *up to $2j$ errors per block.*

Proof. Suppose that the message is the string s, with codeword $t := Gs$ and the transmission is affected by the noise n with weight $w(n) \leq j$. Then, for any codeword $z \in G(\mathbb{F}_2^K)$ one has

$$2j + 1 \leq d(t, z) \leq d(t, r) + d(r, z) = w(n) + d(r, z) \leq j + d(r, z), \tag{1.4}$$

which implies $d(r, z) \geq j + 1$, while by hypothesis $d(r, t) \leq j$. Thus, the minimum-distance decoding allows one to correct up to j errors correctly. Let us now suppose that $1 \leq w(n) \leq 2j$. In this case one has

$$1 \leq d(t, r) \leq 2j, \tag{1.5}$$

and thus r cannot be a different codeword. This implies that the occurrence of less than $2j + 1$ errors is always detected. However, in case $w(n) > j$ correct decoding is not possible. □

Let us now show the general structure of a block code. Most of the analysis is based on the decoding matrix H. For our purposes, it is useful to introduce the *kernel* of a matrix $F : \mathbb{F}_2^M \to \mathbb{F}_2^J$, which is defined as the set $\mathsf{Ker}_F := \{y \in Z_2^M | \, Fy = 0\}$, where $0 \in \mathbb{F}_2^J$ is the string of J bits 0, i.e. $00 \ldots 0$.

Definition 1.6 (Associated parity check and generator matrices)**.** A *canonical parity check matrix* in $M_{N-K\times N}(\mathbb{F}_2)$ is a matrix of the form

$$H = \begin{pmatrix} A & I_{N-K} \end{pmatrix}. \tag{1.6}$$

A *standard generator matrix* in $M_{N\times K}(\mathbb{F}_2)$ is a matrix of the form

$$G = \begin{pmatrix} I_K \\ A \end{pmatrix}. \tag{1.7}$$

Two such matrices with a common $(N-K)\times K$ block A are *associated* to each other. They define a (N,K) block code and set of codewords $\mathsf{C} := G(\mathbb{F}_2^N)$.

The main consequence of the above definition is the following result that we already exploited for the case of the $(7,4)$ block code.

Theorem 1.7. *Let G, H be associated to each other.Then $HG = 0$.*

Proof. If G, H are associated to each other we have

$$HG = \begin{pmatrix} A & I_{N-K} \end{pmatrix} \begin{pmatrix} I_K \\ A \end{pmatrix} = A \oplus A = 0, \tag{1.8}$$

where we used the property of element-wise sum modulo two that $X \oplus X = 0$. □

Actually, we can prove a slightly stronger statement.

Theorem 1.8. *Let G, H be associated. Then $\mathsf{C} = \mathsf{Ker}_H$.*

Proof. By theorem 1.7 we clearly have $\mathsf{C} \subseteq \mathsf{Ker}_H$. We then need to prove that $\mathsf{Ker}_H \subseteq \mathsf{C}$. Indeed, let $y \in \mathsf{Ker}_H$. Let us decompose Hy as follows

$$Hy = \begin{pmatrix} A & I_{N-K} \end{pmatrix} \begin{pmatrix} y_u \\ y_d \end{pmatrix} = Ay_u \oplus y_d = 0, \tag{1.9}$$

where $y_u \in \mathbb{F}_2^N$ and $y_d \in \mathbb{F}_2^{N-K}$. Then, $Ay_u = y_d$, namely one can write

$$y = \begin{pmatrix} y_u \\ Ay_u \end{pmatrix} = \begin{pmatrix} I_N \\ A \end{pmatrix} y_u = Gy_u. \tag{1.10}$$

Thus, we proved that $y \in G(\mathbb{F}_2^N) = \mathsf{C}$. □

Finally, we can prove the following theorem.

Theorem 1.9. *Let G, H define a (K,N) block-code. The minimum distance $d_m(G,H)$ is j iff the parity check matrix H is such that every set of $l < j$ columns $\{h_c^{(i_1)}, h_c^{(i_2)}, \ldots, h_c^{(i_l)}\}$ satisfies $h_c^{(i_1)} \oplus h_c^{(i_2)} \oplus \cdots \oplus h_c^{(i_l)} \neq 0$, while there is a set of j columns $\{h_c^{(i_1)}, h_c^{(i_2)}, \ldots, h_c^{(i_j)}\}$ such that $h_c^{(i_1)} \oplus h_c^{(i_2)} \oplus \cdots \oplus h_c^{(i_j)} = 0$.*

Proof. Let e_i denote the canonical basis vector in $\mathbb{F}_2^N$, then $h_c^{(i)} = He_i$. Now, by theorem 1.8, two strings $y, z \in \mathbb{F}_2^{N-K}$ are codewords, i.e. elements of C, if and only if $y, z \in \mathsf{Ker}_H$. The minimum distance $d_m(G, H)$ is then larger than $j-1$ if and only if $y \oplus e_{i_1} \oplus e_{i_2} \oplus \cdots \oplus e_{i_l} \notin \mathsf{Ker}_H$ for all $y \in \mathsf{Ker}_H$ and for any choice of the l vectors $\{e_{i_k}\}_{k=1}^{l}$ with $l < j$. Thus, the Hamming distance of any two codewords y, z is larger than $j-1$ if and only if $H(e_{i_1} \oplus e_{i_2} \oplus \cdots \oplus e_{i_l}) \neq 0$ for any choice of the l vectors $\{e_{i_k}\}_{k=1}^{l}$, with $l < j$. Now, since the vectors He_i are columns $h_c^{(i)}$ of H, the last condition is equivalent to the first statement of the thesis, namely every set of $l < j$ columns $\{h_c^{(i_1)}, h_c^{(i_2)}, \ldots, h_c^{(i_l)}\}$ satisfies $h_c^{(i_1)} \oplus h_c^{(i_2)} \oplus \cdots \oplus h_c^{(i_l)} \neq 0$. Finally, the minimum distance is precisely j if there are two codewords y, z such that $y = z \oplus e_{i_1} \oplus e_{i_2} \oplus \cdots \oplus e_{i_j}$, and then $H(e_{i_1} \oplus e_{i_2} \oplus \cdots \oplus e_{i_j}) = h_c^{(i_1)} \oplus h_c^{(i_2)} \oplus \cdots \oplus h_c^{(i_j)} = 0$. □

Exercise 1.1

Prove that the $(7, 4)$ block code can detect up to 2 errors, and correct up to 1.

Answer of exercise 1.1

Let us consider the matrix H in equation (1.3). Every couple of columns is different, but if we consider e.g. the columns h_3, h_4, h_5 we have $h_3 \oplus h_4 \oplus h_5 = 0$. Thus by theorem 1.9 we have $d_m(G, H) = 3$. Finally, by theorem 1.5 the $(7, 4)$ block code can detect 2 errors and correct one.

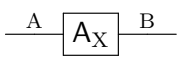
Chapter 2

Lecture 2: Operational Probabilistic Theories

In this lecture we introduce Operational Probabilistic Theories (OPTs). Every OPT represents a possible theory of information. What do we expect a theory of information to look like? In order to answer this question, let us think about our experience. One situation where we deal with information and its processing in our everyday life is when handling a smartphone. A smartphone is a device that manipulates the electromagnetic field in its surroundings and within its silicon circuits, as well as the field of phonons (i.e. vibration modes of the surrounding air, whose excitations are acoustic waves). Information in this case is encoded in the carrier systems—modes of the electromagnetic field and of acoustic vibrations—and processed by suitable operations. A theory of information is a mathematical model aimed at describing such kind of situation. In the first place, we require our theory to provide a way to represent information carrying systems, as well as operations that such systems can undergo.

2.1 Operational theories

In our example, every operation reads the state of a set of *input* systems (e.g. electromagnetic pulses that are absorbed), produces a new state of a set of *output* systems (e.g. re-writes some blocks of the memory), and displays some *outcome* (e.g. a message on the display). In the context of OPTs, we call such an operation a *test*. In classical information processing the state of the output system and the outcome may often coincide, but this is not a necessary condition. A better understanding of what a test is can be given by a quantum measurement: one has an input state, a process that produces an outcome—e.g. a string of digits on a display—and an output state, that possibly depends on the outcome—e.g. the state of the quantum system after a von Neumann-Lüders measurement. An operational theory Θ consists then in i) a collection $\mathsf{T}(\Theta)$ of tests $\mathsf{T}_{\mathrm{X}}^{\mathrm{A}\to\mathrm{B}}$, each labelled by input and output letters from a collection $\mathsf{Sys}(\Theta)$ denoting system types, e.g. $\mathrm{A}\to\mathrm{B}$, and by a finite set of outcomes X; for every pair of types $\mathrm{A},\mathrm{B}\in\mathsf{Sys}(\Theta)$ the set of tests of type $\mathrm{A}\to\mathrm{B}$ is denoted $\mathsf{Test}(\mathrm{A}\to\mathrm{B})$. A test A_{X} of type $\mathrm{A}\to\mathrm{B}$ can be represented by a diagram as the following



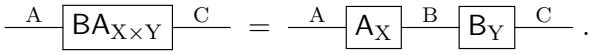
.

In general, an information processing algorithm requires more than a single operation, and is articulated in subsequent steps. Thus, an operational theory is also specified by ii) a rule for sequential composition: the test $\mathsf{T}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ can be followed by the test $\mathsf{R}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{A}' \to \mathrm{B}')$ if $\mathrm{A}' \equiv \mathrm{B}$, thus obtaining the sequential composition $\mathsf{R}_{\mathrm{Y}}\mathsf{T}_{\mathrm{X}} := \mathsf{R}\mathsf{T}_{\mathrm{X}\times\mathrm{Y}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B}')$. Sequential composition must have the following properties.

1. Associativity: given $\mathsf{R}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$, $\mathsf{T}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{B} \to \mathrm{C})$, and $\mathsf{W}_{\mathrm{Z}} \in \mathsf{Test}(\mathrm{C} \to \mathrm{D})$, one has
$$\mathsf{W}_{\mathrm{Z}}(\mathsf{T}_{\mathrm{Y}}\mathsf{R}_{\mathrm{X}}) = (\mathsf{W}_{\mathrm{Z}}\mathsf{T}_{\mathrm{Y}})\mathsf{R}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{D}).$$

2. Identity: for every $\mathrm{A} \in \mathsf{Sys}(\Theta)$, there exists a test with $\mathrm{X} = \{*\}$, denoted by $\mathsf{I}^{\mathrm{A}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{A})$ such that $\mathsf{I}^{\mathrm{B}}\mathsf{R}_{\mathrm{X}} = \mathsf{R}_{\mathrm{X}}\mathsf{I}^{\mathrm{A}} = \mathsf{S}_{\mathrm{X}}$, for every $\mathsf{R}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$.

In terms of diagrams, sequential composition of $\mathsf{A}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ and $\mathsf{B}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{B} \to \mathrm{C})$ is represented as follows



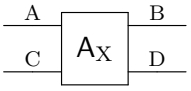
Now, let us consider e.g. the blocks of a hard disk. Every block is a system, but an array of blocks can be treated as a composite system itself, carrying information that is stored in a delocalised way. For example, if we want to encode four symbols, we need two bits. If we want to process information encoded in four symbols, we need to operate jointly on the two cells containing the two bits. More generally, every pair of systems can be composed *in parallel* to constitute composite systems. When operating on composite systems, one can run a test that involves all component systems at once. However, a distinctive feature of composite systems is that one can carry independent tests on each of their components simultaneously. For example, when operating on an array of two bits, one can "reset the first bit to the value 0, and flip the second bit". The result of such an operation is independent of the order of the two operations on the single bits, and these can be also run simultaneously. Such a test is the *parallel composition* of two tests: "reset the first bit to the value 0", and "flip the second bit". To complete the picture of an operational theory we then require iii) a rule $\otimes : (\mathrm{A}, \mathrm{B}) \mapsto \mathrm{AB}$ for composing systems in parallel, and a corresponding rule for tests $\otimes : (\mathsf{S}_{\mathrm{X}}, \mathsf{T}_{\mathrm{Y}}) \mapsto (\mathsf{S} \otimes \mathsf{T})_{\mathrm{X}\times\mathrm{Y}}$, with the following properties.

1. Associativity: $(\mathrm{AB})\mathrm{C} = \mathrm{A}(\mathrm{BC})$.

2. Unit: there is a unique system I such that $\mathrm{IA} = \mathrm{AI} = \mathrm{A}$ for every $\mathrm{A} \in \mathsf{Sys}(\Theta)$.

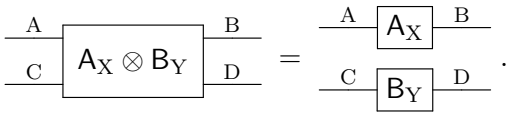
3. For every $\mathsf{R}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ and $\mathsf{T}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{C} \to \mathrm{D})$, one has $\mathsf{R}_{\mathrm{X}} \otimes \mathsf{T}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{AC} \to \mathrm{BD})$. Associativity of $\otimes$ holds:
$$(\mathsf{R}_{\mathrm{X}} \otimes \mathsf{T}_{\mathrm{Y}}) \otimes \mathsf{W}_{\mathrm{Z}} = \mathsf{R}_{\mathrm{X}} \otimes (\mathsf{T}_{\mathrm{Y}} \otimes \mathsf{W}_{\mathrm{Z}}).$$

For composite systems we use diagrams with multiple wires, e.g.



.

In particular, for tests that are parallel compositions we will draw



.

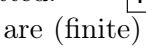
Altogether, sequential and parallel composition must satisfy the following properties

1. For every $\mathsf{A}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$, $\mathsf{B}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{B} \to \mathrm{C})$, $\mathsf{D}_{\mathrm{Z}} \in \mathsf{Test}(\mathrm{D} \to \mathrm{E})$, $\mathsf{E}_{\mathrm{W}} \in \mathsf{Test}(\mathrm{E} \to \mathrm{F})$, one has

$$(\mathsf{B}_{\mathrm{Y}} \otimes \mathsf{E}_{\mathrm{W}})(\mathsf{A}_{\mathrm{X}} \otimes \mathsf{D}_{\mathrm{Z}}) = (\mathsf{B}_{\mathrm{Y}}\mathsf{A}_{\mathrm{X}}) \otimes (\mathsf{E}_{\mathrm{W}}\mathsf{D}_{\mathrm{Z}}). \tag{2.1}$$

2. Swap: for every pair of system types A, B, there exists a test $\mathsf{S}^{\mathrm{AB}} \in \mathsf{Test}(\mathrm{AB} \to \mathrm{BA})$ such that $\mathsf{S}^{\mathrm{BA}}\mathsf{S}^{\mathrm{AB}} = \mathsf{I}^{\mathrm{AB}}$, and $\mathsf{S}^{\mathrm{AB}}(\mathsf{A}_{\mathrm{X}} \otimes \mathsf{B}_{\mathrm{Y}}) = (\mathsf{B}_{\mathrm{Y}} \otimes \mathsf{A}_{\mathrm{X}})\mathsf{S}^{\mathrm{AB}}$. Moreover,

$$\mathsf{S}^{(\mathrm{AB})\mathrm{C}} = (\mathsf{I}^{\mathrm{A}} \otimes \mathsf{S}^{\mathrm{BC}})(\mathsf{S}^{\mathrm{AC}} \otimes \mathsf{I}^{\mathrm{B}}),$$
$$\mathsf{S}^{\mathrm{A}(\mathrm{BC})} = (\mathsf{S}^{\mathrm{AB}} \otimes \mathsf{I}^{\mathrm{C}})(\mathsf{I}^{\mathrm{B}} \otimes \mathsf{S}^{\mathrm{AC}}).$$

The identity test will be often omitted: 

.

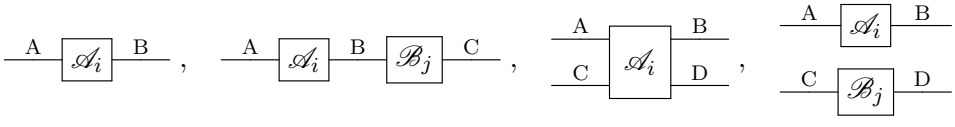
All tests of an operational theory are (finite) collections of *events*: $\mathsf{Test}(\mathrm{A} \to \mathrm{B}) \ni \mathsf{R}_{\mathrm{X}} = \{\mathscr{R}_i\}_{i\in\mathrm{X}}$. Now, if $\mathsf{Test}(\mathrm{A} \to \mathrm{B}) \ni \mathsf{R}_{\mathrm{X}} = \{\mathscr{R}_i\}_{i\in\mathrm{X}}$ and $\mathsf{Test}(\mathrm{B} \to \mathrm{C}) \ni \mathsf{T}_{\mathrm{Y}} = \{\mathscr{T}_j\}_{j\in\mathrm{Y}}$, then

$$\mathsf{Test}(\mathrm{A} \to \mathrm{C}) \ni (\mathsf{TR})_{\mathrm{X}\times\mathrm{Y}} \coloneqq \{\mathscr{T}_j\mathscr{R}_i\}_{(i,j)\in\mathrm{X}\times\mathrm{Y}}.$$

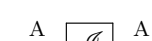
Similarly, for $\mathsf{R}_{\mathrm{X}} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ and $\mathsf{T}_{\mathrm{Y}} \in \mathsf{Test}(\mathrm{C} \to \mathrm{D})$,

$$(\mathsf{R} \otimes \mathsf{T})_{\mathrm{X}\times\mathrm{Y}} \coloneqq \{\mathscr{R}_i \otimes \mathscr{T}_j\}_{(i,j)\in\mathrm{X}\times\mathrm{Y}}.$$

The set of events of tests in $\mathsf{Test}(\mathrm{A} \to \mathrm{B})$ is denoted by $\mathsf{Transf}(\mathrm{A} \to \mathrm{B})$. By the properties of sequential and parallel composition of tests, one can easily derive associativity of sequential and parallel composition of events, as well as the analogue of Eq. (2.1). We will use for events the same diagrammatic notation as for tests:



.

As for the identity test, also the identity event will be often omitted: 

.

2.2 Coarse graining

For every test $\mathsf{T}_\mathrm{X} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ with $\mathsf{T}_\mathrm{X} = \{\mathscr{T}_i\}_{i\in\mathrm{X}}$, and every disjoint partition $\{\mathrm{X}_j\}_{j\in\mathrm{Y}}$ of $\mathrm{X} = \bigcup_{j\in\mathrm{Y}} \mathrm{X}_j$, one can define the *coarse graining* that maps T_X to $\mathsf{T}'_\mathrm{Y} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$, with $\mathsf{T}'_\mathrm{Y} = \{\mathscr{T}'_j\}_{j\in\mathrm{Y}}$. We define $\mathscr{T}_{\mathrm{X}_j} := \mathscr{T}'_j$. This mathematical map represents a test where some outcome is read, and part of the information is forgotten. Consider, e.g. the roll of a die. The system is the die, the state is the upper face. A roll will change the state and produce an outcome which is the reading of the upper face. A coarse graining is e.g. the test corresponding to a roll of the die with a device only announcing whether the upper face is even or odd.

Parallel and sequential composition distribute over coarse graining:

$$\mathscr{T}_{\mathrm{X}_j} \otimes \mathscr{R}_k = (\mathscr{T} \otimes \mathscr{R})_{\mathrm{X}_j\times\{k\}},$$
$$\mathscr{A}_l\mathscr{T}_{\mathrm{X}_j}\mathscr{B}_k = (\mathscr{A}\,\mathscr{T}\,\mathscr{B})_{\{l\}\times\mathrm{X}_j\times\{k\}}.$$

Notice that for every test $\mathsf{T}_\mathrm{X} \in \mathsf{Test}(\mathrm{A} \to \mathrm{B})$ there exists the singleton test $\mathsf{T}'_* := \{\mathscr{T}_\mathrm{X}\}$. One can easily prove that the identity event $\mathscr{I}_\mathrm{A}$, such that $\mathsf{I}^\mathrm{A} = \{\mathscr{I}_\mathrm{A}\}$, satisfies $\mathscr{I}_\mathrm{B}\mathscr{T} = \mathscr{T}\mathscr{I}_\mathrm{A}$ for every event $\mathscr{T} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$. Similarly, for $\mathsf{S}_\mathrm{AB} = \{\mathscr{S}_\mathrm{AB}\}$ we have $\mathscr{S}_\mathrm{BA}\mathscr{S}_\mathrm{AB} = \mathscr{I}_\mathrm{AB}$. The collection of events of a theory Θ will be denoted by $\mathsf{Ev}(\Theta)$.

2.3 Probabilistic theories

Operational theories as we defined them allow one to *describe* a cluster of processes in parallel and in a sequence, but at the present stage they are devoid of any predictive power. In the present section we introduce new structures that make an operational theory predictive. These come in the form of rules for assessing *probabilities* of events. An operational theory is an *operational probabilistic theory* (OPT) if the tests $\mathsf{Test}(\mathrm{I} \to \mathrm{I})$ are probability distributions: $1 \geq \mathscr{T}_i = p_i \geq 0$, so that $\sum_{i\in\mathrm{X}} p_i = 1$, and given two tests $\mathsf{S}_\mathrm{X}, \mathsf{T}_\mathrm{Y} \in \mathsf{Test}(\mathrm{I} \to \mathrm{I})$ with $\mathscr{S}_i = p_i$ and $\mathscr{T}_i = q_i$, the following identities hold

$$\mathscr{S}_i \otimes \mathscr{T}_j = \mathscr{S}_i\mathscr{T}_j := p_iq_j,$$
$$\mathscr{T}_{\mathrm{X}_j} := \sum_{i\in\mathrm{X}_j} p_i,$$

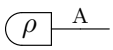
meaning that events in the same test are mutually exclusive and events in different tests of system I are independent. While it is immediate that $1 \in \mathsf{Transf}(\mathrm{I} \to \mathrm{I})$, since 1 is the only singleton test, we will assume that $0 \in \mathsf{Transf}(\mathrm{I} \to \mathrm{I})$.

An OPT Θ is specified by the collections $(\mathsf{Sys}(\Theta), \mathsf{T}(\Theta))$, along with rules to calculate probabilities for every test $\mathsf{Test}(\mathrm{I} \to \mathrm{I})$

2.4 States and effects

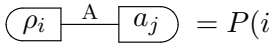
Two special families of of events are those of type $\mathsf{Transf}(\mathrm{I} \to \mathrm{A})$—from a trivial system to a non-trivial one—and $\mathsf{Transf}(\mathrm{A} \to \mathrm{I})$—transforming a non-trivial system to a trivial one. The first type can be interpreted as a process where a system is initialised, i.e. a procedure

of state preparation. Notice that the preparation might not prepare a unique state, but could result in a collection of different preparations $\{\rho_i\}_{i\in \mathrm{X}}$, heralded by some outcome $i \in \mathrm{X}$. We call these tests *preparation-tests*. The second type of test can be interpreted as a measurement whose aftermath is completely disregarded. This kind of test is called an *observation-test*. We will denote the sets $\mathsf{Transf}(\mathrm{I} \to \mathrm{A})$ and $\mathsf{Transf}(\mathrm{A} \to \mathrm{I})$ by the symbols $\mathsf{St}(\mathrm{A})$ and $\mathsf{Eff}(\mathrm{A})$, respectively. Events in $\mathsf{St}(\mathrm{A})$ are called *states*, and denoted by lower-case greek letters, e.g. ρ, while events in $\mathsf{Eff}(\mathrm{A})$ are called *effects*, and denoted by lower-case latin letters, e.g. a. The following diagrammatic notation, where we denote states and effects by the symbols



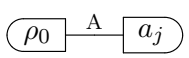
respectively, turns out to be very useful and will be used extensively in the course.

As a consequence of the above definitions, we have that a sequence made of a preparation-test and an observation test is a joint probability distribution:



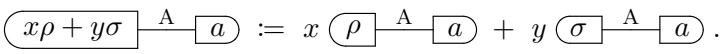
$= P(i, j \mid \rho, a).$

If we consider a single event $\rho_0 \in \mathsf{St}(\mathrm{A})$ in a preparation test, we then have a *functional* that maps every effect $a_j \in \mathsf{Eff}(\mathrm{A})$ to a real number

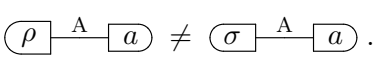
$0 \leq$ 

$\leq 1\ .$

Given two states $\rho, \sigma \in \mathsf{St}(\mathrm{A})$ one can then formally define their linear combination $x\rho + y\sigma$ as the functional on effects $a \in \mathsf{Eff}(\mathrm{A})$ given by



The set $\mathsf{St}(\mathrm{A}) \coloneqq \mathsf{Transf}(\mathrm{I} \to \mathrm{A})$ is then a spanning subset of a real vector space $\mathsf{St}(\mathrm{A})_{\mathbb{R}}$ of real functionals on effects. On the other hand, $\mathsf{Eff}(\mathrm{A})$ is a *separating* set of positive linear functionals on $\mathsf{St}(\mathrm{A})$, namely for every pair of different states $\rho, \sigma \in \mathsf{St}(\mathrm{A})$ there exists an effect $a \in \mathsf{Eff}(\mathrm{A})$ such that



One can prove that, by the above property, $\mathsf{Eff}(\mathrm{A})$ spans the dual space $\mathsf{St}(\mathrm{A})^*_{\mathbb{R}} =: \mathsf{Eff}(\mathrm{A})_{\mathbb{R}}$. The dimension D_{A} of $\mathsf{St}(\mathrm{A})_{\mathbb{R}}$ (which is the same as that of $\mathsf{Eff}(\mathrm{A})_{\mathbb{R}}$) is called *size* of system A. Using the properties of parallel composition, one can easily prove that $D_{\mathrm{AB}} \geq D_{\mathrm{A}} D_{\mathrm{B}}$, and that in any OPT Θ, I is the unique system with unit size $D_{\mathrm{I}} = 1$.

2.5 Transformations

Let us now consider a general event $\mathscr{A} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ of an OPT Θ. By definition of sequential composition, given a state $\rho \in \mathsf{St}(\mathrm{A}) = \mathsf{Transf}(\mathrm{I} \to \mathrm{A})$, we have that

$\in \mathsf{Transf}(\mathrm{I} \to \mathrm{B}) = \mathsf{St}(\mathrm{B}).$

This means that an event $\mathscr{A} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ transforms input states of system A to output states of system B. It is possible to prove that an event defines a linear transformation from the real vector space $\mathsf{St}(\mathrm{A})_{\mathbb{R}}$ to the real vector space $\mathsf{St}(\mathrm{B})_{\mathbb{R}}$. Events will then be often also called *transformations*[1]. A test is then a collection of possible transformations, labelled by outcomes that can be thought of as the reading of a pointer, heralding the occurrence of a specific event within the test. Coarse graining in an OPT corresponds to the *sum* of linear maps:

$$\mathscr{T}_{\mathrm{X}_j} = \sum_{i \in \mathrm{X}_j} \mathscr{T}_i.$$

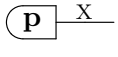
A *singleton test*, i.e. a test with a unique outcome (such as the full coarse graining of any test) is called *deterministic*, and its (unique) event is called a *channel*. A channel $\mathscr{C} \in \mathsf{Transf}(\mathrm{A} \to \mathrm{B})$ is *reversible* if there exists a channel $\mathscr{D} \in \mathsf{Transf}(\mathrm{B} \to \mathrm{A})$ such that $\mathscr{D}\mathscr{C} = \mathscr{I}_{\mathrm{A}}$ and $\mathscr{C}\mathscr{D} = \mathscr{I}_{\mathrm{B}}$.

2.6 Classical information theory

Classical information theory is a special Operational Probabilistic Theory. As such, it is specified by a class of systems and their possible tests. In practice, classical information carrying systems can be thought of as physical memory cells with n internal levels, each one encoding one symbol from an alphabet with n symbols. The theoretical description of classical systems X is thus given by alphabets, such as $\mathrm{X} \leftrightarrow \{x_1, x_2, \ldots, x_n\}$ having an integer number of characters n, every character labelling one level. The *type* of a system X is the cardinality n of the corresponding alphabet. For example $\mathrm{X} = \{0, 1, 2\}$ corresponds to a system of type $n = 3$. For the purpose of information encoding, all systems of the same type are equivalent, independently of the specific alphabet. For example, encoding in the alphabet $\{A, B\}$ is completely equivalent to encoding in $\{0, 1\}$.

The encoding of a symbol in a system corresponds to the preparation of a special *state* of the system. The preparation may occur according to a probabilistic algorithm. Consequently, it is convenient to consider states of a system X of type n as (generally sub-normalised) probability distributions $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$ over n values: $\sum_{i=1}^n p_i \leq 1$. A sub-normalised distribution $\mathbf{p}$ corresponds to a preparation event that might occur within a preparation test, with probability $\mathbb{P}_{\mathrm{X}}[\mathbf{p}] = \sum_{i=1}^n p_i$. The state $\mathbf{p}$ is *deterministic* if and only if it is *normalised*, namely $\sum_{i=1}^n p_i = 1$, i.e. it occurs with certainty when the corresponding preparation test is performed. In the course of these lectures, we will only deal with normalised states.

A state $\mathbf{p}$ of system X can be represented by the following diagram



(2.2)

Notice that not only the particular alphabet is irrelevant for evaluation of the information content of a system—what matters being only the cardinality of the alphabet—but even

[1]The notion of transformation is actually more complicate than the one reported here, but this will be sufficient for the theories of interest in the present course, i.e. classical and quantum theory.

the details of the encoding procedure are irrelevant. For example, is not important if a symbol is produced by flipping a coin or by rolling a die and recording the parity of the upper face. Indeed, the information content of a symbol is fully captured by the probability distribution $\mathbf{p}$ alone.

Every state $\mathbf{p}$ of a system of type n is then a probability distribution over a finite alphabet $\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$, namely a vector with n non-negative real entries bounded by 1. The set of states of system X is then given by

$$\mathsf{St}(\mathrm{X}) = \{\mathbf{p} \in \mathbb{R}^n \mid \sum_{i=1}^{n} p_i \leq 1,\ \forall i\ p_i \geq 0\}, \tag{2.3}$$

while the set of deterministic states is $\mathsf{St}(\mathrm{X})_1 \subseteq \mathsf{St}(\mathrm{X})$, defined as

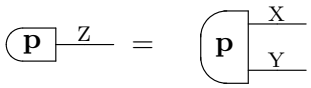
$$\mathsf{St}(\mathrm{X})_1 = \{\mathbf{p} \in \mathbb{R}^n \mid \sum_{i=1}^{n} p_i = 1,\ \forall i\ p_i \geq 0\}. \tag{2.4}$$

The *state space* of a system X, instead, is $\mathsf{St}(\mathrm{X})_{\mathbb{R}} = \mathbb{R}^n$. A basis of such space is given by the vectors $\mathbf{e}_i$ with components $(\mathbf{e}_i)_j = \delta_{ij}$. Notice that for every $1 \leq i \leq n$ one has

$$\mathbf{e}_i \in \mathsf{St}(\mathrm{X})_1 \subseteq \mathsf{St}(\mathrm{X}). \tag{2.5}$$

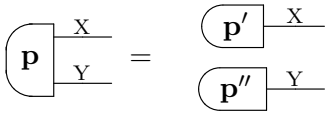
Summarising what we observed so far, from the point of view of classical information theory, the encoding of a symbol is completely described by a random variable.

The first purpose of classical information theory is to quantify the amount of information that a given preparation encodes in a system. The quantification of information typically pertains a file or string. For this reason, it is important to specify the way in which systems compose into composite systems, carrying strings of symbols. Notice that strings of symbols from an alphabet X of type n of a given length L can be thought of as symbols from a new alphabet X^L of type n^L. More generally, composing two systems, say X of type n and Y of type m, *in parallel* allows one to encode messages in a new system Z whose alphabet has mn characters, namely $\mathrm{XY} = \mathrm{Z}$ of type mn. The alphabet of a composite system is made of pairs of symbols (or more generally by strings of symbols, if the system is made of more than two components), each one from the alphabet of the corresponding component system. Their probability distribution can in general exhibit correlations, and will be diagrammatically represented as follows



(2.6)

Factorised distributions, such as $p_{mn} = p'_m p''_n$, do not have correlations, and represent *independent* preparations of the subsystems. We will represent such special preparations as follows



.

Notice that Each symbol $x_{i_j} \in \mathrm{X}_j$, in a string of a composite system made of equivalent components $\mathrm{X}_1\mathrm{X}_2 \dots \mathrm{X}_L = \mathrm{X}^L$, is distributed according to a probability distribution $\mathbf{p}^{(j)}$—the marginal distribution—that can be the same for every symbol, for example the sheer statistics of symbols in a long string (this however is too simplistic a model in most cases, because it completely neglects correlations), or may represent some prior knowledge, e.g. knowing that "the symbol x_j occurs at position j within a text that is written in Greek".

Let us now consider transformations in classical information theory. From the framework of OPTs we know that transformations of type $\mathrm{X} \to \mathrm{Y}$ with X of type n and Y of type m are in correspondence with linear maps from $\mathsf{St}(\mathrm{X})_{\mathbb{R}} = \mathbb{R}^n$ to $\mathsf{St}(\mathrm{Y})_{\mathbb{R}} = \mathbb{R}^m$. A transformation in $\mathsf{Transf}(\mathrm{X} \to \mathrm{Y})$ is then represented by a $m \times n$ real matrix (C_{ij}). But what matrices do actually represent an event? Let us start considering deterministic transformations, namely those that belong to singleton tests. Let C be such a transformation. Then, reminding Eq. (2.5) we have $C\mathbf{e}_i \in \mathsf{St}(\mathrm{Y})_1$. Then $\mathbf{p}_i := C\mathbf{e}_i$, that represents the i-th column of the matrix (C_{ji}), is a probability distribution, namely $p_j = (\mathbf{p})_j = (C\mathbf{e}_i)_j = C_{ji} \geq 0$. Moreover, summing over the row index j one has $\sum_{j=1}^m C_{ji} = 1$. A matrix C with these properties, i.e. $C_{ji} \geq 0$ and $\sum_{j=1}^m C_{ji} = 1$ for all i, is called a *stochastic matrix*.

The columns of a stochastic matrix can be interpreted as conditional probability distributions $C_{ji} = p(j|i)$: they represent the probabilities that the channel provides output symbols j given that the input symbol was i.

Finally, what are the most general events E of the theory, i.e. transformations that occur within a test? First of all, events must map states to states. This implies that the corresponding matrices (E_{ji}) must have positive entries, since E_{ji} is the j-th component of the vector $E\mathbf{e}_i$, which must be a state. Moreover, the general (sub-normalized) states $E\mathbf{e}_i$ must belong to the set $\mathsf{St}(\mathrm{Y})$ characterised by Eq. (2.3), namely

$$\sum_{j=1}^{m} E_{ji} \leq 1.$$

General transformations then correspond to *sub-stochastic matrices*, and a test is a collection of transformations whose sum is a stochastic matrix.

Chapter 3

Lecture 3: Classical systems and Random variables

3.1 Random variables

Since we defined systems of classical information theory as alphabets and their states as random variables, we now take a closer look at the notion of random variable. A random variable X can take values in a set called *range* of X and denoted as $\mathsf{Rng}(\mathrm{X}) = \{x_1, x_2, \ldots, x_n\}$. The random variable is also defined by the probability distribution $\mathbf{p} = \{p(x_1), p(x_2), \ldots, p(x_n)\}$. More formally, in order to define a random variable we need to introduce the notion of *probability space*, i.e. a triple $(\Omega, \mathscr{A}, \mu)$ where $\mathscr{A}$ is a σ-algebra of subsets of the set Ω, namely a family of subsets such that

1. $\emptyset, \Omega \in \mathscr{A}$,
2. $A \in \mathscr{A} \Rightarrow \Omega \setminus A \in \mathscr{A}$,
3. $\{A_i\}_{i\in \mathrm{I}} \subseteq \mathscr{A} \Rightarrow \bigcup_{i\in \mathrm{I}} A_i \in \mathscr{A}$, $\forall \mathrm{I} \subseteq \mathbb{N}$.

and μ is a *probability measure*, namely a function $\mu : \mathscr{A} \to \mathbb{R}$ that satisfies

1. $\forall A \in \mathscr{A}$, $\mu(A) \geq 0$,
2. $\mu(\Omega) = 1$,
3. $\{A_i\}_{i\in \mathrm{I}} \subseteq \mathscr{A}$, $A_i \cap A_j = \emptyset \Rightarrow \mu\left(\bigcup_{i\in \mathrm{I}} A_i\right) = \sum_{i\in \mathrm{I}} \mu(A_i)$, $\forall \mathrm{I} \subseteq \mathbb{N}$.

Then $\mu(A)$ is called *probability* of A. The probability space (as any measurable space) is called *complete* if for all $A \in \mathscr{A}$ with $\mu(A) = 0$ and all $B \subseteq A$ one has $B \in \mathscr{A}$. Every probability space can be completed by including all subsets B of null-probability sets A in the extended σ-algebra $\bar{\mathscr{A}}$ and extending the probability measure μ to μ^* which amounts to 0 on the new sets B.

A random variable is a measurable function X from a complete probability space $(\Omega, \mathscr{A}, \mu)$ to the measurable space $(\mathbb{R}^n, \mathscr{B})$, where $\mathscr{B}$ is the σ-algebra of Borel sets. We recall that a function $\mathrm{X} : \Omega \to \mathbb{R}^n$ is measurable if

$$\mathrm{X}^{-1}(B) \in \mathscr{A}, \quad \forall B \in \mathscr{B}.$$

Notice that a random variable *pushes forward* the probability measure to the Borel sets of $\mathbb{R}^n$, by the probability measure $\mathbb{P}_{\mathrm{X}}$ defined as

$$\mathbb{P}_{\mathrm{X}}(B) := \mu[\mathrm{X}^{-1}(B)]$$

More precisely, we construct $\mathbb{P}_{\mathrm{X}}$ as follows. Let us first define $\mathbf{a} < \mathbf{b}$ if $a_i < b_i$ for all $1 \leq i \leq n$, and then introduce the measurable sets $\mathscr{A} \ni A_{\mathbf{r}} := \{\omega \in \Omega | \mathrm{X}(\omega) < \mathbf{r}\} = \mathrm{X}^{-1}(\{\mathbf{a} \in \mathbb{R}^n | \mathbf{a} < \mathbf{r}\})$. The *cumulative* function $F_{\mathrm{X}}(\mathbf{r}) := \mu(A_{\mathbf{r}}) = \mathbb{P}_{\mathrm{X}}(\mathrm{X} < \mathbf{r})$ defines a measure on $\mathbb{R}^n$, whose completion is the Lebesgue-Stieltjes measure induced by F_{X}.

The case of a discrete probability space is a special case, corresponding to a cumulative function F_{X} that is piecewise constant. In the case of discrete random variables it is sufficient to consider the case where $\mathsf{Rng}(\mathrm{X}) \subseteq \mathbb{N}$, as we will do in the following—unless otherwise stated.

Example 3.1. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ and $\mu(\{f\}) = \frac{1}{6}$ for every face f of the dice. Then we can define $\mathrm{X}(\boxdot) = 1$, $\mathrm{X}(\boxdot) = 2$, $\mathrm{X}(\boxdot) = 3$, $\mathrm{X}(\boxdot) = 4$, $\mathrm{X}(\boxdot) = 5$, $\mathrm{X}(\boxdot) = 6$, and the corresponding cumulative function is

$$F_{\mathrm{X}}(r) = \begin{cases} 0 & -\infty < r < 1, \\ \frac{1}{6} & 1 \leq r < 2, \\ \frac{1}{3} & 2 \leq r < 3, \\ \frac{1}{2} & 3 \leq r < 4, \\ \frac{2}{3} & 4 \leq r < 5, \\ \frac{5}{6} & 5 \leq r < 6, \\ 1 & 6 \leq r < +\infty. \end{cases}$$

Example 3.2. Let $\Omega = \{H, T\}$ with $\mu(\{H\}) = p$ and $\mu(\{T\}) = 1 - p$. Then we can define $\mathrm{X}(H) = 0$, $\mathrm{X}(T) = 1$, and the corresponding cumulative function is

$$F_{\mathrm{X}}(r) = \begin{cases} 0 & -\infty < r < 0, \\ p & 0 \leq r < 1, \\ 1 & 1 \leq r < +\infty, \end{cases}$$

and the σ-algebra of Ω is $\mathscr{A} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.

Let us now consider a measurable function $f : \mathbb{R}^n \to \mathbb{R}^m$. Then, $f(\mathrm{X})$ is another random variable: $\mathrm{Y} \coloneqq f \circ \mathrm{X} : \Omega \to \mathbb{R}^m$. The measurable sets defining the cumulative function in this case are $B_{\mathbf{r}} = \{\omega \in \Omega | f(X(\omega)) < \mathbf{r}\} = X^{-1}[f^{-1}(\{\mathbf{a} \in \mathbb{R}^m | \mathbf{a} < \mathbf{r}\})]$. The cumulative function $F_{f(\mathrm{X})}$ is then $F_{f(\mathrm{X})}(\mathbf{r}) = \mu(B_{\mathbf{r}}) = \mathbb{P}_{\mathrm{X}}[f(\mathrm{X}) < \mathbf{r}] = \mathbb{P}_{f(\mathrm{X})}[\mathrm{Y} < \mathbf{r}]$.

Example 3.3. Let $f(x) = x^2$. Then $F_{f(\mathrm{X})}(y) = \mathbb{P}_{\mathrm{X}}[\mathrm{X}^2 < y]$. In this case we have

$$F_{f(\mathrm{X})}(y) = \begin{cases} 0 & y < 0, \\ \mathbb{P}_{\mathrm{X}}[\mathrm{X}^2 < y] & y \geq 0. \end{cases}$$

Since $\mathbb{P}_{\mathrm{X}}[\mathrm{X}^2 < y] = \mathbb{P}_{\mathrm{X}}[|\mathrm{X}| < \sqrt{y}] = \mathbb{P}_{\mathrm{X}}[-\sqrt{y} < |\mathrm{X}| < \sqrt{y}]$, we conclude that

$$F_{f(\mathrm{X})}(y) = F_{\mathrm{X}}(\sqrt{y}) - F_{\mathrm{X}}(-\sqrt{y}).$$

For a discrete random variable X the probability distribution $\mathbf{p}$ is defined by $p_i := \mathbb{P}_{\mathrm{X}}(\mathrm{X} = x_i)$. With a slight abuse of notation, we will write $\mathbb{P}_{\mathrm{X}}(x)$ instead of $\mathbb{P}_{\mathrm{X}}(\{x\})$, thus we have $p_i = \mathbb{P}_{\mathrm{X}}(\mathrm{X} = x_i) = \mathbb{P}_{\mathrm{X}}(x_i)$. In the following we will interchangeably use the notation $\mathbf{p}$ or $\mathbb{P}_{\mathrm{X}}$, depending on the context.

Thus, summarising, we have

Alphabet	Σ	a	b	...	z
Range	$\mathsf{Rng}(\mathrm{X})$	1	2	...	26
Probability function	$\mathbb{P}_{\mathrm{X}}$	$\mathbb{P}_{\mathrm{X}}(\mathrm{X}=1)$	$\mathbb{P}_{\mathrm{X}}(\mathrm{X}=2)$	...	$\mathbb{P}_{\mathrm{X}}(\mathrm{X}=26)$
Probability distribution	$\mathbf{p}$	p_1	p_2	...	p_{26}

Expectations

Given a random variable X with probability space $(\Omega, \mathscr{A}, \mu)$ and measure space $(\mathbb{R}^n, \mathscr{B})$, we define the *expectation value*, or simply *expectation* (sometimes called *mean* or *average*) of $f(\mathrm{X})$ for a measurable $f : \mathbb{R}^n \to \mathbb{R}^m$ as follows

$$\mathbb{E}[f(\mathrm{X})] = \int_\Omega d\mu(\omega) f[\mathrm{X}(\omega)] := \int_{\mathbb{R}^n} dF_{\mathrm{X}}(\mathbf{r}) f(\mathbf{r}), \tag{3.1}$$

where the latter integral is defined in the Lebesgue-Stieltjes sense, as F_{X} is monotonic and right continuous. Notice that

$$\mathbb{P}_{\mathrm{X}}[A] = \mathbb{E}[\chi_A(\mathrm{X})], \tag{3.2}$$

where $\chi_A(\mathbf{r})$ is the *characteristic function* of the set A, i.e. $\chi_A(\mathbf{r}) = 1$ for $\mathbf{r} \in A$ and $\chi_A(\mathbf{r}) = 0$ for $\mathbf{r} \notin A$. If $f(x) \geq 0$ for every $x \in \mathsf{Rng}(\mathrm{X})$, then clearly $\mathbb{E}[f(\mathrm{X})] \geq 0$. Notice that, given two functions f, g such that $f(\mathbf{r}) \leq g(\mathbf{r})$, one clearly has $\mathbb{E}[f(\mathrm{X})] \leq \mathbb{E}[g(\mathrm{X})]$. Indeed, $g(\mathbf{r}) - f(\mathbf{r}) \geq 0$, and $0 \leq \mathbb{E}[(g-f)(\mathrm{X})] = \mathbb{E}[g(\mathrm{X})] - \mathbb{E}[f(\mathrm{X})]$. An expectation playing a crucial role is the following

$$\sigma_{\mathrm{X}}^2 := \mathbb{E}[(\mathrm{X} - \mathbb{E}(\mathrm{X}))^2] = \mathbb{E}(\mathrm{X}^2) - \mathbb{E}(\mathrm{X})^2. \tag{3.3}$$

The quantity σ_{X} is called *root mean square* of X, and σ_{X}^2 is the *variance* of X.

In the case of a discrete random variable we have

$$\mathbb{E}[f(\mathrm{X})] = \sum_{\omega\in\Omega} \{\mu(\omega) f[X(\omega)]\} = \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} \{\mathbb{P}_{\mathrm{X}}(x_i) f(x_i)\} = \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} \{p_i f(x_i)\}.$$

Independent random variables

Let $\mathscr{A}_1$ and $\mathscr{A}_2$ be two σ-algebras $\mathscr{A}_i \subseteq \mathscr{A}$ in a measurable space $(\Omega, \mathscr{A}, \mu)$. We say that $\mathscr{A}_1$ and $\mathscr{A}_2$ are *independent* if $\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2)$ for all $A_1 \in \mathscr{A}_1$ and $A_2 \in \mathscr{A}_2$. Let now Z be a random variable with range $\mathsf{Rng}(\mathrm{Z}) = \mathsf{Rng}(\mathrm{X}) \times \mathsf{Rng}(\mathrm{Y}) \subseteq \mathbb{R}^m \times \mathbb{R}^n$. We can define the two projections $\Pi_{\mathrm{X}} : \mathsf{Rng}(Z) \to \mathsf{Rng}(\mathrm{X})$ and $\Pi_{\mathrm{Y}} : \mathsf{Rng}(\mathrm{Z}) \to \mathsf{Rng}(\mathrm{Y})$ defined by $\Pi_{\mathrm{X}}(x, y) = x$ and $\Pi_{\mathrm{Y}}(x, y) = y$. We say that X and Y are *independent random*

variables if both projections are measurable and $\mathrm{Z}^{-1}[\Pi_{\mathrm{X}}^{-1}(\mathscr{B}_{\mathrm{X}})]$ and $\mathrm{Z}^{-1}[\Pi_{\mathrm{Y}}^{-1}(\mathscr{B}_{\mathrm{Y}})]$ are independent σ-algebras with respect to the measure μ_{Z}, where $\mathscr{B}_{\mathrm{X}}$ and $\mathscr{B}_{\mathrm{Y}}$ are the Borel σ-algebras in $\mathsf{Rng}(\mathrm{X})$ and $\mathsf{Rng}(\mathrm{Y})$, respectively. The measures $\mathbb{P}_{\mathrm{X}}$ and $\mathbb{P}_{\mathrm{Y}}$ are given by $\mathbb{P}_{\mathrm{X}}(A) := \mathbb{P}_{\mathrm{Z}}[\Pi_{\mathrm{X}}^{-1}(A)]$, and $\mathbb{P}_{\mathrm{Y}}(B) := \mathbb{P}_{\mathrm{Z}}[\Pi_{\mathrm{Y}}^{-1}(B)]$. For discrete random variables this definition reduces to the intuitive notion given in the following.

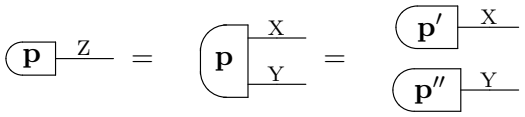
Definition 3.4 (Marginal distribution)**.** Let Z be a discrete random variable with $\mathsf{Rng}(Z) = \mathsf{Rng}(\mathrm{X}) \times \mathsf{Rng}(\mathrm{Y})$. The *marginal distributions* of X and Y are the probability distributions $\mathbb{P}_{\mathrm{X}}(x) := \sum_{y \in \mathsf{Rng}(\mathrm{Y})} \mathbb{P}_{\mathrm{Z}}(x, y)$ and $\mathbb{P}_{\mathrm{Y}}(y) := \sum_{x \in \mathsf{Rng}(\mathrm{X})} \mathbb{P}_{\mathrm{Z}}(x, y)$, respectively.

The definition is generalised to $\mathsf{Rng}(Z) = \mathsf{Rng}(\mathrm{X}^{(1)}) \times \mathsf{Rng}(\mathrm{X}^{(2)}) \times \cdots \times \mathsf{Rng}(\mathrm{X}^{(N)})$ for finite N, as an easy exercise.

Definition 3.5 (Independent random variables)**.** Let Z be a discrete random variable with $\mathsf{Rng}(Z) = \mathsf{Rng}(\mathrm{X}) \times \mathsf{Rng}(\mathrm{Y})$. We say that X and Y are *independent* if $\mathbb{P}_{\mathrm{Z}}(x, y) = \mathbb{P}_{\mathrm{X}}(x)\mathbb{P}_{\mathrm{Y}}(y)$, where $\mathbb{P}_{\mathrm{X}}$ and $\mathbb{P}_{\mathrm{Y}}$ are the marginal distributions of $\mathbb{P}_{\mathrm{Z}}$.

Also this definition can be easily generalised to the case of N independent random variables for arbitrary finite N.

We remind a remark that we already made in the previous lecture, i.e. that random variables Z with $\mathsf{Rng}(\mathrm{Z}) = \mathsf{Rng}(\mathrm{X}) \times \mathsf{Rng}(\mathrm{Y})$ represent states of composite systems in our classical information theory. Independent random variables play a special role in the correspondence between classical states and random variables, since they represent *independent classical preparations*, represented diagrammatically as follows



(3.4)

where $\mathbf{p}$, $\mathbf{p}'$, and $\mathbf{p}''$ denote the probability distributions $\mathbb{P}_{\mathrm{Z}}$, $\mathbb{P}_{\mathrm{X}}$, and $\mathbb{P}_{\mathrm{Y}}$, respectively. On the other hand, a state of the composite system $\mathrm{Z} = \mathrm{XY}$ generally corresponds to a joint probability distribution $\mathbb{P}_{\mathrm{Z}}(x, y) \neq \mathbb{P}_{\mathrm{X}}(x)\mathbb{P}_{\mathrm{Y}}(y)$ exhibiting *correlations*.

3.1.1 Inequalities and probability bounds

In this subsection we introduce some bounds on probabilities that will appear ubiquitously in the proofs of theorems. Some of the bounds regard *i.i.d.* variables, namely a finite number N of variables with the same range and with the same distribution, and with no correlation. These variables correspond to N copies of the same system, all prepared in the same state. Let us give a formal definition.

Definition 3.6 (I.i.d. random variables)**.** Let X be a random variable with range $\mathsf{Rng}(\mathrm{X}) = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}$ and probability distribution $\mathbb{P}_{\mathrm{X}}(\mathrm{X} = x_i) = p_i$. We define the *identical independently distributed* (i.i.d.) random variables $\mathrm{X}^{(1)}\mathrm{X}^{(2)} \ldots \mathrm{X}^{(N)}$, as a single random variable X^N, with range $\mathsf{Rng}(\mathrm{X}^N) = \mathsf{Rng}(\mathrm{X})^{\times N}$ and

$$\mathbb{P}_{\mathrm{X}^N}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) := \mathbb{P}_{\mathrm{X}}(x_{i_1})\mathbb{P}_{\mathrm{X}}(x_{i_2}) \ldots \mathbb{P}_{\mathrm{X}}(x_{i_N}). \tag{3.5}$$

In the following for every $\mathbf{i} := i_1 i_2 \dots i_N$ we define $x_{\mathbf{i}} := x_{i_1} x_{i_2} \dots x_{i_n}$. Thus $\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) = \mathbb{P}_{\mathrm{X}^N}(x_{i_1}, x_{i_2}, \dots, x_{i_n})$.

Lemma 3.7 (Chebyshev inequality 1 - Markov Inequality)**.** *For a non-negative random variable* X *and any positive real number* a*, one has*

$$\mathbb{P}_{\mathrm{X}}[\mathrm{X} \geq a] \leq \frac{\mathbb{E}(\mathrm{X})}{a}. \tag{3.6}$$

Proof. We remind (3.2), which implies that $\mathbb{P}_{\mathrm{X}}[\mathrm{X} \geq a] = \mathbb{E}[\theta(\mathrm{X} - a)]$, where $\theta(x)$ is the Heavyside step function. By definition, the function satisfies the bound

$$\theta(x - a) \leq \frac{x}{a}, \tag{3.7}$$

and thus, taking the expectation of both sides and remembering the properties of expectation values, we obtain the thesis. □

Lemma 3.8 (Chebyshev inequality 2)**.** *For a random variable* X *and any positive real number* a*, one has*

$$\mathbb{P}_{\mathrm{X}}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2 \geq a] \leq \frac{\sigma_{\mathrm{X}}^2}{a}, \quad \sigma_{\mathrm{X}} = \mathbb{E}[(\mathrm{X} - \mathbb{E}(\mathrm{X}))^2]^{\frac{1}{2}}. \tag{3.8}$$

Proof. This is just a special case of Lemma 3.7 for the random variable $(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2$. A common way of stating the Chebyshev bound is also

$$\mathbb{P}_{\mathrm{X}}[|\mathrm{X} - \mathbb{E}[\mathrm{X}]| \geq k\sigma_{\mathrm{X}}] \leq \frac{1}{k^2}.$$ □

Lemma 3.9 (Weak law of large numbers)**.** *Let* $\mathrm{X}^{(1)}, \mathrm{X}^{(2)}, \dots \mathrm{X}^{(N)}$ *be* N *i.i.d. random variables, with mean* $\mathbb{E}(\mathrm{X}) = \bar{x}$ *and variance* $\sigma_{\mathrm{X}}^2 = \sigma^2$*. Then for the random variable* $\mathrm{Z} := \frac{1}{N} \sum_{i=1}^{N} \mathrm{X}^{(i)}$ *and any positive real number* a *one has*

$$\mathbb{P}_{\mathrm{Z}}[(\mathrm{Z} - \bar{x})^2 \geq a] \leq \frac{\sigma^2}{aN}. \tag{3.9}$$

Proof. First, notice that using the definition of i.i.d. random variables one can easily find $\mathbb{E}(\mathrm{Z}) = \bar{x}$. Thus, applying the Chebyshev bound of lemma 3.8 one has

$$\mathbb{P}_{\mathrm{Z}}[(\mathrm{Z} - \bar{x})^2 \geq a] \leq \frac{\sigma_{\mathrm{Z}}^2}{a} \tag{3.10}$$

We then evaluate σ_{Z} as follows.

$$\sigma_{\mathrm{Z}}^2 = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{X}^{(i)} - \bar{x}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}(\mathrm{X}^{(i)} - \bar{x})\right)^2\right] = \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}[(\mathrm{X}^{(i)} - \bar{x})(\mathrm{X}^{(j)} - \bar{x})]$$
$$= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\mathbb{E}[(\mathrm{X}^{(i)} - \bar{x})^2] + \sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}[(\mathrm{X}^{(i)} - \bar{x})(\mathrm{X}^{(j)} - \bar{x})]\right\}$$
$$= \frac{\sigma^2}{N}.$$

Finally, substituting the last result in equation (3.10) one has the thesis. □

Lemma 3.10 (Chernoff bound)**.** *Suppose that* $\mathrm{X}^{(1)}, \mathrm{X}^{(2)} \ldots \mathrm{X}^{(N)}$ *are binary i.i.d. random variables, with range* $\mathsf{Rng}(\mathrm{X}^{(i)}) = \{0, 1\}$ *and* $\mathbb{P}_{\mathrm{X}}(0) = \frac{1}{2} + \varepsilon$*, and* $\mathbb{P}_{\mathrm{X}}(1) = \frac{1}{2} - \varepsilon$*, with* $0 \leq \varepsilon \leq \frac{1}{2}$*. Then for the random variable* $\mathrm{Z} := \sum_{i=1}^{N}\mathrm{X}^{(i)}$ *one has*

$$\mathbb{P}_{\mathrm{Z}}\left[\mathrm{Z} \geq \frac{N}{2}\right] \leq e^{-2\varepsilon^2 N}. \tag{3.11}$$

Proof. By the Markov inequality (3.6) we have

$$\mathbb{P}_{\mathrm{Z}}[\mathrm{Z} \geq c] = \mathbb{P}_{\mathrm{Z}}[e^{t\mathrm{Z}} \geq e^{tc}] \leq \frac{\mathbb{E}(e^{t\mathrm{Z}})}{e^{tc}}, \quad t > 0.$$

Since the variables $\mathrm{X}^{(i)}$ are i.i.d., we have

$$\mathbb{E}(e^{t\mathrm{Z}}) = \mathbb{E}\left(e^{t\sum_{i=1}^{N}\mathrm{X}^{(i)}}\right) = \mathbb{E}\left(\prod_{i=1}^{N}e^{t\mathrm{X}^{(i)}}\right) = \prod_{i=1}^{N}\mathbb{E}(e^{t\mathrm{X}^{(i)}}),$$

and for every variable $\mathrm{X}^{(i)}$ one has

$$\mathbb{E}(e^{t\mathrm{X}^{(i)}}) = \left(\frac{1}{2} + \varepsilon\right) + \left(\frac{1}{2} - \varepsilon\right)e^t. \tag{3.12}$$

This implies that

$$\mathbb{P}_{\mathrm{Z}}[\mathrm{Z} \geq c] \leq f(t), \quad f(t) := \frac{\left[\left(\frac{1}{2} + \varepsilon\right) + \left(\frac{1}{2} - \varepsilon\right)e^t\right]^N}{e^{tc}}. \tag{3.13}$$

For $c < N$, it is a simple exercise to find the minimum of f, which for $c = N/2$ corresponds to $t_0 = \ln\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}$. Since the bound (3.13) holds for every $t > 0$, choosing $t = t_0$ we have

$$\mathbb{P}_{\mathrm{Z}}[\mathrm{Z} \geq \tfrac{N}{2}] \leq \frac{[1 + 2\varepsilon]^N}{\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}\right)^{\frac{N}{2}}} = [1 - 4\varepsilon^2]^{\frac{N}{2}}.$$

We now apply a bound that will be recurrent, and is known as the *fundamental inequality of information theory*, namely $1 - x \leq e^{-x}$, which for $x = 4\varepsilon^2$ gives $1 - 4\varepsilon^2 \leq e^{-4\varepsilon^2}$, and

$$\mathbb{P}_{\mathrm{Z}}[\mathrm{Z} \geq \tfrac{N}{2}] \leq e^{-2\varepsilon^2 N}.$$ □

3.1.2 Convex functions

Let us now recapitulate some elementary properties of convex functions. First of all, we remind the definition of convex set and convex function.

Definition 3.11 (Convex set)**.** Let $C \subseteq \mathbb{R}^n$. The set C is *convex* if for any couple of points $\mathbf{x}_1, \mathbf{x}_2 \in C$ and every $0 \leq \lambda \leq 1$ the point $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ belongs to C.

Definition 3.12 (Epigraph)**.** Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. The *epigraph* of f is the set $\mathsf{Epi}(f) \subseteq \mathbb{R}^{n+1}$ defined as

$$\mathsf{Epi}(f) := \{(\mathbf{r}, y) \in D \times \mathbb{R} | y \geq f(\mathbf{r})\}$$

Definition 3.13 (Convex function)**.** Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. We say that f is *convex* if $\mathsf{Epi}(f)$ is convex.

Notice that if the epigraph of f is convex, given $\mathbf{x}_1, \mathbf{x}_2 \in D$ $\lambda(\mathbf{x}_1, f(\mathbf{x}_1)) + (1-\lambda)(\mathbf{x}_2, f(\mathbf{x}_2)) = (\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2)) \in \mathsf{Epi}(f)$, and thus $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in D$, namely the domain D of f must be convex.

Proposition 3.14. *A function $f : D \to \mathbb{R}$ is convex if and only if D is convex and for every $0 \leq \lambda \leq 1$ and every $\mathbf{x}_1, \mathbf{x}_2 \in D$,*

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2).$$

Proof. Let us prove necessity. Notice that for every $\mathbf{x}_1, \mathbf{x}_2 \in D$ we have $(\mathbf{x}_i, f(\mathbf{x}_i)) \in \mathsf{Epi}(f)$. If $\mathsf{Epi}(f)$ is convex, then

$$\lambda(\mathbf{x}_1, f(\mathbf{x}_1)) + (1-\lambda)(\mathbf{x}_2, f(\mathbf{x}_2)) = (\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2)) \in \mathsf{Epi}(f),$$

and by definition this implies $\lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \geq f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]$. Let us now prove sufficiency. Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \in \mathsf{Epi}(f)$. Then $f(\mathbf{x}_i) \leq y_i$. The following chain of inequalities holds

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \leq \lambda y_1 + (1-\lambda)y_2, \tag{3.14}$$

and then the point $(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda y_1 + (1-\lambda)y_2)$ belongs to $\mathsf{Epi}(f)$, which is then convex. □

Jensen's inequality

As a consequence of the properties of convex functions, Jensen's inequality holds. This crucial result is stated in the following theorem.

Theorem 3.15 (Jensen's inequality)**.** *Let $f : D \to \mathbb{R}$ be a convex function. For every set of points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq D$ and any probability distribution $(p_1, p_2, \ldots, p_n)$ (namely $p_j \geq 0$ for every $1 \leq j \leq n$ and $\sum_{i=1}^n p_i = 1$) the following inequality holds*

$$f\left(\sum_{i=1}^n \{p_i\mathbf{x}_i\}\right) \leq \sum_{i=1}^n \{p_i f(\mathbf{x}_i)\}. \tag{3.15}$$

Proof. The proof is by induction on n. We know that the thesis holds for $n = 2$. Let the thesis hold for $n = n_0$, and consider the case $n = n_0 + 1$. Let us define $q := p_{n_0+1}$, so that the l.h.s. in equation (3.15) can be written as

$$f\left(\sum_{i=1}^{n}\{p_i\mathbf{x}_i\}\right) = f\left[(1-q)\sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}\mathbf{x}_i\right\} + q\mathbf{x}_{n_0+1}\right].$$

Notice that $p_i/(1-q) \geq 0$ and $\sum_{i=1}^{n_0}\{p_i/(1-q)\} = 1$, and since the domain D is convex, the point $\sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}\mathbf{x}_i\right\}$ belongs to D. Thus by proposition 3.14 we have

$$f\left(\sum_{i=1}^{n}\{p_i\mathbf{x}_i\}\right) \leq (1-q)f\left(\sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}\mathbf{x}_i\right\}\right) + qf(\mathbf{x}_{n_0+1}).$$

Now, using the induction hypothesis, we have

$$f\left(\sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}\mathbf{x}_i\right\}\right) \leq \sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}f(\mathbf{x}_i)\right\},$$

and finally

$$f\left(\sum_{i=1}^{n}\{p_i\mathbf{x}_i\}\right) \leq (1-q)\sum_{i=1}^{n_0}\left\{\frac{p_i}{1-q}f(\mathbf{x}_i)\right\} + qf(\mathbf{x}_{n_0+1}) = \sum_{i=1}^{n_0+1}\{p_i f(\mathbf{x}_i)\},$$

which concludes the proof. □

Jensen's inequality can be immediately re-phrased in the language of discrete random variables as follows

Corollary 3.16 (Jensen inequality for random variables)**.** *Let* X *be a random variable with range* $\mathsf{Rng}(\mathrm{X}) = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$*. Let* $f : D \to \mathbb{R}$ *be a convex function, where* D *contains the convex hull of* $\mathsf{Rng}(\mathrm{X})$*. Then one has*

$$f[\mathbb{E}(\mathrm{X})] \leq \mathbb{E}[f(\mathrm{X})]. \tag{3.16}$$

Log-sum inequality

We now apply Jensen's inequality to prove the *log-sum* inequality.

Lemma 3.17 (Log-sum inequality)**.** *Let* $\{a_i\}_{i=1}^n \subseteq \mathbb{R}$ *satisfy* $a_i \geq 0$ *for all* $1 \leq i \leq n$*, and* $\{b_i\}_{i=1}^n \subseteq \mathbb{R}$ *satisfy* $b_i > 0$ *for all* $1 \leq i \leq n$*. Let us also define* $a := \sum_{i=1}^n a_i$ *and* $b := \sum_{i=1}^n b_i$*. Then the following inequality holds*

$$\sum_{i=1}^{n}\left\{a_i \log\frac{a_i}{b_i}\right\} \geq a\log\frac{a}{b}, \tag{3.17}$$

where the logarithm is taken in any basis $q > 1$*.*

Proof. First of all, we define $0 \log 0 := 0$ by continuity. Now, let us define the function $f : [0, 1] \to \mathbb{R}$ as $f(x) := x \log x$. It is easy to verify that f is convex, since it is twice differentiable in $(0, 1)$ and the second derivative is $f''(x) = 1/x \geq 0$. We can now write

$$\sum_{i=1}^{n} \left\{ a_i \log \frac{a_i}{b_i} \right\} = b \sum_{i=1}^{n} \left\{ \frac{b_i}{b} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \right\} = b \sum_{i=1}^{n} \left\{ q_i f\left(\frac{a_i}{b_i} \right) \right\}, \quad q_i := \frac{b_i}{b}.$$

Applying Jensen's inequality to f we then obtain

$$\sum_{i=1}^{n} \left\{ a_i \log \frac{a_i}{b_i} \right\} \geq b f\left(\sum_{i=1}^{n} \left\{ q_i \frac{a_i}{b_i} \right\} \right) = b f\left(\frac{a}{b} \right) = a \log \frac{a}{b}.$$

This completes the proof. □

Chapter 4

Lecture 4: Shannon entropy and typical sequences

4.1 Shannon entropy

In this section we introduce some preliminary notions that will serve the purpose of quantifying the information content of a source. We give some intuitive understanding of the main measures of information content, and derive their most relevant properties. However, we need to keep in mind that the precise theoretical and practical interpretation of such quantifications has to be found in theorems—primarily the two Shannon theorems and the data-processing inequality—which give precise operational meaning to the Shannon entropy and the mutual information.

In the following, we will often consider the probabilities of all the characters in the alphabet Σ as strictly positive. Reasoning in terms of the random variable X that maps the alphabet into $\mathsf{Rng}(\mathrm{X})$, this means that $p_x > 0$ for all $x \in \mathsf{Rng}(\mathrm{X})$, and treat the special case in which $p_x = 0$ for some x by the alphabet Σ' obtained by removing the symbol $\mathrm{X}^{-1}(x)$ from the alphabet Σ, i.e. $\Sigma' := \Sigma \setminus \mathrm{X}^{-1}(x)$. Since the range of X will be $\{1, 2, \ldots, |\Sigma|\}$, in the above case we define a new random variable X' with $\mathsf{Rng}(\mathrm{X}') = \{1, 2, \ldots, |\Sigma| - 1\}$. Finally, keep in mind that in the following we will often use the shorthand $\mathbb{P}_{\mathrm{X}}(x)$ instead of $\mathbb{P}_{\mathrm{X}}[\{x\}]$ or $\mathbb{P}_{\mathrm{X}}[\mathrm{X} = x]$.

4.1.1 Shannon information content and Shannon entropy

Definition 4.1 (Shannon information content)**.** The *Shannon information content* of the discrete random variable X is the function $h_{\mathrm{X}} : \mathsf{Rng}(\mathrm{X}) \to \mathbb{R}$ defined as

$$h_{\mathrm{X}}(x) := \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x)} \ . \tag{4.1}$$

When there will be no ambiguity in the specification of the random variable, we will omit it, and write $h(x)$ instead of $h_{\mathrm{X}}(x)$. The Shannon information content was originally called by Shannon *self-information*, and it is often referred to as *surprisal*. Indeed, the smaller $\mathbb{P}_{\mathrm{X}}(x)$, the larger is the surprisal upon occurrence of the event x.

Notice that since $0 < \mathbb{P}_{\mathrm{X}}(x) \leq 1$, one has $h_{\mathrm{X}}(x) \geq 0$ for every $x \in \mathsf{Rng}(\mathrm{X})$.

Definition 4.2 (Shannon entropy)**.** The Shannon entropy of a discrete random variable X is defined as

$$H(\mathrm{X}) := \sum_{x\in\mathsf{Rng}(\mathrm{X})} \left\{ \mathbb{P}_{\mathrm{X}}(x) \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x)} \right\} = - \sum_{x\in\mathsf{Rng}(\mathrm{X})} \{ \mathbb{P}_{\mathrm{X}}(x) \log_2 P_{\mathrm{X}}(x) \} \ . \tag{4.2}$$

Notice that $H(\mathrm{X}) = \sum_{x\in\mathsf{Rng}(\mathrm{X})} \{\mathbb{P}_{\mathrm{X}}(\mathrm{X}=x) h_{\mathrm{X}}(x)\} = \mathbb{E}[h_{\mathrm{X}}(\mathrm{X})]$, namely the Shannon entropy is equal to the Shannon information content averaged over all possible outcomes. We will sometimes write $H(\mathbf{p})$ instead of $H(\mathrm{X})$, when we need to emphasise the dependence of the Shannon entropy on the probability distribution. Indeed, notice that the definition never involves the range of X, only depending on the probability measure $\mathbb{P}_{\mathrm{X}}$. This is a property that we expect from a quantity that measures the information content of a source: it must not depend on the symbols used for the encoding.

Being the expectation of the nonnegative function $h_{\mathrm{X}}(\mathrm{X})$, the Shannon entropy is non-negative $H(\mathrm{X}) \geq 0$. In the following we adopt the convention $0 \log_2 0 = 0$, which extends the function $x \log_2 x$ by continuity in 0, and thus $H(\mathrm{X}) = 0$ for probability distributions $\mathbf{p}$ that are concentrated at one value $x_i \in \mathsf{Rng}(\mathrm{X})$, i.e. $p_j = \delta_{i,j}$.

Intuitive meaning of the information content

The actual meaning of the Shannon entropy will be given by the first Shannon theorem, in terms of compression ratio of a source. However, we can already provide a preliminary analysis of its most intuitive features.

Remark 1 (Additivity). What has $h_{\mathrm{X}}(x)$ to do with the information content? First of all, notice that if one has two independent random variables X and Y, namely with probability distributions that are factorised $\mathbb{P}_{\mathrm{XY}}(x,y) = \mathbb{P}_{\mathrm{X}}(x)\mathbb{P}_{\mathrm{Y}}(y)$, the function $h_{\mathrm{X,Y}}(x,y)$ of the corresponding states is, by definition, additive, and this is what we expect from a quantity that measures the information content: when we read two independent messages the amount of information that we gain is the sum of the two amounts of information gained by reading each of the two messages separately. The additivity property is clearly inherited by the Shannon entropy. Therefore, for two *independent* variables X and Y, we have

$$h_{\mathrm{X,Y}}(x,y) = h_{\mathrm{X}}(x) + h_{\mathrm{Y}}(y), \quad H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y}).$$

Example 4.3 (Battleship)**.** Consider a simplified battleship game, where a player must locate a submarine on a grid of 8×8 squares. Notice that all the cells of the checkerboard can be addressed by strings of 6 bits. At each round, the player can query whether the target is located in one square. They receive two possible answers: h (hit) and m (missed). At the first round the probability of finding the submarine at any square is $p_i = \frac{1}{64}$, and the Shannon information content of the event "hit" is

$$h_{(1)}(h) = \log_2 64 = 6 \text{ bits}.$$

Indeed, we have actually gained 6 bits of information, namely the 6 bits needed to specify the exact location of the submarine in the grid. On the other hand, the Shannon

information content of the event "missed" is

$$h_{(1)}(m) = \log_2 \frac{64}{63} = 0.0227 \text{ bits.}$$

Let us now play the second round. Now we have $p = \frac{1}{63}$ for all the remaining squares and $p = 0$ for the queried one. Then,

$$h_{(2)}(m) = \log_2 \frac{63}{62} = 0.0230 \text{ bits.}$$

Now, suppose that the player misses the submarine for 32 rounds. Then they gained

$$h_{(1)}(m) + h_{(2)}(m) + \cdots + h_{(32)}(m) = \log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{33}{32} = \log_2 2 = 1 \text{ bits.}$$

Indeed, the board is now divided into two parts: the queried squares and the remaining ones, that can be addressed by one bit. We know that the submarine is located in the second half, i. e. we have one bit of information. Now, it is easy to calculate that whenever the player hits the submarine, say at round k, the total information gained is

$$\begin{aligned} &h_{(1)}(m) + h_{(2)}(m) + \cdots + h_{(k-1)}(m) + h_{(k)}(h) \\ =&\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{64-k+1}{64-k} + \log_2(64-k) \\ =&\log_2 64 = 6 \text{ bits,} \end{aligned}$$

independently of k. Notice that 6 is the number of binary questions (questions with yes/no answer) needed to identify a square on the grid, e.g. by subsequent bipartitions.

Example 4.4 (Binary Shannon entropy)**.** For a binary random variable X with range $\mathsf{Rng}(\mathrm{X}) = \{0, 1\}$ and probability distribution $p(0) = p$ and $p(1) = 1 - p$, the Shannon entropy is a function of the parameter p. More precisely

$$H_2(p) := H(\{p, 1-p\}) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}. \tag{4.3}$$

From the plot of $H_2(p)$ in Fig. 4.1 we can see that the binary Shannon entropy is a concave function of p, with maximum at $p = \frac{1}{2}$, where $H_2(p) = 1$ bits. Indeed, the information that we gain from tossing a fair coin is 1 bit. On the other hand, for a biased coin the information gain from one toss is less than one bit, as one can understand analysing the extreme case in which the outcome "head" is certain: In this case we gain no information at all from one toss. More generally one can prove that $H(\mathbf{p})$ is a concave function of $\mathbf{p}$ also for $\mathbf{p} \in \mathbb{R}^n$:

$$H(q\mathbf{p}_0 + (1-q)\mathbf{p}_1) \geq qH(\mathbf{p}_0) + (1-q)H(\mathbf{p}_1).$$

This means that when mixing two random variables with probabilities q and $1 - q$, one obtains a new random variable with an information content which at least equals the average information of each variable. In other words, we have more ignorance about the

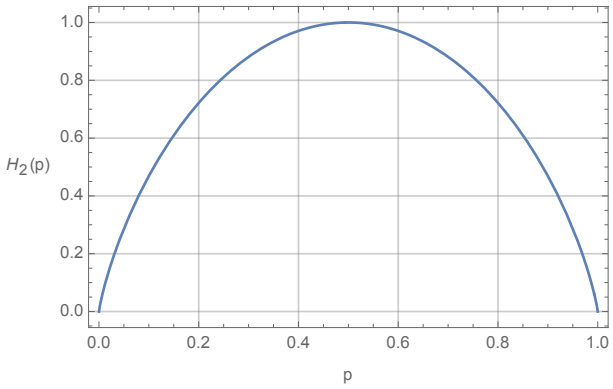


Figure 4.1 The binary entropy function $H_2(p)$ in Eq. (4.3).

new variable than the average ignorance about the original ones. For example, one can think of the new variable as obtained by the following procedure. First, a referee samples a binary variable C with $\mathsf{Rng}(\mathrm{C}) = \{0, 1\}$ and probabilities $(q, 1-q)$. Then, if $c = 0$, they sample the variable B_0 with $\mathsf{Rng}(\mathrm{B}_0) = \{0, 1\}$ and distribution $\mathbf{p}_0$, while if $c = 1$ they sample B_1 with $\mathsf{Rng}(\mathrm{B}_1) = \{0, 1\}$ and distribution $\mathbf{p}_1$. Finally, after erasing the outcome c, they produce the outcome of the second sampled variable B_c. It is then clear that in this scenario, once we read the final outcome b_c, distributed according to $q\mathbf{p}_0 + (1-q)\mathbf{p}_1$, we ignore the value of c, i.e. which of the two variables B_0 or B_1 was actually sampled in the given instance.

From what we have seen, we can say that the Shannon entropy $H(\mathrm{X})$ measures our prior uncertainty about X, or else $H(\mathrm{X})$ measures the amount of information that one gains from knowledge of X.

4.1.2 Relative entropy

We now introduce a quantity that has no immediate operational meaning, but turns out to be a powerful tool for proofs. This is the *relative entropy*, also known as *Kullback-Leibler divergence*.

Definition 4.5 (Kullback-Leibler divergence or relative entropy)**.** Let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions for random variables whose range has cardinality n. Their *Kullback-Leibler divergence*, denoted as $H(\mathbf{p}\|\mathbf{q})$, is defined as

$$H(\mathbf{p}\|\mathbf{q}) := \begin{cases} \sum_{i=1}^{n} \left\{ p_i \log_2 \frac{p_i}{q_i} \right\} & q_i > 0 \ \forall i \text{ such that } p_i > 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.4}$$

where we adopt the convention $0 \log_2 0 = 0 \log_2 \frac{0}{0} = 0 \log_2 \frac{0}{q} = 0$.

The relative entropy $H(\mathrm{X}\|\mathrm{Y})$ of two random variables X and Y having probability distributions $\mathbf{p}$ and $\mathbf{q}$, respectively, is defined as the Kullback-Leibler divergence $H(\mathbf{p}\|\mathbf{q})$. Intuitively speaking, the Kullback-Leibler divergence $H(\mathbf{p}\|\mathbf{q})$ measures how close the two

probability distributions $\mathbf{p}$ and $\mathbf{q}$ are. Nevertheless, it does not satisfy the requirements for a distance.

The main property of the Kullback-Leibler divergence is given by the following lemma

Lemma 4.6 (Gibbs' inequality)**.** *The Kullback-Leibler divergence $H(\mathbf{p}\|\mathbf{q})$ is nonnegative and vanishes only if $\mathbf{p} = \mathbf{q}$.*

Proof. If $q_i = 0$ for some i with $p_i > 0$, the inequality is trivially satisfied. We then focus on the case where $q_i > 0$ for all i such that $p_i > 0$. Let us use the fundamental inequality of information theory in the following form

$$\ln \frac{1}{x} \geq 1 - x,$$

and remind that $\ln y = \log_2 y \ln 2$, which implies

$$\log_2 \frac{1}{x} \geq \frac{1-x}{\ln 2}.$$

Now, we can write

$$H(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \geq \frac{1}{\ln 2} \sum_{i=1}^{n} \left\{ p_i \left(1 - \frac{q_i}{p_i} \right) \right\} = \frac{1}{\ln 2} \sum_{i=1}^{n} \{p_i - q_i\} = 0.$$

Since the only inequality that we used is the fundamental inequality of information theory, which is tight only for $x = 1$, we conclude that also $H(\mathbf{p}\|\mathbf{q}) = 0$ only if $\frac{p_i}{q_i} = 1$ for all i, namely $\mathbf{p} = \mathbf{q}$. □

Exercise 4.1

Let $M \in M_{m\times n}(\mathbb{R})$ be a stochastic (or Markov matrix), namely $M_{ij} \geq 0$ and $\sum_{i=1}^{m} M_{ij} = 1$ for all j. Prove that

$$H(M\mathbf{p}\|M\mathbf{q}) \leq H(\mathbf{p}\|\mathbf{q}) \tag{4.5}$$

Answer of exercise 4.1

Fix a value of i and apply the log-sum inequality to the expression

$$\sum_l M_{il} p_l \log \frac{\sum_k M_{ik} p_k}{\sum_s M_{is} q_s} \leq \sum_l M_{il} p_l \log \frac{M_{il} p_l}{M_{il} q_l} = \sum_l M_{il} p_l \log \frac{p_l}{q_l}.$$

Taking now the sum over i on both sides, we obtain $H(M\mathbf{p}\|M\mathbf{q}) \leq H(\mathbf{p}\|\mathbf{q})$.

The second property that we prove provides an interpretation of the Kullback-Leibler divergence as a sort of distance between probability distributions, despite it is not symmetric and therefore it cannot define a metric.

4.2 Bound on Shannon entropy

The relative entropy is not useful by itself, but only because other entropic quantities can be expressed in terms of it. As we will see, it is mostly a proving tool. The bounds that we prove here are important examples.

Lemma 4.7. *The Shannon entropy of a random variable* X *satisfies the following bound*

$$H(\mathrm{X}) \leq \log_2 |\mathsf{Rng}(\mathrm{X})|. \tag{4.6}$$

Moreover, $H(\mathrm{X}) = \log_2 |\mathsf{Rng}(\mathrm{X})|$ *only for the uniform distribution* $\mathbb{P}_{\mathrm{X}}(x) = \frac{1}{|\mathsf{Rng}(\mathrm{X})|}$ *for all* $x \in \mathsf{Rng}(\mathrm{X})$.

Proof. Let us set $r := |\mathsf{Rng}(\mathrm{X})|$, and define the random variable Y with $\mathsf{Rng}(\mathrm{Y}) = \mathsf{Rng}(\mathrm{X})$ and $\mathbb{P}_{\mathrm{Y}}(x) = \frac{1}{r}$ for all $x \in \mathsf{Rng}(\mathrm{X})$. If we evaluate the relative entropy $H(\mathrm{X}\|\mathrm{Y})$ we find that

$$\begin{aligned} H(\mathrm{X}\|\mathrm{Y}) &= \sum_{x \in \mathsf{Rng}(\mathrm{X})} \left\{ \mathbb{P}_{\mathrm{X}}(x) \log_2 \frac{\mathbb{P}_{\mathrm{X}}(x)}{\mathbb{P}_{\mathrm{Y}}(x)} \right\} = -H(\mathrm{X}) + \sum_{x \in \mathsf{Rng}(\mathrm{X})} \{ \mathbb{P}_{\mathrm{X}}(x) \log_2 r \} \\ &= -H(X) + \log_2 r. \end{aligned}$$

Since, by Gibbs' inequality, $H(\mathrm{X}\|\mathrm{Y}) \geq 0$, we have the first statement of the thesis. Moreover, equality holds if and only if the two probability distributions are the same, namely $\mathbb{P}_{\mathrm{X}}(x) = \frac{1}{|\mathsf{Rng}(\mathrm{X})|}$. □

4.2.1 Raw bit content

Definition 4.8 (Raw bit content)**.** The *raw bit content* of a random variable X is defined as $H_0(\mathrm{X}) := \log_2 |\mathsf{Rng}(\mathrm{X})|$.

If we neglect the "surprisal" due to our prior expectations about the outcome of a random variable, a plain count of the number of bits needed for a perfect lossless encoding of the alphabet of X is precisely its raw bit content. Notice that the raw bit content is additive: For the variable Z = XY one has

$$H_0(\mathrm{X}, \mathrm{Y}) = \log_2(|\mathsf{Rng}(\mathrm{X})|\,|\mathsf{Rng}(\mathrm{Y})|) = H_0(\mathrm{X}) + H_0(\mathrm{Y}).$$

The meaning of the raw bit content is then very intuitive: We need one string for every character in our alphabet, which means $|\mathsf{Rng}(\mathrm{X})|$ different strings. Since N bits correspond to an alphabet with 2^N symbols, $|\mathsf{Rng}(\mathrm{X})|$ strings require $H_0(\mathrm{X})$ bits for an exact encoding. The question is then: how is it possible to compress a string without loosing information?

4.3 Lossy and lossless compressors

Formally, a compression scheme $(\mathscr{E}, \mathscr{D})$ is defined in terms of a couple of channels: the encoding $\mathscr{E} : \mathsf{Rng}(\mathrm{X}) \to \bigcup_{N \in \mathbb{N}} \{0,1\}^N$ and the decoding $\mathscr{D} : \bigcup_{N \in \mathbb{N}} \{0,1\}^N \to \mathsf{Rng}(\mathrm{X})$. The length $l[\mathscr{E}(x)]$ of the encoded character $x \in \mathsf{Rng}(\mathrm{X})$ is denoted as $l_{\mathscr{E}}(x)$.

Clearly, by the simple counting argument that we used to derive the raw bit content, it is easy to see that it is impossible to make a reversible compression that reduces the size of every possible string. There are then only two kinds of compressors:

1. **Lossy compressor**. A lossy compressor compresses some strings to the same fixed length, but there are strings which after compression would be confusable. In some applications (e.g. image compression) the occurrence of confusable strings can be afforded. We will denote by δ the probability that the string belongs to the confusable set—the so-called probability of failure. The lossy compressor is practically useful if the probability of failure δ can be made small.

2. **Lossless compressor**. A lossless compressor maps all strings to a different encoding. Hence, while shortening some strings, it necessarily makes some other strings longer. The compressor is still useful if the probability δ that a string is lengthened is very small, and the probability $1-\delta$ that it is shortened is high, as long as the *average* length of encoded strings is small.

In the present chapter we will focus on *lossy* compressors. In the next chapter we will analyse *lossless* compressors.

Example 4.9. Consider an alphabet of eight characters $A = \{a, b, c, d, e, f, g, h\}$, not equally distributed. A single-character string is described by the random variable X with

$$\mathsf{Rng}(\mathrm{X}) = \{1, 2, 3, 4, 5, 6, 7, 8\} \tag{4.7}$$

$$\mathbf{p} = \{\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{3}{16}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}\}. \tag{4.8}$$

The raw bit content is $H_0(\mathrm{X}) = 3$ bits, but if we allow for a probability $\delta = \frac{1}{16}$ of loss, we can compress the subset $\{a, b, c, d\} \subset A$, loosing $\{e, f, g, h\}$. Then it is possible to compress with a small error probability to only 2 bits.

4.3.1 Smallest δ-sufficient set and essential bit content

The idea in the above example is captured by the notion of *essential bit content*, defined in the following. First of all, if we want a small probability of error—say $p_{\mathrm{err}} \leq \delta$—we must identify all possible subsets of our alphabet with probability larger than $1-\delta$. A choice of such set will correspond to the set of strings that will be shortened without loss.

Definition 4.10 (δ-sufficient sets)**.** Given a random variable X with probability distribution $\mathbb{P}_{\mathrm{X}}$, for $0 \leq \delta \leq 1$ the family of *δ-sufficient sets* of X is the set

$$\mathscr{S}_\delta(\mathrm{X}) := \{\mathsf{S} \subseteq \mathsf{Rng}(\mathrm{X}) \mid \mathbb{P}_{\mathrm{X}}[\mathrm{X} \in \mathsf{S}] \geq 1-\delta\}. \tag{4.9}$$

Now, every choice of a δ-sufficient set will lead to a small error probability (in other words, the probability of confusable strings will be smaller than or equal to δ). However, in order to have an efficient compressor, among all δ-sufficient sets we would like to find the smallest one. Indeed, the smaller the set, the smaller is the number of bits that we will use to encode it.

Definition 4.11 (Smallest δ-sufficient set)**.** Given a random variable X with probability distribution $\mathbb{P}_{\mathrm{X}}$, for $0 \leq \delta \leq 1$ the *smallest δ-sufficient set* of X is the set

$$\mathsf{S}_\delta(\mathrm{X}) := \arg \min_{\mathsf{S} \in \mathscr{S}_\delta(\mathrm{X})} |\mathsf{S}| \tag{4.10}$$

In practice the subset $\mathsf{S}_\delta(\mathrm{X})$ can be constructed by i) sorting the outcomes in order of decreasing probability, and ii) including elements in $\mathsf{S}_\delta(\mathrm{X})$ starting from the most likely, until the total probability becomes larger than $1-\delta$. Finding the smallest δ-sufficient set is not computationally easy, because it requires sorting the elements of $\mathsf{Rng}(\mathrm{X})$ according to their probability. However, once the sorting has been completed, the smallest δ-sufficient set can be nicely visualized, as in the following Fig. 4.2 where $\mathsf{Rng}(\mathrm{X}) = \{1, 2, \ldots, 7\}$ and $\mathbb{P}_{\mathrm{X}}[x] = N/2^x$, with $N = 128/127$.

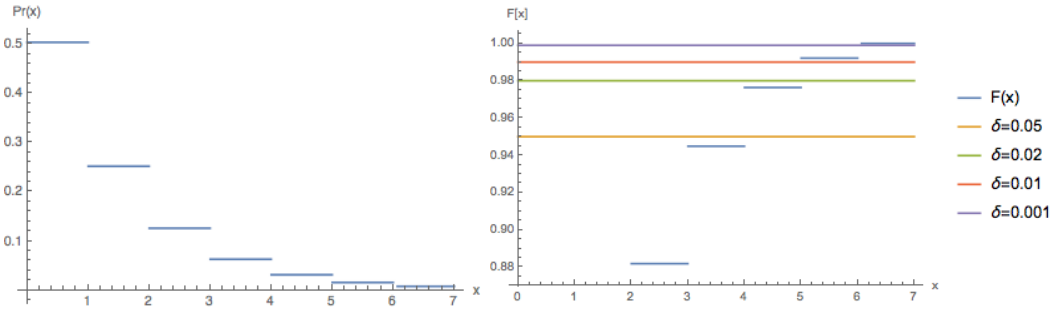


Figure 4.2 Plot on the left: Probability distribution $\mathbb{P}_{\mathrm{X}}(x)$ over the sorted range of X. Right: Cumulative distribution F_{X} over the sorted range of X, evidencing the smallest δ-sufficient set: The first level above the line corresponding to a value of δ corresponds to the maximum value of x in $\mathsf{S}_\delta(\mathrm{X})$. For example, for $\delta = 0.05$ one has $\mathsf{S}_\delta = \{1, 2, 3, 4, 5\}$

We are finally in position to define the quantity that measures the number of bits needed for compression of the smallest δ-sufficient set.

Definition 4.12 (Essential bit content)**.** Given a random variable X with probability distribution $\mathbb{P}_{\mathrm{X}}$, for $0 \leq \delta \leq 1$ the *essential bit content* of X is defined as

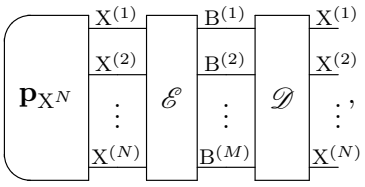
$$H_\delta(\mathrm{X}) := \log_2 |\mathsf{S}_\delta(\mathrm{X})|. \tag{4.11}$$

Remark 2. Notice that the notation for the raw bit content $H_0(\mathrm{X})$ is consistent with that for the essential bit content: indeed, the definition of raw bit content coincides with that of essential bit content in the special case $\delta = 0$.

From the above definitions we see that for affordable risk δ we can systematically achieve a compression length $H_\delta(\mathrm{X})$. In Example 4.9 we considered single-character strings. In the following we will more practically consider strings $x_{\mathbf{i}} := x_{i_1} x_{i_2} \ldots x_{i_N}$ of N characters, described by the random variable X^N made of N i.i.d. random variables equal to X. Then, for N very large, if X is not uniformly distributed, the least probable strings will be exponentially less probable than the least probable symbols, and this will make compression more and more reliable.

4.4 Source coding

A *lossy compression scheme* for strings of N i.i.d. random variables X^N is a pair of channels $(\mathscr{E}, \mathscr{D})$ with $\mathscr{E} : \mathsf{Rng}(\mathrm{X})^N \to \mathsf{Rng}(\mathrm{Y})$, and $\mathscr{D} : \mathsf{Rng}(\mathrm{Y}) \to \mathsf{Rng}(\mathrm{X}^N)$, with $\mathsf{Rng}(\mathrm{Y}) \subseteq \{0,1\}^M$ for some integer M. Diagrammatically:



where $\mathrm{B}^{(i)}$ is binary for every i. While X^N are i.i.d., the same is not true for B^M, thus we better think of the random variable Y.

Given a compression scheme $(\mathscr{E}, \mathscr{D})$, we define its *ratio* as follows.

Definition 4.13 (Data compression ratio)**.** Let X^N be N i.i.d. random variables, and let $(\mathscr{E}, \mathscr{D})$ be a compression scheme for X^N. The *compression ratio* of the scheme $(\mathscr{E}, \mathscr{D})$ is

$$R_{\mathscr{E}}(\mathrm{X}) := \frac{\log_2 |\mathscr{E}[\mathsf{Rng}(\mathrm{X}^N)]|}{N} \,. \tag{4.12}$$

The compression ratio quantifies the number of bits needed per character in a string. It is clear that this quantity accounts for the compression performances of the scheme $(\mathscr{E}, \mathscr{D})$. On the other hand, the performances of the compression scheme are also evaluated by the error probability, namely the probability that a string is compressed to a confusable string. The first Shannon theorem provides the smallest value for the compression ratio of a compression scheme for which the error probability can be made arbitrarily small.

4.4.1 Typical sets and asymptotic equipartition

The proof of Shannon's source coding theorem requires the notion of typical set, along with its properties.

Definition 4.14 (Typical set)**.** Given a discrete random variable X with $\mathsf{Rng}(\mathrm{X}) = \{x_1, x_2, \ldots, x_n\}$ distributed with probability $\{p_1, p_2, \ldots, p_n\}$, we denote by $\mathsf{T}_{N,\varepsilon}(\mathrm{X}) \subseteq \mathsf{Rng}(\mathrm{X}^N)$ the *typical set to tolerance* ε defined as the following set of strings of length N

$$\mathsf{T}_{N,\varepsilon}(\mathrm{X}) := \left\{ x_{\mathbf{i}} \in \mathsf{Rng}(\mathrm{X}^N) \;\middle|\; \left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}})} - H(\mathrm{X}) \right| \leq \varepsilon \right\} \tag{4.13}$$

The elements of $\mathsf{T}_{N,\varepsilon}(\mathrm{X})$ are called *typical strings*.

Using the weak law of large numbers we now see that the typical sequences have probabilities distributed according to the so-called asymptotic equipartition principle, which essentially states that the event $\mathrm{X}^N = x_{\mathbf{i}} = x_{i_1} x_{i_2} \ldots x_{i_N}$ almost certainly belongs to a subset having only $2^{NH(\mathrm{X})}$ elements, each one having probability close to $2^{-NH(\mathrm{X})}$. More precisely the following theorem holds.

Theorem 4.15 (Asymptotic equipartition principle)**.** *Let* X *be a random variable. The following three facts hold*

1. *The probability of strings* $x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}$ *is bounded as*

$$2^{-N[H(\mathrm{X})+\varepsilon]} \leq \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) \leq 2^{-N[H(\mathrm{X})-\varepsilon]}. \tag{4.14}$$

2. *The probability of the event* $x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})$ *is lower bounded as*

$$\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}] \geq 1 - \frac{\sigma^2_{\mathrm{Y}}}{\varepsilon^2 N}, \tag{4.15}$$

   where σ^2_{Y} *is the variance of the random variable* $\mathrm{Y} := h_{\mathrm{X}}(\mathrm{X})$.

3. *For every* $\delta > 0$*, there exists* N_0 *such that for* $N > N_0$ *the cardinality of the typical set is bounded as*

$$(1-\delta)2^{N[H(\mathrm{X})-\varepsilon]} \leq |\mathsf{T}_{N,\varepsilon}(\mathrm{X})| \leq 2^{N[H(\mathrm{X})+\varepsilon]}. \tag{4.16}$$

Proof. Item 1: This is just a way of rewriting the defining condition of the typical set, indeed

$$\left|\frac{1}{N}\log_2 \frac{1}{\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}})} - H(\mathrm{X})\right| \leq \varepsilon \;\Leftrightarrow\; -N[H(\mathrm{X})+\varepsilon] \leq \log_2 \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) \leq -N[H(\mathrm{X})-\varepsilon].$$

Item 2: Let us define the random variables $\mathrm{Y}(\mathrm{X})$ with range $(h_{\mathrm{X}}(x))_{x\in\mathsf{Rng}(\mathrm{X})}$ and probability distribution $\mathbb{P}_{\mathrm{Y}}[h_{\mathrm{X}}(x)] = \mathbb{P}_{\mathrm{X}}(x)$, and $\mathrm{Z}(\mathrm{X}^N) := \frac{1}{N}\sum_{i=1}^N \mathrm{Y}(\mathrm{X}^{(i)})$. Then we have

$$\mathrm{Z}(x_{\mathbf{i}}) = \frac{1}{N}\log_2 \frac{1}{\mathbb{P}_{\mathrm{Z}}(x_{\mathbf{i}})} = \frac{1}{N}\sum_{j=1}^N \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x_{i_j})},$$

$$\mathbb{E}[\mathrm{Z}] = \frac{1}{N}\sum_{j=1}^N \mathbb{E}\left[\log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x_{i_j})}\right] = H(\mathrm{X}).$$

The variable Z can then be used to rewrite the definition of typical set as follows

$$\mathsf{T}_{N,\varepsilon} = \{x_{\mathbf{i}} |\ [\mathrm{Z}(x_{\mathbf{i}}) - \mathbb{E}(\mathrm{Z})]^2 \leq \varepsilon^2\}.$$

Using the weak law of large numbers (lemma 3.9), we have

$$\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})] = \mathbb{P}_{\mathrm{Z}}[(\mathrm{Z}(x_{\mathbf{i}}) - \mathbb{E}(\mathrm{Z}))^2 \leq \varepsilon^2] \geq 1 - \frac{\sigma^2_{\mathrm{Y}}}{N\varepsilon^2}.$$

Item 3: Given a value of δ, take N_0 such that $N_0 \geq \frac{\sigma^2_{\mathrm{Y}}}{\delta\varepsilon^2}$. Then, by item 2, for $N > N_0$ the probability of the typical set $\mathsf{T}_{N,\varepsilon}$ is bounded as

$$(1-\delta) \leq \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})] \leq 1. \tag{4.17}$$

Thus, using the definition of typical set and the upper bound in equation (4.17), one can prove the following chain of inequalities

$$1 \geq \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) \geq \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} 2^{-N[H(\mathrm{X}+\varepsilon)]} = |\mathsf{T}_{N,\varepsilon}(\mathrm{X})| 2^{-N[H(\mathrm{X}+\varepsilon)]}.$$

Thus $|\mathsf{T}_{N,\varepsilon}(\mathrm{X})| \leq 2^{N[H(\mathrm{X}+\varepsilon)]}$. Conversely, from the lower bound in equation (4.17) one can prove the following chain of inequalities

$$(1-\delta) \leq \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) \leq \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} 2^{-N[H(\mathrm{X})-\varepsilon]} = |\mathsf{T}_{N,\varepsilon}(\mathrm{X})| 2^{-N[H(\mathrm{X})-\varepsilon]}.$$

Thus $(1-\delta)2^{N[H(\mathrm{X})-\varepsilon]} \leq |\mathsf{T}_{N,\varepsilon}(\mathrm{X})|$. □

4.4.2 Shannon source coding theorem

Using the asymptotic equipartition theorem, we can now prove the first Shannon theorem, known as *source coding theorem*. The theorem states that the essential bit content of N i.i.d. random variables X^N can be made arbitrarily close to $NH(\mathrm{X})$ for any value δ of the error probability, by taking long enough sequences. Since the essential bit content is the number of bits per character—or better, int he case of interest, per length N-string—in a reliable compression scheme with probability δ of error, the theorem provides the operational interpretation of the Shannon entropy, that is the smallest reliable asymptotic compression ratio for an information source represented by the random variable X. In other words, given a classical source whose state corresponds to a random variable X with probability distribution $\mathbb{P}_{\mathrm{X}}$, one can find a family $(\mathscr{E}^N, \mathscr{D}^N)$ of encodings for strings of length N such that the error probability for these schemes is arbitrarily small for large N, while their compression ratio gets arbitrarily close to $H(\mathrm{X})$.

Let us now see the precise statement and the proof of the theorem.

Theorem 4.16 (Shannon's source coding theorem)**.** *Let* X *a random variable with Shannon entropy* $H(\mathrm{X})$*. Given* $\varepsilon > 0$ *and* $0 < \delta < 1$*, there exists a positive integer* N_0 *such that for* $N > N_0$ *one has*

$$\left| \frac{1}{N} H_\delta(\mathrm{X}^N) - H(\mathrm{X}) \right| \leq \varepsilon \tag{4.18}$$

Proof. The statement is equivalent to the following chain of inequalities

$$H(\mathrm{X}) - \varepsilon \leq \frac{1}{N} H_\delta(\mathrm{X}^N) \leq H(\mathrm{X}) + \varepsilon$$

The proof is split in two parts, in which we prove the two inequalities separately.

– *Proof of* $H(\mathrm{X}) - \varepsilon \leq \frac{1}{N} H_\delta(\mathrm{X}^N)$. Let $\mathsf{S} \subseteq \mathsf{Rng}(\mathrm{X}^N)$ be any set with cardinality $|\mathsf{S}| \leq 2^{N[H(\mathrm{X})-2\alpha]}$. We can bound the probability measure of such set as follows

$$\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S}] = \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S} \cap \mathsf{T}_{N,\alpha}(\mathrm{X})] + \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S} \cap \bar{\mathsf{T}}_{N,\alpha}(\mathrm{X})],$$

where $\bar{\mathsf{T}}_{N,\alpha}(\mathrm{X}) := \mathsf{Rng}(\mathrm{X}^N) \setminus \mathsf{T}_{N,\alpha}(\mathrm{X})$ is the complement of the typical set $\mathsf{T}_{N,\alpha}(\mathrm{X})$. Since the probability of sequences in the typical set $\mathsf{T}_{N,\alpha}(\mathrm{X})$ is bounded by definition by $2^{-N[H(\mathrm{X})-\alpha]}$, one has

$$\begin{aligned}\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S} \cap \mathsf{T}_{N,\alpha}(\mathrm{X})] &\leq \sum_{x_{\mathbf{i}} \in \mathsf{S} \cap \mathsf{T}_{N,\alpha}(\mathrm{X})} 2^{-N[H(\mathrm{X})-\alpha]} \leq \sum_{x_{\mathbf{i}} \in \mathsf{S}} 2^{-N[H(\mathrm{X})-\alpha]} \\ &\leq 2^{N[H(\mathrm{X})-2\alpha]} 2^{-N[H(\mathrm{X})-\alpha]} = 2^{-N\alpha}.\end{aligned}$$

On the other hand, by item 2 of the asymptotic equipartition theorem 4.15, one has

$$\begin{aligned}\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S} \cap \bar{\mathsf{T}}_{N,\alpha}(\mathrm{X})] &\leq \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \bar{\mathsf{T}}_{N,\alpha}(\mathrm{X})] = 1 - \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{T}_{N,\alpha}(\mathrm{X})] \\ &\leq 1 - \left(1 - \frac{\sigma^2}{N\alpha^2}\right) = \frac{\sigma^2}{N\alpha^2},\end{aligned}$$

where σ^2 is the variance of $\log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x)}$. Collecting the bounds on the two terms, we then have

$$\mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{S}] \leq 2^{-N\alpha} + \frac{\sigma^2}{N\alpha^2}.$$

Setting $2\alpha = \varepsilon$ and $N_0 \geq \max\{\frac{8\sigma^2}{\varepsilon^2\gamma}, \frac{1}{2\varepsilon}\log_2 \frac{2}{\gamma}\}$, we conclude that for $N \geq N_0$, any set with cardinality $|\mathsf{S}| \leq 2^{N[H(\mathrm{X})-\varepsilon]}$ has probability smaller than γ, and for suitably small γ it cannot be δ-sufficient. Then the smallest δ-sufficient set must have cardinality $|\mathsf{S}_\delta(\mathrm{X}^N)|$ larger than $2^{N[H(\mathrm{X})-\varepsilon]}$. This implies that $\frac{1}{N}H_\delta(\mathrm{X}^N) = \frac{1}{N}\log_2 |\mathsf{S}_\delta(\mathrm{X}^N)| \geq H(\mathrm{X}) - \varepsilon$

– *Proof of* $\frac{1}{N}H_\delta(\mathrm{X}^N) \leq H(\mathrm{X}) + \varepsilon$. From item 2 of the asymptotic equipartition theorem 4.15, setting $N > N_0$ in equation 4.15 and $N_0 \geq \frac{\sigma^2}{\varepsilon^2\delta}$, we observe that the typical set is δ-sufficient. The smallest δ-sufficient set has then cardinality not larger than the typical set, namely $|\mathsf{S}_\delta(\mathrm{X}^N)| \leq |\mathsf{T}_{N,\varepsilon}(\mathrm{X})| \leq 2^{N[H(\mathrm{X})+\varepsilon]}$. Taking the logarithm we obtain $\frac{1}{N}H_\delta(\mathrm{X}^N) \leq H(\mathrm{X}) + \varepsilon$. □

Intuitive consequences of the Shannon's source coding theorem are the following.

– For sufficiently large N the function $H_\delta(\mathrm{X}^N)$ is essentially constant versus δ, for $0 < \delta < 1$.

– N i.i.d. random variables with Shannon entropy $H(\mathrm{X})$ can be compressed asymptotically for $N \to \infty$ into more than $NH(\mathrm{X})$ bits with negligible risk of information loss, whereas, if we try to compress them into fewer than $NH(\mathrm{X})$ bits, it is certain that some information will be lost.

Indeed, the first part of the proof tells us that even for very small probability of error δ, the number of bits per symbol $\frac{1}{N}H_\delta(\mathrm{X}^N)$ required to specify a string $x_{\mathbf{i}}$ of N symbols does not need to exceed $H(\mathrm{X}) + \varepsilon$ bits. In other words, for an arbitrarily small tolerance for error, the number of bits required drops from $H_0(\mathrm{X})$ to $H(\mathrm{X}) + \varepsilon$. On the other hand, even if we tolerate an error δ very close to 1, the average number of bits per

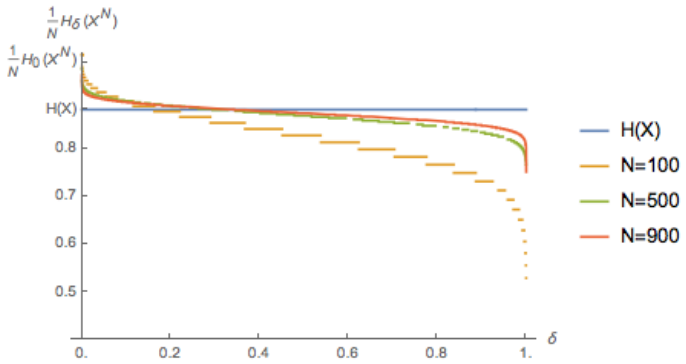


Figure 4.3 The essential bit content $1/N H_\delta(\mathrm{X}^N)$ of N i.i.d. binary random variables X with probability distribution $\{p, 1-p\}$, as a function of the tolerated error probability δ. The plot represents the cases $N = 100, 500, 900$ for $p = 0.7$

symbol needed to specify $x_{\mathbf{i}}$ must still be at least $H(\mathrm{X}) - \varepsilon$: The information becomes immediately unreliable if we take compression ratios smaller than $H(\mathrm{X})$. See figure 4.4.2.

We have thus far analysed the problem of source coding from a very general point of view, and Shannon's theorem summarises all the results that we can derive in the asymptotic case, exploiting the properties of typical sets. Shannon's theorem is proved by showing that, for large N, the smallest δ-sufficient set gets increasingly closer to the typical set. However, the ratio at which this phenomenon occurs is not analysed in detail, and thus a compression algorithm based on coding the typical set and messing up the remaining sequences may have performances that converge very slowly to the asymptotic regime. In the next lecture we will study a special class of coding algorithms, called *symbol codes*, that in the case of finite N already get close to the optimal compression ratio, represented by the Shannon entropy.

Chapter 5

Lecture 5: Symbol codes and stream codes

5.1 Symbol codes

Let $\Sigma = \{x_1, x_2, \dots, x_n\}$ be an alphabet. The set Σ^N denotes the set of length-N strings of characters from Σ, namely $\Sigma^N := \{x_{i_1} x_{i_2} \dots x_{i_N} |\ x_{i_j} \in \Sigma\}$. We then define

$$\Sigma^+ := \bigcup_{N \geq 1} \Sigma^N, \quad \Sigma^* := \{\lambda\} \cup \Sigma^N,$$

where λ denotes the empty string.

Example 5.1. For $\Sigma = \{0, 1\}$, we have

$$\Sigma^3 = \{000, 001, 010, 011, 100, 101, 110, 111\},$$
$$\Sigma^+ = \{0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, \dots\}$$

Let us now consider a classical state of a system of type n, namely a random variable X with $|\mathsf{Rng}(\mathrm{X})| = n$. in this case we have

$$\mathsf{Rng}(\mathrm{X})^+ = \{x_1, x_2, \dots, x_n, x_1x_1\,,\, x_1x_2, \dots, x_nx_n\,,\, x_1x_1x_1\,,\, x_1x_1x_2, \dots\}.$$

A symbol code is a code that maps strings in $\mathsf{Rng}(\mathrm{X})^+$ to strings of bits with variable length, by juxtaposing the codewords of the corresponding symbols.

Definition 5.2 (Symbol code)**.** Let X be a random variable. A *symbol code* for X is a map $\mathscr{E} : \mathsf{Rng}(\mathrm{X}) \to \{0,1\}^+$. The image of a *symbol* $x \in \mathsf{Rng}(\mathrm{X})$, denoted as $\mathscr{E}(x) \in \{0,1\}^+$, is a *codeword*, and the number of bits in $\mathscr{E}(x)$ is its length, denoted by $l_{\mathscr{E}}(x) = l(x)$. The codebook is the set of codewords $\mathscr{E}[\mathsf{Rng}(\mathrm{X})]$

Given a symbol code for X, one can use it to encode strings in $\mathsf{Rng}(\mathrm{X})^+$, and this procedure defines the extended code.

Definition 5.3 (Extended code)**.** Given a symbol code $\mathscr{E}$ for X, we define its *extended code* as the map $\mathscr{E}^+ : \mathsf{Rng}(\mathrm{X})^+ \to \{0,1\}^+$, defined by concatenating codewords:

$$\mathscr{E}^+(x_{i_1} x_{i_2} \dots x_{i_k}) = \mathscr{E}(x_{i_1})\mathscr{E}(x_{i_2}) \dots \mathscr{E}(x_{i_k}). \tag{5.1}$$

x	$\mathscr{E}_0(x)$	$l(x)$	$\mathbb{P}_{\mathrm{X}}(x)$
1	00	2	$\frac{1}{2}$
2	01	2	$\frac{1}{4}$
3	10	2	$\frac{1}{4}$

Example 5.4. Let $\Omega = \{a, b, c\}$, $\mathsf{Rng}(\mathrm{X}) = \{1, 2, 3\}$ with $\mathbf{p} = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$. We can define the following symbol code

The following two examples are based on $\Omega = \{a, b, c, d\}$, with $\mathsf{Rng}(\mathrm{X}) = \{1, 2, 3, 4\}$ with $\mathbf{p} = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$.

x	$\mathscr{E}_1(x)$	$l(x)$	$\mathbb{P}_{\mathrm{X}}(x)$
1	0	1	$\frac{1}{2}$
2	10	2	$\frac{1}{4}$
3	110	3	$\frac{1}{8}$
4	111	3	$\frac{1}{8}$

x	$\mathscr{E}_2(x)$	$l(x)$	$\mathbb{P}_{\mathrm{X}}(x)$
1	0	1	$\frac{1}{2}$
2	10	2	$\frac{1}{4}$
3	100	3	$\frac{1}{8}$
4	111	3	$\frac{1}{8}$

Given a symbol code, we can define its expected length, that assesses its compression rate.

Definition 5.5 (Expected length)**.** Given a symbol code $\mathscr{E}$ for a random variable X with $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$, the *expected length* $L(\mathscr{E}, \mathrm{X})$ of the code is defined as

$$L(\mathscr{E}, \mathrm{X}) := \sum_{i=1}^{n} \{\mathbb{P}_{\mathrm{X}}(x_i) l(x_i)\}. \tag{5.2}$$

As an example, consider the codes $\mathscr{E}_1$ and $\mathscr{E}_2$ introduced above, for which $L(\mathscr{E}_1, \mathrm{X}) = L(\mathscr{E}_2, \mathrm{X}) = 7/4$ bit/character. The length is not the only feature of symbol codes we are interested in. In particular, we need the code to be uniquely decodable. The following definitions will allow us to find the features that grant fulfilment of this requirement.

Definition 5.6 (Non-singular code)**.** A symbol code $\mathscr{E}$ for X is *non-singular* if for every two symbols $x \neq y$ in $\mathsf{Rng}(\mathrm{X})$ one has $\mathscr{E}(x) \neq \mathscr{E}(y)$

The above condition is not sufficient yet to grant decodability. We need the following notion.

Definition 5.7 (Unique decodability)**.** A symbol code $\mathscr{E}$ for X is *uniquely decodable* if for every two strings $x \neq y$ in $\mathsf{Rng}(\mathrm{X})^+$ one has $\mathscr{E}^+(x) \neq \mathscr{E}^+(y)$

Example 5.8. The codes $\mathscr{E}_0$, $\mathscr{E}_1$ and $\mathscr{E}_2$ are non-singular. However, only $\mathscr{E}_0^+$ and $\mathscr{E}_1^+$ are uniquely decodable. Indeed, consider the following strings for the code $\mathscr{E}_2$: $x = acbad$ and $y = ababad$. We have

$$\mathscr{E}_2^+[\mathrm{X}^{(5)}(acbad)] = 0100100111,$$
$$\mathscr{E}_2^+[\mathrm{X}^{(6)}(ababad)] = 0100100111.$$

How can we ensure unique decodability? We will now define a property that provides a sufficient condition.

Definition 5.9 (Prefix)**.** Let $p, w \in \{0,1\}^+$. We say that p is a *prefix* for w if there is $t \in \{0,1\}^*$ such that $w = pt$.

Definition 5.10 (Prefix code)**.** We say that a code $\mathscr{E}$ for X has the *prefix property*, or simply is *prefix*, if no codeword is a prefix for another codeword.

As an example, the codes $\mathscr{E}_0$ and $\mathscr{E}_1$ are prefix. Moreover, the code $\{0, 101\}$ for $\Omega = \{a, b\}$ is prefix, while $\{1, 101\}$ is not prefix, but it is uniquely decodable. The prefix property is sufficient for unique decodability, but it may seem too restrictive, as it is not necessary. However, the Kraft-McMillan theorem will show that we have no loss in performances if we restrict to prefix codes; on the other hand, prefix codes have the property of *self-punctuation*, namely one can decode the message on the fly (the so-called instantaneous decoding), without waiting for the end of transmission. Moreover, the prefix property is simply checkable, as it can be checked on $\mathsf{Rng}(\mathrm{X})$.

Lemma 5.11. *If the symbol code $\mathscr{E}$ is prefix the extended code $\mathscr{E}^+$ is uniquely decodable.*

Proof. If $\mathscr{E}$ is prefix then $\mathscr{E}(x_{i_1}) = \mathscr{E}(x_{i_2})t$ iff $x_{i_1} = x_{i_2}$ and $t = \lambda$. Suppose now that $\mathscr{E}^+$ is not uniquely decodable, and let $x \neq y$ be two strings with $\mathscr{E}^+(x) = \mathscr{E}^+(y)$. Let $j := \min\{k | x_{i_k} \neq y_{i_k}\}$. Then we have

$$x = x_{i_1}x_{i_2}\dots x_{i_j}x_{i_{j+1}}\dots \quad y = x_{i_1}x_{i_2}\dots y_{i_j}y_{i_{j+1}}\dots$$
$$\mathscr{E}(x_{i_1})\mathscr{E}(x_{i_2})\dots\mathscr{E}(x_{i_j})\mathscr{E}(x_{i_{j+1}})\dots = \mathscr{E}(x_{i_1})\mathscr{E}(x_{i_2})\dots\mathscr{E}(y_{i_j})\mathscr{E}(y_{i_{j+1}})\dots.$$

This implies that

$$\mathscr{E}(x_{i_j})\mathscr{E}(x_{i_{j+1}})\dots = \mathscr{E}(y_{i_j})\mathscr{E}(y_{i_{j+1}})\dots.$$

We now have three possibilities:

1. $l_{\mathscr{E}}(x_{i_j}) = l_{\mathscr{E}}(y_{i_j})$: in this case, it must be $\mathscr{E}(x_{i_j}) = \mathscr{E}(y_{i_j})$.
2. $l_{\mathscr{E}}(x_{i_j}) > l_{\mathscr{E}}(y_{i_j})$: in this case, it must be $\mathscr{E}(x_{i_j}) = \mathscr{E}(y_{i_j})s$.
3. $l_{\mathscr{E}}(x_{i_j}) < l_{\mathscr{E}}(y_{i_j})$: in this case, it must be $\mathscr{E}(x_{i_j})s' = \mathscr{E}(y_{i_j})$.

In all three cases, by the prefix property of $\mathscr{E}$ it must be $x_{i_j} = y_{i_j}$, contrarily to the hypothesis. □

The prefix property is checkable on $\mathsf{Rng}(\mathrm{X})$, contrarily to unique decodability: The code $\mathscr{E}_2$ in our example 5.4 indeed is uniquely decodable on $\mathsf{Rng}(\mathrm{X})$, however $\mathscr{E}_2^+(ba) = \mathscr{E}_2^+(c)$.

5.2 Kraft-McMillan theorem

In this section we prove the most important result about symbol codes, that provides a necessary and sufficient condition for unique decodability, and allows one also to prove that for every uniquely decodable symbol code, there is a prefix one with the same codeword lengths.

Theorem 5.12 (Kraft-McMillan theorem)**.** *Let* X *be a random variable.*

1. *A uniquely decodeable binary code* $\mathscr{E}$ *for* X *must satisfy McMillan's bound*

$$\sum_{x_i \in \mathsf{Rng}(\mathrm{X})} 2^{-l(x_i)} \leq 1. \tag{5.3}$$

2. *Conversely, if the bound 5.3 is satisfied by a set of lengths* $\{l_1, l_2, \ldots, l_{|\mathsf{Rng}(\mathrm{X})|}\}$*, then there exists a prefix code* $\mathscr{E}$ *for* X *with* $l_{\mathscr{E}}(x_i) = l_i$*.*

Proof. Let us start form item 1 Define $S := \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} 2^{-l(x_i)} = \sum_{i=1}^{R} 2^{-l_i}$, where we set $|\mathsf{Rng}(\mathrm{X})| = R$ and $l_i := l(x_i)$. Then we have

$$S^N = \sum_{i_1=1}^{R} \sum_{i_2=1}^{R} \cdots \sum_{i_N=1}^{R} 2^{-(l_{i_1} + l_{i_2} + \cdots + l_{i_N})}. \tag{5.4}$$

Since $l_{i_1} + l_{i_2} + \cdots + l_{i_N} = l(x_{i_1} x_{i_2} \ldots x_{i_N})$ for some string $x_{\mathbf{i}} = x_{i_1} x_{i_2} \ldots x_{i_N} \in \mathsf{Rng}(\mathrm{X}^N)$, we can re-write equation (5.4) as follows

$$S^N = \sum_{l=Nl_m}^{Nl_M} \{2^{-l} N(l)\}, \tag{5.5}$$

where $l_m := \min_{x_i \in \mathsf{Rng}(\mathrm{X})} l(x_i)$, $l_M := \max_{x_i \in \mathsf{Rng}(\mathrm{X})} l(x_i)$, and $N(l) := |\{x_{\mathbf{i}} \in \mathsf{Rng}(\mathrm{X}^N) | l(x_{\mathbf{i}}) = l\}|$ is the number of strings $x_{\mathbf{i}} \in \mathsf{Rng}(\mathrm{X}^N)$ with $l(x_{\mathbf{i}}) = l$. Since the number of bit strings of length l is 2^l, unique decodability enforces $N(l) \leq 2^l$. Thus,

$$S^N \leq \sum_{Nl_m}^{Nl_M} \{2^{-l} 2^l\} = N(l_M - l_m) + 1 \leq N(l_M - l_m + 1) = Nk, \quad k > 0.$$

Since the bound must hold for every N, it must be $S \leq 1$. Indeed, $S > 1$ would imply that, for sufficiently large N, one would have $S^N > Nk$. Let us now consider item 2. Let M denote the number of *different* values of the considered lengths—e.g., if $R = 5$ and $l_1 = 1, l_2 = 3, l_3 = 3, l_4 = 3, l_5 = 6$, then $M = 3$. Consider now the set of different lengths $\lambda_1 < \lambda_2 \ldots < \lambda_M$, and let N_k denote the number of lengths l_i with the same value λ_k—e.g., in the same example as before, $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 6$, and $N_1 = 1$, $N_2 = 3$, and $N_3 = 1$. We can express the hypothesis of item 2 as

$$S := \sum_{k=1}^{M} N_k 2^{-\lambda_k} \leq 1.$$

We now prove that one can construct a prefix code with N_k codewords of length λ_k for every $1 \leq k \leq M$, as follows. Consider any integer j with $1 \leq j \leq M$. One has

$$N_j 2^{-\lambda_j} + \sum_{k=m}^{j-1} \{N_k 2^{-\lambda_k}\} \leq S \leq 1,$$

and then

$$N_j \leq 2^{\lambda_j} - \sum_{k=1}^{j-1} \{N_k 2^{\lambda_j - \lambda_k}\}. \tag{5.6}$$

The proof now proceeds by induction on $j < M$. For $j = 1$ one has $N_1 \leq 2^{\lambda_1}$, namely the number N_1 is smaller than the number of strings of length l_1, and one can thus encode N_1 different symbols in such strings. Clearly, no string of length λ_1 can be a prefix for another string of the same length. Suppose now that one can choose N_k codewords of length λ_k for every $k < j$, without using any assigned codeword as a prefix for another one. Since $2^{\lambda_j - \lambda_k}$ is the number of strings of length λ_j having a fixed prefix of length λ_k, $\sum_{k=1}^{j-1} N_k 2^{\lambda_j - \lambda_k}$ is precisely the number of strings of length λ_j with prefixes that have already been assigned as shorter codewords. If we subtract this number from the total number 2^{λ_j} of strings of length λ_j, inequality 5.6 ensures that we are left with a number of available strings of length λ_j that is not smaller than the number of strings we need to encode. Since this holds for every j, McMillan's bound grants the possibility to construct a prefix code. □

Remark 3. Given a uniquely decodable code $\mathscr{E}$, its codeword lengths must satisfy McMillan's inequality. However, by Kraft-McMillan's theorem, this means that we can construct a prefix code with exactly the same lengths. This is the reason why it is not restrictive to consider prefix codes instead of uniquely decodable ones.

An important class of symbol codes is that of codes saturating McMillan's bound.

Definition 5.13 (Complete code)**.** A uniquely decodable symbol code $\mathscr{E}$ for X is called *complete* if its lengths satisfy

$$\sum_{x_i \in \mathsf{Rng}(\mathrm{X})} 2^{-l(x_i)} = 1. \tag{5.7}$$

5.3 Source coding theorem for symbol codes

The task of compression consists in minimising the expected length of a code. For symbol codes, the figure of merit to be minimised is then

$$L(\mathscr{E}, \mathrm{X}) := \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} \{\mathbb{P}_{\mathrm{X}}(x_i) l(x_i)\}.$$

In this section, we will prove that the expected length is bounded from below by the Shannon entropy of X.

Theorem 5.14 (Lower bound for the expected code length)**.** *The expected length* $L(\mathscr{E}, \mathrm{X})$ *of a uniquely decodable symbol code* $\mathscr{E}$ *is bounded from below by* $H(\mathrm{X})$*.*

Proof. Let us define the two vectors

$$\mathbf{p} := (p_1, p_2, \ldots, p_R), \quad p_i := \mathbb{P}_{\mathrm{X}}(x_i)$$
$$\mathbf{q} := (q_1, q_2, \ldots, q_R), \quad q_i := \frac{2^{-l(x_i)}}{Z},$$

where $l(x_i)$ are the lengths for $\mathscr{E}$ and the *partition function* Z is $Z := \sum_{i=1}^{R} 2^{-l(x_i)} \leq 1$. The probabilities $\mathbf{q}$ are called *implicit probabilities* of $\mathscr{E}$. Let us now apply Gibbs' inequality $0 \leq H(\mathbf{p}\|\mathbf{q})$ as follows

$$\sum_{i=1}^{R} \left\{ p_i \log_2 \frac{1}{q_i} \right\} \geq \sum_{i=1}^{R} \left\{ p_i \log_2 \frac{1}{p_i} \right\},$$

obtaining

$$\sum_{i=1}^{R} \{p_i l(x_i)\} = \sum_{i=1}^{R} \left\{ p_i \log_2 \frac{1}{q_i} \right\} - \log_2 Z \geq \sum_{i=1}^{R} \left\{ p_i \log_2 \frac{1}{p_i} \right\} = H(\mathrm{X}). \tag{5.8}$$

Notice that equality is achieved under two conditions: $Z = 1$ (the code is complete) and $q_i = p_i$, namely $l(x_i) = \log_2 \frac{1}{p_i}$. □

Corollary 5.15 (Optimal source code-lengths)**.** *The expected length of the code is minimum only if the code-length of a symbol equals its Shannon information content*

$$l(x_i) = \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x_i)} \quad \textit{(Shannon's lengths).} \tag{5.9}$$

Conversely, any choice of code-lengths $l(x_i)$ *defines a probability distribution* $\mathbb{P}_L(x_i) = 2^{-l(x_i)}/Z$ *for which the code-lengths would be optimal. If in addition the code is complete, then the expected length will equal the Shannon entropy* $H(\mathrm{X})$*, and* $\mathbb{P}_{\mathrm{X}}(x_i) = 2^{-l(x_i)}$*.*

Notice however that the Shannon length $l(x_i)$ is generally not integer. Then, the question is: How close can we get to the lower bound for the expected length? This is provided by the following theorem

Theorem 5.16 (Source coding theorem for symbol codes)**.** *For any random variable* X *there exists a prefix code* $\mathscr{E}$ *with expected length satisfying the source coding theorem for symbol codes*

$$H(\mathrm{X}) \leq L(\mathscr{E}, \mathrm{X}) < H(\mathrm{X}) + 1. \tag{5.10}$$

Proof. In general, the Shannon length $l(x_i) = -\log_2 \mathbb{P}_{\mathrm{X}}(x_i)$ cannot be exactly achieved because it is not an integer. Therefore, we use lengths slightly larger than the optimum, namely

$$l(x_i) = \left\lceil \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}}(x_i)} \right\rceil, \tag{5.11}$$

where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. This choice of lengths satisfies McMillan's bound 5.3, since

$$\sum_{i=1}^{R} 2^{-l(x_i)} = \sum_{i=1}^{R} 2^{-\left\lceil \log_2 \frac{1}{p_i} \right\rceil} \leq \sum_{i=1}^{R} 2^{-\log_2 \frac{1}{p_i}} = \sum_{i=1}^{R} p_i = 1. \tag{5.12}$$

Then, according to theorem 5.12, there exists a prefix code with lengths $l(x_i)$. Moreover, we have

$$L(\mathscr{E}, \mathrm{X}) = \sum_{i=1}^{R} \left\{ p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil \right\} < \sum_{i=1}^{R} \left\{ p_i \left(\log_2 \frac{1}{p_i} + 1 \right) \right\} = H(X) + 1. \tag{5.13}$$

The lower bound was already proved in theorem 5.14. □

Remark 4. For complete codes, $Z = 1$, and by equation 5.8 one immediately sees that in that case the expected code length exceeds the Shannon entropy $H(\mathrm{X})$ exactly by the Kullback-Leibler divergence $H(\mathbf{p}\|\mathbf{q})$ between the actual distribution and that of implicit probabilities $\mathbf{q}$.

Chapter 6

Lecture 6: Huffman coding and stream codes

6.1 Optimal code: Huffman's coding

Given a probability distribution $\mathbf{p}$ for the random variable X, can we construct the optimal symbol code? The answer is: Yes. We now show Huffman's code—constructed by a simple protocol—which is a prefix symbol code $\mathscr{E}$ minimising the expected length $L(\mathscr{E}, \mathrm{X})$. The code is obtained by the following algorithm.

Huffman code $\mathscr{E}_H(\mathrm{X})$

1. Start by defining the random variable $\mathrm{X}_0 := \mathrm{X}$, with symbols sorted from the one with highest probability ($x_1^{(0)}$) to the one with smallest probability ($x_R^{(0)}$). Assign the empty string $\lambda \in \{0,1\}^*$ to all symbols $x_i^{(0)} \in \mathsf{Rng}(\mathrm{X}_0)$.

2. If there is a unique symbol $x_1^{(n)}$, stop the algorithm. Otherwise take the two least probable symbols $x_{R-n-1}^{(n)}$, $x_{R-n}^{(n)}$ in the alphabet X_n, and append two different bits, 0 and 1, to the left of all strings associated to them.

3. Combine the two symbols $\{x_{R-n-1}^{(n)}, x_{R-n}^{(n)}\}$ into a single one $x_{R-n-1}^{\prime(n)} := \{x_{R-n-1}^{(n)}\} \cup \{x_{R-n}^{(n)}\}$, and define the new variable X_{n+1} with $x_j^{(n+1)} = x_j^{(n)}$ for $j < R-n-1$, and $x_{R-n-1}^{(n)} = x_{R-n-1}^{\prime(n)}$, and $p_j^{(n+1)} = p_j^{(n)}$ for $j < R-n-1$, and $p_{R-n-1}^{(n+1)} = p_{R-n-1}^{(n)} + p_{R-n}^{(n)}$. Sort X_{n+1} from the symbol with highest probability to the one with the lowest. Iterate the procedure from step 2.

Notice that by construction the Huffman algorithm associates the n-th longest codeword to the n-th least probable symbol. This feature is crucial in the proof of optimality of Huffman codes.

Example 6.1. Consider the alphabet:

$$\mathrm{X} = \begin{cases} \Omega = \{\mathsf{a}, & \mathsf{b}, & \mathsf{c}, & \mathsf{d}, & \mathsf{e}\} \\ \mathbf{p} = \{0.25, & 0.25, & 0.2, & 0.15, & 0.15\} \end{cases}$$

The Huffman coding is obtained as follows

1.

$x_i^{(0)}$	$\mathbf{p}^{(0)}$	$\mathscr{E}_H(x)$	$x_i^{(1)}$	$\mathbf{p}^{(1)}$
a	0.25	λ	{d, e}	0.30
b	0.25	λ	a	0.25
c	0.20	λ	b	0.25
d	0.15	0	c	0.20
e	0.15	1		

2.

$x_i^{(1)}$	$\mathbf{p}^{(1)}$	$\mathscr{E}_H(x)$	$x_i^{(2)}$	$\mathbf{p}^{(2)}$
{d, e}	0.30	{0, 1}	{b, c}	0.45
a	0.25	λ	{d, e}	0.30
b	0.25	0	a	0.25
c	0.20	1		

3.

$x_i^{(2)}$	$\mathbf{p}^{(2)}$	$\mathscr{E}_H(x)$	$x_i^{(3)}$	$\mathbf{p}^{(3)}$
{b, c}	0.45	{0, 1}	{d, e, a}	0.55
{d, e}	0.30	{00, 01}	{b, c}	0.45
a	0.25	1		

4.

$x_i^{(3)}$	$\mathbf{p}^{(3)}$	$\mathscr{E}_H(x)$
{d, e, a}	0.55	{000, 001, 01}
{b, c}	0.45	{10, 11}

Thus we end up with the following code

x_i	p_i	$\mathscr{E}_H(x_i)$	$h(x_i)$
a	0.25	01	2.0
b	0.25	10	2.0
c	0.20	11	2.3
d	0.15	000	2.7
e	0.15	001	2.7

The expected length is $L = 2.40$ bits to be compared with the Shannon entropy $H(\mathrm{X}) = 2.2855$ bits.

Example 6.2. Consider the alphabet:

$$\mathrm{X} = \begin{cases} \Omega = \{\mathsf{a}, & \mathsf{b}, & \mathsf{c}, & \mathsf{d}, & \mathsf{e}, & \mathsf{f}, & \mathsf{g}\} \\ \mathbf{p} = \{0.01, & 0.24, & 0.05, & 0.20, & 0.47, & 0.01, & 0.02\} \end{cases}$$

The Huffman coding is obtained as follows

1.

$x_i^{(0)}$	$p_i^{(0)}$	$\mathscr{E}_H(x)$	$x_i^{(1)}$	$p_i^{(1)}$
e	0.47	λ	e	0.47
b	0.24	λ	b	0.24
d	0.20	λ	d	0.20
c	0.05	λ	c	0.05
g	0.02	λ	g	0.02
a	0.01	1	{a, f}	0.02
f	0.01	0		

2.

$x_i^{(1)}$	$p_i^{(1)}$	$\mathscr{E}_H(x)$	$x_i^{(2)}$	$p_i^{(2)}$
e	0.47	λ	e	0.47
b	0.24	λ	b	0.24
d	0.20	λ	d	0.20
c	0.05	λ	c	0.05
g	0.02	0	{g, a, f}	0.04
{a, f}	0.02	{10, 11}		

3.

$x_i^{(2)}$	$p_i^{(2)}$	$\mathscr{E}_H(x)$	$x_i^{(3)}$	$p_i^{(3)}$
e	0.47	λ	e	0.47
b	0.24	λ	b	0.24
d	0.20	λ	d	0.20
c	0.05	0	{c, g, a, f}	0.09
{g, a, f}	0.04	{10, 110, 111}		

4.

$x_i^{(3)}$	$p_i^{(3)}$	$\mathscr{E}_H(x)$	$x_i^{(4)}$	$p_i^{(4)}$
e	0.47	λ	e	0.47
b	0.24	λ	{d, c, g, a, f}	0.29
d	0.20	0	b	0.24
{c, g, a, f}	0.09	{10, 110, 1110, 1111}		

5.

$x_i^{(4)}$	$p_i^{(4)}$	$\mathscr{E}_H(x)$	$x_i^{(5)}$	$p_i^{(5)}$
e	0.47	λ	{d, c, g, a, f, b}	0.53
{d, c, g, a, f}	0.29	{00, 010, 0110, 01110, 01111}	e	0.47
b	0.24	1		

6.

$x_i^{(5)}$	$p_i^{(5)}$	$\mathscr{E}_H(x)$
{d, c, g, a, f, b}	0.53	{000, 0010, 00110, 001110, 001111, 01}
e	0.47	1

Thus we end up with the following code

x_i	p_i	$\mathscr{E}_H(x_i)$	$h(x_i)$
a	0.01	001110	6.64
b	0.24	01	2.06
c	0.05	0010	4.32
d	0.20	000	2.32
e	0.47	1	1.09
f	0.01	001111	6.64
g	0.02	00110	5.64

The expected length is $L = 1.97$ bits to be compared with the Shannon entropy $H(\mathrm{X}) = 1.93$ bits.

From the examples we can see that the lengths of the code are close to the Shannon information content—sometimes smaller, sometimes larger. The Huffman code achieves an expected length that satisfies $H(\mathrm{X}) \leq L(\mathscr{E}, \mathrm{X}) < H(\mathrm{X}) + 1$, as proven in theorem 6.5. This overhead between 0 and 1 bits per symbol can be in practice a big one, when $H(\mathrm{X})$ is itself 1 bit. Indeed, in practice one can achieve better compression not by *symbol codes*, but using *stream codes*.

We can prove optimality of the Huffman code by the following argument. First, we need a definition and a lemma.

Definition 6.3 (Sibling strings)**.** Two binary strings $x, y \in \{0,1\}^+$ are *sibling* if there exists $p \in \{0,1\}^+$ such that $x = p0$ and $y = p1$.

Lemma 6.4. *The optimal uniquely decodable code $\mathscr{E}_O$ must have the longest codewords occurring in pairs of sibling strings, with one pair encoding the two least probable symbols x_a, x_b.*

Proof. Since the optimal code can be chosen to be prefix without loss of generality, the longest codewords must occur in sibling pairs whose length we denote by l_M. Indeed, if one of the longest codewords had no sibling, since the code $\mathscr{E}_B$ is prefix, one could shorten it without spoiling the prefix property, contradicting the optimality hypothesis. Suppose now that none of the longest sibling strings encodes one of the two least probable symbols—say x_a. Let then $l(x_a) < l(x_c) = l_M$, with $p(x_a) < p(x_c)$. If we now build a different code $\mathscr{E}_C$, exchanging the codewords for x_a, x_c, the difference of the expected lengths is

$$\Delta L := L(\mathscr{E}_C, \mathrm{X}) - L(\mathscr{E}_B, \mathrm{X}) = (p_c l_a + p_a l_c) - (p_c l_c + p_a l_a) = (p_c - p_a)(l_a - l_c) < 0.$$

This implies that the new code is better, contrarily to the optimality hypothesis. Thus, the best code must associate two sibling longest codewords to the two least probable symbols. □

Optimality of Huffman codes can now be proved by induction on the number of symbols in the alphabet $\mathsf{Rng}(\mathrm{X})$ as follows.

Theorem 6.5 (Optimality of the Huffman code)**.** *The code constructed by Huffman's algorithm is optimal.*

Proof. The proof proceeds by induction. Indeed, for an alphabet with a unique symbol the encoding results in a code with $L(\mathscr{E}_H, \mathrm{X}) = 0$, which is clearly optimal. Suppose now that the Huffman algorithm provides the optimal code $\mathscr{E}_{H,n}$ for alphabets with n symbols, and consider an alphabet $\mathsf{Rng}(\mathrm{X})$ with $n+1$ symbols. By the above lemma the optimal code $\mathscr{E}_O$ for X must associate two sibling strings of maximal length to the least probable symbols $x_n^{(0)}, x_{n+1}^{(0)}$. Let us then form the random variable X_1 with $x_j^{(1)} = x_j^{(0)}$ for $1 \leq j \leq n-1$ and $x_n^{(1)} := \{x_n^{(0)}, x_{n+1}^{(0)}\}$. The probabilities of X_1 are $p(x_i^{(1)}) = p(x_i^{(0)})$ for $1 \leq i \leq n-1$ and $p(x_n^{(1)}) := p(x_n^{(0)}) + p(x_{n+1}^{(0)})$. Consider now the code $\mathscr{E}_{O,n}$ obtained from $\mathscr{E}_O$ by merging the symbols $x_n^{(0)}, x_{n+1}^{(0)}$, and associating to them the string $\mathscr{E}_{O,n}(x_n^{(1)}) := [\mathscr{E}_O(x_n^{(0)})]' = [\mathscr{E}_O(x_{n+1}^{(0)})]'$, where w' means the string w without the last bit. The code $\mathscr{E}_{O,n}$ is a code for the n-symbol random variable X_1. Now, one has

$$\begin{aligned} L(\mathscr{E}_O, \mathrm{X}) &= \sum_{i=1}^{n-1} p(x_i^{(0)}) l_{\mathscr{E}_O}(x_i^{(0)}) + p(x_n^{(0)}) l_{\mathscr{E}_O}(x_n^{(0)}) + p(x_{n+1}^{(0)}) l_{\mathscr{E}_O}(x_{n+1}^{(0)}) \\ &= \sum_{i=1}^{n-1} p(x_i^{(1)}) l_{\mathscr{E}_{O,n}}(x_i^{(1)}) + p(x_n^{(1)})[l_{\mathscr{E}_{O,n}}(x_n^{(1)}) + 1] \\ &= L(\mathscr{E}_{O,n}, \mathrm{X}_1) + p(x_n^{(0)}) + p(x_{n+1}^{(0)}). \end{aligned}$$

If we now consider the Huffman code $\mathscr{E}_{H,n+1}$, by an identical calculation we obtain

$$L(\mathscr{E}_{H,n+1}, \mathrm{X}) = L(\mathscr{E}_{H,n}, \mathrm{X}_1) + p(x_n^{(0)}) + p(x_{n+1}^{(0)}),$$

where $L(\mathscr{E}_{H,k}, \mathrm{Y})$ denotes the expected length of the Huffman code for the k-symbol random variable Y. Finally, by the induction hypothesis we have that $L(\mathscr{E}_{O,n}, \mathrm{X}_1) \geq L(\mathscr{E}_{H,n}, \mathrm{X}_1)$, which implies $L(\mathscr{E}_O, \mathrm{X}) \geq L(\mathscr{E}_{H,n+1}, \mathrm{X})$. Since $\mathscr{E}_O$ is optimal by hypothesis, it must be $L(\mathscr{E}_O, \mathrm{X}) = L(\mathscr{E}_{H,n+1}, \mathrm{X})$, thus proving the optimality of the Huffman code. □

6.2 Stream codes and redundancy

We have seen that symbol codes obey a source coding theorem that envisages a possible overhead of 1 bit/symbol. This may seem a little cost, however one should consider that e.g. a text written in the English language has an essential bit content of about 1 bit/symbol. This means that an overhead of 1 bit/symbol is huge. One possible way around the problem consists in treating long strings as symbols, namely encoding

$\mathsf{Rng}(\mathrm{X}^N)$ instead of $\mathsf{Rng}(\mathrm{X})$. However, even this strategy can be very inefficient. Indeed, the Huffman algorithm constructs a full encoding table, which takes an exponential time to complete. In particular, sorting strings according to their decreasing probability is the part of the algorithm that takes exponential time versus the number of symbols. This makes the encoding necessarily rigid, since it is not convenient to update the table once the probabilities are updated. For this reason in practical applications where the i.i.d. hypothesis for symbols within long strings is not at all justified, it is convenient to go beyond symbol codes.

A better strategy with respect to the construction of Huffman's code tables for every update of probabilities is one which can be easily adapted to longer block lengths without an exponential computational complexity. This is the case of *stream codes*. We will analyse two main cases of stream codes: the arithmetic code and the Lempel-Ziv code. Arithmetic coding compresses data on the basis of a probabilistic modelling of the source. Since 1999 this is the best compression method for text files, and it is used also by some advanced image compressors. The Lempel-Ziv coding is a *universal* method, which is designed to compress any source. Even though it is not so effective as the arithmetic codes, it has a reasonable efficiency on most files, and it is proved to reach the Shannon limit asymptotically. The Lempel-Ziv algorithm is used in the popular `compress` and `gzip` routines.

The redundancy of the English language is a well known fact: in cross-puzzles, or in the game *hangman*, one can guess quite effectively what is the next letter in an incomplete word. The redundancy is not just due to the fact that characters are non equally distributed, but in particular it is due to the dependence of probabilities of symbols on the preceding character. More generally, entire strings of symbols have a likelihood that depends on the context. It is this kind of redundancy that can be exploited in compressing a text file. Indeed, the hang man example is particularly suited to our purpose, since we will now see how a similar guessing game can be used for compression.

Compression as a guessing game

Consider the following simple game in which each consecutive character of a sentence should be guessed by a player, and the number of attempts is recorded by a referee for each character. We restrict to the 27 characters of the English alphabet including only upper-case letters `A`,`B`,`C`,...,`Z` and the space '-'. For example, we report a sentence and the corresponding list of attempts below

```
I LOVE YOU
11 8211 311
```

Notice how easy it is to guess the right letter in just a single attempt, e.g. after `Y`. How can the above game be used for compression? Suppose that there is an exact twin of the guessing player, who makes exactly the same guesses as the first player when they are in the same context. Suppose that now there is a second referee, who claims to play the same game with the twin, as if they knew the sentence. However, all the second referee knows is the sequence of numbers of attempts from the first twin. The referee would

ask the second twin to guess the letter, and would say 'yes, that's right!' whenever the number of guesses equals the number in the sequence. Since the second player is a perfect twin of the first one, exactly the same string would be obtained. It is easy to recognise that the first player has the role of the *encoder*, whereas his twin is the *decoder*. Indeed, we have not reduced the size of the alphabet, however, since the number of attempts is a small number most of the times, the new file has much less redundancy than the original one, and will be much easier to compress.

Do we have perfectly identical twins in practice? Yes: they would be simulated by a computer program with a rule for generating attempts for any possible preceding string (up to some maximum window length). This would be just a particular probabilistic language model—in our case this is the L-th order Markov model with maximum window length L. However, any probabilistic model can be used, more or less effectively, depending on the adequateness of the probabilistic model.

6.3 Arithmetic code

Let us consider a random variable X with $\mathsf{Rng}(\mathrm{X}) = \{x_1, x_2, \ldots, x_n\}$, including the 'end of file' symbol $x_n := \odot$. Let us now consider the variable $\mathrm{X}^{(N)}$, with $\mathsf{Rng}(\mathrm{X}^{(N)}) = \mathsf{Rng}(\mathrm{X})^N$, but *not* i.i.d. Since the joint probabilities $P[x_{\mathbf{i}}^{(N)}] := \mathbb{P}_{\mathrm{X}^{(N)}}[x_{i_1}, x_{i_2}, \ldots, x_{i_N}]$ of strings are not factorised, we need an efficient algorithm to compute them, in the first place. This is achieved through the computation of *conditional probabilities* $P[x_{i_k}|x_{\mathbf{i}}^{(k-1)}]$, where $x_{\mathbf{i}}^{(k)}$ denotes the string made of the first k symbols of $x_{\mathbf{i}}^{(N)}$, namely $x_{\mathbf{i}}^{(k)} = x_{\mathbf{i}}^{(k-1)} x_{i_k}$. The i-th character in the string will be denoted by the random variable $\mathrm{X}^{(i)}$. The same algorithm will be used to compute conditional probabilities, both at the encoding and at the decoding stage. This algorithm models the twin guessers.

Encoding and decoding

Every binary transmission can be regarded as the transmission of rational number belonging to the interval $[0, 1)$: the general string $x_{\mathbf{i}}^{(k)}$ made of k binary digits is just the binary form of the fractional part of the number. For example the string 0101 corresponds to $\frac{1}{4} + \frac{1}{16} = 0.3125$. Let us denote the above map as

$$\mathsf{Bin} : \{0,1\}^+ \to [0,1),$$
$$\mathsf{Bin} : x_{\mathbf{i}}^{(k)} \mapsto \sum_{j=1}^{k} x_{i_j} 2^{-j}.$$

One can easily verify that the above map is bijective.

We now consider the following variation of the map Bin, that associates binary sequences $b \in \{0,1\}^+$ to *binary subintervals* of $[0, 1)$ in the following way:

$$0 \to [0, .5),\ 1 \to [.5, 1),$$

$$00 \to [0, .25),\ 01 \to [.25, .5),\ 10 \to [.5, .75),\ 11 \to [.75, 1),$$

and rewriting the intervals in binary form

$$0 \to [0, .1),\ 1 \to [.1, 1),$$

$$00 \to [0, .01),\ 01 \to [.01, .1),\ 10 \to [.1, .11),\ 11 \to [.11, 1),$$

and so on. For example one has

$$01011 \to [.01011, .01011 + .00001) = [.01011, .01100).$$

Let us define the set BI of binary subintervals of $[0, 1)$, namely

$$\mathsf{BI} := \{[\, 0.x_{i_1}x_{i_2}\ldots x_{i_k}\,,\, 0.x_{i_1}x_{i_2}\ldots x_{i_k} + 0.\underbrace{00\ldots0}_{k-1 \text{ zeros}}1\,) \mid x_{\mathbf{i}}^{(k)} \in \{0,1\}^+\}$$

We will denote the above bijective map as

$$\mathsf{Bintv} : \{0,1\}^+ \leftrightarrow \mathsf{BI},$$

where BI denotes the set of *binary intervals* in $[0, 1)$ defined above.

How can we encode information using a mapping on intervals? The idea is the following. We can divide the interval $[0, 1)$ also into n subintervals of lengths equal to the probabilities $\mathbb{P}_{\mathrm{X}^{(N)}}[\mathrm{X}^{(1)} = x_{i_1}]$, $i = 1, 2, \ldots, n$. We will denote these intervals by the symbols $I_{x_{i_1}}$. Then we can subdivide each interval x_{i_1} into subintervals denoted as $I_{x_{i_1}x_{i_2}}$, with lengths $|I_{x_{i_1}x_{i_2}}|$ proportional to $\mathbb{P}_{\mathrm{X}^{(N)}}[\mathrm{X}^{(2)} = x_{i_2}|\mathrm{X}^{(1)} = x_{i_1}]$, or $P[x_{i_2}|x_{i_1}]$ in short notation. Indeed the length $|I_{x_{i_1}x_{i_2}}|$ of the interval $I_{x_{i_1}x_{i_2}}$ will be precisely the joint probability

$$|I_{x_{i_1}x_{i_2}}| = \mathbb{P}_{\mathrm{X}^{(N)}}[X^{(1)} = x_{i_1}, X^{(2)} = x_{i_2}] = \mathbb{P}_{\mathrm{X}^{(N)}}[X^{(1)} = x_{i_1}]\mathbb{P}_{\mathrm{X}^{(N)}}[X^{(2)} = x_{i_2}|X^{(1)} = x_{i_1}].$$

Upon iterating the process, we will have a division into tiny intervals with lengths given by

$$|I_{x_{\mathbf{i}}^{(k)}}| = \mathbb{P}_{\mathrm{X}^{(N)}}[x_{\mathbf{i}}^{(k)}] = \mathbb{P}_{\mathrm{X}^{(N)}}[x_{\mathbf{i}}^{(k-1)}]\mathbb{P}_{\mathrm{X}^{(N)}}[x_{i_k}|x_{\mathbf{i}}^{(k-1)}]. \tag{6.1}$$

The above map is one-to-one by construction, and we will denote it by

$$\mathsf{Intv}_{\mathrm{X}} : \mathsf{Rng}(\mathrm{X})^+ \leftrightarrow \mathsf{I}(\mathrm{X}),$$

where $\mathsf{I}(\mathrm{X})$ denotes the set of intervals $I_{x_{\mathbf{i}}^{(k)}}$ in $[0, 1)$ defined above. Notice that the set $\mathsf{BI} \equiv \mathsf{I}(\mathrm{B})$ can be regarded as a special case of $\mathsf{I}(\mathrm{X})$ corresponding to the random variable $\mathrm{X} = \mathrm{B}$ with $\mathsf{Rng}(\mathrm{B}) = \{0, 1\}$ and uniform probability distribution. Similarly,

$$\mathsf{Bintv} \equiv \mathsf{Intv}_{\mathrm{B}} : \mathsf{Rng}(\mathrm{B})^+ \leftrightarrow \mathsf{I}(\mathrm{B}).$$

Now, consider the map

$$\mathsf{Inc}_{\mathrm{X}} : \mathsf{I}(\mathrm{X}) \to \mathsf{I}(\mathrm{B}), \quad \mathsf{Inc}_{\mathrm{X}}(I_{\mathbf{i}}) := \arg \max_{\substack{I_{b_{\mathbf{j}}} \in \mathsf{I}(\mathrm{B}) \\ I_{b_{\mathbf{j}}} \subseteq I_{x_{\mathbf{i}}}}} |I_{b_{\mathbf{j}}}|$$

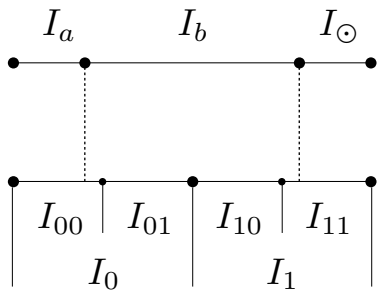


Figure 6.1 An example of interval $I_{x_\mathbf{i}} = I_b$ with two largest binary subintervals $I_{b_{\mathbf{j}_1}} = I_{01}$ and $I_{b_{\mathbf{j}_2}} = I_{10}$.

which maps the interval $I_{x_\mathbf{i}} \in \mathsf{I}(\mathrm{X})$ to the largest binary interval $I_{b_\mathbf{j}}$ contained in $I_{x_\mathbf{i}}$. However, the map Inc_X is generally not well defined, because given a random variables $\mathrm{X}^{(N)}$, it may happen that for some string $x_\mathbf{i}$ the interval $I_{x_\mathbf{i}}$ contains *two* largest intervals $I_{b_{\mathbf{j}_1}}$ and $I_{b_{\mathbf{j}_2}}$ of the same size. See e.g. Fig. 6.1. Notice that if the largest subinterval $I_{b_\mathbf{j}}$ contained in $I_{x_\mathbf{i}}$ is not unique, there can only be two largest included intervals. Indeed, if there are three included intervals $I_{b_{\mathbf{j}_1}}$, $I_{b_{\mathbf{j}_2}}$ and $I_{b_{\mathbf{j}_3}}$ of size 2^{-k}, then either $I_{b_{\mathbf{j}_1}} \cup I_{b_{\mathbf{j}_2}}$ or $I_{b_{\mathbf{j}_2}} \cup I_{b_{\mathbf{j}_3}}$ would be an interval in $\mathsf{I}(\mathrm{B})$, included in $I_{x_\mathbf{i}}$, of size 2^{-k+1}, and thus the three intervals $I_{b_{\mathbf{j}_1}}$, $I_{b_{\mathbf{j}_2}}$ and $I_{b_{\mathbf{j}_3}}$ could be the largest intervals of $\mathsf{I}(\mathrm{B})$ included in $I_{x_\mathbf{i}}$. Since the inclusion map Inc_X is precisely the one that we will use for arithmetic coding, we complete the definition by the rule that whenever there are two largest binary intervals $I_{b_{\mathbf{j}_1}}$ and $I_{b_{\mathbf{j}_2}}$ in $\mathsf{I}(\mathrm{B})$ included in $I_{x_\mathbf{i}}$, then

$$\mathsf{Inc}_\mathrm{X}(I_{x_\mathbf{i}}) := I_{\min\{b_{\mathbf{j}_1}, b_{\mathbf{j}_2}\}}.$$

Even though this case might seem rather exceptional, it will play a crucial role in the evaluation of the compression rate of the arithmetic coding. Having defined the map Inc_X, one can easily prove that it is invertible on its range, with

$$\mathsf{Inc}_\mathrm{X}^{-1} : \mathsf{I}(\mathrm{B}) \to \mathsf{I}(\mathrm{X}), \quad \mathsf{Inc}_\mathrm{X}^{-1}(I_{b_\mathbf{j}}) := \arg \min_{\substack{I_{x_\mathbf{i}} \in \mathsf{I}(\mathrm{X}) \\ I_{x_\mathbf{i}} \supseteq I_{b_\mathbf{j}}}} |I_{x_\mathbf{i}}|.$$

Now, to encode the string $x_\mathbf{i}$ we just locate the corresponding interval, and send the binary string of the largest interval (unique by construction) that lies in $I_{x_\mathbf{i}}$. This corresponds to the one-to-one map

$$\mathscr{E}(x_\mathbf{i}) = \mathsf{Intv}_\mathrm{B}^{-1} \circ \mathsf{Inc}_\mathrm{X} \circ \mathsf{Intv}_\mathrm{X}(x_\mathbf{i}). \tag{6.2}$$

The fact that the encoding is a one-to-one map guarantees us that we can decode by just using the inverse map $\mathscr{D} := \mathscr{E}^{-1}$. In practice the decoding can be performed on the fly very efficiently by the following routines.

Encoding

The probabilities $P[x_i^{(1)}]$ for all symbols, including the separation and end-of-file characters, are included in the model. At step k, the encoder just evaluates the conditional probabilities $P[x_{i_k}|x_{\mathbf{i}}^{(k-1)}]$ for the particular string $x_{\mathbf{i}}^{(k-1)}$ that has been input up to step $k-1$. Then, they calculate the boundaries of the interval $I_{x_{\mathbf{i}}^{(k)}}$. Considering now binary intervals, they find the largest one included in $I_{x_{\mathbf{i}}^{(k)}}$, say $I_{b_{\mathbf{j}}^{(k')}}$. Normally, $I_{b_{\mathbf{j}}^{(k')}}$ is a unique binary subinterval of $I_{b_{\mathbf{j}}^{(k'-1)}}$, thus providing the next encoded bit. It may happen, however, that $I_{b_{\mathbf{j}}^{(k)}}$ is not univocally identified. In this case, they proceed to the step $k+1$. If the last symbol x_{i_N} is such that the largest binary interval included in $I_{x_{\mathbf{i}}^{(N)}}$ is not unique, they adopt the convention described above, taking the leftmost interval.

Decoding

The decoder uses the same algorithm as the one used by the encoder and calculates the probabilities $P[x_i^{(1)}]$ for all symbols. The decoder examines the first bit, and if the corresponding binary interval fits in the interval corresponding to any of the $x_i^{(1)}$'s, it starts decoding the first symbol. Otherwise they consider the binary interval corresponding to the first two bits, and so on increasing the number of bits, until an enclosing interval $I_{x_{i_1}}$ is found. Then they compute the probabilities $P[x_{i_2}|x_{i_1}^{(1)}]$ for all $x_{i_2} \in \mathsf{Rng}(X)$, and deduce the boundaries of the intervals $x_{\mathbf{i}}^{(2)}$. Next, they examine the subsequent encoded bits up to identification of the second Interval $I_{x_{\mathbf{i}}^{(2)}}$, from which they can decode the second symbol x_{i_2}. They then compute the probabilities $P[x_{i_3}|x_{\mathbf{i}}^{(2)}]$ for all $x_{i_3} \in \mathsf{Rng}(X)$, and deduce the boundaries of the intervals $I_{x_{i_3}}$. The algorithm is repeated until the 'end-of-file' character $\odot$ is decoded.

Compression rate

Arithmetic coding is very nearly optimal. Indeed, by construction the encoding map gives $\mathscr{E}(x_{\mathbf{i}}^{(N)}) = b_{\mathbf{j}}^{(k)}$ with $|I_{b_{\mathbf{j}}^{(k)}}| = \frac{1}{2^k} \leq |I_{x_{\mathbf{i}}^{(N)}}| = \mathbb{P}[x_{\mathbf{i}}^{(N)}]$. Thus, the number k of bits used to encode the string $x_{\mathbf{i}}^{(N)}$ is bounded by

$$k = \log_2 \frac{1}{|I_{b_{\mathbf{j}}^{(k)}}|} \geq \log_2 \frac{1}{\mathbb{P}[x_{\mathbf{i}}^{(N)}]},$$

which is precisely the Shannon information content of the encoded string. In order to find an upper bound to $l(x_{\mathbf{i}}^{(N)})$, one has to find a lower bound to $|I_{b_{\mathbf{j}}^{(k)}}|$, which requires the analysis of the situation in Fig. 6.1: in the case when there are two largest binary subintervals of $I_{x_{\mathbf{i}}^{(N)}}$, indeed one has $|I_{b_{\mathbf{j}}^{(k)}}| \geq |I_{x_{\mathbf{i}}^{(N)}}|/4$, which thus implies

$$k = \log_2 \frac{1}{|I_{b_{\mathbf{j}}^{(k)}}|} \leq \log_2 \frac{4}{|I_{x_{\mathbf{i}}^{(N)}}|} = \log_2 \frac{1}{\mathbb{P}[x_{\mathbf{i}}^{(N)}]} + 2,$$

thus leading to a possible overhead of two bits. Notice however that the overhead is now computed over the full string, namely we have $2/N$ bits overhead per symbol, to be compared with 1 bit/symbol of the Huffman code.

Complexity

Also the complexity of probability tables construction for the arithmetic coding is by far smaller than that of Huffman's code. Indeed, in order to communicate a string of N letters both the encoder and the decoder needed to compute only $N|\mathsf{Rng}(X)|$ conditional probabilities, namely the probabilities of all possible letters, only in those (N) contexts that are actually encountered. This must be compared to the complexity of using Huffman's code with large blocks (N-symbols), where all sequences that may occur in principle must be considered, and all their probabilities must be evaluated. Their number amounts to $|\mathsf{Rng(X)}|^N$, which grows exponentially as a function of the length N, instead of the linear growth required for the arithmetic code. Also, notice the flexibility of the algorithm: it can be used with any alphabet and any probability distribution, depending on the case.

The probabilistic model

For the purpose of evaluating the probabilities it is common to use a Bayesian model, in which the conditional probabilities are given by the Dirichlet rule

$$P[x_{i_k}|x_{\mathbf{i}}^{(k-1)}] = \frac{F_{x_{i_k}}(x_{\mathbf{i}}^{(k-1)}) + \alpha}{\sum_{i_l}[F_{x_{i_l}}(x_{\mathbf{i}}^{(k-1)}) + \alpha]}, \tag{6.3}$$

$0 < \alpha \leq 1$, where $F_{x_{i_l}}(x_{\mathbf{i}}^{(k-1)})$ denotes the number of times that x_{i_l} occurs in the string $x_{\mathbf{i}}^{(k-1)}$. In other words, the probability is updated based on the number of occurrences in the previous steps. The parameter α sets an updating responsivity. The most responsive update rule is obtained by $\alpha = 0$. The special case $\alpha = 1$ is the most conservative model, and is known as Laplace rule.

6.4 Lempel-Ziv algorithm

Differently from the arithmetic coding, the Lempel-Ziv algorithm does not rely on probabilistic modelling. Roughly speaking, the compression method works on binary strings, and basically replaces a substring with a pointer to an earlier occurrence of the same substring, that has been suitably recorded in a *dictionary*. The incoming binary string $x_{\mathbf{i}}^{(N)}$ is parsed into substrings, in such a way that every substring is composed by a prefix that has already been included in the dictionary plus an extra bit. For example, if the k-th substring is $= x^{(j)}x_{i_k}$, then the encoder outputs $(s(x^{(j)}), x_{i_k})$, namely the pointer $s(x^{(j)})$ to the previously recorded substring $x^{(j)}$ plus the tail bit x_{i_k}. In the output string the parentheses and comma separators are omitted. The algorithm starts setting by convention $x_0 = \lambda$ and $x^{(0)} = \lambda$, namely for the first string the output is (λ, x_{i_1}).

The algorithm can be expressed in detail as follows.

1. Append the empty string λ in front of the input string $x_{\mathbf{i}}^{(N)}$, as follows $x_{\mathbf{i}}^{(N)} \mapsto \lambda x_{\mathbf{i}}^{(N)}$.

2. Parse $\lambda x_{\mathbf{i}}^{(N)}$ into substrings, from left to right, in such a way that the j-th substring did not appear in the first $j-1$ instances.

3. Assign a binary pointer to every string, using ordered binary strings.

4. Re-write every substring—whose general length is denoted by l—as a couple, where the first element is the $l-1$-bits prefix string, and the second element is the l-th bit. Substitute the prefix with its pointer to the dictionary. When processing the k-th string, express the pointer using $\lceil \log_2 k \rceil$ bits.

5. Remove all parentheses and commas, thus obtaining the encoded string $\mathscr{E}(x_{\mathbf{i}}^{(N)})$

The decoding algorithm works exactly in the same way, apart form the fact that the dictionary must be decoded.

1. Append the empty string λ in front of the input string $x_{\mathbf{i}}^{(N)}$, as follows $\mathscr{E}(x_{\mathbf{i}}^{(N)}) \mapsto \lambda\mathscr{E}(x_{\mathbf{i}}^{(N)})$.

2. Parse $\lambda\mathscr{E}(x_{\mathbf{i}}^{(N)})$ into substrings, from left to right, in such a way that the j-th substring has a prefix of $\lceil \log_2 j \rceil$ bits and a tail of 1 bit.

3. Assign a string to the binary pointer k by decoding the k-th couple, reading the string in the dictionary slot addressed by the prefix, and adding to it the tail bit.

4. Juxtapose all the dictionary entries from left to right thus obtaining the decoded string $x_{\mathbf{i}}^{(N)}$

Let us see how the Lempel-Ziv algorithm works in some specific examples.

Example 6.6. Consider the input string

$$\texttt{1011010100010}\ldots$$

1. We initialise the string appending the empty string λ to the left, thus obtaining $\lambda\texttt{1011010100010}\ldots$.

2. We start parsing the string into an ordered *dictionary* of substrings, where the j-th string has not appeared in the first $j-1$ instances, as follows:

$$\lambda, \texttt{1}, \texttt{0}, \texttt{11}, \texttt{01}, \texttt{010}, \texttt{00}, \texttt{10}, \ldots$$

3. We include the empty substring λ as the first substring in the dictionary, and assign ordered binary pointers to the substrings in the dictionary, following the order in which they appear.

source string: $\texttt{1011010100010}\ldots$.

$x^{(j)}$	λ	1	0	11	01	010	00	10
$s(x^{(j)})$	λ	1	10	11	100	101	110	111

4. We re-write every substring as a couple, expressing the pointer by $\lceil \log_2 k \rceil$ bits, where k is the number of strings previously included in the dictionary.

source string: 1011010100010...

$x^{(j)}$	λ	1	0	11	01	010	00	10
$s(x^{(j)})$	λ	1	10	11	100	101	110	111
(pointer,bit)		(λ,1)	(0,0)	(01,1)	(10,1)	(100,0)	(010,0)	(001,0)

5. After removing parentheses and commas, we end up with the following encoded output: 100011101100001000010....

In this example the output is actually longer than the input, because there is not much redundancy in the source string.

Example 6.7. Consider the input string 000000000000100000000000.

1. Initialised string: λ000000000000100000000000

2. Parsed string

$$\lambda 0, 00, 000, 0000, 001, 00000, 000000$$

3. Dictionary

source string: 000000000000100000000000....

$x^{(j)}$	λ	0	00	000	0000	001	00000	000000
$s(x^{(j)})$	λ	1	10	11	100	101	110	111

4. Re-writing

source string: 000000000000100000000000....

$x^{(j)}$	λ	0	00	000	0000	001	00000	000000
$s(x^{(j)})$	λ	1	10	11	100	101	110	111
(pointer,bit)		(λ,0)	(1,0)	(10,0)	(11,0)	(010,1)	(100,0)	(110,0)

5. Encoded output: 010100110010110001100....

Exercise 6.1

Decode the string 00101011101100100100011010101000011.

Answer of exercise 6.1

1. Initialised string λ00101011101100100100011010101000011

2. Parsed string

$$\lambda, 0, 01, 010, 111, 0110, 0100, 1000, 1101, 01010, 00011$$

3. Dictionary

source string: 00101011101100100100011010101000011...

	(λ,0)	(0,1)	(01,0)	(11,1)	(011,0)	(010,0)	(100,0)	(110,1)	(0101,0)	(0001,1)
λ	1	10	11	100	101	110	111	1000	1001	1010
λ	0	1	00	001	000	10	0010	101	0000	01

4. Encoded output: 0100001000100010101000001

Even though the Lempel-Ziv coding doesn't rely on any probabilistic model, it is possible to show that for an ergodic source (i.e. memoryless on a sufficiently long scale) asymptotically it compresses up to the entropy of the source. The effectiveness of the Lempel-Ziv relies on the fact that in practice many files contain repetition of particularly short sequences of characters, a kind of redundancy which the algorithm is suited to.

Summary on codes

We have seen that symbol codes employ a variable-length code for each symbol of the source alphabet. The optimal code lengths are determined by the probabilities of the different symbols. The Huffman's code constructs an optimal symbol code for a given alphabet and probability distribution. The encoding is lossless, i.e. it is uniquely decodeable, and if the source has the assumed probability distribution, the expected length will stay in the interval $[H, H+1)$, clearly with possible statistical fluctuations, that can give lengths longer or shorter than the expected length. If the source is not matched to the assumed probability distribution, then the expected length is increased by the relative entropy between the source distribution and the code's implicit distribution. For source with small entropy the code uses at least one bit per source symbol, and essentially no compression is achieved. One can, however, achieve compression by using bigger alphabets made of blocks of characters.

Stream codes, compared with symbol codes, are not constrained to emit at least one bit per input symbol, hence one can encode large strings into a smaller number of bits. Among stream codes, the arithmetic ones combine a probabilistic model with an encoding algorithm that identifies each string with a subinterval of $[0, 1)$ of size equal to the probability of that string under the model. Such coding is almost optimal in the sense that asymptotically the compressed length of a input string equals its Shannon information content, namely the expected length averaged over many strings achieves the Shannon limit. On the opposite side, Lempel-Ziv codes are independent of any probability model, and essentially compress based on the principle of building up a database of strings and encoding the strings into a list of pointers to the database (the length of the pointer will be logarithmically shorter than the string itself). Also the Lempel-Ziv coding achieves the Shannon compression limit asymptotically.

Both arithmetic and Lempel-Ziv encoding are very fragile to any alteration of the encoded string: a single bit error will completely destroy the remaining part of the compressed file. Therefore, if we need to communicate or store over noisy media, it will be essential to use error-correcting codes.

Chapter 7

Lecture 7: Mutual information and Markov chains

In the previous section we have considered encoding in order to achieve compression of information. However, we have implicitly assumed that the channel between the encoding and the decoding was noise-free. In real-life instances, channels are noisy, and even the tiniest error can ruin dramatically the recovery of the encoded file—as we have seen, especially when using the most efficient stream codes.

The aim of the next lectures is then to make the noisy channel effectively behave like a noiseless one. This will be achieved by error-correction coding. Therefore, the communication scheme will be the following. The input stream is first compressed, then encoded with redundancy for later error-correction. The output will then enter the noisy physical channel. At the output of the channel a first decoder will recover errors, and finally a second decoder will decompress the message and restore the original string. This communication scheme is summarised in Fig. 7.1.

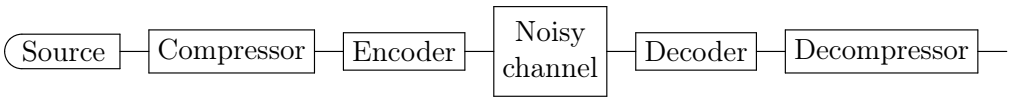


Figure 7.1 Schematic representation of a reliable and efficient communication channel

The problem with errors is that we do not know where they occurred in a given received string. For example one may think that a probability of bit-flip $f = 0.1$ in a symmetric binary channel would affect the rate of transmission of information by 10%, but this is not true: the situation is much worse. Indeed, for $f = 1/2$ there will be no transmitted information at all. A quantification of the transmitted information will be given by the mutual information between input and output. In the next lectures we will see that the channel capacity—i.e. the mutual information maximised over all prior probabilities of the input alphabet—provides the maximum rate at which we can transmit reliably through a noisy channel by a suitable encoding: this is the main content of the famous Shannon's noisy channel theorem.

7.1 Entropies for bi- and multi-variate random variables

7.1.1 Joint entropy and conditional entropy

In the remainder of this lecture we will consider the joint random variable $\mathrm{Z} = (\mathrm{X}, \mathrm{Y})$ whose range contains ordered couples (x_i, y_j). As already discussed in section 3.1, one can define the two marginal random variables X and Y with marginal probabilities $p(x_i) = \sum_{y_j \in \mathsf{Rng}(\mathrm{Y})} p(x_i, y_j)$ and $p(y_j) = \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} p(x_i, y_j)$. Notice that generally $p(x_i, y_j) \neq p(x_i)p(y_j)$, i.e. the two marginal random variables need not be independent. More generally, we will consider multivariate random variables $\mathrm{Z} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_N)$. The marginal distribution for the M-tuple $(\mathrm{X}_{i_1}, \mathrm{X}_{i_2}, \ldots \mathrm{X}_{i_M})$, obtained by excluding the random variables $\mathrm{X}_{l_{M+1}}, \mathrm{X}_{l_{M+2}}, \ldots \mathrm{X}_{l_N}$ is defined by

$$p(x_{j_{i_1}}, x_{j_{i_2}}, \ldots, x_{j_{i_M}}) := \sum_{\substack{x_{j_{l_{M+1}}} \in \mathsf{Rng}(\mathrm{X}_{l_{M+1}}) \\ x_{j_{l_{M+2}}} \in \mathsf{Rng}(\mathrm{X}_{l_{M+2}}) \\ \vdots \\ x_{j_{l_N}} \in \mathsf{Rng}(\mathrm{X}_{l_N})}} p(x_{j_1}, x_{j_2}, \ldots, x_{j_N}).$$

Similarly, conditional probabilities are defined through Bayes'rule

$$p(x_{j_{i_1}}, x_{j_{i_2}}, \ldots, x_{j_{i_M}} | x_{j_{l_{M+1}}}, x_{j_{l_{M+2}}}, \ldots, x_{j_{l_N}}) := \frac{p(x_{j_1}, x_{j_2}, \ldots, x_{j_N})}{p(x_{j_{l_{M+1}}}, x_{j_{l_{M+2}}}, \ldots, x_{j_{l_N}})}.$$

Definition 7.1 (Joint entropy)**.** The joint entropy of two random variables X and Y, denoted as $H(\mathrm{X}, \mathrm{Y})$, is defined as

$$H(\mathrm{X}, \mathrm{Y}) := \sum_{\substack{x_i \in \mathsf{Rng}(\mathrm{X}) \\ y_j \in \mathsf{Rng}(\mathrm{Y})}} \mathbb{P}_{\mathrm{X},\mathrm{Y}}[\mathrm{X} = x_i, \mathrm{Y} = y_j] \log_2 \frac{1}{\mathbb{P}_{\mathrm{X},\mathrm{Y}}[\mathrm{X} = x_i, \mathrm{Y} = y_j]} \tag{7.1}$$

$$= \sum_{\substack{x_i \in \mathsf{Rng}(\mathrm{X}) \\ y_j \in \mathsf{Rng}(\mathrm{Y})}} p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} \tag{7.2}$$

The joint entropy is defined to be symmetric: $H(\mathrm{Y}, \mathrm{X}) := H(\mathrm{X}, \mathrm{Y})$. Indeed, $\mathbb{P}_{\mathrm{X},\mathrm{Y}}[\mathrm{Y} = y_i, \mathrm{X} = x_j] = \mathbb{P}_{\mathrm{X},\mathrm{Y}}[\mathrm{X} = x_j, \mathrm{Y} = y_i] = p(x_i, y_j)$. Now that we have defined the joint entropy of the random variable X, Y, we will refer to the entropy $H(\mathrm{X})$ as the *marginal entropy* of X. The above definitions are straightforwardly extended to the case of multivariate random variables. We will then have

$$H(X_1, X_2, \ldots, X_N) := \sum_{\substack{x_{1_{i_1}} \in \mathsf{Rng}(\mathrm{X}_1) \\ x_{2_{i_2}} \in \mathsf{Rng}(\mathrm{X}_2) \\ \vdots \\ x_{N_{i_N}} \in \mathsf{Rng}(\mathrm{X}_N)}} p(x_{1_{i_1}}, x_{2_{i_2}}, \ldots, x_{N_{i_N}}) \log_2 \frac{1}{p(x_{1_{i_1}}, x_{2_{i_2}}, \ldots, x_{N_{i_N}})}.$$

Theorem 7.2 (Subadditivity of the Shannon entropy)**.** *For every couple of random variables* X *and* Y *the following inequality holds*

$$H(\mathrm{X},\mathrm{Y}) \leq H(\mathrm{X}) + H(\mathrm{Y}), \quad \textit{(with equality holding iff } \mathrm{X} \textit{ and } \mathrm{Y} \textit{ are independent).} \tag{7.3}$$

Proof. One has

$$\begin{aligned} H(p(x_i,y_j)\|p(x_i)p(y_j)) =& - H(p(x_i,y_j)) - \sum_{\substack{x_i\in\mathsf{Rng}(\mathrm{X})\\ y_j\in\mathsf{Rng}(\mathrm{Y})}} p(x_i,y_j)\log_2(p(x_i)p(y_j)) \\ =& - H(\mathrm{X},\mathrm{Y}) + H(\mathrm{X}) + H(\mathrm{Y}). \end{aligned}$$

The statement is then just an application of Gibbs' inequality. Notice that the equal sign holds iff $p(x_i,y_j) = p(x_i)p(y_j)\ \forall x_i, y_j$, namely iff X and Y are independent. □

The theorem then just asserts that the information content of the joint probability $p(x_i,y_j)$ is smaller than or equal to the sum of the two information contents of the two marginals. This can be intuitively understood, considering that the marginal distributions forget about correlations. For example, if we know that the face of a rolled fair die is even we have 1 bit of information. If we know that the face is ⚃, we have $\log_2 6 = 1 + \log_2 3 = 2.58$ bits. However, if we know both facts together, we do not have 3.58 bits. A more rigorous account for this observation is captured by the result that *conditioning reduces entropy.*

Definition 7.3 (Conditional entropy)**.** The conditional entropy $H(\mathrm{X}|\mathrm{Y}=y_j)$ of X given $\mathrm{Y}=y_j$ is the entropy of the conditional probability distribution

$$p(x_i|y_j) = \mathbb{P}_{\mathrm{X}|y_j}[\mathrm{X}=x_i|\mathrm{Y}=y_j] := \frac{\mathbb{P}_{\mathrm{X},\mathrm{Y}}[\mathrm{X}=x_i,\mathrm{Y}=y_j]}{\mathbb{P}_{\mathrm{Y}}[\mathrm{Y}=y_j]}.$$

Explicitly, using the notation $\mathbf{p}_{|y_j} = (p(x_1|y_j), p(x_2|y_j), \dots p(x_n|y_j))$ one has

$$H(\mathrm{X}|\mathrm{Y}=y_j) := H(\mathbf{p}_{|y_j}) = \sum_{x_i\in\mathsf{Rng}(\mathrm{X})} p(x_i|y_j)\log_2\frac{1}{p(x_i|y_j)}.$$

The conditional entropy $H(\mathrm{X}|\mathrm{Y})$ of X given Y is the expectation of $H(\mathrm{X}|\mathrm{Y}=y_j)$ over $\mathsf{Rng}(\mathrm{Y})$, namely

$$\begin{aligned} H(\mathrm{X}|\mathrm{Y}) :=& \mathbb{E}[H(\mathrm{X}|\mathrm{Y}=y_j)] = \sum_{y_j\in\mathsf{Rng}(\mathrm{Y})} p(y_j)H(\mathbf{p}_{|y_j}) = \sum_{y_j\in\mathsf{Rng}(\mathrm{Y})} p(y_j)H(\mathrm{X}|\mathrm{Y}=y_j) \\ =& \sum_{\substack{x_i\in\mathsf{Rng}(\mathrm{X})\\ y_j\in\mathsf{Rng}(\mathrm{Y})}} p(x_i,y_j)\log_2\frac{1}{p(x_i|y_j)}, \end{aligned}$$

where we used Bayes' rule $p(x_i,y_j) = p(y_j)p(x_i|y_j)$.

We stress that the above definition can be applied also to multivariate random variables $\mathrm{Z} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_N)$, provided that one subdivides the X_i's in two multivariate random variables $\mathrm{Y}_1 = (\mathrm{X}_{i_1}, \mathrm{X}_{i_2}, \ldots, \mathrm{X}_{i_k})$ and $\mathrm{Y}_2 = (\mathrm{X}_{j_1}, \mathrm{X}_{j_2}, \ldots, \mathrm{X}_{j_l})$, with $\{1, 2, \ldots, N\} = \{i_1, i_2, \ldots, i_k\} \cup \{j_1, j_2, \ldots, j_l\}$, and then

$$H(\mathrm{X}_{i_1}, \mathrm{X}_{i_2}, \ldots \mathrm{X}_{i_k} | \mathrm{X}_{j_1}, \mathrm{X}_{j_2}, \ldots, \mathrm{X}_{j_l}) := H(\mathrm{Y}_1|\mathrm{Y}_2).$$

Notice that since $H(\mathrm{X}|\mathrm{Y} = y_j) \geq 0$ for all y_j, also $H(\mathrm{X}|\mathrm{Y}) \geq 0$.

Lemma 7.4 (Chain rules for entropies)**.** *For the information content and for the conditional entropy one has the chain rules*

$$\log_2 \frac{1}{p(x_i, y_j)} = \log_2 \frac{1}{p(x_i)} + \log_2 \frac{1}{p(y_j|x_i)} = \log_2 \frac{1}{p(y_j)} + \log_2 \frac{1}{p(x_i|y_j)}, \tag{7.4}$$

$$H(\mathrm{X}, \mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y}|\mathrm{X}) = H(\mathrm{Y}) + H(\mathrm{X}|\mathrm{Y}). \tag{7.5}$$

Proof. The identity for the information content is a trivial consequence of its definition. The second identity is obtained from the first one by just taking the expectation value with $\mathbb{P}_{\mathrm{X},\mathrm{Y}}$ on both sides. □

Corollary 7.5. *The following bound holds*

$$H(\mathrm{X}, \mathrm{Y}) \geq H(\mathrm{X}). \tag{7.6}$$

Proof. From the chain rule $H(\mathrm{X}, \mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y}|\mathrm{X})$ and non negativity of $H(\mathrm{Y}|\mathrm{X})$ we can easily establish the bound $H(\mathrm{X}, \mathrm{Y}) \geq H(\mathrm{X})$. □

Theorem 7.6 (Conditioning reduces entropy)**.** *The conditional entropy is bounded as follows*

$$0 \leq H(\mathrm{X}|\mathrm{Y}) \leq H(\mathrm{X}), \tag{7.7}$$

and is exactly zero if and only if $\mathrm{X} = f(\mathrm{Y})$*, whereas it achieves the upper bound if and only if* X *and* Y *are independent.*

Proof. Non negativity of $H(\mathrm{X}|\mathrm{Y})$ follows from its definition. Now, if $\mathrm{X} = f(\mathrm{Y})$, then $p(x_i = f(y_j)|y_j) = 1$ and $p(x_i \neq f(y_j)|y_j) = 0$, hence $H(\mathrm{X}|\mathrm{Y} = y_j) = H(\mathbf{p}_{|y_j}) = 0$ for every $y_j \in \mathsf{Rng}(\mathrm{Y})$, and $H(\mathrm{X}|\mathrm{Y}) = 0$. Conversely, if $0 < p(x_i|y_j) < 1$ strictly for some x_i, y_j, then $\log_2 \frac{1}{p(x_i|y_j)} > 0$, and consequently $H(\mathrm{X}|\mathrm{Y}) > 0$ strictly. Thus, $H(\mathrm{X}|\mathrm{Y}) = 0$ implies that $p(x_i|y_j) = \delta_{x_i, x(y_j)}$, namely $x_i = f(y_j)$. Finally, the upper bound along with its achievement follow from the subadditivity of Shannon entropy in Theorem 7.2, and the chain rules of Lemma 7.4. Indeed, one has $H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X}, \mathrm{Y}) - H(\mathrm{Y})$, and then

$$H(\mathrm{X}|\mathrm{Y}) - H(\mathrm{X}) = H(\mathrm{X}, \mathrm{Y}) - H(\mathrm{X}) - H(\mathrm{Y}) \leq 0,$$

with equality holding iff X and Y are independent. □

The above theorem has the immediate interpretation that on average *conditioning reduces the entropy*, it does not reduce it if the conditional variable is independent of the conditioning one, and it reduces it to zero when the conditional variable is functionally dependent on the conditioning one.

The conditional entropy $H(\mathrm{X}|\mathrm{Y})$ has a nice interpretation in quantifying the quality of inference of X from Y, in the sense that, intuitively speaking the smaller the conditional entropy, the better the inference: indeed, the uncertainty about X is decreased by the knowledge of Y. The intuition is strengthened by considering the limiting case $H(\mathrm{X}|\mathrm{Y}) = 0$ where one has $\mathrm{X} = f(\mathrm{Y})$. This interpretation is made more rigorous by Fano's inequality, that we will prove in the next lecture.

Notice that the inequality $H(\mathrm{Y}|\mathrm{X}) \leq H(\mathrm{Y})$ holds only in average. Indeed, occurrence of a special outcome $\mathrm{Y} = y_j$ may increase our uncertainty about X, namely there may exist y_j such that

$$H(\mathrm{X}|\mathrm{Y} = y_j) > H(\mathrm{X}).$$

This is the case, for example, for the following distribution for two bits X, Y.

	$p(0,y)$	$p(1,y)$
$p(x,0)$	$\frac{1}{8}$	$\frac{1}{8}$
$p(x,1)$	0	$\frac{3}{4}$

In this case the marginal distribution for Y is obtained by summing the entries in the two rows: $\mathbb{P}_{\mathrm{Y}}[\mathrm{Y} = 0] = \frac{1}{4}$ and $\mathbb{P}_{\mathrm{Y}}[\mathrm{Y} = 1] = \frac{3}{4}$. Thus, $\mathbb{P}_{\mathrm{X}|y=0}[\mathrm{X} = 0|\mathrm{Y} = 0] = \frac{1}{2}$, $\mathbb{P}_{\mathrm{X}|y=0}[\mathrm{X} = 1|\mathrm{Y} = 0] = \frac{1}{2}$, while $\mathbb{P}_{\mathrm{X}|y=1}[\mathrm{X} = 0|\mathrm{Y} = 1] = 0$, $\mathbb{P}_{\mathrm{X}|y=1}[\mathrm{X} = 1|\mathrm{Y} = 1] = 1$. The marginal distribution for X, instead, is obtained by summing the entries in the two columns. In this way we obtain $\mathbb{P}_{\mathrm{X}}[\mathrm{X} = 0] = \frac{1}{8}$, $\mathbb{P}_{\mathrm{X}}[\mathrm{X} = 1] = \frac{7}{8}$. One can then easily compute

$$\begin{aligned}
&H(\mathrm{X}) = H_2(\tfrac{1}{8}) = 0.54\\
&H(\mathrm{X}|\mathrm{Y} = 0) = H_2(\tfrac{1}{2}) = 1,\\
&H(\mathrm{X}|\mathrm{Y} = 1) = H_2(0) = 0,\\
&H(\mathrm{X}|\mathrm{Y}) = 0.25.
\end{aligned}$$

While, clearly, $H(\mathrm{X}|\mathrm{Y}) = 0.25 < H(\mathrm{X}) = 0.54$, for the case $\mathrm{Y} = 0$ we have $H(\mathrm{X}|\mathrm{Y} = 0) = 1 > H(\mathrm{X})$.

7.1.2 Mutual information

We are now in position to define the mutual information.

Definition 7.7 (Mutual information)**.** The *mutual information between* X *and* Y*, denoted by* $I(\mathrm{X} : \mathrm{Y})$*, is defined as*

$$\begin{aligned}
I(\mathrm{X} : \mathrm{Y}) =& H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X})\\
=& H(\mathrm{X}) + H(\mathrm{Y}) - H(\mathrm{X},\mathrm{Y}).
\end{aligned} \tag{7.8}$$

Theorem 7.8 (Properties of the mutual information)**.** *The mutual information satisfies the following properties:*

1. $I(\mathrm{X}:\mathrm{Y}) \geq 0$*;*
2. $I(\mathrm{X}:\mathrm{Y}) = I(\mathrm{Y}:\mathrm{X})$*;*
3. $I(\mathrm{X}:\mathrm{Y}) \leq H(\mathrm{Y})$ *(equality holds iff* $\mathrm{Y} = f(\mathrm{X})$*).*

Proof. Non negativity of mutual information is a direct consequence of subadditivity of the Shannon entropy. The symmetry is a direct consequence of the definition and of symmetry of $H(\mathrm{X},\mathrm{Y})$. Finally the last item is just a restatement of Theorem 7.6. Indeed,

$$H(\mathrm{X}) + H(\mathrm{Y}) - H(\mathrm{X},\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X}) \leq H(\mathrm{Y}),$$

with equality holding iff $H(\mathrm{Y}|\mathrm{X}) = 0$. □

Inferring the output from the input

The mutual information $I(\mathrm{X}:\mathrm{Y})$ represents the amount of information that one can infer about X by knowing Y. This can be seen with the following steps

1. Before reading Y our knowledge about X is given by the marginal probability distribution $p(x_i)$, and uncertainty about X is quantified by the Shannon entropy $H(\mathrm{X})$.
2. We now read Y, getting outcome $\mathrm{Y} = y_j$. We know the conditional probability $p(x_i|y_j)$. For example, this is the case if X and Y represent the input and the output of a channel, since the probability $p(y_j|x_i)$ of reading y_j at the output for input x_i characterises the channel: it is the so-called *transfer matrix*. Hence, the probability that the input was $\mathrm{X} = x_i$ given that we read $\mathrm{Y} = y_j$ is given by the Bayes rule

$$p(x_i|y_j) = \frac{p(x_i,y_j)}{p(y_j)} = \frac{p(x_i)p(y_j|x_i)}{\sum_{x_k\in\mathsf{Rng}(X)} p(x_k)p(y_j|x_k)}. \tag{7.9}$$

   In the process we have updated the *prior* probability to the *posterior* one:

$$\mathbf{p} \longrightarrow \mathbf{p}_{|y_j}.$$

3. How much information have we gained about X from reading $\mathrm{Y} = y_j$? This is quantified by the "decrease" (possibly negative) of our uncertainty from $H(\mathrm{X}) \equiv H(\mathbf{p})$ to $H(\mathrm{X}|\mathrm{Y} = y_j) \equiv H(\mathbf{p}_{|y_j})$, namely the difference $H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y} = y_j)$.
4. Upon averaging over all possible outcomes $y_j \in \mathsf{Rng}(\mathrm{Y})$ we get

$$\sum_{y_j\in\mathsf{Rng}(\mathrm{Y})} p(y_j)[H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y} = y_j)] = H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y}) = I(\mathrm{X}:\mathrm{Y}),$$

   namely, the amount of information that we gained in average on X from Y is given by the mutual information $I(\mathrm{X}:\mathrm{Y})$.

From the above considerations, we see that the mutual information quantifies the amount of information that the output of a channel conveys about the input. Our aim is now to find ways of using the channel in a reliable way, with negligible error probability in recovering the transmitted bits. Shannon's noisy channel theorem shows that this is possible, and the mutual information exactly quantifies the amount of information that can be reliably transmitted.

Regarding the practical problem of evaluating numerically the mutual information for a channel, we notice that using $I(\mathrm{X}:\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X})$ instead of $I(\mathrm{X}:\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y})$ is more convenient, as it avoids evaluating the posterior probabilities in Eq. (7.9).

7.2 Markov chains

In the analysis of channel coding, an important role is played by Markov chains. This notion is conceptually very relevant, as it will allow us to prove the *data processing theorem*, and it turns out to be technically useful to characterise the extreme cases where the inequalities of strong subadditivity hold tightly.

Definition 7.9 (Markov chain)**.** We say that the multivariate random variable

$$\mathrm{Z} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, \mathrm{X}_n, \mathrm{X}_{n+1}, \ldots)$$

is a Markov chain, denoted as

$$\mathrm{X}_1 \to \mathrm{X}_2 \to \mathrm{X}_3 \to \ldots \to \mathrm{X}_n \to \mathrm{X}_{n+1} \to \ldots,$$

if for all $n \geq 1$

$$\begin{aligned} &\mathbb{P}(\mathrm{X}_{n+1} = x_{i_{n+1}} | \mathrm{X}_n = x_{i_n}, \mathrm{X}_{n-1} = x_{i_{n-1}}, \ldots, \mathrm{X}_1 = x_{i_1}) = \\ &\mathbb{P}(\mathrm{X}_{n+1} = x_{i_{n+1}} | \mathrm{X}_n = x_{i_n}). \end{aligned} \tag{7.10}$$

Lemma 7.10. *Given three random variables* X, Y *and* Z*, one has that* $\mathrm{X} \to \mathrm{Y} \to \mathrm{Z}$ *iff* $\mathrm{Z} \to \mathrm{Y} \to \mathrm{X}$*.*

Proof. Indeed, one has

$$p(x_i|y_j, z_k) = \frac{p(x_i, y_j, z_k)}{p(y_j, z_k)} = \frac{p(z_k|x_i, y_j)p(x_i, y_j)}{p(z_k|y_j)p(y_j)}.$$

Now, using the definition of Markov chain, $p(z_k|x_i, y_j) = p(z_k|y_j)$, and then

$$p(x_i|y_j, z_k) = \frac{p(z_k|y_j)p(x_i, y_j)}{p(z_k|y_j)p(y_j)} = p(x_i|y_j). \qquad \square$$

For this reason Markov chains are also denoted as

$$X_1 \leftrightarrow X_2 \leftrightarrow X_3 \leftrightarrow \ldots \leftrightarrow X_n \leftrightarrow X_{n+1} \leftrightarrow \ldots,$$

The "Markovianity" property is the same as conditional independence. More precisely, we have

Lemma 7.11 (Conditional independence)**.** $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$ *if and only if*

$$p(x_i, z_k|y_j) = p(x_i|y_j)p(z_k|y_j). \tag{7.11}$$

Proof. If $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$ one has

$$p(x_i, z_k|y_j) = \frac{p(x_i, y_j, z_k)}{p(y_j)} = \frac{p(z_k|x_i, y_j)p(x_i, y_j)}{p(y_j)} = \frac{p(z_k|y_j)p(x_i, y_j)}{p(y_j)} = p(x_i|y_j)p(z_k|y_j).$$

On the other hand, if $p(x_i, z_k|y_j) = p(x_i|y_j)p(z_k|y_j)$ then one has

$$p(x_i|y_j, z_k) = \frac{p(x_i, y_j, z_k)}{p(y_j, z_k)} = \frac{p(x_i, z_k|y_j)p(y_j)}{p(y_j, z_k)} = \frac{p(x_i|y_j)p(z_k|y_j)p(y_j)}{p(y_j, z_k)} = p(x_i|y_j). \quad \square$$

From the above lemma the following theorem follows.

Theorem 7.12 (Strong subadditivity of the joint entropy)**.** *The following inequality holds*

$$H(\mathrm{X}, \mathrm{Y}, \mathrm{Z}) + H(\mathrm{Y}) \leq H(\mathrm{X}, \mathrm{Y}) + H(\mathrm{Y}, \mathrm{Z}), \quad (\textit{equality holds iff } \mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}). \tag{7.12}$$

Proof. If we treat $p(x_i, z_k|y_j)$ as a usual joint distribution, where y_j is just a parameter, we can apply subadditivity of Shannon entropy to obtain

$$H(\mathrm{X}|\mathrm{Y} = y_j) + H(\mathrm{Z}|\mathrm{Y} = y_j) - H(\mathrm{X}, \mathrm{Z}|\mathrm{Y} = y_j) \geq 0.$$

Averaging over y_j one has

$$\begin{aligned} &H(\mathrm{X}|\mathrm{Y}) + H(\mathrm{Z}|\mathrm{Y}) - H(\mathrm{X}, \mathrm{Z}|\mathrm{Y}) \\ &\quad = \sum_{y_j \in \mathsf{Rng}(\mathrm{Y})} p(y_j)[H(\mathrm{X}|\mathrm{Y} = y_j) + H(\mathrm{Z}|\mathrm{Y} = y_j) - H(\mathrm{X}, \mathrm{Z}|\mathrm{Y} = y_j)] \geq 0. \end{aligned} \tag{7.13}$$

Then one has

$$H(\mathrm{X}, \mathrm{Z}|\mathrm{Y}) \leq H(\mathrm{X}|\mathrm{Y}) + H(\mathrm{Z}|\mathrm{Y}),$$

namely

$$H(\mathrm{X}, \mathrm{Y}, \mathrm{Z}) - H(\mathrm{Y}) \leq H(\mathrm{X}, \mathrm{Y}) - H(\mathrm{Y}) + H(\mathrm{Y}, \mathrm{Z}) - H(\mathrm{Y}).$$

Considering equation (7.13), one easily realises that equality holds iff $H(\mathrm{X}|\mathrm{Y} = y_j) + H(\mathrm{Z}|\mathrm{Y} = y_j) - H(\mathrm{X}, \mathrm{Z}|\mathrm{Y} = y_j) = 0$ for every y_j, and by the subadditivity of Shannon entropy in theorem 7.2 this happens iff

$$p(x_i, z_k|y_j) = p(x_i|y_j)p(z_k|y_j),$$

namely iff conditional independence holds, which in turn is equivalent of the condition that $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ is a Markov chain (see Lemma 7.11). $\square$

From the above theorem we can generalize the assertion from theorem 7.6 that conditioning reduces the entropy

Corollary 7.13 (Conditioning reduces the entropy)**.** *The following bound holds*

$$H(\mathrm{X}|\mathrm{Y},\mathrm{Z}) \leq H(\mathrm{X}|\mathrm{Y}), \tag{7.14}$$

with the equality iff $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$*.*

Proof. In fact, using the definition of conditional entropy, the statement can be rewritten as follows

$$H(\mathrm{X},\mathrm{Y},\mathrm{Z}) - H(\mathrm{Y},\mathrm{Z}) \leq H(\mathrm{X},\mathrm{Y}) - H(\mathrm{Y}),$$

which is just strong subadditivity of theorem 7.12. The case of equality is just the definition of Markov chain.

Alternatively, one can just observe that the statement is the thesis of theorem 7.6 for $p(x_i, z_k|y_j)$, averaged over y_j. □

7.3 Data Processing theorem

We now prove the data processing theorem for Markov chains, and other preliminary results that will be used in the proof of Shannon's noisy channel coding theorem. In particular, we will prove Fano's inequality, which provides a quantitative evidence for the intuitive notion that "the better is the guess of Y given the knowledge of X, the smaller is $H(\mathrm{Y}|\mathrm{X})$".

Theorem 7.14 (Data-processing theorem)**.** *If* $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$ *is a Markov chain the following chain of bounds holds*

$$I(\mathrm{X}:\mathrm{Z}) \leq I(\mathrm{X}:\mathrm{Y}) \leq H(\mathrm{X}), \tag{7.15}$$

with the last bound saturated iff $\mathrm{X} = f(\mathrm{Y})$*. Equivalently,*

$$H(\mathrm{X}|\mathrm{Z}) \geq H(\mathrm{X}|\mathrm{Y}) \geq 0, \tag{7.16}$$

with the last bound saturated iff $\mathrm{X} = f(\mathrm{Y})$

Proof. Equivalence of equations (7.15) and (7.16) is obtained by considering that

$$\begin{aligned} &I(\mathrm{X}:\mathrm{Z}) \leq I(\mathrm{X}:\mathrm{Y}) \;\Leftrightarrow\; H(\mathrm{X}) - I(\mathrm{X}:\mathrm{Z}) \geq H(\mathrm{X}) - I(\mathrm{X}:\mathrm{Y}) \\ &I(\mathrm{X}:\mathrm{Y}) \leq H(\mathrm{X}) \;\Leftrightarrow H(\mathrm{X}) - I(\mathrm{X}:\mathrm{Y}) \geq 0, \end{aligned}$$

and reminding the definition of $I(\mathrm{X}:\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y})$. The second bound, along with conditions for its saturation is thus already proved in theorem 7.6, as $H(\mathrm{X}|\mathrm{Y}) \geq 0$ and $H(\mathrm{X}|\mathrm{Y}) = 0$ iff $\mathrm{X} = f(\mathrm{Y})$. Let us now consider the first bound. Since $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$ is a Markov chain, by the strong subadditivity theorem 7.12, and in particular by corollary 7.13, we have

$$H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y},\mathrm{Z}).$$

This allows us to rewrite the first bound, in the form of equation 7.16, as

$$H(\mathrm{X}|\mathrm{Z}) \geq H(\mathrm{X}|\mathrm{Y},\mathrm{Z}), \tag{7.17}$$

which is equivalent to strong subadditivity [see equation (7.14)]. □

Notice that, despite $X \leftrightarrow Y \leftrightarrow Z$ is a Markov chain, the bound in Eq. (7.17) cannot be saturated, because this would require the Markov condition to hold with the ordering $X \leftrightarrow Z \leftrightarrow Y$.

The data-processing theorem states that *forgetful* data-processing can only destroy information. This is a relevant and non trivial assertion, even though we have to be careful with both words "data processing" and "information". Since the theorem holds for a Markov chain of random variables, we can regard the "data-processing" also as the input-output correlation in an analog device, such as an amplifier, or a telephone line, etc.

Example 7.15 (Old vinyl music)**.** As an example, consider the case of amplification of vibrations by the stylus of a gramophon. At first sight one can think that the amplifier increases the amount of information encoded on the vinyl record. However, by the data processing theorem it can only degrade it. Why does the amplifier work, then? The stylus vibrations are too faint to be heard, and the amplifier enhances them above our audibility threshold. However, there is more information on the tracks of the vinyl record than that coming to our ears.

Example 7.16 (Recovering information from noise by digitalisation)**.** Can one digitalise the sound from a gramophon, and then enhance it by a smart software on a computer? Of course, this is possible, however one cannot recover the information lost in the noisy amplifier: The post-processing software is just a forgetful data-processing device, because it is completely independent of the original track. This way of processing corresponds to a step in a Markov chain. According to the data-processing theorem, it will be useless in re-gaining the information lost. However, one may hear a better sound (e.g. after applying a Dolby filter), just because the processing optimises the sound for the characteristics of the human ear.

Example 7.17 (Saturable amplifiers)**.** There is no linear analog signal amplifier: for high input power, the output will saturate. The high-power signal cut due to saturation cuts high-energy fluctuations, thus effectively improving the signal-to-noise ratio (i.e. $\mathbb{E}(Y^2)/\sigma_Y^2$). Does this mean that saturation is a positive feature? Of course not. Indeed, improving the signal to noise ratio does not increase the mutual information with the source if the output signal has little to do with the input. The data-processing theorem immediately helps clarifying this point.

Chapter 8

Lecture 8: Channel capacity and joint typicality

8.1 Fano's inequality

We now prove a key result for noisy channel coding: Fano's inequality. This result will provide a more quantitative account of the intuition that conditional entropy measures the guessing uncertainty of the random variable X upon knowledge of Y, and will be used in the proof of the Shannon noisy channel theorem. First, let us prove the following chaining rules for conditional entropy.

Theorem 8.1 (Chaining rules)**.** *The following rules hold for every* $N \geq 1$

$$H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_N|\mathrm{Y}) = \sum_{i=1}^{N} H(\mathrm{X}_i|\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{i-1}), \tag{8.1}$$

where $\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{i-1}$ *is by definition* Y *for* $i = 1$.

Proof. The proof is by induction. For $N = 1$ the statement is trivial, since it reduces to $H(\mathrm{X}_1, |\mathrm{Y}) = H(\mathrm{X}_1, |\mathrm{Y})$. Let us now suppose it is true for $N = n$. Then

$$\begin{aligned} H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n+1}|\mathrm{Y}) =& H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n+1}, \mathrm{Y}) - H(\mathrm{Y}) \\ =& H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n+1}, \mathrm{Y}) - H(\mathrm{X}_1, \mathrm{X}_2 \ldots, \mathrm{X}_n, \mathrm{Y}) \\ &+ H(\mathrm{X}_1, \mathrm{X}_2 \ldots, \mathrm{X}_n, \mathrm{Y}) - H(\mathrm{Y}) \\ =& H(\mathrm{X}_{n+1}|\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n) + H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n|\mathrm{Y}). \end{aligned}$$

If we now use the induction hypothesis, we have

$$\begin{aligned} H(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n+1}|\mathrm{Y}) =& H(\mathrm{X}_{n+1}|\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n) \\ &+ \sum_{j=1}^{n} H(\mathrm{X}_j|\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{j-1}) \\ =& \sum_{j=1}^{n+1} H(\mathrm{X}_j|\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{j-1}) \end{aligned}$$

□

We can now prove Fano's theorem. The situation to which the theorem refers is the following. Suppose that there are two random variables, X and Y, X representing Alice's source. Bob has only access to Y, e.g. because it is the output of a noisy channel fed with input X, and he tries to guess the value of X. Whatever strategy Bob uses, his guess $\hat{\mathrm{X}}$ is a random variable that represents the third element of a Markov chain $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \hat{\mathrm{X}}$, because Bob does not have direct access to X, and he can only produce his guess out of a—possibly random—algorithm that processes Y. Thus, $p(\hat{x}|y,x) = p(\hat{x}|y)$.

Theorem 8.2 (Fano's inequality)**.** *Let* X, Y *be a pair of random variables. Let* $\hat{\mathrm{X}}$ *be a third random variable, representing a guess of the value of* X *provided that one knows* Y*. Clearly,* $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \hat{\mathrm{X}}$*. The following bound holds*

$$H(\mathrm{X}|\mathrm{Y}) \leq H(\mathrm{X}|\hat{\mathrm{X}}) \leq H_2(p_\mathrm{e}) + p_\mathrm{e} \log_2(|\mathsf{Rng}(\mathrm{X})| - 1), \tag{8.2}$$

where p_e *denotes the error probability in the guess, namely* $p_e := 1 - \mathbb{P}_{\mathrm{X},\hat{\mathrm{X}}}[\mathrm{X} = \hat{\mathrm{X}}]$*.*

Proof. Let us define the binary random variable E with $\mathsf{Rng}(\mathrm{E}) = \{0,1\}$ as a function of X and $\hat{\mathrm{X}}$ as follows

$$\mathrm{E} := \begin{cases} 0 & \hat{x} = x, \\ 1 & \hat{x} \neq x. \end{cases}$$

One then has

$$\begin{aligned} H(\mathrm{E}) =& \mathbb{P}_\mathrm{E}(\mathrm{E}=0) \log_2 \frac{1}{\mathbb{P}_\mathrm{E}(\mathrm{E}=0)} + \mathbb{P}_\mathrm{E}(\mathrm{E}=1) \log_2 \frac{1}{\mathbb{P}_\mathrm{E}(\mathrm{E}=1)} \\ =& (1-p_\mathrm{e}) \log_2 \frac{1}{1-p_\mathrm{e}} + p_\mathrm{e} \log_2 \frac{1}{p_\mathrm{e}} = H_2(p_\mathrm{e}). \end{aligned}$$

Applying the chaining rule to $H(\mathrm{E}, \mathrm{X}|\hat{\mathrm{X}})$ one obtains

$$H(\mathrm{E}, \mathrm{X}|\hat{\mathrm{X}}) = H(\mathrm{E}|\mathrm{X}, \hat{\mathrm{X}}) + H(\mathrm{X}|\hat{\mathrm{X}}).$$

Since $\mathrm{E} = f(\mathrm{X}, \hat{\mathrm{X}})$, by theorem 7.6 the conditional entropy $H(\mathrm{E}|\mathrm{X}, \hat{\mathrm{X}})$ is null, thus

$$H(\mathrm{E}, \mathrm{X}|\hat{\mathrm{X}}) = H(\mathrm{X}|\hat{\mathrm{X}}).$$

On the other hand, applying a different chaining rule, we have

$$H(\mathrm{E}, \mathrm{X}|\hat{\mathrm{X}}) = H(\mathrm{X}|\mathrm{E}, \hat{\mathrm{X}}) + H(\mathrm{E}|\hat{\mathrm{X}}).$$

By theorem 7.6, conditioning reduces entropy, then $H(\mathrm{E}|\hat{\mathrm{X}}) \leq H(\mathrm{E}) = H_2(p_\mathrm{e})$. Thus, we have

$$H(\mathrm{E}, \mathrm{X}|\hat{\mathrm{X}}) \leq H(\mathrm{X}|\mathrm{E}, \hat{\mathrm{X}}) + H_2(p_\mathrm{e}).$$

Moreover, we can bound the first term on r.h.s. by the observing that

$$H(\mathrm{X}|\mathrm{E}, \hat{\mathrm{X}}) = p_\mathrm{e} H(\mathrm{X}|\hat{\mathrm{X}}, \mathrm{E}=1) + (1-p_\mathrm{e}) H(\mathrm{X}|\hat{\mathrm{X}}, \mathrm{E}=0),$$

and since $\mathbb{P}(\mathrm{X} = x|\hat{\mathrm{X}} = \hat{x}, \mathrm{E} = 0) = \delta_{x,\hat{x}}$, we have $H(\mathrm{X}|\hat{\mathrm{X}}, \mathrm{E} = 0) = 0$. Then we can write

$$\begin{aligned} H(\mathrm{X}|\mathrm{E}, \hat{\mathrm{X}}) =& p_{\mathrm{e}} H(\mathrm{X}|\hat{\mathrm{X}}, \mathrm{E} = 1) \\ =& p_{\mathrm{e}} \sum_{\hat{x} \in \mathsf{Rng}(\mathrm{X})} p(\hat{x}|1) \sum_{x \in \mathsf{Rng}(\mathrm{X}) \setminus \{\hat{x}\}} p(x|\hat{x}, 1) \log_2 \frac{1}{p(x|\hat{x}, 1)} \\ \leq& p_{\mathrm{e}} \sum_{\hat{x} \in \mathsf{Rng}(\mathrm{X})} p(\hat{x}|1) \log_2 [|\mathsf{Rng}(\mathrm{X})| - 1] \\ =& p_{\mathrm{e}} \log_2 [|\mathsf{Rng}(\mathrm{X})| - 1]. \end{aligned}$$

We then proved that

$$H(\mathrm{X}|\hat{\mathrm{X}}) \leq H_2(p_{\mathrm{e}}) + p_{\mathrm{e}} \log_2 [|\mathsf{Rng}(\mathrm{X})| - 1].$$

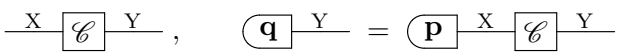
As to the first inequality in the thesis, we use the fact that $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \hat{\mathrm{X}}$ is a Markov chain, and by the data processing theorem

$$I(\mathrm{X} : \hat{\mathrm{X}}) \leq I(\mathrm{X} : \mathrm{Y}) \;\Leftrightarrow\; H(\mathrm{X}|\mathrm{Y}) \leq H(\mathrm{X}|\hat{\mathrm{X}}). \qquad \square$$

8.2 Discrete memoryless channels

In section 2.6 we introduced classical systems, their types, their states and their transformations. Transformations describe transmission media, or more generally any information processing algorithm, which is a sequence of physical or logical operations, producing an *output* state when fed with an *input* state. Also transformations have a type: If the input is the state of a system of type n and the output is a system of type m the transformation has type $n \to m$.

Transformations allow us to determine the state of the output system provided that we know the state of the input one. As we discussed in lecture 3, states of a system X of type n represent the (probabilistic) preparation procedure for the system—a register of cells with n levels. States of a system of type n are thus probability distributions for a random variable X with $|\mathsf{Rng}(\mathrm{X})| = n$. A transformation of type $n \to m$ turns states of system X into state of system Y of type m. The diagrammatic representation of a transformation $\mathscr{C}$ of type $n \to m$ is the following



The probability distribution $\mathbf{p}$, corresponding to the state of system X describes a random preparation algorithm that produces an excitation of the j-th level of the cell X with probability p_j. A transformation represents a second algorithm that reads the content of the cell X and randomly prepares system Y in a new state $\mathbf{q}$.

Clearly, the transformation algorithm cannot behave differently depending on the probability distribution $\mathbf{p}$. For example, suppose that X is a bit B, prepared by reading the outcome of a coin toss. Let Y be a second bit B′. A transformation $\mathscr{C}$ of type $2 \to 2$ represents a transmission of the state of B to the state of B′, which cannot be sensitive to the bias p of the coin used to prepare the input system B. The transformation will just

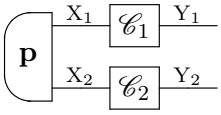
read the content of the input cell, and upon reading 0 performs a protocol, while upon reading 1 performs another one.

The above requirement on the transformation $\mathscr{C}$, of behaving in a fixed way, independently of the distribution $\mathbf{p}$, is mathematically expressed through the property of *linearity*: The state $\mathbf{q}$ of Y_m is a linear function of the state $\mathbf{p}$ of X_n. Then, the transformation $\mathscr{C}$ corresponds to a $m \times n$ matrix $Q = [Q_{i,j}]$, such that $q_i = \sum_{j=1}^{n} Q_{i,j} p_j$. In order to respect positivity and (sub-)normalisation, the linear transformation must have positive elements $Q_{ij} \geq 0$, and every column must be (sub-)normalised: $\sum_{i=1}^{m} Q_{i,j} \leq 1$ for every j.

A *deterministic* transformation is a normalised one, i.e. $\sum_{i=1}^{m} Q_{i,j} = 1$ for every j. Normalised transformations are usually referred to as *channels*, and the corresponding matrices are *stochastic* matrices, whose entries $Q_{i,j}$ can be interpreted as a conditional probability distribution $Q_{i,j} = p(y_i|x_j)$, and the output state $\mathbf{q} = (q(y_1), q(y_2), \dots q(y_m))$ is obtained as the marginal distribution

$$q(y_i) = \sum_{j=1}^{n} q(y_i|x_j) p(x_j).$$

The parallel action of a channel $\mathscr{C}_1$ from X_1 to Y_1 and a channel $\mathscr{C}_2$ from X_2 to Y_2, as in the following diagram



is represented by the tensor product of matrices $Q^{(1)}$ and $Q^{(2)}$, namely

$$p(y_{i_1}, y_{i_2}) = \sum_{j_1, j_2} Q^{(1)}_{i_1,j_1} Q^{(2)}_{i_2,j_2} p(x_{j_1}, x_{j_2}).$$

The memoryless channel is then defined as follows.

Definition 8.3 (Discrete memoryless channel)**.** A discrete memoryless channel $\mathscr{C}$ of type $n \to m$ is characterised by an input and an output system X of type n and Y of type m, respectively, and by a complete set of conditional probabilities $Q_{i,j} = p(y_i|x_j)$ for all $x_j \in \mathsf{Rng}(X)$ and $y_i \in \mathsf{Rng}(Y)$, such that $\mathbf{p}_\mathrm{Y} = Q\mathbf{p}_\mathrm{X}$..

The definition can be generalised to the case of N uses of the same channel $\mathscr{C}$ as follows.

Definition 8.4 (Extended channel)**.** For a discrete memoryless channel $\mathscr{C}$ we consider the channel $\mathscr{C}^N$ corresponding to N uses of the channel $\mathscr{C}$, having input and output systems of type n^N and m^N with ranges $\mathsf{Rng}(\mathrm{X})^N$ and $\mathsf{Rng}(\mathrm{Y})^N$, respectively, and with conditional probability given by

$$p(y^{(N)}_{\mathbf{j}}|x^{(N)}_{\mathbf{i}}) = \prod_{j=1}^{N} p(y_{i_j}|x_{i_j}). \tag{8.3}$$

Alternatively, the transfer matrix of $\mathscr{C}^N$ is given by the N-th tensor power $Q^{\otimes N}$ of the transfer matrix Q, where

$$(Q^{\otimes N})_{\mathbf{i},\mathbf{j}} = Q_{i_1,j_1} Q_{i_2,j_2} \dots Q_{i_N,j_N}, \tag{8.4}$$

$\mathbf{i} = (i_1, i_2, \dots, i_N)$, and $\mathbf{j} = (j_1, j_2, \dots, j_N)$ denoting poly-indices.

Remark 5. This model of channel actually hides an important underlying assumption: All the uses of the channel are considered to be independent and identical, pretty much in the same way as N uses of a source are assumed to prepare N i.i.d. random variables. This assumption is based on the hypothesis that the channel has no internal memory, allowing it to count the uses. Otherwise, every use could end up in a different behaviour, possibly influenced by the previous input states. These models of channels have been studied, also in the quantum case, under the name of *channels with memory* or *memory channels*, for obvious reasons. However, in the present course we will focus on the case of *memoryless channels*.

Let us now have a look at a few examples of discrete memoryless channels.

Example 8.5 (Binary symmetric channel)**.** The input and output system types are 2, thus $\mathsf{Rng}(\mathrm{X}) = \mathsf{Rng}(\mathrm{Y}) = \{0, 1\}$. The transfer matrix is given by

$$Q = \begin{pmatrix} 1-f & f \\ f & 1-f \end{pmatrix},$$

where $0 < f < 1/2$ is the error probability for every input.

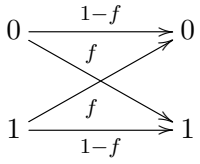


Figure 8.1 Graphical representation of the binary symmetric channel.

Example 8.6 (Binary erasure channel)**.** The input system type is 2, thus the input random variable X is a bit with $\mathsf{Rng}(X) = \{0, 1\}$, while the output system type is 3, namely Y has a range with cardinality 3: we choose $\mathsf{Rng}(\mathrm{Y}) = \{0, ?, 1\}$, which makes the physical interpretation of the channel clear. The transfer matrix of the binary erasure channel is given by

$$Q = \begin{pmatrix} 1-f & 0 \\ f & f \\ 0 & 1-f \end{pmatrix},$$

where $0 < f < 1/2$ describes the probability of losing the input, and is independent of the particular input. In other words, there is a probability of loss, however errors are

always detected because they produce the outcome "?". This makes the error probability exactly null, while it decreases the rate, because in general more than one transmission is required in order to deliver the input bit.

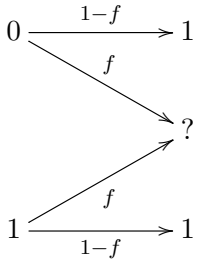


Figure 8.2 Graphical representation of the binary erasure channel.

Example 8.6 is particularly interesting, as it represents the simplest case of nonideal channel with no error.

Example 8.7 (Binary Z channel)**.** In this case $\mathsf{Rng}(X) = \mathsf{Rng}(Y) = \{0, 1\}$ and the transfer matrix is

$$Q = \begin{pmatrix} 1 & f \\ 0 & 1-f \end{pmatrix}$$

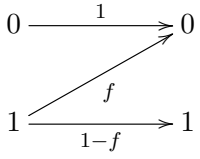


Figure 8.3 Graphical representation of the binary Z channel.

Example 8.8 (Noisy typewriter)**.** In this case $\mathsf{Rng}(\mathrm{X}) = \mathsf{Rng}(\mathrm{Y}) = \{\mathsf{A}, \mathsf{B}, \ldots, \mathsf{Z}, -\}$ One can think of the letters as if they were arranged on a circle, and when the typist presses one button $\mathrm{X} = x_i$ then one of the three letters consisting in the input x_i itself and its two neighbours $x_{i\oplus 1}$, $x_{i\ominus 1}$ is produced, where $\oplus$ and $\ominus$ denote sum and difference modulo 27. So if the button B is pressed, the output can be A, B or C, each one with probability

$\frac{1}{3}$. The transfer matrix is given by

$$Q = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 & \dots & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \dots & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & \dots & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \tag{8.5}$$

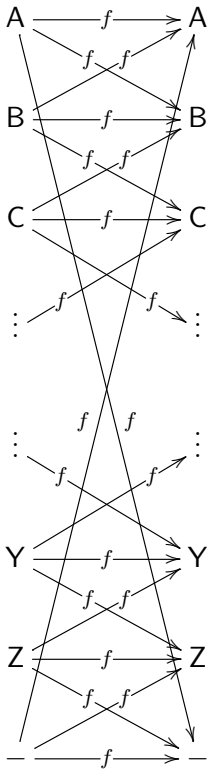


Figure 8.4 Graphical representation of the noisy typewriter channel, where $f = 1/3$.

8.3 Capacity of a channel

The mutual information between input and output of a channel $\mathscr{C}$ is not just a function of the channel Q—from now on we assume that the channel is completely specified by its transfer matrix—but depends also on the *prior* probability distribution $\mathbb{P}_{\mathrm{X}}$ of the random variable representing the state of the input system X. Indeed, by Eq. (8.3), we have

$$\mathbb{P}_{\mathrm{X,Y}}(x_i, y_j) = Q_{j,i}\mathbb{P}_{\mathrm{X}}(x_i). \tag{8.6}$$

In order to define a quantity that only depends on Q, we then maximise the mutual information over such prior distribution: This quantity is the so-called *channel capacity*.

Definition 8.9 (Channel capacity)**.** For a discrete memoryless channel Q, the channel capacity is defined as

$$C(Q) = \max_{\mathbb{P}_{\mathrm{X}}} I(\mathrm{X} : \mathrm{Y}). \tag{8.7}$$

Exercise 8.1

Evaluate the channel capacity of the noisy typewriter channel.

Answer of exercise 8.1

We first evaluate the mutual information for an arbitrary prior. The conditional probabilities are given by the only non-null elements

$$\mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}(\mathrm{Y}=x|\mathrm{X}=x) = \mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}(\mathrm{Y}=x\oplus 1|\mathrm{X}=x) = \mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}(\mathrm{Y}=x\ominus 1|\mathrm{X}=x) = \frac{1}{3},$$

which gives

$$\begin{aligned} H(\mathrm{Y}|\mathrm{X}=x) \equiv H(\mathbf{p}_{|x}) =& \sum_{i=1}^{27} \mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}(\mathrm{Y}=y_i|\mathrm{X}=x) \log_2 \frac{1}{\mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}(\mathrm{Y}=y_i|\mathrm{X}=x)} \\ =& 3 \times \frac{1}{3}\log_2 3 = \log_2 3. \end{aligned}$$

independently of x, hence $H(\mathrm{Y}|\mathrm{X}) = \log_2 3$. The mutual information is then bounded as

$$I(\mathrm{X} : \mathrm{Y}) = H(\mathrm{Y}) - \log_2 3 \leq \log_2 27 - \log_2 3 = \log_2 9.$$

The question is now whether we can saturate the bound for a suitable choice of $\mathbb{P}_{\mathrm{X}}$, namely whether we can make the distribution $\mathbb{P}_{\mathrm{Y}}$ uniform. The marginal probability $\mathbb{P}_{\mathrm{Y}}$ is given by

$$\mathbb{P}_{\mathrm{Y}}(y) = \sum_i \mathbb{P}_{\mathrm{Y}|\mathrm{X}=x_i}(\mathrm{Y}=y|\mathrm{X}=x_i)\mathbb{P}_{\mathrm{X}}(x_i) = \frac{1}{3}\left(\mathbb{P}_{\mathrm{X}}(y) + \mathbb{P}_{\mathrm{X}}(y\oplus 1) + \mathbb{P}_{\mathrm{X}}(y\ominus 1)\right),$$

which can be made uniform, e.g. for uniform $\mathbb{P}_{\mathrm{X}}$. In such case, then, $H(\mathrm{Y}) = \log_2 27$, and we have

$$C(Q) = \max_{\mathbb{P}_{\mathrm{X}}}\{H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X})\} = \log_2 9.$$

For the noisy typewriter channel we can actually achieve the capacity by coding only on a *non-confusable set* of characters, picking out one symbol out of three, e.g. $\mathsf{B}, \mathsf{E}, \mathsf{H}\ldots$ as in figure 8.5. In this way, when the receiver gets A, B or C he/she knows for sure that the letter B was transmitted, and similarly for all the remaining groups of three subsequent letters. The communication by non-confusable sets is then equivalent to a new ideal, noiseless channel for a system of type 9, namely with a range of nine letters. Thus, the new channel has capacity $\log_2 9$.

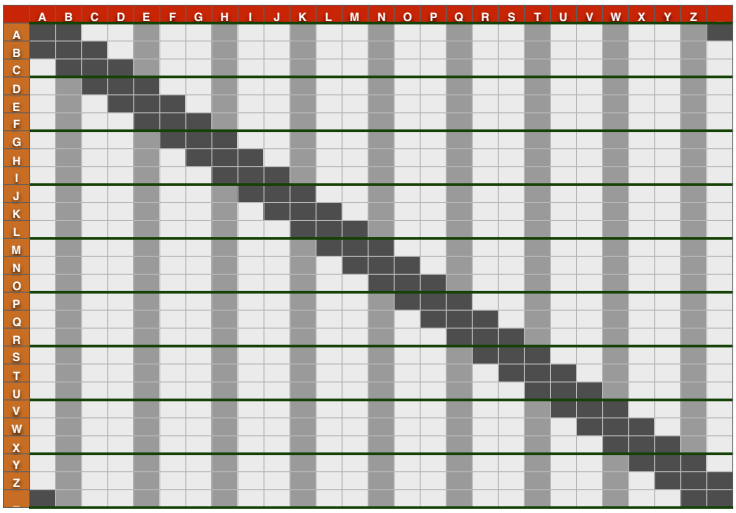


Figure 8.5 A non-confusable subset of inputs for the noisy typewriter.

Exercise 8.2

Evaluate the channel capacity of the binary symmetric channel.

Answer of exercise 8.2

We evaluate the mutual information through the following expression

$$I(\mathrm{X}:\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X}).$$

The output marginal probability is given by

$$\begin{aligned}\mathbb{P}_{\mathrm{Y}}(\mathrm{Y}=0) =& \mathbb{P}_{\mathrm{Y}|\mathrm{X}=0}(\mathrm{Y}=0|\mathrm{X}=0)\mathbb{P}_{\mathrm{X}}(0) + \mathbb{P}_{\mathrm{Y}|\mathrm{X}=1}(\mathrm{Y}=0|\mathrm{X}=1)\mathbb{P}_{\mathrm{X}}(1)\\ =& (1-f)p_0 + f(1-p_0),\end{aligned}$$

where $p_0 := \mathbb{P}_{\mathrm{X}}(0)$. Therefore, we have

$$H(\mathrm{Y}) = H_2((1-f)p_0 + (1-p_0)f).$$

On the other hand we have

$$H(\mathrm{Y}|\mathrm{X}) = \sum_x p_x H(\mathrm{Y}|\mathrm{X}=x) = \sum_x p_x H_2(f) = H_2(f),$$

hence, overall we have

$$I(\mathrm{X}:\mathrm{Y}) = H_2((1-f)p_0 + (1-p_0)f) - H_2(f).$$

Since $H_2(x) \leq 1$ is maximal for $x = 1/2$, the channel capacity is achieved for $p_0 = \frac{1}{2}$, and is given by

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

According to the result of Exercise 8.2, for $f = 0.1$ one has $C = 0.53$. According to the Shannon noisy-channel theorem, which will be presented in the next chapter, it is possible to have reliable communication at the rate equal to the channel capacity with vanishingly small probability of error for sufficiently long strings. Therefore, as already mentioned when discussing the majority vote error-correction, if you have a disk drive whose error probability is that of a binary symmetric channel with $f = 0.1$, in order to achieve a bit error probability 10^{-15} using a majority-voting code you would need *sixty* disks, whereas according to the second Shannon theorem *only two* disks are needed.

Exercise 8.3

Evaluate the channel capacity of the binary erasure channel.

Answer of exercise 8.3

The mutual information $I(\mathrm{X} : \mathrm{Y})$ is evaluated as

$$I(\mathrm{X} : \mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X}).$$

The conditional probability $\mathbb{P}_{\mathrm{Y}|\mathrm{X}=x}$ is given by

$$\begin{aligned}
&\mathbb{P}_{\mathrm{Y}|\mathrm{X}=0}(\mathrm{Y}=0|\mathrm{X}=0) = 1-f && \mathbb{P}_{\mathrm{Y}|\mathrm{X}=1}(\mathrm{Y}=0|\mathrm{X}=1) = 0\\
&\mathbb{P}_{\mathrm{Y}|\mathrm{X}=0}(\mathrm{Y}=?|\mathrm{X}=0) = f && \mathbb{P}_{\mathrm{Y}|\mathrm{X}=1}(\mathrm{Y}=?|\mathrm{X}=1) = f\\
&\mathbb{P}_{\mathrm{Y}|\mathrm{X}=0}(\mathrm{Y}=1|\mathrm{X}=0) = 0 && \mathbb{P}_{\mathrm{Y}|\mathrm{X}=1}(\mathrm{Y}=1|\mathrm{X}=1) = 1-f.
\end{aligned}$$

Then, the conditional entropy is

$$H(\mathrm{Y}|\mathrm{X}=0) = H(\mathrm{Y}|\mathrm{X}=1) = H_2(f), \;\Rightarrow\; H(\mathrm{Y}|\mathrm{X}) = H_2(f).$$

Now, the joint probability $\mathbb{P}_{\mathrm{X},\mathrm{Y}}$ is given by

$$\begin{aligned}
&\mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=0,\mathrm{X}=0) = p_0(1-f) && \mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=0,\mathrm{X}=1) = 0\\
&\mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=?,\mathrm{X}=0) = p_0 f && \mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=?,\mathrm{X}=1) = p_1 f\\
&\mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=1,\mathrm{X}=0) = 0 && \mathbb{P}_{\mathrm{Y},\mathrm{X}}(\mathrm{Y}=1,\mathrm{X}=1) = p_1(1-f),
\end{aligned}$$

And the marginal probability $\mathbb{P}_{\mathrm{Y}}$ is then

$$\mathbb{P}_{\mathrm{Y}}(0) = p_0(1-f), \qquad \mathbb{P}_{\mathrm{Y}}(?) = f, \qquad \mathbb{P}_{\mathrm{Y}}(1) = p_1(1-f).$$

Finally, this allows us to calculate $H(\mathrm{Y})$ that amounts to

$$\begin{aligned}
H(\mathrm{Y}) &= -p_0(1-f)\log_2[p_0(1-f)] - f\log_2 f - p_1(1-f)\log_2[p_1(1-f)]\\
&= (1-f)H_2(p_0) + H_2(f) = (1-f)H(\mathrm{X}) + H_2(f).
\end{aligned}$$

Finally, we have

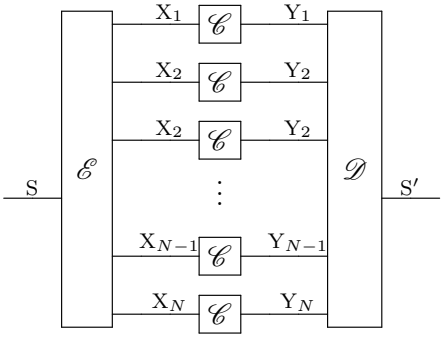
$$I(\mathrm{X} : \mathrm{Y}) = (1-f)H(\mathrm{X}).$$

The optimisation is now straightforward: The optimal choice is the uniform distribution for X, which yields

$$C(Q) = (1 - f).$$

Notice that in the case of a binary erasure channel, the fact that we know when the errors occur makes the channel capacity, namely the rate as we will show in Shannon's noisy channel coding theorem, exactly equal to the complement of the error probability.

When we analysed the noisy typewriter channel, we realised that it allows for a set of non-confusable characters. The basic idea of channel coding is that all extended channels—namely channels consisting in multiple uses of the same discrete memoryless channel—get increasingly close to a sort of noisy typewriter channel, as the number of uses increases. This makes it possible to find sets of strings that are approximately non-confusable, reaching channel capacity with vanishingly small error probability. The scheme representing extended channels is the following



In the remainder we introduce the tools used for the analysis of extended channels.

8.4 Block codes

Definition 8.10 (Block code)**.** A (N, K) block code for a channel $\mathscr{C}$ with input system X and output Y is a map $\mathscr{E} : \mathsf{S} \to \mathsf{Rng}(\mathrm{X})^N$, where S is a the set $\mathsf{S} = \{1, 2, \ldots, 2^K\}$ of cardinality $|\mathsf{S}| = 2^K$. The images of the elements $s \in \mathsf{S}$ under $\mathscr{E}$ are 2^K *codewords* denoted by

$$\{x_{\mathbf{i}^{(1)}}, x_{\mathbf{i}^{(2)}}, \ldots, x_{\mathbf{i}^{(2^K)}}\},\ x_{\mathbf{i}^{(s)}} \in \mathsf{Rng}(X)^N, \tag{8.8}$$

each codeword of length N.

The rate of a code is defined as

Definition 8.11 (Rate of a code)**.** Let $\mathscr{E}$ be a (N, K) block code for a channel $\mathscr{C}$. The *rate of the code* is $R := K/N$ bits per channel use.

Using a code $\mathscr{E}$ one encodes the "signal" $s \in \mathsf{S}$ into the codeword $\mathscr{E}(s) = x_{\mathbf{i}^{(s)}}$. Notice that the number of codewords 2^K is an integer, whereas K (the number of bits necessary to specify s) generally is not. Notice also that "per use" refers to N, which is not a base 2 logarithm. The reader should keep in mind such asymmetry in the definition of *rate* as number of bits per channel use.

Definition 8.12 (Decoder)**.** A decoder for a (N, K) block code $\mathscr{E}$ is a map $\mathscr{D} : \mathsf{Rng}(\mathrm{Y})^N \to \mathsf{S}'$ from the set of length-N strings of channel outputs Y^N, to the set $\mathsf{S}' = \{0, 1, 2, \ldots, 2^K\}$ of codeword labels $\hat{s} \in \mathsf{S}'$. The extra label $\hat{s} = 0$ is useful for indicating a "failure".

In Fig. 8.6 the optimal decoder for a binary erasure channel is illustrated: in this particular case the decoding is error-free, and there is only a small probability of failure. The rate is $R = K/N = 1/2$.

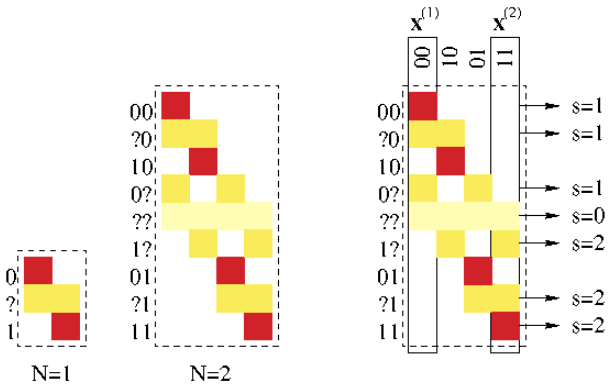


Figure 8.6 (a) Extended channel of the binary erasure channel with $f = 0.15$: each column corresponds to a possible input and each row to a possible output. (b) A block code consisting of the two codewords 00 and 11. (c) The optimal decoder for this code: notice that the decoding is error-less, and there is only a small probability of failure.

8.5 Error probabilities

In the following we will need the precise definitions of different kinds of error probability. We will treat the input signal s and the output s' as the value of two random variables S and S′, with ranges $\mathsf{Rng}(\mathrm{S}) = \mathsf{S}$ and $\mathsf{Rng}(\mathrm{S}') = \mathsf{S}' = \mathsf{S} \cup \{0\}$.

Definition 8.13 (Probability of error for a codeword)**.** For given channel $\mathscr{C}$, code $\mathscr{E}$, and decoder $\mathscr{D}$, the *probability of error for the codeword* $s \in \mathsf{S}$ is given by

$$\begin{aligned} p_s(\mathscr{E}) =& \mathbb{P}_{\mathrm{S}'|\mathrm{S}=s}(\mathrm{S}' \neq s | \mathrm{S} = s, \mathscr{E}, \mathscr{D}) \\ =& \sum_{s' \neq s} \sum_{y_{\mathbf{j}} \in \mathsf{Rng}(\mathrm{Y})^N} \underbrace{p(s'|y_{\mathbf{j}})}_{\mathscr{D}} \underbrace{p(y_{\mathbf{j}}|\overbrace{x_{\mathbf{i}(s)}}^{\mathscr{E}})}_{Q} . \end{aligned} \tag{8.9}$$

Definition 8.14 (Probability of block error)**.** For a given channel $\mathscr{C}$, and for given code $\mathscr{E}$, decoder $\mathscr{D}$, and prior probability $\mathbb{P}_{\mathrm{S}}(s)$ of the signal, the *probability of block error* is given by

$$p_{\mathrm{B}}(\mathscr{E}) = \sum_{s \in \mathsf{S}} \mathbb{P}_{\mathrm{S}}(s) \mathbb{P}_{\mathrm{S}'|\mathrm{S}=s}(\mathrm{S}' \neq s | \mathrm{S} = s, \mathscr{E}, \mathscr{D}). \tag{8.10}$$

Definition 8.15 (Maximum probability of block error)**.** For a given channel $\mathscr{C}$, and for given code $\mathscr{E}$, decoder $\mathscr{D}$, and prior probability $\mathbb{P}_{\mathrm{S}}(s)$ of the signal, the *maximum probability of block error* is

$$p_{\mathrm{BM}}(\mathscr{E}) = \max_s \mathbb{P}_{\mathrm{S}'|\mathrm{S}=s}(\mathrm{S}' \neq s|\mathrm{S} = s, \mathscr{E}, \mathscr{D}), \tag{8.11}$$

8.6 Optimal decoder

Given a channel $\mathscr{C}$, a prior distribution $\mathbb{P}_{\mathrm{S}}(s)$, and an encoding $\mathscr{E}$, the optimal decoder is uniquely identified.

Definition 8.16 (Optimal decoder)**.** The optimal decoder for a channel code $\mathscr{E}$ is the one that minimises the probability of block error. It decodes an output $y_{\mathbf{j}}$ as the symbol s_0 that has maximum posterior probability $p(s_0|y_{\mathbf{j}})$, where the prior $p(s_0)$ is $p(s_0) := \mathbb{P}_{\mathrm{S}}(\mathrm{S} = s_0)$

$$p(s_0|y_{\mathbf{j}}) = \frac{p(y_{\mathbf{j}}|s_0)p(s_0)}{\sum_t p(y_{\mathbf{j}}|t)p(t)}, \quad s_{\mathrm{opt}}(y_{\mathbf{j}}) := \arg\max_s p(s|y_{\mathbf{j}}). \tag{8.12}$$

When the maximum is not unique, then $s_{\mathrm{opt}}(y_{\mathbf{j}}) := 0$. In other words, $\mathscr{D}$ has transfer matrix $p(s'|y_{\mathbf{j}}) = \mathbb{P}_{\mathrm{S}'|Y^N=y_{\mathbf{j}}}[\mathrm{S}' = s'|Y^N = y_{\mathbf{j}}] := \delta_{s',s_{\mathrm{opt}}(y_{\mathbf{j}})}$

For uniform prior probability distribution $\mathbb{P}_{\mathrm{S}}(s)$ the optimal decoder is also the *maximum likelihood decoder*, namely it maps an output $y_{\mathbf{j}}$ to the input s that has maximum likelihood $P(y_{\mathbf{j}}|s)$.

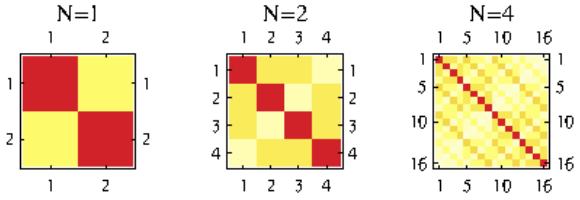


Figure 8.7 Extended channel of the binary symmetric channel with $f = 0.15$. Each column corresponds to a possible input and each row to a possible output.

Example 8.17 (Block decoders for the binary symmetric channel)**.** In Fig. 8.7 the Q matrix (columns correspond to possible inputs and rows to possible outputs) of the extended channel of the binary symmetric channel with $f = 0.15$ is reported for $N = 1, 2$ and 4. Notice how the extended channel looks more and more similar to a noisy typewriter channel. The same phenomenon occurs essentially for any noisy channel for increasing N. For $N = 2$ we can devise an optimal decoding for $s \in \{1, 2\}$ with codewords $x^{(1)} = 00$ and $x^{(2)} = 11$ corresponding to $\hat{s}(00) = 1, \hat{s}(11) = 1$ and $\hat{s}(01) = \hat{s}(10) = 0$. We can see that in this way we achieve a rate $R = .5$ with an average block error probability $2f - f^2 = .28$.

For $N = 4$ instead we can devise an optimal decoding for $s \in \{1, 2\}$ with codewords $x^{(1)} = 0000$, $x^{(2)} = 1111$, and decoding 1 if the majority of bits is 0, and 2 if the

majority is 1 (and all other codewords corresponding to the failure value $\hat{s} = 0$). In this way we can achieve a rate $R = 1/4 = .25$ bits with an average error probability $1 - (1 + 3f)(1 - f)^3 = .11$. We can see that by increasing N we have improved the error probability, however we had to decrease the rate. As we will see, according to the second Shannon theorem, for sufficiently large N it possible to achieve a rate equal to the channel capacity with a vanishingly small probability of error. In the present case the channel capacity is $C = 1 - H_2(f) = .61$ (see Example 8.2).

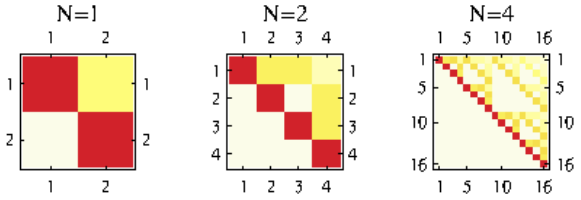


Figure 8.8 Extended channel of the binary Z channel with $f = 0.15$. Each column corresponds to a possible input and each rows to a possible output.

8.7 Joint typicality

As in the case of Shannon's first theorem, also in the case of noisy channel coding the proof exploits the property of typical sets. Differently from the case of source coding, however, here two random variables are involved: The input and the output. We then need an extension of the notion of typical sets, given by the following definition.

Definition 8.18 (Jointly typical sequences)**.** A pair of sequences $(x_{\mathbf{i}}, y_{\mathbf{j}})$ that are values of N i.i.d. pairs of random variables $(\mathrm{X}, \mathrm{Y})^N$ are called jointly typical to tolerance ε if $x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})$, $y_{\mathbf{j}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{Y})$, and $(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{T}_{N,\epsilon}(\mathrm{X}, \mathrm{Y})$, the third condition meaning that

$$\left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}})} - H(\mathrm{X}, \mathrm{Y}) \right| = \left| \frac{1}{N} \sum_{l=1}^{N} \log_2 \frac{1}{\mathbb{P}_{(\mathrm{X},\mathrm{Y})}(x_{i_l}, y_{j_l})} - H(\mathrm{X}, \mathrm{Y}) \right| \leq \varepsilon. \tag{8.13}$$

Definition 8.19 (Jointly typical set)**.** We denote the set of all jointly typical sequences for X and Y by $\mathsf{J}_{N,\varepsilon}(\mathrm{X}, \mathrm{Y})$. This set is the *Jointly typical set.*

Having introduced the notion of joint typicality, we can now prove the joint typicality theorem.

Theorem 8.20 (Joint typicality)**.** *Consider the values $x_{\mathbf{i}}, y_{\mathbf{j}}$ of $(\mathrm{X}, \mathrm{Y})^N = (\mathrm{X}^N, \mathrm{Y}^N)$, where the couples are i.i.d., namely they are distributed according to the probability distribution $\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}}) = \prod_{k=1}^{N} \mathbb{P}_{(\mathrm{X},\mathrm{Y})}(x_{i_k}, y_{j_k})$. One has*

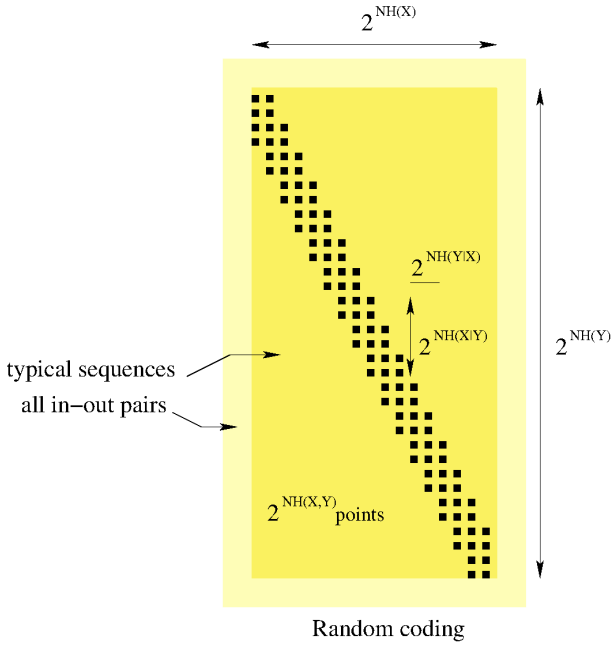


Figure 8.9 The jointly typical set. The horizontal direction represents $\mathsf{Rng}(\mathrm{X})^N$, the set of all input strings of length N, and the vertical one represents $\mathsf{Rng}(\mathrm{Y})^N$, the set of all output strings of length N. The outer box contains all input-output pairs. The inside box contains all pairs made of typical sequences. Dots represent jointly typical pairs of sequences $(x_{\mathbf{i}}, y_{\mathbf{j}})$. The total number of jointly-typical sequences is approximately $2^{NH(\mathrm{X},\mathrm{Y})}$.

1. *For every $\delta > 0$ there exists N_0 such that for every $N \geq N_0$ the probability $\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})]$ of the jointly typical set is bounded as*

$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] \geq 1 - \delta. \tag{8.14}$$

2. *The number of jointly typical sequences $|\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})|$ is close to $2^{NH(\mathrm{X},\mathrm{Y})}$, in the sense that*

$$(1-\delta)2^{N(H(\mathrm{X},\mathrm{Y})-\varepsilon)} \leq |\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})| \leq 2^{N(H(\mathrm{X},\mathrm{Y})+\varepsilon)} \tag{8.15}$$

3. *Let X' and Y' be independent random variables with $\mathsf{Rng}(\mathrm{X}') = \mathsf{Rng}(\mathrm{X})$, $\mathsf{Rng}(\mathrm{Y}') = \mathsf{Rng}(\mathrm{Y})$, and distributed as the marginals X and Y of (X,Y), namely*

$$\mathbb{P}_{\mathrm{X}',\mathrm{Y}'}(x_i, y_j) = \mathbb{P}_{\mathrm{X}}(x_i)\mathbb{P}_{\mathrm{Y}}(y_j). \tag{8.16}$$

Then the probability $\mathbb{P}_{(\mathrm{X}',\mathrm{Y}')}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})]$ that $(\mathrm{X}',\mathrm{Y}')$ occurs in the jointly typical set of (X,Y) is bounded as

$$\mathbb{P}_{(\mathrm{X}',\mathrm{Y}')^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] \leq 2^{-N(I(\mathrm{X}:\mathrm{Y})-3\varepsilon)}. \tag{8.17}$$

Proof. The proof of items 1 and 2 are exactly the same as the analogous ones in the asymptotic equipartition theorem 4.15. Let

$$\mathsf{A} := \left\{ (x_{\mathbf{i}}, y_{\mathbf{j}}) \mid \left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}})} - H(\mathrm{X}) \right| > \varepsilon \right\},$$
$$\mathsf{B} := \left\{ (x_{\mathbf{i}}, y_{\mathbf{j}}) \mid \left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{\mathrm{Y}^N}(y_{\mathbf{j}})} - H(\mathrm{Y}) \right| > \varepsilon \right\},$$
$$\mathsf{C} := \left\{ (x_{\mathbf{i}}, y_{\mathbf{j}}) \mid \left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}})} - H(\mathrm{X},\mathrm{Y}) \right| > \varepsilon \right\}.$$

By definition, we have that

$$\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y}) = \mathsf{Rng}(\mathrm{X},\mathrm{Y})^N \setminus [\mathsf{A} \cup \mathsf{B} \cup \mathsf{C}],$$

and thus

$$\begin{aligned}\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] &= 1 - \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in (\mathsf{A} \cup \mathsf{B} \cup \mathsf{C})] \\ &\geq 1 - \{\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{A}] + \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{B}] + \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{C})]\}.\end{aligned}$$

From the asymptotic equipartition theorem 4.15 we have that, for sufficiently large N_0, for every $N \geq N_0$

$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{A}] = 1 - \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in T_{N,\varepsilon}(\mathrm{X})] \leq \frac{\sigma^2_{\mathrm{X}}}{\varepsilon^2 N_0},$$
$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{B}] = 1 - \mathbb{P}_{\mathrm{X}^N}[y_{\mathbf{j}} \in T_{N,\varepsilon}(\mathrm{Y})] \leq \frac{\sigma^2_{\mathrm{Y}}}{\varepsilon^2 N_0},$$
$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{C}] \leq \frac{\sigma^2_{(\mathrm{X},\mathrm{Y})}}{\varepsilon^2 N_0} = \frac{\sigma^2_{\mathrm{X}}}{\varepsilon^2 N_0} + \frac{\sigma^2_{\mathrm{Y}}}{\varepsilon^2 N_0},$$

where σ^2_{W} is the variance of $\log_2 \frac{1}{\mathbb{P}_{\mathrm{W}}(\mathrm{W}=w)}$. Thus, for $N \geq N_0$, and upon suitable choice of N_0,

$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{A}] + \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{B}] + \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{C}] \leq \delta.$$

Finally, this implies that

$$\mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] \geq 1 - \delta.$$

As to item 2, by definition of jointly typical sequence we have that for $(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})$

$$2^{-N[H(\mathrm{X},\mathrm{Y})+\varepsilon]} \leq \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}}) \leq 2^{-N[H(\mathrm{X},\mathrm{Y})-\varepsilon]},$$

and then

$$\begin{aligned}1 &= \sum_{(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{Rng}(\mathrm{X},\mathrm{Y})^N} \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}}) \\ &\geq \sum_{(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})} \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}}) \geq |\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})| 2^{-N[H(\mathrm{X},\mathrm{Y})+\varepsilon]},\end{aligned}$$

which implies the upper bound in item 2. Analogously, since

$$\begin{aligned} 1-\delta \leq \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] \\ = \sum_{(x_{\mathbf{i}},y_{\mathbf{j}})\in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})} \mathbb{P}_{(\mathrm{X},\mathrm{Y})^N}(x_{\mathbf{i}}, y_{\mathbf{j}}) \leq |\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})| 2^{-N[H(\mathrm{X},\mathrm{Y})-\varepsilon]}, \end{aligned}$$

we also have the lower bound. Item 3 is proved through the following steps

$$\begin{aligned} \mathbb{P}_{(\mathrm{X}',\mathrm{Y}')}[(x_{\mathbf{i}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})] = & \sum_{(x_{\mathbf{i}},y_{\mathbf{j}})\in \mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})} \mathbb{P}_{\mathrm{X}'^N}(x_{\mathbf{i}}) \mathbb{P}_{\mathrm{Y}'^N}(y_{\mathbf{j}}) \\ \leq & |\mathsf{J}_{N,\varepsilon}(\mathrm{X},\mathrm{Y})| 2^{-N[H(\mathrm{X})-\varepsilon]} 2^{-N[H(\mathrm{Y})-\varepsilon]} \\ \leq & 2^{N[H(\mathrm{X},\mathrm{Y})+\varepsilon]-N[H(\mathrm{X})+H(\mathrm{Y})-2\varepsilon]} = 2^{-N[I(\mathrm{X}:\mathrm{Y})-3\varepsilon]}. \ \square \end{aligned}$$

Chapter 9

Lecture 9: Shannon noisy channel coding theorem

9.1 Random coding and typical-set decoding

In this section we describe a coding scheme, named *random coding*, along with a decoder that is sub-optimal, and defined as *typical set decoding*. Surprisingly, these very lame procedures for identification of $(\mathscr{E}, \mathscr{D})$ are sufficient to provide a proof of the noisy channel coding theorem.

For a given prior probability distribution for the alphabet $\mathbb{P}_{\mathrm{X}}$, a random (N, K) code $\mathscr{E}$ is generated as follows

1. Generate 2^K codewords of a (N, K) code at random, sampling from the probability distribution
$$\mathbb{P}(x_{\mathbf{i}}) = \prod_{n=1}^{N} \mathbb{P}_{\mathrm{X}}(x_{i_n}). \tag{9.1}$$
2. The code is shared by sender and receiver.
3. The signal s is chosen with uniform probability distribution $\mathbb{P}_{\mathrm{S}}(s) = \frac{1}{2^K}$. Upon transmission of $x_{\mathbf{i}(s)}$, the output probability distribution is given by
$$\mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x_{\mathbf{i}(s)}}(y_{\mathbf{j}}|x_{\mathbf{i}(s)}) = \prod_{n=1}^{N} p(y_{j_n}|x_{i_n^{(s)}}). \tag{9.2}$$
4. Decode by the *typical-set decoding*, i.e. decode $\mathscr{D}(y_{\mathbf{j}}) = \hat{s}$ if $(x_{\mathbf{i}(\hat{s})}, y_{\mathbf{j}})$ is jointly typical *and* there is no other $\hat{s}'$ such that $(x_{\mathbf{i}(\hat{s}')}, y_{\mathbf{j}})$ is jointly typical—otherwise declare failure $(\hat{s} = 0)$.

A random coding is illustrated in Fig. 9.1.
For random coding and typical set decoding, the following result holds.

Lemma 9.1. *The average block error $\langle p_{\mathrm{B}} \rangle$ over all codes, defined as*
$$\langle p_{\mathrm{B}} \rangle := \sum_{\mathscr{E}} \mathbb{P}(\mathscr{E}) p_{\mathrm{B}}(\mathscr{E}) = \sum_{\mathscr{E}} \mathbb{P}(\mathscr{E}) \sum_{s \in \mathsf{S}} \mathbb{P}_{\mathrm{S}}(s) P(s' \neq s | s, \mathscr{E}), \tag{9.3}$$

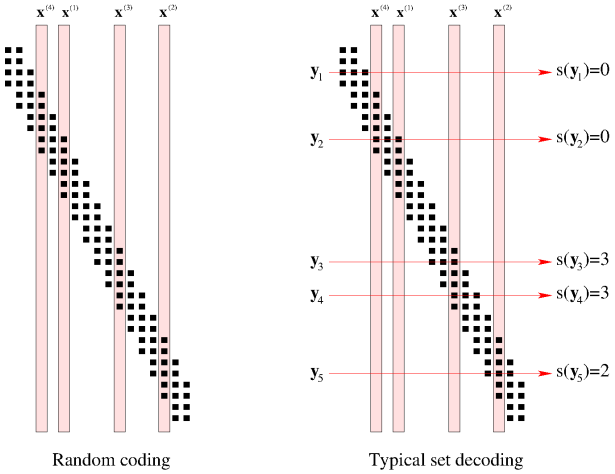


Figure 9.1 (a) A random code. (b) Typical-set decoding. A sequence that is not jointly-typical with any codeword such as y_1 is decoded to $\hat{s} = 0$. A sequence that is jointly-typical with a single codeword $x^{(s)}$ (e.g. y_3 and y_4 is decoded to $\hat{s} = s$. A sequence that is jointly-typical with more than one codewords such as y_2 is decoded to $\hat{s} = 0$.

where $P_{\mathrm{S}}(s) = 1/2^K$*, is equal to*

$$\langle p_{\mathrm{B}} \rangle = \sum_{\mathscr{E}} \mathbb{P}(\mathscr{E}) P(s' \neq 1 | s = 1, \mathscr{E}). \tag{9.4}$$

Proof. We can write a code $\mathscr{E}$ as a $2^K \times N$ matrix B, the *codebook matrix* whose lines are the codewords $x_{\mathbf{i}(s)}$, namely $B_{s,j} = x_{i_j(s)}$. Now, since codewords are randomly extracted with i.i.d. characters distributed according to $\mathbb{P}_{\mathrm{X}}$, for every code $\mathscr{E}$, the probability

$$\mathbb{P}(\mathscr{E}) = \prod_{s=1}^{2^K} \prod_{j=1}^{N} \mathbb{P}_{\mathrm{X}}(x_{i_j(s)}), \tag{9.5}$$

is independent of permutations of the rows. We can then substitute a code $\mathscr{E}$ with the couple (B_0, τ), where B_0 is a fixed *reference* codebook with the same codewords as $\mathscr{E}$, and τ is the permutation of S by which the rows of the codebook matrix B_0 are mapped to those of the codebook matrix B for $\mathscr{E}$. Notice that in general one can have a codebook B_0 with repeated instances. In this case, the non-trivial permutations $\tilde{\tau}$ are not all the possible permutations of 2^K elements. In particular, if the codebook B_0 contains the m words $x_{\mathbf{i}(1)}, x_{\mathbf{i}(2)}, \ldots, x_{\mathbf{i}(m)}$ repeated $n_1, n_2, \ldots n_m$ times, respectively, then for every relevant permutation $\tilde{\tau}$ there are $n_1! n_2! \ldots n_m!$ permutations τ of the 2^K elements that are equivalent to $\tilde{\tau}$, as can be verified by simple combinatorial arguments. This implies that the number of relevant permutations $\tilde{\tau}$ is $2^K!/(n_1! n_2! \ldots n_m!)$. Now, re-expressing $\mathscr{E}$ as a couple $(B_0, \tilde{\tau})$, according to equation 9.5, one has that the probability of every code with the same reference codebook B_0 is the same. Thus, by the counting of relevant

permutations $\tilde{\tau}$ for a given B_0, we have

$$\mathbb{P}(\mathscr{E}) = \mathbb{P}(B_0, \tilde{\tau}) = \mathbb{P}(B_0)\mathbb{P}(\tilde{\tau}|B_0) = \frac{n_1!n_2!\dots n_m!}{2^K!}\mathbb{P}(B_0). \tag{9.6}$$

Moreover, for a general code $\mathscr{E}$ the decoding of the symbol s depends on the set of strings $y_{\mathbf{j}}$ that are jointly typical with $x_{\mathbf{i}(s)}$. Notice that the set of strings $y_{\mathbf{j}}$ that are jointly typical with $x_{\mathbf{i}}$ is independent of the particular label s associated to the string $x_{\mathbf{i}}$. Let us then consider two codes $\mathscr{E}$ and $\mathscr{E}'$ with the same set of codewords, and differing only by the (non-trivial) permutation $\tilde{\sigma}$ of labels, namely $x_{\mathbf{i}} = \mathscr{E}'(s) = \mathscr{E}[\tilde{\sigma}(s)]$. One has

$$\mathbb{P}_{\mathrm{S}'|\mathrm{S}=s}(\mathrm{S}' \neq s|\mathrm{S} = s, \mathscr{E}') = \mathbb{P}_{\mathrm{S}'|\mathrm{S}=\tilde{\sigma}(s)}(\mathrm{S}' \neq \tilde{\sigma}(s)|\mathrm{S} = \tilde{\sigma}(s), \mathscr{E}). \tag{9.7}$$

Now, considering the code $\mathscr{E}_0$ corresponding to the reference codebook B_0, $\mathscr{E}_0 \leftrightarrow (B_0, \mathrm{id})$, if we associate $\mathscr{E} \leftrightarrow (B_0, \tilde{\tau})$ we have $\mathscr{E}(s) = \mathscr{E}_0[\tilde{\tau}(s)]$, and thus $\mathscr{E}'(t) = \mathscr{E}[\tilde{\sigma}(t)] = \mathscr{E}_0[\tilde{\tau}\tilde{\sigma}(t)]$. This implies that $\mathscr{E}' \leftrightarrow (B_0, \tilde{\tau}\tilde{\sigma})$, and one can rewrite the identity in equation (9.7) as

$$\mathbb{P}_{\mathrm{S}'|\mathrm{S}=s}(\mathrm{S}' \neq s|\mathrm{S} = s, B_0, \tilde{\tau}\tilde{\sigma}) = \mathbb{P}_{\mathrm{S}'|\mathrm{S}=\tilde{\sigma}(s)}(\mathrm{S}' \neq \tilde{\sigma}(s)|\mathrm{S} = \tilde{\sigma}(s), B_0, \tilde{\tau}).$$

In the following, in order to make the notation lighter, we omit the parameters of the probability distribution. Since the number of permutations sending 1 to s is $(2^K - 1)!$, one clearly has

$$\begin{aligned}
\frac{1}{2^K}\sum_{s\in\mathsf{S}}\mathbb{P}(\mathrm{S}' \neq s|\mathrm{S} = s, B_0, \tilde{\tau}) &= \frac{1}{2^K}\sum_{s\in\mathsf{S}}\frac{1}{(2^K-1)!}\sum_{\sigma(1)=s}\mathbb{P}(\mathrm{S}' \neq \sigma(1)|\mathrm{S} = \sigma(1), B_0, \tilde{\tau}) \\
&= \frac{1}{2^K!}\sum_{\sigma}\mathbb{P}(\mathrm{S}' \neq \sigma(1)|\mathrm{S} = \sigma(1), B_0, \tilde{\tau}) \\
&= \frac{n_1!n_2!\dots n_m!}{2^K!}\sum_{\tilde{\sigma}}\mathbb{P}(\mathrm{S}' \neq \tilde{\sigma}(1)|\mathrm{S} = \tilde{\sigma}(1), B_0, \tilde{\tau}) \\
&= \frac{n_1!n_2!\dots n_m!}{2^K!}\sum_{\tilde{\sigma}}\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, B_0, \tilde{\tau}\tilde{\sigma}),
\end{aligned}$$

where in the second last equality we used the fact that for every non-trivial permutation $\tilde{\sigma}$ there are $n_1!n_2!\dots n_m!$ equivalent permutations σ. We can then rewrite equation 9.3 as follows

$$\begin{aligned}
\langle p_{\mathrm{B}}\rangle &= \sum_{B_0,\tilde{\tau}}\mathbb{P}(B_0, \tilde{\tau})\frac{n_1!n_2!\dots n_m!}{2^K!}\sum_{\tilde{\sigma}}\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, B_0, \tilde{\tau}\tilde{\sigma}) \\
&= \sum_{B_0,\tilde{\tau}}\mathbb{P}(B_0, \tilde{\tau})\frac{n_1!n_2!\dots n_m!}{2^K!}\sum_{\tilde{\sigma}'}\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, B_0, \tilde{\sigma}') \\
&= \sum_{B_0}\mathbb{P}(B_0)\frac{n_1!n_2!\dots n_m!}{2^K!}\sum_{\tilde{\sigma}'}\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, B_0, \tilde{\sigma}') \\
&= \sum_{B_0,\tilde{\sigma}'}\mathbb{P}(B_0, \tilde{\sigma}')\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, B_0, \tilde{\sigma}') \\
&= \sum_{\mathscr{E}}\mathbb{P}(\mathscr{E})\mathbb{P}(\mathrm{S}' \neq 1|\mathrm{S} = 1, \mathscr{E}),
\end{aligned}$$

where in the second identity we used the fact that $\sum_{\tilde{\sigma}} f(\tilde{\tau}\tilde{\sigma}) = \sum_{\tilde{\sigma}'} f(\tilde{\sigma}')$, and in the fourth we used the fact that $\mathbb{P}(B_0, \tilde{\sigma}) = \mathbb{P}(B_0)\frac{n_1!n_2!\dots n_m!}{2^K!}$ as in equation (9.6). □

9.2 The second theorem of Shannon's

In the proof of the noisy-channel coding theorem we will use the following lemma.

Lemma 9.2 (Expurgation)**.** *For a* uniformly distributed *random variable* X*, and a non-negative (measurable) function* $f : \mathsf{Rng}(\mathrm{X}) \to \mathbb{R}_+$*, the subset* $\mathsf{A} \subseteq \mathsf{Rng}(\mathrm{X})$ *of values* $x_i \in \mathsf{Rng}(\mathrm{X})$ *with* $f(x_i) < 2\mathbb{E}[f(\mathrm{X})]$*, namely*

$$\mathsf{A} := \{x_i \in \mathsf{Rng}(\mathrm{X}) | f(x_i) < 2\mathbb{E}[f(\mathrm{X})]\},$$

has probability $\mathbb{P}_{\mathrm{X}}[x \in \mathsf{A}] \geq \frac{1}{2}$ *and cardinality* $|\mathsf{A}| \geq \frac{1}{2}|\mathsf{Rng}(\mathrm{X})|$.

Proof. Using the Markov inequality (3.8)

$$\mathbb{P}_{\mathrm{X}}[f(\mathrm{X}) \geq a] \leq \frac{\mathbb{E}[f(\mathrm{X})]}{a},$$

with $a = 2\mathbb{E}[f(\mathrm{X})]$, one has

$$\mathbb{P}_{\mathrm{X}}[x_i \in \mathsf{A}] = \mathbb{P}_{\mathrm{X}}[f(\mathrm{X}) < 2\mathbb{E}[f(\mathrm{X})]] = 1 - \mathbb{P}_{\mathrm{X}}[f(\mathrm{X}) \geq 2\mathbb{E}[f(\mathrm{X})]] \geq \frac{1}{2}.$$

Moreover, since X is uniformly distributed, one has

$$\frac{1}{2} \leq \mathbb{P}_{\mathrm{X}}[x_i \in \mathsf{A}] = \sum_{x_i \in \mathsf{A}} \mathbb{P}_{\mathrm{X}}[x_i] = \sum_{x_i \in \mathsf{A}} \frac{1}{|\mathsf{Rng}(\mathrm{X})|} = \frac{|\mathsf{A}|}{|\mathsf{Rng}(\mathrm{X})|},$$

from which we conclude $|\mathsf{A}| \geq \frac{1}{2}|\mathsf{Rng}(\mathrm{X})|$. □

We will now prove the celebrated noisy-channel coding theorem of Shannon. The proof will use all the following previously analyzed ingredients:

1. joint typicality: item (3) of Theorem 8.20;
2. random coding and typical-set decoding: lemma 9.1;
3. the Fano inequality in Theorem 8.2;
4. the data-processing Theorem 7.14;
5. the expurgation technique (see Lemma 9.2);
6. the definition of channel capacity for a single use;
7. the probability of errors defined in 8.13 and 8.14.

Theorem 9.3 (Noisy-channel coding theorem)**.** *Let* $\mathscr{C}$ *be a discrete memoryless channel with transfer matrix* Q*. Then we have the following.*

1. *For every* $R_0 < C(Q)$*, for every* $\varepsilon > 0$ *there exists* N_0 *such that for* $N \geq N_0$ *there exists a block code* $\mathscr{E}$ *of length* N *and rate* R *with* $R_0 < R < C(Q)$*, and a decoding* $\mathscr{D}$ *such that the maximal probability of block error is* $p_{\mathrm{BM}}(\mathscr{E}) < \varepsilon$*.*
2. *Conversely, for any block code* $\mathscr{E}$ *with rate* $R > C(Q)$*, the probability* $p_{\mathrm{BM}}(\mathscr{E})$ *is bounded from below by* $k > 0$*, independently of* N*.*

Sketch of the proof

Before giving the proof of the theorem, we want to focus attention on the sketch of the logical sequence that constitutes the proof of item 1. The point is to prove existence of a code that has maximum probability of block error bounded by ε. Here are the steps for this proof.

1. Since it is very difficult to evaluate the probability of block error $p_{\mathrm{B}}(\mathscr{E}) := \sum_{s\in\mathsf{Rng}(\mathrm{S})} \mathbb{P}(\mathrm{S}' \neq s|\mathrm{S}=s,\mathscr{E})\mathbb{P}_{\mathrm{S}}(s)$ for a particular code, while it is easy to evaluate its average over all randomly chosen codes $\langle p_{\mathrm{B}}\rangle = \sum_{\mathscr{E}} \mathbb{P}(\mathscr{E})p_{\mathrm{B}}(\mathscr{E})$, Shannon first bounds the above average.

2. Clearly, there must exist a code $\mathscr{E}$ for which $p_{\mathrm{B}}(\mathscr{E}) \leq \langle p_{\mathrm{B}}(\mathscr{E})\rangle$.

3. However, we need to bound the maximal block error probability, given by the expression $p_{BM}(\mathscr{E}) := \max_s \mathbb{P}(\mathrm{S}' \neq s|\mathrm{S}=s,\mathscr{E})$, and the fact that we have found a code $\mathscr{E}$ for which $p_B(\mathscr{E})$ is bounded does not mean that $p_{BM}(\mathscr{E})$ is bounded. This code can be modified by throwing away the worst 50% of its codewords using the Expurgation of Lemma 9.2, yielding bounded maximal error without affecting the rate by any significant amount.

The proof

Now let us see the real proof of Shannon's noisy channel coding theorem.

Proof. We use random coding and typical set decoding. We remind that this implies extracting the code at random, with i.i.d. characters distributed according to the distribution $\mathbb{P}_{\mathrm{X}}(x_{i_k^{(s)}})$, independent of s, k.

Let us then evaluate the probability of block error averaged over all codes, namely

$$\langle p_{\mathrm{B}}\rangle = \sum_{\mathscr{E}} \mathbb{P}(\mathscr{E})p_{\mathrm{B}}(\mathscr{E})$$

The typical set decoding is subject to two different kinds of errors: *type 1:* the output $y_{\mathbf{j}}$ is not jointly typical with the input $x_{\mathbf{i}(s)}$. This case includes both the cases when $y_{\mathbf{j}}$ is jointly typical with no codeword, and when $(x_{\mathbf{i}(s')}, y_{\mathbf{j}})$ is jointly typical with $s' \neq s$; *type 2:* $(x_{\mathbf{i}(s)}, y_{\mathbf{j}})$ may or may not be jointly typical, but there is another codeword s' in $\mathscr{E}$ that is actually jointly typical with $y_{\mathbf{j}}$. Notice that the two types of errors are not independent, as it may happen that both $(y_{\mathbf{j}}, x_{\mathbf{i}(s)}) \notin J_{N,\varepsilon}(\mathrm{X},\mathrm{Y})$ and $(y_{\mathbf{j}}, x_{\mathbf{i}(s')}) \in J_{N,\varepsilon}(\mathrm{X},\mathrm{Y})$ for $s' \neq s$. According to lemma 9.1, we can use the expression of eq. 9.4 for the average block error probability, and restrict to the case $s = 1$ without loss of generality. The situation is thus the following: one generates the code by random coding, and then the output $y_{\mathbf{j}}$ is

generated from input $x_{\mathbf{i}^{(1)}}$. The joint probability for this sequence of events is

$$\begin{aligned} p(\mathscr{E}, y_{\mathbf{j}}|\mathrm{S}=1) &= \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}(x_{\mathbf{i}^{(1)}}, x_{\mathbf{i}^{(2)}}, \dots, x_{\mathbf{i}^{(2^K)}}, y_{\mathbf{j}}) \\ &= P_{\mathrm{X}^N}(x_{\mathbf{i}^{(1)}})\mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x_{\mathbf{i}^{(1)}}}(y_{\mathbf{j}}|x_{\mathbf{i}^{(1)}}) \prod_{j=2}^{2^K} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(j)}}), \\ \mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x^{(1)}}(y_{\mathbf{j}}|x^{(1)}) &= \prod_{l=1}^{N} p(y_{j_l}|x_{i_l^{(1)}}). \end{aligned} \tag{9.8}$$

We then define the following sets related to the two error types

$$\begin{aligned} \mathsf{A} &:= \{(x_{\mathbf{i}^{(1)}}, \dots, x_{\mathbf{i}^{(2^K)}}, y_{\mathbf{j}}) \mid (x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) \notin \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})\}, \\ \mathsf{B} &:= \{(x_{\mathbf{i}^{(1)}}, \dots, x_{\mathbf{i}^{(2^K)}}, y_{\mathbf{j}}) \mid (x_{\mathbf{i}^{(s)}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y}), s \neq 1\}. \end{aligned}$$

Since the two types of error are not independent, it may happen that $\mathsf{A} \cap \mathsf{B} \neq \emptyset$. Anyway, we can bound the average block error probability as

$$\begin{aligned} \langle p_B \rangle &= \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}[\mathsf{A} \cup \mathsf{B}] \\ &\leq \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}[\mathsf{A}] + \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}[\mathsf{B}] \\ &= \langle p_B^1 \rangle + \langle p_B^2 \rangle, \end{aligned}$$

where we used the *union bound* $\mathbb{P}(\mathsf{C} \cup \mathsf{D}) \leq \mathbb{P}(\mathsf{C}) + \mathbb{P}(\mathsf{D})$, holding for general probability distributions and general sets C, D, and $\langle p_B^n \rangle$ denotes the contribution from type n error.

Let us first evaluate $\langle p_B^1 \rangle$. According to item 1 of theorem 8.20, for a given code $\mathscr{E}$ the probability that the input $x_{\mathbf{i}^{(1)}}$ is jointly typical with the output $y_{\mathbf{j}}$ with $x_{\mathbf{i}^{(1)}}$ and $y_{\mathbf{j}}$ distributed according to the distribution $\mathbb{P}_{\mathrm{X}^N,\mathrm{Y}^N}(x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) = \mathbb{P}_{\mathrm{X}}(x_{\mathbf{i}^{(1)}})\mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x_{\mathbf{i}^{(1)}}}(y_{\mathbf{j}}|x_{\mathbf{i}^{(1)}})$—which is obtained as the marginal of the joint distribution in equation (9.8)—converges asymptotically to 1 for $N \to \infty$. More precisely, for any desired $\delta > 0$ there exists a block length N_δ such that, for $N \geq N_\delta$, it holds that $\mathbb{P}_{\mathrm{X}^N,\mathrm{Y}^N}[(x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) \notin \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})] < \delta$. Thus one has $\langle p_B^1 \rangle = \mathbb{P}_{\mathrm{X}^N,\mathrm{Y}^N}[\mathsf{A}] = \mathbb{P}_{\mathrm{X}^N,\mathrm{Y}^N}[(x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) \notin \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})] < \delta$ for $N \geq N_\delta$.

Let us then evaluate the second term $\langle p_B^2 \rangle$. One has

$$\begin{aligned} \langle p_B^2 \rangle &= \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}[\bigcup_{s\neq 1} \{(x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) | (x_{\mathbf{i}^{(s)}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})\}] \\ &\leq \sum_{s\neq 1} \mathbb{P}_{\mathrm{X}^{2^K N},\mathrm{Y}^N}[\{(x_{\mathbf{i}^{(1)}}, y_{\mathbf{j}}) | (x_{\mathbf{i}^{(s)}}, y_{\mathbf{j}}) \in \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})\}]. \end{aligned}$$

Since the average is over all random codes, we can rewrite the last expression as

$$\begin{aligned}
\langle p_B^2\rangle \leq \sum_{s\neq 1} \sum_{x_{\mathbf{i}^{(1)}}\in \mathsf{Rng}(\mathrm{X}^N)} \overbrace{\sum_{x_{\mathbf{i}^{(2)}}\in \mathsf{Rng}(\mathrm{X}^N)} \cdots \sum_{x_{\mathbf{i}^{(2^K)}}\in \mathsf{Rng}(\mathrm{X}^N)}}^{2^K-2 \text{ sums}} & \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(1)}})\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(2)}})\ldots\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(2^K)}}) \\
\times \left\{ \sum_{(x_{\mathbf{i}^{(s)}},y_{\mathbf{j}})\in \mathsf{J}_{N,\beta}(\mathrm{X},\mathrm{Y})} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(s)}})\mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x_{\mathbf{i}^{(1)}}}(y_{\mathbf{j}}|x_{\mathbf{i}^{(1)}}) \right\} & \\
= \sum_{s\neq 1} \sum_{x_{\mathbf{i}^{(1)}}\in \mathsf{Rng}(\mathrm{X}^N)} \sum_{(x_{\mathbf{i}^{(s)}},y_{\mathbf{j}})\in \mathsf{J}_{N,\beta}(\mathrm{X}^N,\mathrm{Y}^N)} & \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(s)}})\mathbb{P}_{\mathrm{Y}^N|\mathrm{X}^N=x_{\mathbf{i}^{(1)}}}(y_{\mathbf{j}}|x_{\mathbf{i}^{(1)}})\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(1)}}) \\
= \sum_{s\neq 1} \sum_{x_{\mathbf{i}^{(1)}}\in \mathsf{Rng}(\mathrm{X}^N)} \sum_{(x_{\mathbf{i}^{(s)}},y_{\mathbf{j}})\in \mathsf{J}_{N,\beta}(\mathrm{X}^N,\mathrm{Y}^N)} & \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(s)}})\mathbb{P}_{\mathrm{Y}^N,\mathrm{X}^N}(y_{\mathbf{j}},x_{\mathbf{i}^{(1)}}) \\
= \sum_{s\neq 1} \sum_{(x_{\mathbf{i}^{(s)}},y_{\mathbf{j}})\in \mathsf{J}_{N,\beta}(\mathrm{X}^N,\mathrm{Y}^N)} & \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(s)}})\mathbb{P}_{\mathrm{Y}^N}(y_{\mathbf{j}}).
\end{aligned}$$

According to item 3 of Theorem 8.20 we can bound every term of the sum over s as

$$\sum_{(x_{\mathbf{i}^{(s)}},y_{\mathbf{j}})\in \mathsf{J}_{N,\beta}(\mathrm{X}^N,\mathrm{Y}^N)} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}^{(s)}})\mathbb{P}_{\mathrm{Y}^N}(y_{\mathbf{j}}) \leq 2^{-N(I(\mathrm{X}:\mathrm{Y})-3\beta)},$$

and we thus have

$$\langle p_B^2\rangle \leq \sum_{s=2}^{2^K} 2^{-N(I(X:Y)-3\beta)} = (2^K-1)2^{-N(I(\mathrm{X}:\mathrm{Y})-3\beta)} < 2^{-N(I(\mathrm{X}:\mathrm{Y})-R-3\beta)},$$

since there are $2^K - 1 = 2^{NR} - 1$ terms in the sum. Therefore, the total average block error probability is bounded as

$$\langle p_B\rangle < \delta + 2^{-N(I(X:Y)-R-3\beta)}.$$

Now, if the following condition holds

$$R < I(X:Y) - 3\beta, \tag{9.9}$$

for sufficiently large N we can make

$$\langle p_B\rangle < 2\delta, \tag{9.10}$$

Now, if we choose $\mathbb{P}_{\mathrm{X}}$ to be that particular probability distribution maximising the mutual information, condition (9.9) becomes

$$R < C(Q) - 3\beta. \tag{9.11}$$

Since the error probability averaged over all codes is smaller than 2δ, there exists a code $\mathscr{E}_0$ with mean probability of block error $p_B(\mathscr{E}_0) < 2\delta$. However, such code may have

maximum probability of error $p_{BM}(\mathscr{E}_0) > 2\delta$. In order to show that a code exists having $p_{BM}(\mathscr{E})$ arbitrarily small, we use the *expurgation* technique to build up a new $\mathscr{E}'$ from $\mathscr{E}_0$, by removing the half of the codewords that are most likely to produce errors, i.e. those with the highest probability of error. Since the codewords s are uniformly distributed by the hypothesis of random coding, if we apply Lemma 9.2 to the random variable $p_s(\mathscr{E}_0)$, the remaining codewords must have probability of error $p_s(\mathscr{E}') = p_s(\mathscr{E}_0) \leq 2p_B(\mathscr{E}_0) < 4\delta$ for every s, and thus $p_{BM}(\mathscr{E}') < 4\delta$. The expurgated code $\mathscr{E}'$ has $2^{NR'} = 2^{NR-1}$ codewords, namely the rate has been reduced from R to $R' = R - \frac{1}{N}$, which is a negligible reduction for large N. In conclusion, for $R < C(Q)$ we have proved the existence of a code with rate $R' = R - \frac{1}{N} < C(Q) - 3\beta$ and with maximum probability of error $p_{BM} < 4\delta$. The statement of the theorem is finally obtained by setting $R := (R_0 + C(Q))/2 - 1/N \geq R_0$, $\delta = \varepsilon/4$, $\beta < (C(Q) - R)/3$, and N sufficiently large to satisfy the remaining conditions. We have proved in this way the first part of the theorem.

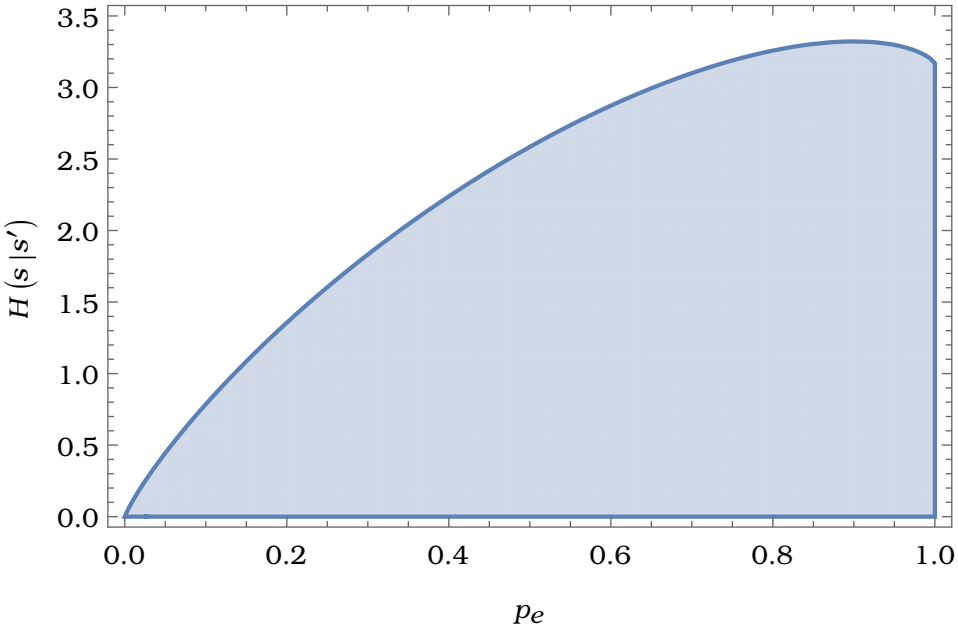


Figure 9.2 Plot of the Fano bound in Eq. (9.13) for $S = 10$.

To prove the second part of the theorem, consider any coding-decoding, and denote by p_e the corresponding error probability. The Data Processing Theorem 7.14 for the Markov chain $s \leftrightarrow x_{\mathbf{i}(s)} \leftrightarrow y_{\mathbf{j}} \leftrightarrow s'$ guarantees that

$$I(s : s') \leq I(s : y_{\mathbf{j}}) \leq I(x_{\mathbf{i}(s)} : y_{\mathbf{j}}) \leq NC(Q), \tag{9.12}$$

where the last bound follows by the definition of channel capacity. According to the Fano inequality (8.2)

$$H_2(p_e) + p_e \log_2(|\mathsf{S}| - 1) \geq H(s|s'), \tag{9.13}$$

For error probability $p_e = 0$ one has $H(s|s') = 0$, and $I(s : s') = H(s)$. Thus, for uniform probability distribution $p(s) = 1/2^K$ one has $I(s : s') = \log_2 |\mathsf{S}| = K = NR$ (see the

definition 8.10 of block code), and inequality (9.12) must be satisfied in this special case, namely

$$NR \leq NC(Q) \quad \Leftrightarrow \quad R \leq C(Q). \tag{9.14}$$

On the other hand, if $R - C(Q) > 0$ one has

$$H(s) - H(s|s') \leq NC(Q) \;\Rightarrow\; H(s|s') \geq H(s) - NC(Q) = N(R - C(Q)) > 0.$$

Thus, by Fano's theorem

$$\begin{aligned} N(R - C(Q)) &\leq H(s|s') \\ &\leq H_2(p_e) + p_e \log_2(|\mathsf{S}| - 1) \\ &\leq 1 + p_e \log_2 |\mathsf{S}| \\ &= 1 + p_e K \\ &= 1 + p_e NR, \end{aligned}$$

and finally one has

$$p_e \geq \frac{N(R - C(Q)) - 1}{NR} = 1 - \frac{C(Q)}{R} - \frac{1}{NR}, \tag{9.15}$$

which for large enough N implies $p_e > k$ with $k > 0$ independent of N. □

Notice the precise form of the statement of the theorem. In the first part—the *direct* statement—it is not claimed that the maximum rate is the channel capacity, but that for every $R < C(Q)$ and for any $\epsilon > 0$ there exists a rate $R' \geq R$ such that the maximum probability of block error is $p_{BM} < \epsilon$. Indeed, in principle, rates larger than the channel capacity are allowed, as long as a non-vanishing error probability is accepted. The second part—the *converse* statement—specifies this situation in terms of the allowed bit error probability. The extension of information theory at non-zero error probability is the objective of the so-called *rate distortion theory*.

9.3 Communication above capacity [extra]

What can we say about error probability in a communication with rate above capacity? In general we have:

$$R \leq \frac{C(Q)}{1 - H_2(p_b)},$$

where p_b is called *bit error probability*. What is p_b precisely? It is a measure of communication error that we now define. In a (N, K) block code one encodes 2^K strings, namely $K = NR$ bits. Let us then represent the input and output symbols $s, s' \in \mathsf{S}$ by the binary representation

$$b : \mathsf{S} \to \{0,1\}^K :: s \mapsto \mathbf{b}_s = b_{s1} b_{s2} \ldots b_{sK}.$$

Whenever a communication error occurs, one has $\mathbf{b}_s \neq \mathbf{b}_{s'}$, namely there exists $1 \leq l \leq K$ such that $b_{sl} \neq b_{s'l}$. We can then define p_l as the probability of occurrence of such error. We then define the bit error probability as follows.

Definition 9.4 (Bit error probability)**.** The bit error probability for a (N, K) block code is defined as

$$p_b := \frac{1}{K} \sum_{l=1}^{K} p_l. \tag{9.16}$$

We now prove the following lemma that will be used to bound the attainable rates above capacity.

Lemma 9.5. *Let the input symbols of* $\mathbf{b}_s$ *be i.i.d., and let* $\mathbf{b}_{s'}$ *denote the output string for a* (N, K) *block code for channel* $\mathscr{C}$*. Then the following bound holds:*

$$I(\mathbf{b}_s : \mathbf{b}_{s'}) \geq \sum_{l=1}^{K} I(b_{sl} : b_{s'l}). \tag{9.17}$$

Proof. Let us recall that

$$\begin{aligned} I(\mathbf{b}_s : \mathbf{b}_{s'}) = H(\mathbf{b}_s) - H(\mathbf{b}_s|\mathbf{b}_{s'}) &= \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(\mathbf{b}_s|\mathbf{b}_{s'})}{p(\mathbf{b}_s)} \\ &= \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(\mathbf{b}_s|\mathbf{b}_{s'})}{p(b_{s1})p(b_{s2}) \dots p(b_{sK})}. \end{aligned}$$

On the other hand, one has

$$\begin{aligned} \sum_{l=1}^{K} I(b_{sl} : b_{s'l}) &= \sum_{l=1}^{K} \sum_{b_{sl}, b_{s'l}} p(b_{sl}, b_{s'l}) \log_2 \frac{p(b_{sl}|b_{s'l})}{p(b_{sl})} \\ &= \sum_{l=1}^{K} \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(b_{sl}|b_{s'l})}{p(b_{sl})} \\ &= \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} \sum_{l=1}^{K} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(b_{sl}|b_{s'l})}{p(b_{sl})} \\ &= \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(b_{s1}|b_{s'1})p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K})}{p(b_{s1})p(b_{s2}) \dots p(b_{sK})}. \end{aligned}$$

Now, one can write

$$
\begin{aligned}
I(\mathbf{b}_s : \mathbf{b}_{s'}) - \sum_{l=1}^{K} I(b_{sl} : b_{s'l}) &= \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(\mathbf{b}_s|\mathbf{b}_{s'})}{p(b_{s1}|b_{s'1}) p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K})} \\
&= - \sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(\mathbf{b}_s, \mathbf{b}_{s'}) \log_2 \frac{p(b_{s1}|b_{s'1}) p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K})}{p(\mathbf{b}_s|\mathbf{b}_{s'})} \\
&= -\mathbb{E}_{\mathbf{b}_s, \mathbf{b}_{s'}} \left(\log_2 \frac{p(b_{s1}|b_{s'1}) p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K})}{p(\mathbf{b}_s|\mathbf{b}_{s'})} \right) \\
&\geq - \log_2 \mathbb{E}_{\mathbf{b}_s, \mathbf{b}_{s'}} \left(\frac{p(b_{s1}|b_{s'1}) p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K})}{p(\mathbf{b}_s, \mathbf{b}_{s'})} p(\mathbf{b}_{s'}) \right) \\
&= - \log_2 \left[\sum_{\mathbf{b}_s, \mathbf{b}_{s'}} p(b_{s1}|b_{s'1}) p(b_{s2}|b_{s'2}) \dots p(b_{sK}|b_{s'K}) p(\mathbf{b}_{s'}) \right] = 0. \qquad \square
\end{aligned}
$$

We are now in position to prove the main result in this section, that provides a lower bound for the rate above the channel capacity as a function of the bit error probability.

Theorem 9.6. *For a discrete memory channel $\mathscr{C}$ with capacity $C(Q)$, every block code with bit error probability p_b must have a rate R that satisfies the bound*

$$
R \leq \frac{C(Q)}{1 - H_2(p_b)}. \tag{9.18}
$$

Proof. In the first place, by the second Shannon theorem one has the following inequalities

$$
I(\mathbf{b}_s : \mathbf{b}_{s'}) = I(s : s') \leq I(x_{\mathbf{i}^{(s)}} : y_{\mathbf{j}}) \leq NC(Q).
$$

By lemma 9.5 we then have

$$
I(\mathbf{b}_s : \mathbf{b}_{s'}) \geq \sum_{l=1}^{K} I(b_{sl} : b_{s'l}) = \sum_{l=1}^{K} \{ H(b_{sl}) - H(b_{sl}|b_{s'l}) \}.
$$

Now, since the symbols s are uniformly distributed, one has $H(b_{sl}) = H_2(1/2) = 1$. Finally, by Fano's inequality one has

$$
H(b_{sl}|b_{s'l}) \leq H_2(p_l) + p_l \log_2 1 = H_2(p_l).
$$

This provides the bound

$$
I(\mathbf{b}_s : \mathbf{b}_{s'}) \geq NR \left[1 - \frac{1}{NR} \sum_{l=1}^{K} H_2(p_l) \right].
$$

Now, by convexity of the function $-H_2(x)$, Jensen inequality gives

$$
-\frac{1}{NR} \sum_{l=1}^{K} H_2(p_l) \geq -H_2(p_b).
$$

Recollecting all the inequalities, we obtain

$$
NC(Q) \geq NR[1 - H_2(p_b)]. \qquad \square
$$

The inequality in Eq. (9.18) can be recast inverting the function H_2 (on the range $0 \leq p_b \leq 1/2$), obtaining

$$p_b \geq H_2^{-1}\left(1 - \frac{C(Q)}{R}\right). \tag{9.19}$$

The plot of the above function in the plane (R, p_e) is given in Fig. 9.3. All the points

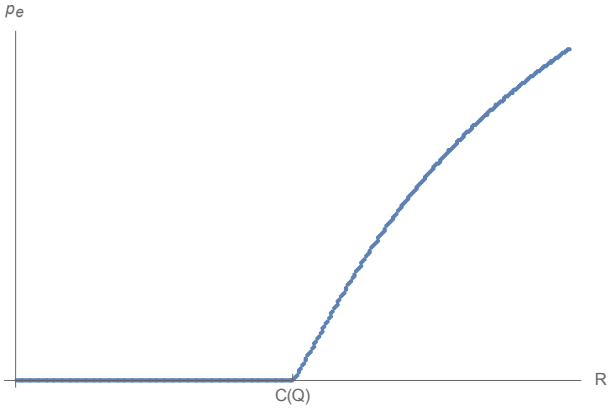


Figure 9.3 The plot of the boundary of the attainable region of block codes for a discrete memoryless channel with capacity $C(Q)$. Points above the blue line represent pairs (R, p_e) of values of rate and maximal block error probability that can be achieved by a suitable (N, K) block code.

above the bound are actually attainable.

Part II

Quantum information theory

Chapter 10

Lecture 10: Elements of linear algebra

We will use in the following the theory of finite dimensional Hilbert spaces. We review here the basic facts about linear algebra on complex Hilbert spaces, avoiding the most advanced topics and sticking to the structures that are required in the development of the quantum theory of information.

10.1 Hilbert spaces

A complex Hilbert space is a vector space $\mathcal{H}$ on $\mathbb{C}$, equipped with a sesquilinear form $(,) : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ with the following properties

$$\begin{aligned}
&(\varphi, \psi) = (\psi, \varphi)^* \\
&(\varphi, a\psi + b\eta) = a(\varphi, \psi) + b(\varphi, \eta) \\
&(\varphi, \varphi) \geq 0, \quad (\varphi, \varphi) = 0 \ \Leftrightarrow \ \varphi = 0,
\end{aligned}$$

and closed in the metric $d(\varphi, \psi) := \|\varphi - \psi\|$ induced by the sesquilinear form by $\|\eta\| := (\eta, \eta)^{\frac{1}{2}}$. The dual space $\mathcal{H}^*$ of continuous linear functionals on $\mathcal{H}$ is isomorphic to $\mathcal{H}$ by the Riesz-Fréchet representation theorem

$$D : \mathcal{H}^* \to \mathcal{H} \quad \tilde{\eta} \mapsto D\tilde{\eta} : \ \tilde{\eta}(\psi) = (D\tilde{\eta}, \psi).$$

We will adopt the Dirac notation, using *kets* $|\psi\rangle$ to denote vectors $\psi \in \mathcal{H}$ and *bras* $\langle\varphi|$ to denote vectors in $\mathcal{H}^*$, thus

$$(\varphi, \psi) = \langle\varphi|\psi\rangle. \tag{10.1}$$

It is often useful to represent vectors by expanding them on an *orthonormal basis*, i.e. a collection of orthonormal vectors $\{e_i\}_{i=1}^d$ such that

$$\langle e_i | e_j \rangle = \delta_{i,j}.$$

Indeed, one can prove that such a collection is a basis, namely it is *complete*

$$\psi \in \mathcal{H} \ \Rightarrow \ \exists \{\psi_i\}_{i=1}^d \in \mathbb{C}^d \quad |\psi\rangle = \sum_{i=1}^d \psi_i |e_i\rangle,$$

and linearly independent

$$\sum_{i=1}^{d} \eta_i |e_i\rangle = 0 \;\Leftrightarrow\; \eta_i = 0 \;\forall i.$$

Thus, every complex Hilbert space $\mathcal{H}$ of dimension d is isomorphic to $\mathbb{C}^d$. Notice that for $d = \infty$ the sums are replaced by series, and convergence is meant in the norm induced by the sesquilinear product. In the following we will restrict attention to the *finite dimensional* case $d < \infty$.

10.1.1 Subspaces

Let now $\mathcal{K} \subseteq \mathcal{H}$ be a subset closed under linear combinations, namely

$$\varphi, \psi \in \mathcal{K} \quad \rightarrow \quad \forall\, a, b \in \mathbb{C}\; a|\varphi\rangle + b|\psi\rangle \in \mathcal{K}.$$

We call such $\mathcal{K}$ *subspace* of $\mathcal{H}$ [1]. The dimension if $\mathcal{K}$ is the maximum cardinality of a set of linearly independent vectors in $\mathcal{K}$. Clearly, 0 belongs to every subspace $\mathcal{K}$. Special cases are $\mathcal{K} = \{0\}$ and $\mathcal{K} = \mathcal{H}$. For every subspace $\mathcal{K} \subseteq \mathcal{H}$ one can define the set

$$\mathcal{K}^\perp := \{\psi \in \mathcal{H} | \langle\varphi|\psi\rangle = 0 \;\forall \varphi \in \mathcal{K}\}.$$

One can easily prove that $\mathcal{K}^\perp$ is itself a subspace, and $(\mathcal{K}^\perp)^\perp = \mathcal{K}$. Actually, $\mathcal{K}^\perp$ is a subspace even if $\mathcal{K}$ is not. Thus, in general $(\mathcal{K}^\perp)^\perp$ is the *linear closure* of the subset $\mathcal{K}$, namely the smallest subspace that contains $\mathcal{K}$.

Given two subspaces $\mathcal{K}_1, \mathcal{K}_2 \subseteq \mathcal{H}$ such that for every $\psi \in \mathcal{H}$ there exists a unique pair $(\varphi_1, \varphi_2) \in \mathcal{K}_1 \times \mathcal{K}_2$ such that $\psi = \varphi_1 + \varphi_2$, we write $\mathcal{H} = \mathcal{K}_1 \oplus \mathcal{K}_2$. It is easy to prove, in particular, that for every subspace $\mathcal{K} \subseteq \mathcal{H}$ one has

$$\mathcal{H} = \mathcal{K} \oplus \mathcal{K}^\perp.$$

10.2 Linear operators on complex Hilbert spaces

Let $\mathcal{H}_\mathrm{A}, \mathcal{H}_\mathrm{B}$ be two Hilbert spaces with dimension $d_\mathrm{A}, d_\mathrm{B}$, respectively. The set of *linear operators* $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B})$ from $\mathcal{H}_\mathrm{A}$ to $\mathcal{H}_\mathrm{B}$ is the set of linear maps $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$, namely

$$\begin{aligned} &T|\psi\rangle \in \mathcal{H}_\mathrm{B} \quad \forall\; \psi \in \mathcal{H}_\mathrm{A}, \\ &T(a|\psi\rangle + b|\varphi\rangle) = aT|\psi\rangle + bT|\varphi\rangle. \end{aligned}$$

Two linear operators $S, T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ are equal if and only if $T|\psi\rangle = S|\psi\rangle$ for all $\psi \in \mathcal{H}_\mathrm{A}$. Clearly, given any orthonormal basis $\{e_i\}_{i=1}^{d_\mathrm{A}}$ we have

$$T = S \quad \Leftrightarrow \quad T|e_i\rangle = S|e_i\rangle, \quad \forall 1 \leq i \leq d_\mathrm{A}.$$

[1] In the infinite dimensional case this definition corresponds to a linear manifold. A subspace in this case needs to be topologically closed, in addition to being closed under linear combinations. This means that Cauchy sequences in $\mathcal{K}$ must have their limit within $\mathcal{K}$. The results for subspaces in the finite dimensional case hold exactly also in the infinite dimensional case, but they might fail if manifolds are considered instead of subspaces.

A relevant class of linear operators in $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B})$ is that of operators of the form

$$T_{\eta,\varphi}|\psi\rangle := |\eta\rangle\langle\varphi|\psi\rangle,$$

for $\eta \in \mathcal{H}_\mathrm{B}$ and $\varphi \in \mathcal{H}_\mathrm{A}$. We can consistently denote $T_{\eta,\varphi}$ as $|\eta\rangle\langle\varphi|$, so that $(|\eta\rangle\langle\varphi|)|\psi\rangle = |\eta\rangle\langle\varphi|\psi\rangle$. Linear operators $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B})$ form a complex vector space, provided that we define

$$(aT + bS)|\psi\rangle := aT|\psi\rangle + bS|\psi\rangle.$$

A special linear operator in $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ is the *identical operator*, or *identity* I_A:

$$I_\mathrm{A}|\psi\rangle = |\psi\rangle \quad \forall\ \psi \in \mathcal{H}_\mathrm{A}.$$

Notice that given the orthonormal basis $\{e_i\}_{i=1}^{d_\mathrm{A}}$ one has

$$\sum_{i=1}^{d_\mathrm{A}} |e_i\rangle\langle e_i| = I_\mathrm{A}. \tag{10.2}$$

Two linear operators $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ and $S : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{C}$ can be composed, obtaining the liner operator $R := ST : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{C}$. The product satisfies the following properties

$$\begin{aligned}
&(AB)C = A(BC),\\
&A(B + C) = AB + AC,\\
&(A + B)C = AC + BC.
\end{aligned}$$

The above properties imply that $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ is an *associative algebra*. Notice that, for any $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ one has $I_\mathrm{B}T = TI_\mathrm{A} = T$, thus

$$T = I_\mathrm{B}TI_\mathrm{A} = \sum_{i=1}^{d_\mathrm{B}}\sum_{j=1}^{d_\mathrm{A}} |e_i\rangle\langle e_i|T|f_j\rangle\langle f_j| = \sum_{i=1}^{d_\mathrm{B}}\sum_{j=1}^{d_\mathrm{A}} T_{i,j}|e_i\rangle\langle f_j|, \tag{10.3}$$

where $\{e_i\}_{i=1}^{d_\mathrm{A}}$ and $\{f_j\}_{j=1}^{d_\mathrm{B}}$ are orthonormal bases in $\mathcal{H}_\mathrm{A}$ and $\mathcal{H}_\mathrm{B}$, respectively. Thus, chosen a basis in every space, operators $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ are in one-to-one correspondence with matrices $\mathbb{M}_{d_\mathrm{B}\times d_\mathrm{A}}(\mathbb{C})$ as $T \leftrightarrow (T_{i,j})$.

The matrix representation of operators satisfies the following properties

$$\begin{aligned}
(aA + bB)_{i,j} &= aA_{i,j} + bB_{i,j}\\
(AB)_{i,j} &= \sum_{k=1}^{d_\mathrm{B}} A_{i,k}B_{k,j}
\end{aligned}$$

For every operator $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ we define the *adjoint* $T^\dagger : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A}$, defined by the following equation

$$|\eta\rangle = T^\dagger|\psi\rangle \quad \Leftrightarrow \quad \langle\eta|\varphi\rangle = \langle\psi|T|\varphi\rangle, \quad \forall\ \varphi \in \mathcal{H}_\mathrm{A}. \tag{10.4}$$

By definition one then has

$$\begin{aligned}\langle\psi|T|\varphi\rangle^* &= \langle\eta|\varphi\rangle^* \\ &= \langle\varphi|\eta\rangle \\ &= \langle\varphi|T^\dagger|\psi\rangle.\end{aligned}$$

The matrix corresponding to the adjoint of T has then matrix elements

$$T^\dagger_{i,j} = T^*_{j,i}$$

This implies the following identities

$$\begin{aligned}&(aT_1 + bT_2)^\dagger = a^*T_1^\dagger + b^*T_2^\dagger, \quad \forall ab, \in \mathbb{C},\ T_1, T_2 : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B} \\ &(AB)^\dagger = B^\dagger A^\dagger, \quad \forall B : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B},\ A : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{C}, \\ &(A^\dagger)^\dagger = A, \quad \forall A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}.\end{aligned}$$

Finally, notice that applying the definition in equation (10.4) one can easily prove that $T^\dagger_{\eta,\varphi} = T_{\varphi,\eta}$, i.e.

$$(|\eta\rangle\langle\varphi|)^\dagger = |\varphi\rangle\langle\eta|. \tag{10.5}$$

Let us now consider $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$, and define the *kernel* of T as the set

$$\mathsf{Ker}(T) := \{\psi \in \mathcal{H}_\mathrm{A} |\ T|\psi\rangle = 0\}.$$

It is straightforward to prove that $\mathsf{Ker}(T)$ is a subspace of $\mathcal{H}_\mathrm{A}$ [2]. We then define the *support* as

$$\mathsf{Supp}(T) := \mathsf{Ker}(T)^\perp,$$

and then $\mathcal{H}_\mathrm{A} = \mathsf{Ker}(T) \oplus \mathsf{Supp}(T)$. Similarly, we define the *range* of T as

$$\mathsf{Rng}(T) := \{\varphi \in \mathcal{H}_\mathrm{B} |\ \exists\psi \in \mathcal{H}_\mathrm{A},\ |\varphi\rangle = T|\psi\rangle\}.$$

Also $\mathsf{Rng}(T)$ is a subspace [3], and we can define the *co-kernel* as $\mathsf{Co-Ker}(T) := \mathsf{Rng}(T)^\perp$. It is remarkable that

$$\dim(\mathsf{Rng}(T)) = \dim(\mathsf{Supp}(T)).$$

The above dimension is called *rank* of T, denoted as $\mathsf{rank}(T) := \dim(\mathsf{Supp}(T))$. Notice that $\varphi \in \mathsf{Ker}(T^\dagger)$ if and only if

$$T^\dagger|\varphi\rangle = 0 \quad \Leftrightarrow \quad \langle\varphi|T|\psi\rangle = 0 \quad \forall\ \psi \in \mathcal{H}_\mathrm{A},$$

[2] In the infinite dimensional case, this is true of bounded operators T (whose domain can always be extended to all $\mathcal{H}_\mathrm{A}$). Indeed, being a linear operator bounded if and only if it is continuous, then $\mathsf{Ker}(T) = T^{-1}(0) = \mathcal{H}_\mathrm{A} \setminus T^{-1}(\mathcal{H}_\mathrm{B} \setminus \{0\})$, which is (topologically) closed, being the complement of the open set $T^{-1}(\mathcal{H}_\mathrm{B} \setminus \{0\})$.

[3] In the infinite dimensional case, again this is true if T is bounded.

which implies $\mathsf{Ker}(T^\dagger) = \mathsf{Rng}(T)^\perp$, namely $\mathsf{Supp}(T^\dagger) = \mathsf{Rng}(T)$. An operator $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ is *invertible* iff it is injective, namely $T|\psi\rangle = 0$ iff $|\psi\rangle = 0$. In other words, T is invertible iff $\mathsf{Ker}(T) = \{0\}$, equivalently $\mathsf{Supp}(T) = \mathcal{H}_\mathrm{A}$. In this case, we can define $T^{-1} : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A}$ such that $T^{-1}T = I_\mathrm{A}$. Indeed, let $\varphi \in \mathsf{Rng}(T)$. Suppose that $|\varphi\rangle = T|\psi\rangle = T|\eta\rangle$, which implies $T(|\psi\rangle - |\eta\rangle) = 0$. Then, it must be $|\psi\rangle = |\eta\rangle$. Thus, there is a unique ψ such that $|\varphi\rangle = T|\psi\rangle$. One can then define $T^{-1}|\varphi\rangle = |\psi\rangle$, and e.g. $\mathsf{Ker}(T^{-1}) = \mathsf{Rng}(T)^\perp$.

Polarisation identity

Let $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$. Then, for every $\psi, \varphi \in \mathcal{H}_\mathrm{A}$ one has

$$\langle\psi|T|\varphi\rangle = \frac{1}{4}\sum_{l=\pm1,\pm i} l\langle\varphi + l\psi|T|\varphi + l\psi\rangle, \tag{10.6}$$

where $|\varphi + l\psi\rangle = |\varphi\rangle + l|\psi\rangle$, and thus $\langle\varphi + l\psi| = \langle\varphi| + l^*\langle\psi|$. Equation (10.6) implies that

$$S = T \quad \Leftrightarrow \quad \langle\psi|T|\psi\rangle = \langle\psi|S|\psi\rangle \quad \forall\ \psi \in \mathcal{H}_\mathrm{A}$$

10.2.1 Unitary operators

A special class of invertible operators in $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ is that of *unitary operators*. An operator U is unitary if

$$U^\dagger U = UU^\dagger = I_\mathrm{A}.$$

By the polarisation identity, it is immediate to observe that the following conditions are equivalent to unitarity

$$\forall\psi \in \mathcal{H}_\mathrm{A}, \quad \|U\psi\| = \|\psi\|, \quad \|U^\dagger\psi\| = \|\psi\|.$$

Since by definition a unitary operator preserves the sesquilinear products, being

$$\langle\varphi|U^\dagger U|\psi\rangle = \langle\varphi|\psi\rangle,$$

a unitary operator maps orthonormal bases to orthonormal bases. Viceversa, an operator that maps an orthonormal basis to another is unitary. Indeed, we can write

$$U = \sum_{i=1}^{d_\mathrm{A}} |f_i\rangle\langle e_i|,$$

and since $|\varphi\rangle\langle\psi|^\dagger = |\psi\rangle\langle\varphi|$, we have

$$U^\dagger U = \sum_{i,i'=1}^{d_\mathrm{A}} |e_{i'}\rangle\langle f_{i'}|f_i\rangle\langle e_i| = \sum_{i=1}^{d_\mathrm{A}} |e_i\rangle\langle e_i| = I_\mathrm{A}.$$

Thus, every change of orthonormal basis is represented by a unitary operator, and viceversa every unitary operator sends an orthonormal basis to another one.

Trace

The quantities that are invariant under change of basis, like the determinant, are intrinsic properties of the operators. One of these quantities that plays a distinguished role in quantum theory is the trace $\mathrm{Tr}[T]$ of $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$, defined as the unique linear functional on the space $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ such that

$$\mathrm{Tr}[|\eta\rangle\langle\varphi|] = \langle\varphi|\eta\rangle. \tag{10.7}$$

One can easily prove that

$$\mathrm{Tr}[T] := \sum_{i=1}^{d_\mathrm{A}} \langle e_i|T|e_i\rangle, \tag{10.8}$$

where $\{e_i\}_{i=1}^{d_\mathrm{A}}$ is an orthonormal basis in $\mathcal{H}_\mathrm{A}$. Let us now consider N operators $T_1 : \mathcal{H}_{\mathrm{A}_1} \to \mathcal{H}_{\mathrm{A}_2}$, $T_2 : \mathcal{H}_{\mathrm{A}_2} \to \mathcal{H}_{\mathrm{A}_3}, \ldots$, $T_N : \mathcal{H}_{\mathrm{A}_N} \to \mathcal{H}_{\mathrm{A}_1}$. Then the following property of the trace, called *invariance under cyclic permutations* holds

$$\mathrm{Tr}[T_N T_{N-1} \ldots T_2 T_1] = \mathrm{Tr}[T_{N-1} \ldots T_2 T_1 T_N].$$

Indeed, one has

$$\begin{aligned}
\mathrm{Tr}[T_N T_{N-1} \ldots T_2 T_1] &= \sum_{e_i=1}^{d_{\mathrm{A}_1}} \langle e_i|T_N T_{N-1} \ldots T_2 T_1|e_i\rangle \\
&= \sum_{i=1}^{d_{\mathrm{A}_1}} \sum_{j=1}^{d_{\mathrm{A}_N}} \langle e_i|T_N|f_j\rangle\langle f_j|T_{N-1} \ldots T_2 T_1|e_i\rangle \\
&= \sum_{e_i=1}^{d_{\mathrm{A}_1}} \sum_{j=1}^{d_{\mathrm{A}_N}} \langle f_j|T_{N-1} \ldots T_2 T_1|e_i\rangle\langle e_i|T_N|f_j\rangle \\
&= \sum_{j=1}^{d_{\mathrm{A}_N}} \langle f_j|T_{N-1} \ldots T_2 T_1 T_N|f_j\rangle \\
&= \mathrm{Tr}[T_{N-1} \ldots T_2 T_1 T_N].
\end{aligned}$$

Now, since every change of basis is represented by a unitary operator $U : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$, we can prove that the trace does not depend on the basis in which equation (10.8) is expressed. Indeed, let us write the trace of $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ in a different basis $\{f_i\}$ with $|f_i\rangle = U|e_i\rangle$

$$\begin{aligned}
\sum_{i=1}^{d_\mathrm{A}} \langle f_i|T|f_i\rangle &= \sum_{i=1}^{d_\mathrm{A}} \langle e_i|U^\dagger T U|e_i\rangle \\
&= \mathrm{Tr}[U^\dagger T U] \\
&= \mathrm{Tr}[T U U^\dagger] \\
&= \mathrm{Tr}[T].
\end{aligned}$$

10.2.2 Selfadjoint operators

A special class of operators $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ that is relevant to quantum theory is that of *selfadjoint operators*, namely those T such that

$$T^\dagger = T$$

Clearly, a necessary and sufficient condition for T to be selfadjoint is that

$$\langle\varphi|T|\psi\rangle = \langle\psi|T|\varphi\rangle^*, \quad \forall\ \psi, \varphi \in \mathcal{H}_\mathrm{A}.$$

As to the matrix elements, T is selfadjoint if and only if

$$T_{i,j} = T^*_{j,i}$$

in any basis. The identity I_A is thus selfadjoint. Notice that the adjoint is basis independent, however it can be obtained as the composition of two involutive maps that, on the contrary, are both basis dependent: the *transpose*

$$\tau : \mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}) \to \mathcal{L}(\mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A}), \qquad \tau : T \mapsto T^T, \qquad (T^T)_{i,j} := T_{j,i},$$

and the *complex conjugate*

$$* : \mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}) \to \mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}), \qquad * : T \mapsto T^*, \qquad (T^*)_{i,j} := (T_{i,j})^*.$$

Notice that, since $\mathsf{Supp}(T^\dagger) = \mathsf{Rng}(T)$, for a selfadjoint T it holds that

$$\begin{aligned} &\mathsf{Rng}(T) = \mathsf{Supp}(T), \\ &\mathsf{Ker}(T) = \mathsf{Co} - \mathsf{Ker}(T). \end{aligned}$$

Remark 6. The trace of T is equal to the trace of T^T. Indeed,

$$\mathrm{Tr}[T] = \sum_{i=1}^{d_\mathrm{A}} T_{i,i}, \tag{10.9}$$

and $(T^T)_{i,i} = T_{i,i}$. Similarly, it is straightforward to verify that $\mathrm{Tr}[T^*] = \mathrm{Tr}[T]^*$, and thus $\mathrm{Tr}[T^\dagger] = \mathrm{Tr}[T^*] = \mathrm{Tr}[T]^*$.

Every operator $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ can be decomposed in a *real* and *imaginary* part, both selfadjoint—the names come from the analogy with complex numbers. The decomposition is the following

$$T = A + iB, \quad A = A^\dagger,\ B = B^\dagger, \tag{10.10}$$

$$A := \frac{1}{2}(T + T^\dagger), \quad B := \frac{1}{2i}(T - T^\dagger). \tag{10.11}$$

Positive operators

A very relevant subclass of selfadjoint operators for quantum theory is that of *positive semidefinite* or *non-negative definite* operators. These are defined as follows

Definition 10.1 (Positive semidefinite operator)**.** An operator $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is *positive semidefinite* if

$$\langle\psi|P|\psi\rangle \geq 0 \quad \forall\psi \in \mathcal{H}_\mathrm{A}. \tag{10.12}$$

We denote positive semidefiniteness of P as $P \geq 0$.

It is immediate to verify that the set $\mathcal{P}(\mathrm{A}) := \{P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A} | P \geq 0\}$ is a *cone*. Indeed, for any $\lambda \geq 0$ and $P \in \mathcal{P}(\mathrm{A})$ one has $\lambda P \in \mathcal{P}(\mathrm{A})$.

Substituting strict inequality in the definition, one obtains the definition of *positive definite* operators $P > 0$. Notice that selfadjointness is not required in the definition. Indeed, it comes as a consequence of the definition itself, through the polarisation identity

Lemma 10.2. *If an operator $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is positive semidefinite, then it is selfadjoint.*

Proof. Let us evaluate through the polarisation identity (10.6) the general quantity

$$\langle\psi|P|\varphi\rangle = \frac{1}{4}\sum_{l=\pm1,\pm i} l\langle\varphi + l\psi|P|\varphi + l\psi\rangle.$$

If $P \geq 0$ then we have

$$\langle\psi|P|\varphi\rangle^* = \frac{1}{4}\sum_{l=\pm1,\pm i} l^*\langle\varphi + l\psi|P|\varphi + l\psi\rangle,$$

and since

$$\begin{aligned} |\varphi + l\psi\rangle &= |\varphi\rangle + l|\psi\rangle \\ &= l(|\psi\rangle + l^*|\varphi\rangle) \\ &= l|\psi + l^*\varphi\rangle, \end{aligned}$$

and $\langle\varphi + l\psi| = l^*\langle\psi + l^*\varphi|$, we have

$$\begin{aligned} \langle\psi|P|\varphi\rangle^* &= \frac{1}{4}\sum_{l=\pm1,\pm i} l^*\langle\psi + l^*\varphi|P|\psi + l^*\varphi\rangle \\ &= \frac{1}{4}\sum_{l'=\pm1,\pm i} l'\langle\psi + l'\varphi|P|\psi + l'\varphi\rangle \\ &= \langle\varphi|P|\psi\rangle. \end{aligned}$$

□

In the following, "positive semidefinite" will be often replaced by "non-negative", or with an even worse abuse of terminology, "positive".

Projections

A further restriction of the class of positive semidefinite operators, that also plays a major role in quantum theory, is that of *projections*.

Definition 10.3 (Projection)**.** An operator $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is a *projection* if $P = P^\dagger$ and $P^2 = P$.

First of all, we prove that a projection is indeed positive semidefinite.

Lemma 10.4. *Let $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be a projection. Then $P \geq 0$.*

Proof. Let us consider the expectation

$$\begin{aligned}\langle\psi|P|\psi\rangle &= \langle\psi|P^2|\psi\rangle \\ &= \langle\psi|P^\dagger P|\psi\rangle \\ &= \|P\psi\|^2 \geq 0, \quad \forall\psi \in \mathcal{H}_\mathrm{A}.\end{aligned}$$

□

Moreover, if P is a projection, also $I_\mathrm{A} - P$ is, since $(I_\mathrm{A} - P)^\dagger = (I_\mathrm{A} - P)$, and $(I_\mathrm{A} - P)^2 = (I_\mathrm{A} - P)$. A projection P projects any vector in the subspace $\mathsf{Rng}(P) = \mathsf{Supp}(P)$. Indeed, suppose that $\psi \in \mathsf{Supp}(P) = \mathsf{Rng}(P)$. Then there exists $\varphi \in \mathcal{H}_\mathrm{A}$ such that

$$|\psi\rangle = P|\varphi\rangle,$$

and then

$$\begin{aligned}P|\psi\rangle &= P^2|\varphi\rangle \\ &= P|\varphi\rangle \\ &= |\psi\rangle\end{aligned}$$

Thus, since $\mathcal{H}_\mathrm{A} = \mathsf{Supp}(P) \oplus \mathsf{Ker}(P)$, if one decomposes a general vector $\psi \in \mathcal{H}_\mathrm{A}$ accordingly, it is

$$\psi = \psi_S + \psi_K,$$

and by definition of kernel one has

$$\begin{aligned}P|\psi\rangle &= P|\psi_S\rangle + P|\psi_K\rangle \\ &= P|\psi_S\rangle \\ &= |\psi_S\rangle.\end{aligned}$$

By the above observation, it is then clear $\mathsf{Supp}(P) = \mathsf{Ker}(I_\mathrm{A} - P)$, and thus also $\mathsf{Ker}(P) = \mathsf{Supp}(I_\mathrm{A} - P)$. This implies that for any projection P there exists $I_\mathrm{A} - P$ such that P projects on $\mathsf{Supp}(P)$ and $I_\mathrm{A} - P$ projects on $\mathsf{Supp}(P)^\perp = \mathsf{Ker}(P)$. This statement can be reversed as follows.

Lemma 10.5. *Let $\mathcal{K} \subseteq \mathcal{H}_\mathrm{A}$ be a subspace. Then there exists a projection P with $\mathsf{Supp}(P) = \mathcal{K}$.*

Proof. Let us construct the operator P as follows. First, let $\psi \in \mathcal{H}$, and consider the decomposition

$$|\psi\rangle = |\psi_{\mathcal{K}}\rangle + |\psi_{\mathcal{K}^\perp}\rangle.$$

Then we define P by

$$P|\psi\rangle \coloneqq |\psi_{\mathcal{K}}\rangle, \quad \forall\ \psi \in \mathcal{H}.$$

The operator P is well defined, since for every $\psi \in \mathcal{H}_{\mathrm{A}}$ the decomposition $\psi = \psi_{\mathcal{K}} + \psi_{\mathcal{K}^\perp}$ is unique, and moreover P is linear, since

$$\begin{aligned} a\psi + b\varphi &= a\psi_{\mathcal{K}} + a\psi_{\mathcal{K}^\perp} + b\varphi_{\mathcal{K}} + b\varphi_{\mathcal{K}^\perp} \\ &= (a\psi_{\mathcal{K}} + b\varphi_{\mathcal{K}}) + (a\psi_{\mathcal{K}^\perp} + b\varphi_{\mathcal{K}^\perp}), \end{aligned}$$

which implies

$$\begin{aligned} P(a|\psi\rangle + b|\varphi\rangle) &= a|\psi_{\mathcal{K}}\rangle + b|\varphi_{\mathcal{K}}\rangle \\ &= aP|\psi\rangle + bP|\varphi\rangle. \end{aligned}$$

Finally, P is a projection. Indeed, P is positive semidefinite, since

$$\begin{aligned} \langle\psi|P|\psi\rangle &= (\langle\psi_{\mathcal{K}}| + \langle\psi_{\mathcal{K}^\perp}|)|\psi_{\mathcal{K}}\rangle \\ &= \langle\psi_{\mathcal{K}}|\psi_{\mathcal{K}}\rangle \geq 0, \end{aligned}$$

and thus $P^\dagger = P$. Moreover,

$$\begin{aligned} P^2|\psi\rangle &= P|\psi_{\mathcal{K}}\rangle \\ &= |\psi_{\mathcal{K}}\rangle \\ &= P|\psi\rangle. \end{aligned}$$

□

Let now $\eta \in \mathsf{Ker}(P)$. Then $\eta_{\mathcal{K}} = 0$, i.e. $\eta \in \mathcal{K}^\perp$. Viceversa, if $\eta \in \mathcal{K}^\perp$ then $P|\eta\rangle = 0$. This shows that $\mathsf{Ker}(P) = \mathcal{K}^\perp$, i.e. $\mathsf{Supp}(P) = \mathcal{K}$.

In conclusion, this result shows that there is a one-to-one correspondence between subspaces of $\mathcal{H}_{\mathrm{A}}$ and projections in $\mathcal{L}(\mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}})$.

10.2.3 Normal operators and the spectral theorem

All the classes of operators $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ that we considered so far belong to a single general class, the *normal operators*. What makes normal operators special is that they are the most general class of operators that satisfy the *spectral theorem*.

Definition 10.6 (Normal operator)**.** An operator $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ is *normal* if it commutes with its adjoint, namley

$$[T, T^\dagger] = 0. \tag{10.13}$$

Notice that the *commutator* of $T, S : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is defined as

$$[S, T] := ST - TS = -[T, S].$$

We say that two operators *commute* if their commutator is null.

Lemma 10.7. *An operator $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is normal iff its real and imaginary parts $A = \frac{1}{2}(T + T^\dagger)$ and $B = \frac{1}{2i}(T - T^\dagger)$ commute.*

Proof. Let $T = A + iB$ with $A = A^\dagger$ and $B = B^\dagger$. Then

$$\begin{aligned}[T, T^\dagger] &= [A + iB, A - iB] \\ &= [A, A] + [B, B] + i[B, A] - i[A, B] \\ &= 2i[B, A].\end{aligned}$$

Thus, $[T, T^\dagger] = 0$ iff $[A, B] = 0$. □

Definition 10.8 (Diagonalisable operator)**.** We say that $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is diagonalisable if there are projections P_k satisfying $P_k P_{k'} = \delta_{kk'} P_k$ and $\{\lambda_k\} \subseteq \mathbb{C}$ with $\lambda_{k_1} \neq \lambda_{k_2}$ such that

$$T = \sum_k \lambda_k P_k, \quad I_\mathrm{A} = \sum_k P_k. \tag{10.14}$$

The support of P_k in equation (10.14) is called *eigenspace* of T.

Lemma 10.9. *Let $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be diagonalisable. The decomposition of equation* (10.14) *for T is unique.*

Proof. If $T = 0$, then $T = 0 I_\mathrm{A}$, and any other decomposition would be forbidden by the condition $\lambda_{k_1} \neq \lambda_{k_2}$. Let then $T \neq 0$, and

$$T = \sum_k \lambda_k P_k = \sum_j \mu_j Q_j.$$

Then

$$P_k T Q_j = \lambda_k P_k Q_j = \mu_j P_k Q_j.$$

Thus, either $P_k Q_j = 0$ or $\mu_j = \lambda_k$. Now, $P_k Q_j$ cannot be null for every j, k, otherwise $\lambda_k P_k = P_k T = \sum_j \mu_j P_k Q_j = 0$ for every k, which means $T = 0$. Then there must exist j such that $P_k Q_j \neq 0$. Such a value of j is unique, otherwise $\mu_{j_1} = \lambda_k = \mu_{j_2}$, contrarily to the hypotheses. Let us then define the function $\varphi(k)$ such that $\lambda_k = \mu_{\varphi(k)}$. The function is injective, otherwise $\varphi(k_1) = \varphi(k_2)$ would imply $\lambda_{k_1} = \mu_{\varphi(k_1)} = \mu_{\varphi(k_2)} = \lambda_{k_2}$. Thus,

$$\begin{aligned}\lambda_k P_k &= P_k T \\ &= \mu_{\varphi(k)} P_k Q_{\varphi(k)} \\ &= \lambda_k P_k Q_{\varphi(k)} \\ &= T Q_{\varphi(k)} \\ &= \mu_{\varphi(k)} Q_{\varphi(k)}.\end{aligned}$$

Finally, this implies $Q_{\varphi(k)} = P_k$ and $\mu_{\varphi(k)} = \lambda_k$, for $\lambda_k \neq 0$. Finally, if there is k_0 such that $\lambda_{k_0} = \mu_{\varphi(k_0)} = 0$, one has

$$\begin{aligned} P_{k_0} &= I - \sum_{k \neq k_0} P_k \\ &= I - \sum_{k \neq k_0} Q_{\varphi(k)} \\ &= Q_{\varphi(k_0)}. \end{aligned}$$

□

Definition 10.10. Let $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$. We say that $0 \neq \psi \in \mathcal{H}_{\mathrm{A}}$ is an *eigenvector* of T with *eigenvalue* λ if

$$T|\psi\rangle = \lambda|\psi\rangle. \tag{10.15}$$

The set of eigenvalues of T is called *spectrum* of T, and denoted as $\mathsf{Spec}(T)$.

Let T be diagonalisable. By the Gram-Schmidt construction, one can take a basis $\{\varphi_i^{(k)}\}_{i=1}^{\dim \mathcal{K}_k}$ for each eigenspace $\mathcal{K}_k := \mathsf{Supp}(P_k)$, and thus $\{\psi_j\}_{j=1}^{d_{\mathrm{A}}} = \bigcup_k \{\varphi_i^{(k)}\}_{i=1}^{\dim \mathcal{K}_k}$ constitutes an orthonormal basis for $\mathcal{H}_{\mathrm{A}}$. Moreover, one can easily verify that for every k

$$P_k = \sum_{i=1}^{\dim \mathcal{K}_k} |\varphi_i^{(k)}\rangle\langle\varphi_i^{(k)}|.$$

Thus, an operator T is diagonalisable iff

$$T = \sum_k \sum_{i=1}^{\dim \mathcal{K}_k} \lambda_k |\varphi_i^{(k)}\rangle\langle\varphi_i^{(k)}|,$$

which implies $T|\varphi_i^{(k)}\rangle = \lambda_k|\varphi_i^{(k)}\rangle$. Namely, if T is diagonalisable there is an orthonormal basis of eigenvectors of T, and viceversa if there is an orthonormal basis $\{\psi_i\}_{i=1}^{d_{\mathrm{A}}}$ of eigenvectors of T one has

$$T|\psi_i\rangle = \lambda_i|\psi_i\rangle,$$

and one can easily verify that this implies

$$T = \sum_{i=1}^{d_{\mathrm{A}}} \lambda_i |\psi_i\rangle\langle\psi_i|.$$

Finally, an eigenspace $\mathcal{K}_k$ contains all eigenvectors corresponding to the same eigenvalue λ_k. The dimension $\dim \mathcal{K}_k = \mathrm{Tr}[P_k]$ is called *degeneracy* of λ_k.

We now state the spectral theorem without proving it, as follows.

Theorem 10.11 (Spectral theorem)**.** *The operator $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ is diagonalisable iff T is normal.*

Corollary 10.12. *Unitary and selfadjoint operators are diagonalisable.*

Lemma 10.13. *Let T be diagonalisable. The spectrum $\mathsf{Spec}(T^\dagger)$ of the adjoint operator of T is the complex conjugate of $\mathsf{Spec}(T)$, and the eigenvectors are the same.*

Proof. Let

$$T = \sum_k \lambda_k P_k.$$

Then

$$T^\dagger = \sum_k \lambda_k^* P_k,$$

and clearly $\mathsf{Spec}(T^\dagger) = \mathsf{Spec}(T)^*$. Moreover, the projections for the decomposition of T and $T^\dagger$ are the same, thus they have the same eigenvectors. □

Corollary 10.14. *1. A normal operator U is unitary iff $\mathsf{Spec}(U)$ is contained in the unit circle C in the complex plane $\mathbb{C}$.*

2. A normal operator T is selfadjoint iff $\mathsf{Spec}(T)$ is contained in the real axis.

3. A normal operator P is positive semidefinite or positive definite iff $\mathsf{Spec}(U)$ is non-negative or positive, respectively.

4. A normal operator P is a projection iff $\mathsf{Spec}(P)$ is contained in $\{0, 1\}$.

Proof. 1. Let U be normal, and $\{\psi_i\}_{i=1}^{d_\mathrm{A}}$ be an orthonormal basis of eigenvectors $U|\psi_i\rangle = \lambda_i|\psi_i\rangle$. Then by lemma 10.13 U is unitary iff

$$\begin{aligned} |\psi_i\rangle &= U^\dagger U|\psi_i\rangle \\ &= \lambda_i U^\dagger|\psi_i\rangle \\ &= \lambda_i\lambda_i^*|\psi_i\rangle \\ &= |\lambda_i|^2|\psi_i\rangle. \end{aligned}$$

Thus, U is unitary iff $\mathsf{Spec}(U) \subseteq \mathsf{C}$.

2. Let T be normal, and $\{\psi_i\}_{i=1}^{d_\mathrm{A}}$ be an orthonormal basis of eigenvectors $T|\psi_i\rangle = \lambda_i|\psi_i\rangle$. Then T is selfadjoint iff

$$\begin{aligned} \lambda_i|\psi_i\rangle &= T|\psi_i\rangle \\ &= T^\dagger|\psi_i\rangle \\ &= \lambda_i^*|\psi_i\rangle, \end{aligned}$$

Thus, T is selfadjoint iff $\mathsf{Spec}(U) \subseteq \mathbb{R}$.

3. Let P be normal, and $\{\psi_i\}_{i=1}^{d_\mathrm{A}}$ be an orthonormal basis of eigenvectors $P|\psi_i\rangle = \lambda_i|\psi_i\rangle$. Then P is positive semidefinite iff

$$\begin{aligned} 0 &\leq \langle\psi_i|P|\psi_i\rangle \\ &= \lambda_i\langle\psi_i|\psi_i\rangle \end{aligned}$$

   Thus, P is positive semidefinite iff $\mathsf{Spec}(P) \geq 0$. For positive definite P, it is sufficient to replace the first inequality by its strict version.

4. Let P be normal and $\{\psi_i\}_{i=1}^{d_\mathrm{A}}$ be an orthonormal basis of eigenvectors $P|\psi_i\rangle = \lambda_i|\psi_i\rangle$. Then P is a projection iff

$$\begin{aligned} \lambda_i|\psi_i\rangle &= P|\psi_i\rangle \\ &= P^2|\psi_i\rangle \\ &= \lambda_i^2|\psi_i\rangle. \end{aligned}$$

   Thus, P is a projection iff $\mathsf{Spec}(P) \subseteq \{0, 1\}$. □

The above results allow us to prove two relevant properties of positive operators. These results are related to the observation that, for an arbitrary $A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$, one has

$$A^\dagger A \geq 0.$$

Indeed, we have

$$\langle\psi|A^\dagger A|\psi\rangle = \|A\psi\|^2 \geq 0, \quad \forall\psi \in \mathcal{H}_\mathrm{A}. \tag{10.16}$$

The above result can also be reversed: P is positive iff $P = A^\dagger A$. Let us now see how.

Theorem 10.15 (Existence and uniqueness of the square root)**.** *The operator $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is positive iff there exists a positive operator $S : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ such that $S^2 = T$. Such an operator S is unique.*

Proof. Let $S \geq 0$ and $T = S^2$. Then $T = S^\dagger S \geq 0$. On the other hand, let $T \geq 0$. Then

$$T = \sum_k \lambda_k P_k, \quad \lambda_k \geq 0 \ \forall\ k.$$

Let us define

$$S := \sum_k \sqrt{\lambda_k} P_k.$$

Then clearly $S \geq 0$ and $S^2 = T$. Let us now prove uniqueness of S. Indeed, suppose that $S_1^2 = S_2^2 = T \geq 0$ and $S_i \geq 0$ for both $i = 1$ and $i = 2$. Then

$$T = \sum_k (\lambda_k^{(1)})^2 P_k^{(1)} = \sum_k (\lambda_k^{(2)})^2 P_k^{(2)}.$$

However, by the uniqueness of the diagonalisation of T proved in lemma 10.9, it must be $P_k^{(2)} = P_{\varphi(k)}^{(1)}$ and since $S_i \geq 0$ for $i = 1, 2$, the eigenvalues of S_i must be positive for $i = 1, 2$, then $\lambda_k^{(2)} = \lambda_{\varphi(k)}^{(1)}$. □

The unique positive operator $S \geq 0$ such that $S^2 = T$ for $T \geq 0$ will be denoted by $\sqrt{T}$ or $T^{\frac{1}{2}}$.

Corollary 10.16. *The operator $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is positive iff there exists $\mathcal{H}_\mathrm{B}$ and $A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ such that $P = A^\dagger A$.*

Corollary 10.17. *Let $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be positive semidefinite. Then*

$$\langle \psi | P | \psi \rangle = 0 \tag{10.17}$$

iff $\psi \in \mathsf{Ker}(P)$.

Proof. Since $P = A^\dagger A$, we have

$$\langle \psi | P | \psi \rangle = 0 \;\Leftrightarrow\; 0 = \langle \psi | A^\dagger A | \psi \rangle = \|A\psi\|^2 = 0$$

which is equivalent to $A|\psi\rangle = 0$. Finally, this implies $A^\dagger A|\psi\rangle = P|\psi\rangle = 0$. The converse is trivial. □

Corollary 10.18. *Let $P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be positive. Then*

$$\mathrm{Tr}[P] = 0 \tag{10.18}$$

iff $P = 0$

Proof. By definition of trace, given any orthonormal basis $\{\varphi_i\}_{i=1}^{d_\mathrm{A}}$ one has

$$\mathrm{Tr}[P] = \sum_{i=1}^{d_\mathrm{A}} \langle \varphi_i | P | \varphi_i \rangle, \tag{10.19}$$

and since $P \geq 0$ implies that $\langle \varphi_i | P | \varphi_i \rangle \geq 0$ for every i, the trace is null iff $\langle \varphi_i | P | \varphi_i \rangle = 0$ for every i. By corollary 10.17, this is equivalent to say that for every basis $\{\varphi_i\}_{i=1}^{d_\mathrm{A}}$, every φ_i belongs to $\mathsf{Ker}(P)$. Thus, $\mathsf{Ker}(P) = \mathcal{H}_\mathrm{A}$ and $P = 0$. □

Notice that the set $\mathcal{P}(\mathrm{A}) := \{P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A} \mid P \geq 0\}$ is complete in $\mathcal{L}(\mathcal{H}_\mathrm{A})$. Indeed, the following lemmas hold.

Lemma 10.19. *Let $A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be selfadjoint. Then $A = A_+ - A_-$ where $A_\pm \geq 0$ and $A_+ A_- = A_- A_+ = 0$.*

Proof. If $A = A^\dagger$ then A is diagonalisable, and has real eigenvalues $\{\lambda_k\}$. If we divide the spectrum as $\{\lambda_k\} = \mathsf{S}_+ \cup \mathsf{S}_-$ such that $\lambda_k \in \mathsf{S}_+$ if $\lambda_k \geq 0$ and $\lambda_k \in \mathsf{S}_-$ if $\lambda_k < 0$, we have

$$\begin{aligned} A &= A_+ - A_- \\ A_+ &:= \sum_{\lambda_k \in \mathsf{S}_+} \lambda_k P_k \geq 0 \\ A_- &:= \sum_{\lambda_k \in \mathsf{S}_-} |\lambda_k| P_k \geq 0. \end{aligned}$$

Moreover, the projections on eigenspaces satisfy $P_k P_{k'} = 0$ for $\lambda_k \in \mathsf{S}_+$ and $\lambda_{k'} \in \mathsf{S}_-$. Thus, $A_+ A_- = A_- A_+ = 0$. □

Moreover, every operator $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ can be expressed as in equation (10.11) as $T = A + iB$, thus we have the following corollary.

Corollary 10.20. *Let $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$. Then T can be expanded as a linear combination of positive operators.*

Proof. It is sufficient to write

$$\begin{aligned} T &= A + iB \\ &= A_+ - A_- + iB_+ - iB_-. \end{aligned}$$ □

Finally, since for every positive operator $A \geq 0$ one has $\mathrm{Tr}[A] = 0$ iff $A = 0$, we have $A = \rho_A \,\mathrm{Tr}[A]$, with $\rho_A \geq 0$ and $\mathrm{Tr}[\rho_A] = 1$. Thus, positive operators with unit trace are complete in $\mathcal{L}(\mathcal{H}_{\mathrm{A}})$, as one can write

$$\begin{aligned} T &= A_+ - A_- + iB_+ - iB_- \\ &= \mathrm{Tr}[A_+]\rho_{A_+} - \mathrm{Tr}[A_-]\rho_{A_-} + i\,\mathrm{Tr}[B_+]\rho_{B_+} - i\,\mathrm{Tr}[B_-]\rho_{B_-}. \end{aligned}$$

Definition 10.21 (Modulus of an operator)**.** Let $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$. Then we define the *modulus* of T as the operator

$$|T| := (T^\dagger T)^{\frac{1}{2}}. \tag{10.20}$$

Remark 7. The above definition of $\sqrt{T}$ is just a special case of functional calculus for normal operators T, which can be consistently defined. Indeed, let $T = \sum_k \lambda_k P_k$. Now, for a general function $f : \mathbb{C} \to \mathbb{C}$ we can define

$$f(T) := \sum_k f(\lambda_k) P_k.$$

One can easily verify that $(g \circ f)(T) = g[f(T)]$. Moreover, functional calculus can be defined for selfadjoint operators also for functions $f : \mathbb{R} \to \mathbb{R}$. Finally, for positive semidefinite operators P one can define $f(P)$ even for functions $f : [0, \infty) \to \mathbb{R}$. This is actually the case of interest here, as we will extensively use the function $f(x) := x \log_2 x$ which can be defined on non-negative reals.

10.3 Isometric operators

The last class of operators that we introduce is that of *isometric operators* or *isometries*.

Definition 10.22 (Isometric operator)**.** Let $V : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ with $d_{\mathrm{B}} \geq d_{\mathrm{A}}$. We say that V is an *isometric operator* or *isometry* if

$$V^\dagger V = I_{\mathrm{A}}. \tag{10.21}$$

Remark 8. By the polarisation identity, it is immediate to observe that the following condition is equivalent to isometry

$$\forall \psi \in \mathcal{H}_{\mathrm{A}}, \quad \|V\psi\| = \|\psi\|.$$

Since by definition an isometric operator preserves the sesquilinear products, being

$$\langle \varphi | V^\dagger V | \psi \rangle = \langle \varphi | \psi \rangle,$$

an isometric operator maps orthonormal bases to orthonormal *sets*, generally not complete.

Remark 9. Notice that, differently from the case of unitary operators we do not require $VV^\dagger = I_{\mathrm{B}}$. In the general case, however, $VV^\dagger$ is a projection. Indeed, $(VV^\dagger)^\dagger = VV^\dagger$, and $(VV^\dagger)(VV^\dagger) = V(V^\dagger V)V^\dagger = VV^\dagger$. Since for an orthonormal basis $\{\varphi_i\}_{i=1}^{d_{\mathrm{A}}}$ one has that $\{V\varphi_i\}_{i=1}^{d_{\mathrm{A}}}$ is an orthonormal set, this set will be a basis for $\mathsf{Rng}(V)$. Moreover, let $\psi \in \mathsf{Rng}(V)$. Then there exists $\varphi \in \mathcal{H}_{\mathrm{A}}$ such that $\psi = V\varphi$. Thus

$$\begin{aligned} VV^\dagger |\psi\rangle &= VV^\dagger V |\varphi\rangle \\ &= V|\varphi\rangle \\ &= |\psi\rangle. \end{aligned}$$

Since $VV^\dagger$ is a projection, it follows that $\mathsf{Rng}(V) \subseteq \mathsf{Supp}(VV^\dagger)$. Moreover, let $\eta \in \mathsf{Rng}(V)^\perp$. Then for all $\psi \in \mathcal{H}_{\mathrm{A}}$

$$\begin{aligned} 0 &= \langle \eta | V | \psi \rangle \\ &= \langle \psi | V^\dagger | \eta \rangle^*, \end{aligned}$$

namely $\eta \in \mathsf{Ker}(V^\dagger)$, and thus $\eta \in \mathsf{Ker}(VV^\dagger)$. Then $\mathsf{Rng}(V)^\perp \subseteq \mathsf{Ker}(VV^\dagger)$, and finally $\mathsf{Supp}(VV^\dagger) \subseteq \mathsf{Rng}(V)$. This implies that $\mathsf{Supp}(VV^\dagger) = \mathsf{Rng}(V)$, and $VV^\dagger$ is the projection on $\mathsf{Rng}(V)$. Thus, $\dim \mathsf{Supp}(VV^\dagger) = \dim \mathsf{Rng}(V) = d_{\mathrm{A}}$, and if $d_{\mathrm{B}} = d_{\mathrm{A}}$ one has $VV^\dagger = I_{\mathrm{B}}$, namely V is unitary.

Remark 10. The adjoint $V^\dagger$ of an isometry V is called co-isometry, and is a special case of a *partial isometry*. In general, an operator $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ is called a *partial isometry* if $V^\dagger V = P_{\mathsf{Supp}(V)}$.

10.4 Polar decomposition

Theorem 10.23 (Polar decomposition)**.** *Let $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$, with $d_{\mathrm{A}} \leq d_{\mathrm{B}}$. Then there exists an isometry $S : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ such that*

$$T = S|T|. \tag{10.22}$$

On the other hand, if $d_{\mathrm{B}} < d_{\mathrm{A}}$ there exists a co-isometry $S : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ such that equation (10.22) *holds, and the projection $S^\dagger S$ satisfies $S^\dagger S \geq P_{\mathsf{Supp}(|T|)}$, namely S is isometric on* $\mathsf{Supp}(|T|)$.

Proof. For $T = 0$ the thesis is trivial. Let us then consider $T \neq 0$. We observe that for every $\psi \in \mathcal{H}_{\mathrm{A}}$

$$\begin{aligned} \|T\psi\|^2 &= \langle \psi | T^\dagger T | \psi \rangle \\ &= \langle \psi | (|T|^2) | \psi \rangle \\ &= \| (|T|\psi) \|^2. \end{aligned} \tag{10.23}$$

Let us now define $S_0 : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ with $\mathsf{Ker}(S_0) = \mathsf{Ker}(|T|)$, and for $\psi \in \mathsf{Rng}(|T|) = \mathsf{Supp}(|T|)$, namely $\psi = |T|\varphi$

$$S_0|\psi\rangle := T|\varphi\rangle.$$

We need to prove that the operator S_0 is well defined. For this purpose, let

$$|\psi\rangle = |T||\varphi_0\rangle = |T||\varphi_1\rangle.$$

Then by equation (10.23) we have

$$\begin{aligned} 0 &= \| \{ |T|(\varphi_0 - \varphi_1) \} \|^2 \\ &= \|T(\varphi_0 - \varphi_1)\|^2, \end{aligned}$$

and finally $S_0|\psi\rangle = T|\varphi_0\rangle = T|\varphi_1\rangle$. Thus, S_0 is well defined. We now show that S_0 is linear. Indeed, let $\psi = a\psi_1 + b\psi_2$. Then

$$\begin{aligned} P_{\mathsf{Supp}(|T|)}|\psi\rangle &= aP_{\mathsf{Supp}(|T|)}|\psi_1\rangle + bP_{\mathsf{Supp}(|T|)}|\psi_2\rangle \\ &= a|T||\varphi_1\rangle + b|T||\varphi_2\rangle \\ &= |T|(a|\varphi_1\rangle + b|\varphi_2\rangle), \end{aligned}$$

and thus

$$\begin{aligned} S_0(a|\psi_1\rangle + b|\psi_2\rangle) &= S_0|\psi\rangle \\ &= S_0 P_{\mathsf{Supp}(|T|)}|\psi\rangle \\ &= T(a|\varphi_1\rangle + b|\varphi_2\rangle) \\ &= aS_0|\psi_1\rangle + bS_0|\psi_2\rangle. \end{aligned}$$

Since $|T|$ is selfadjoint, $\mathsf{Supp}(|T|) = \mathsf{Rng}(|T|)$, therefore for any $\psi \in \mathcal{H}_{\mathrm{A}}$ there exists φ such that $P_{\mathsf{Supp}(|T|)}\psi = |T|\varphi$. By equation (10.23) we then have for every $\psi \in \mathcal{H}_{\mathrm{A}}$

$$\begin{aligned} \langle \psi | (S_0^\dagger S_0 - P_{\mathsf{Supp}(|T|)}) | \psi \rangle &= \langle \psi | P_{\mathsf{Supp}(|T|)} (S_0^\dagger S_0 - P_{\mathsf{Supp}(|T|)}) P_{\mathsf{Supp}(|T|)} | \psi \rangle \\ &= \langle \varphi | |T| (S_0^\dagger S_0 - P_{\mathsf{Supp}(|T|)}) |T| | \varphi \rangle \\ &= 0. \end{aligned}$$

Finally, by the polarisation identity this implies that $S_0^\dagger S_0 = P_{\mathsf{Supp}(|T|)}$. Since $\mathsf{Rng}(|T|) = \mathsf{Supp}(|T|)$, we have $k := \dim \mathsf{Ker}(|T|) = d_{\mathrm{A}} - \dim \mathsf{Supp}(|T|) \leq d_{\mathrm{B}} - \dim \mathsf{Supp}(|T|) = d_{\mathrm{B}} - \dim \mathsf{Rng}(S_0) = \dim \mathsf{Co} - \mathsf{Ker}(S_0)$. Thus, for an orthonormal basis $\{f_j\}_{j=1}^k$ of $\mathsf{Ker}(|T|)$

one can find k orthonormal vectors $\{f'_j\}_{j=1}^k$ in $\mathsf{Co}-\mathsf{Ker}(S_0)$. We then construct the operator

$$S := S_0 + S_1, \quad S_1 := \sum_{i=1}^{k} |f'_i\rangle\langle f_i|.$$

Now, it is immediate to see that $S_1^\dagger S_0 = 0$, since for every $\psi \in \mathcal{H}$ one has $\langle f'_i|S_0|\psi\rangle = 0$ for all $1 \leq i \leq k$, and thus

$$S_1^\dagger S_0 = \sum_{i=1}^{k} |f_i\rangle\langle f'_i|S_0 = 0.$$

Moreover, $S_1^\dagger S_1 = I_{\mathrm{A}} - P_{\mathsf{Supp}(|T|)}$, since

$$S_1^\dagger S_1 = \sum_{i=1}^{k} |f_i\rangle\langle f_i| = P_{\mathsf{Ker}(|T|)}.$$

Then one can easily verify that

$$\begin{aligned} S^\dagger S &= (S_0 + S_1)^\dagger (S_0 + S_1) \\ &= S_0^\dagger S_0 + S_0^\dagger S_1 + S_1^\dagger S_0 + S_1^\dagger S_1 \\ &= S_0^\dagger S_0 + S_1^\dagger S_1 = I_{\mathrm{A}}. \end{aligned}$$

In conclusion, S is isometric and for every $\psi \in \mathcal{H}_{\mathrm{A}}$ one has

$$S|T||\psi\rangle = T|\psi\rangle.$$

Now, if $d_{\mathrm{B}} < d_{\mathrm{A}}$ we can use the first part of the theorem to prove that $T^\dagger = S|T^\dagger|$. Since $|T^\dagger|^2 = TT^\dagger$, we have

$$\begin{aligned} |T|^2 &= T^\dagger T \\ &= S|T^\dagger|^2 S^\dagger \\ &= S|T^\dagger|S^\dagger S|T^\dagger|S^\dagger, \end{aligned}$$

and then, since $\langle\psi|S|T^\dagger|S^\dagger|\psi\rangle \geq 0$ for every $\psi \in \mathcal{H}_{\mathrm{A}}$, one has

$$S|T^\dagger|S^\dagger = |T|.$$

Thus, since $T = |T^\dagger|S^\dagger$, we conclude that

$$T = S^\dagger S|T^\dagger|S^\dagger = S^\dagger|T|.$$

Notice that since $SS^\dagger|T| = S|T^\dagger|S^\dagger = |T|$, one has $SS^\dagger \geq P_{\mathsf{Supp}(|T|)}$. □

Corollary 10.24. *Let $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ and $B : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{C}}$. Then*

$$A^\dagger A = B^\dagger B \tag{10.24}$$

iff $A = VQ$ and $B = WQ$, with $Q := |A| = |B|$ and $P_{\mathsf{Supp}(Q)} \leq V^\dagger V \leq I_{\mathrm{A}}$ and $P_{\mathsf{Supp}(Q)} \leq W^\dagger W \leq I_{\mathrm{A}}$.

10.4.1 Singular value decomposition

Theorem 10.25 (Singular value decomposition)**.** *Let $T : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm B}$, with $d_{\rm A} \leq d_{\rm B}$. Then there exists an isometry $V : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm B}$ and a unitary U such that*

$$T = V\Sigma U, \tag{10.25}$$

where $\Sigma_{i,j} = \sigma_i \delta_{i,j}$, and $\{\sigma_i\} = \mathsf{Spec}(|T|)$. On the other hand, if $d_{\rm B} < d_{\rm A}$ there exists a co-isometry $V : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm B}$ such that equation (10.25) *holds and $V^\dagger V \geq P_{\mathsf{Supp}(\Sigma)}$.*

Proof. The proof uses the polar decomposition

$$T = S|T|,$$

and since $|T| \geq 0$ one has

$$|T| = \sum_{i=1}^{d_{\rm A}} \sigma_i |\psi_i\rangle\langle\psi_i|,$$

with $\sigma_i \geq 0$ for all values of i. Upon defining $U|\psi_i\rangle = |e_i\rangle$, where $\{e_i\}_{i=1}^{d_{\rm A}}$ is the canonical basis, we have $U|T|U^\dagger = \Sigma$, and then

$$T = SU^\dagger \Sigma U.$$

Finally, the thesis is proved by setting $V := SU^\dagger$. □

Corollary 10.26. *For every $T : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm B}$ one has*

$$T = \sum_{i=1}^{d_{\rm A}} \sigma_i |\varphi_i\rangle\langle\psi_i|, \tag{10.26}$$

where $\{\sigma_i\} = \mathsf{Spec}(|T|) \subseteq \mathbb{R}_+$ and $\langle\varphi_i|\varphi_j\rangle = \langle\psi_i|\psi_j\rangle = \delta_{i,j}$.

10.5 Tensor product

We will now construct the tensor product of two Hilbert spaces, and review a few of its properties. We also introduce the *double-ket* notation, a very useful tool for calculations, which makes it very practical to use tensor product spaces.

Let $\mathcal{H}_{\rm A} \simeq \mathbb{C}^{d_{\rm A}}$ and $\mathcal{H}_{\rm B} \simeq \mathbb{C}^{d_{\rm B}}$. We formally introduce free linear span $\mathcal{F}$ of formal vectors $\psi \otimes \varphi$ with $(\psi, \varphi) \in \mathcal{H}_{\rm A} \times \mathcal{H}_{\rm B}$, namely

$$\mathcal{F} = \left\{ \sum_l c_l (\psi_l \otimes \varphi_l) \mid (\psi_l, \varphi_l) \in \mathcal{H}_{\rm A} \times \mathcal{H}_{\rm B},\ \{c_l\} \subseteq \mathbb{C} \right\}.$$

We define the following equivalence relations in $\mathcal{F}$

$$(a\psi_1 + b\psi_2) \otimes \varphi \sim a(\psi_1 \otimes \varphi) + b(\psi_2 \otimes \varphi),$$
$$\psi \otimes (a\varphi_1 + b\varphi_2) \sim a(\psi \otimes \varphi_1) + b(\psi \otimes \varphi_2).$$

We then define the tensor product $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ as the quotient $\mathcal{F}/\sim$. Notice that, taking $b = 0$ one can easily prove that $[(a\psi) \otimes \varphi] = [\psi \otimes (a\varphi)] = [a(\psi \otimes \varphi)]$. Thus, the null element in $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ is the class that contains $0 \otimes \varphi$ for any $\varphi \in \mathcal{H}_\mathrm{B}$ and $\psi \otimes 0$ for any $\psi \in \mathcal{H}_\mathrm{A}$. The sesquilinear product needs to be defined only on classes of *factorised vectors*, represented by $|\psi\rangle \otimes |\varphi\rangle := [\psi \otimes \varphi]$, and then just extended by sesquilinearity. We then set

$$(\langle\psi_1| \otimes \langle\varphi_1|)(|\psi_2\rangle \otimes |\varphi_2\rangle) := \langle\psi_1|\psi_2\rangle \, \langle\varphi_1|\varphi_2\rangle.$$

If we now consider two orthonormal bases $\{e_i\}_{i=1}^{d_\mathrm{A}} \subseteq \mathcal{H}_\mathrm{A}$ and $\{f_j\}_{j=1}^{d_\mathrm{B}} \subseteq \mathcal{H}_\mathrm{B}$, one can easily prove that the set $\{e_i \otimes f_j\}$ is orthonormal in $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$. Moreover, it is complete by construction. Thus $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B} \simeq \mathbb{C}^{d_\mathrm{A} d_\mathrm{B}}$. Moreover, by construction one has

$$|\psi\rangle \otimes |\varphi\rangle = \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} \psi_i \varphi_j (|e_i\rangle \otimes |f_j\rangle) \tag{10.27}$$

10.5.1 Linear operators

The linear operators in $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ are simply the linear operators on $\mathbb{C}^{d_\mathrm{A} d_\mathrm{B}}$. However, due to the construction, there is a special class of linear operators in $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$: *factorised operators* of the form $A \otimes B$, defined by

$$(A \otimes B)(\psi \otimes \varphi) := (A\psi) \otimes (B\psi).$$

Notice that being $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ spanned by factorised vectors this definition is sufficient to specify the action of $A \otimes B$ on the whole $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$. Given two factorised operators $A \otimes B$ and $C \otimes D$, one can then easily verify that

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \tag{10.28}$$

and that the tensor product distributes over sums:

$$(A + B) \otimes C = A \otimes C + B \otimes C, \quad A \otimes (B + C) = A \otimes B + A \otimes C. \tag{10.29}$$

We can also observe that

$$|\psi_1\rangle\langle\psi_2| \otimes |\varphi_1\rangle\langle\varphi_2| = (|\psi_1\rangle \otimes |\varphi_1\rangle)(\langle\psi_2| \otimes \langle\varphi_2|).$$

Reminding the equality in (10.2), the identity operator I_AB on $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ is then equal to

$$\begin{aligned} I_\mathrm{AB} &= \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} (|e_i\rangle \otimes |f_j\rangle)(\langle e_i| \otimes \langle f_j|) \\ &= \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} |e_i\rangle\langle e_i| \otimes |f_j\rangle\langle f_j| \\ &= I_\mathrm{A} \otimes I_\mathrm{B}. \end{aligned} \tag{10.30}$$

The definition of $A \otimes B$ can be straightforwardly generalised to the case where $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{C}}$ and $B : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{D}}$. It is immediate to verify that two operators of the form $A \otimes I$ and $I \otimes B$ with $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ and $B : \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{D}}$ commute, and satisfy the identity

$$(A \otimes I_{\mathrm{D}})(I_{\mathrm{A}} \otimes B) = (I_{\mathrm{B}} \otimes B)(A \otimes I_{\mathrm{C}}) = A \otimes B \tag{10.31}$$

By the same procedure that we adopted in equation (10.3), we can then see that the matrix elements of $T : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}} \otimes \mathcal{H}_{\mathrm{D}}$ are defined by

$$\begin{aligned} T &= I_{\mathrm{CD}} T I_{\mathrm{AB}} \\ &= \sum_{i,j} \sum_{i',j'} (|e_i\rangle\langle e_i| \otimes |f_j\rangle\langle f_j|) T (|e_{i'}\rangle\langle e_{i'}| \otimes |f_{j'}\rangle\langle f_{j'}|) \\ &= \sum_{i,j} \sum_{i',j'} T_{(i,j),(i',j')} |e_i\rangle\langle e_{i'}| \otimes |f_j\rangle\langle f_{j'}|. \end{aligned}$$

Thus, it holds that

$$(A \otimes B)_{(i,j),(i',j')} = A_{i,j} B_{i',j'}.$$

As the set of operators between any pair of Hilbert spaces, also $\mathcal{L}(\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}} \otimes \mathcal{H}_{\mathrm{D}})$ is a complex vector space. In this space, factorised operators are complete. Indeed, we have the following lemma.

Lemma 10.27. *Let $T : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}} \otimes \mathcal{H}_{\mathrm{D}}$. Then there exist two finite sets $\{A_i\}_{i \in \mathrm{X}} \subseteq \mathcal{L}(\mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{C}})$ and $\{B_i\}_{i \in \mathrm{X}} \subseteq \mathcal{L}(\mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{D}})$ such that*

$$T = \sum_{i \in \mathrm{X}} A_i \otimes B_i. \tag{10.32}$$

Proof. Indeed, one has

$$\begin{aligned} T &= \sum_{i,j} \sum_{i',j'} T_{(i,j),(i',j')} |e_i\rangle\langle e_{i'}| \otimes |f_j\rangle\langle f_{j'}| \\ &= \sum_{i,i'} |e_i\rangle\langle e_{i'}| \otimes B_{i,i'}, \\ B_{i,i'} &:= \sum_{j,j'} T_{(i,j),(i',j')} |f_j\rangle\langle f_{j'}|. \end{aligned}$$

□

Before concluding this section, we introduce a relevant unitary operator on $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$, called *swap*.

Definition 10.28 (Swap operator)**.** For every tensor product $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$ with $\mathcal{H}_{\mathrm{A}} \simeq \mathcal{H}_{\mathrm{B}}$ we define the *swap* operator

$$E(|\psi\rangle \otimes |\varphi\rangle) = |\varphi\rangle \otimes |\psi\rangle. \tag{10.33}$$

Lemma 10.29. *The swap operator E is selfadjoint.*

Proof. Let us write the adjoint equation for a factorised basis $|\psi_i\rangle \otimes |\varphi_j\rangle$

$$\begin{aligned}\langle\theta_{i,j}|(|\psi\rangle \otimes |\varphi\rangle) &= (\langle\psi_i| \otimes \langle\varphi_j|)E(|\psi\rangle \otimes |\varphi\rangle) \\ &= (\langle\psi_i| \otimes \langle\varphi_j|)(|\varphi\rangle \otimes |\psi\rangle) \\ &= \langle\psi_i|\varphi\rangle\langle\varphi_j|\psi\rangle \\ &= (\langle\varphi_j| \otimes \langle\psi_i|)(|\psi\rangle \otimes |\varphi\rangle).\end{aligned}$$

Thus, the equation is satisfied by $|\theta_{i,j}\rangle = |\varphi_j\rangle \otimes |\psi_i\rangle$, namely $E^\dagger(|\psi_i\rangle \otimes |\varphi_j\rangle) = |\varphi_j\rangle \otimes |\psi_i\rangle$, and finally $E^\dagger = E$. □

Lemma 10.30. *The square of E is the identity; $E^2 = I_{\mathrm{AB}}$.*

Proof. It is sufficient to prove the statement on factorised states. However, in this case the proof is trivial. □

Corollary 10.31. *The swap operator is unitary.*

Proof. It is sufficient to collect the two previous results, getting

$$\begin{aligned}I_{\mathrm{AB}} &= E^2 \\ &= E^\dagger E \\ &= EE^\dagger.\end{aligned}$$ □

10.6 The partial trace

Notice that $\langle\psi|$ can be thought of as a special kind of operator from $\mathcal{H}_{\mathrm{A}}$ to $\mathcal{H}_{\mathrm{I}} \simeq \mathbb{C}$, namely $d_{\mathrm{I}} = 1$. Clearly, $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{I}} = \mathcal{H}_{\mathrm{I}} \otimes \mathcal{H}_{\mathrm{A}} = \mathcal{H}_{\mathrm{A}}$ Similarly, $|\psi\rangle$ can be thought of as a special kind of operator from $\mathcal{H}_{\mathrm{I}}$ to $\mathcal{H}_{\mathrm{A}}$. Thus, for $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}}$ and $\psi \in \mathcal{H}_{\mathrm{C}}$, we can define the following factorised operators

$$T \otimes \langle\psi| : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{B}}, \quad T \otimes \langle\psi|\,(|\eta\rangle \otimes |\varphi\rangle) := (\langle\psi|\varphi\rangle)T|\eta\rangle, \tag{10.34}$$

$$T \otimes |\psi\rangle : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{C}}, \quad (T \otimes |\psi\rangle)|\eta\rangle := (T|\eta\rangle) \otimes |\psi\rangle. \tag{10.35}$$

We can thus define the *partial trace* as follows.

Definition 10.32 (Partial trace)**.** Let $T : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{C}}$. Then we define its *partial trace* on $\mathcal{H}_{\mathrm{A}}$, denoted as $\mathrm{Tr}_{\mathrm{A}}[T]$, as follows

$$\mathrm{Tr}_{\mathrm{A}}[T] := \sum_{i=1}^{d_{\mathrm{A}}} (\langle e_i| \otimes I_{\mathrm{C}})T(|e_i\rangle \otimes I_{\mathrm{B}}). \tag{10.36}$$

Lemma 10.33. *Let $T : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}}$, and write $T = \sum_i A_i \otimes B_i$. The partial trace of T on $\mathcal{H}_{\mathrm{A}}$ is given by*

$$\mathrm{Tr}_{\mathrm{A}}[T] = \sum_i \mathrm{Tr}[A_i]B_i. \tag{10.37}$$

Proof. Applying the definition one has

$$\begin{aligned}\mathrm{Tr_A}[T] &= \sum_i \sum_{j=1}^{d_\mathrm{A}} (\langle e_j| \otimes I_\mathrm{C})(A_i \otimes B_i)(|e_j\rangle \otimes I_\mathrm{B}) \\ &= \sum_i \sum_{j=1}^{d_\mathrm{A}} (\langle e_j|A_i|e_j\rangle) I_\mathrm{C} B_i I_\mathrm{B} \\ &= \sum_i \mathrm{Tr}[A_i] B_i.\end{aligned}$$

□

By the above lemma and by the independence of the trace on the basis, we have the following corollary

Corollary 10.34. *The partial trace on $\mathcal{H}_\mathrm{A}$ is independent of the basis $\{e_i\}_{i=1}^{d_\mathrm{A}}$ in the defining equation* (10.36).

Moreover, using equation (10.28) and the decomposition of T in equation (10.32), one can easily prove the following corollary.

Corollary 10.35. *The following identity holds for operators T, X, Y defined on suitable Hilbert spaces*

$$\mathrm{Tr_A}[(I_\mathrm{A} \otimes X)T(I_\mathrm{A} \otimes Y)] = X \,\mathrm{Tr_A}[T]Y. \tag{10.38}$$

Lemma 10.36. *The partial traces* $\mathrm{Tr_A}$ *and* $\mathrm{Tr_B}$ *commute on operators $T : \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$, and satisfy the following identity*

$$\mathrm{Tr_A}[\mathrm{Tr_B}[T]] = \mathrm{Tr_B}[\mathrm{Tr_A}[T]] = \mathrm{Tr}[T]. \tag{10.39}$$

Proof. The thesis follows straightforwardly from the definition. Indeed, reminding equation 10.31, along with the fact that $\mathcal{H}_\mathrm{X} \otimes \mathcal{H}_\mathrm{I} = \mathcal{H}_\mathrm{I} \otimes \mathcal{H}_\mathrm{X} = \mathcal{H}_\mathrm{X}$ and $I_\mathrm{I} = 1$, we have

$$\begin{aligned}(I_\mathrm{A} \otimes |\psi\rangle)|\varphi\rangle &= (|\varphi\rangle \otimes I_\mathrm{B})|\psi\rangle = |\varphi\rangle \otimes |\psi\rangle \\ \langle\varphi|(I_\mathrm{A} \otimes \langle\psi|) &= \langle\psi|(\langle\varphi| \otimes I_\mathrm{B}) = \langle\varphi| \otimes \langle\psi|.\end{aligned}$$

Thus the thesis follows. □

10.7 The Vec isomorphism: double-ket notation

In the reminder we will often simplify the notation by posing

$$|\psi\rangle|\varphi\rangle := |\psi\rangle \otimes |\varphi\rangle \tag{10.40}$$

The isomorphism between $\mathcal{H}$ and $\mathcal{H}^*$ and the consequent isomorphism between $\mathcal{H} \to \mathcal{K}$ and $\mathcal{K} \otimes \mathcal{H}^*$ is well known in tensor analysis, and it is extensively used under the operation of *lowering* or *raising* indices. Here we introduce it with a special formalism that is extremely convenient for calculations.

The formalism is based on the following way of denoting vectors $\Psi \in \mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}$

$$|\Psi\rangle\!\rangle := \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} \Psi_{j,i} |f_j\rangle|e_i\rangle. \tag{10.41}$$

Definition 10.37. Let $\{e_i\}_{i=1}^{d_\mathrm{A}}$ and $\{f_j\}_{j=1}^{d_\mathrm{B}}$ be orthonormal bases in $\mathcal{H}_\mathrm{A}$ and $\mathcal{H}_\mathrm{B}$, respectively. Let

$$A = \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} A_{j,i} |f_j\rangle\langle e_i| \tag{10.42}$$

$$|\Psi\rangle\!\rangle = \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} \Psi_{j,i} |f_j\rangle|e_i\rangle \tag{10.43}$$

Then we define the isomorphism $\mathsf{Vec} : \mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}) \to \mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}$ along wiht its inverse $\mathsf{Vec}^{-1} : \mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A} \to \mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B})$ as follows

$$\mathsf{Vec}: \qquad A \mapsto |A\rangle\!\rangle := \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} A_{j,i} |f_j\rangle|e_i\rangle, \tag{10.44}$$

$$\mathsf{Vec}^{-1}: \qquad |\Psi\rangle\!\rangle \mapsto \Psi := \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} \Psi_{j,i} |f_j\rangle\langle e_i| \tag{10.45}$$

Theorem 10.38. *The following identity holds for $C : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$, $A : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{C}$ and $B : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{D}$.*

$$A \otimes B|C\rangle\!\rangle = |ACB^T\rangle\!\rangle \tag{10.46}$$

Moreover, for every pair $A, B : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ one has

$$\mathrm{Tr}_\mathrm{A}[|A\rangle\!\rangle\langle\!\langle B|] = AB^\dagger, \tag{10.47}$$

$$\mathrm{Tr}_\mathrm{B}[|A\rangle\!\rangle\langle\!\langle B|] = A^T B^*, \tag{10.48}$$

$$\mathrm{Tr}[|A\rangle\!\rangle\langle\!\langle B|] = \langle\!\langle B|A\rangle\!\rangle = \mathrm{Tr}[B^\dagger A] = \mathrm{Tr}[B^* A^T]. \tag{10.49}$$

Proof. First of all, we prove the identity of equation (10.46). Indeed, by definition we have

$$|C\rangle\!\rangle = \sum_{i=1}^{d_{\rm A}} \sum_{j=1}^{d_{\rm B}} C_{j,i} |f_j\rangle |e_i\rangle.$$

Thus, if we apply $A \otimes B$ on both sides we obtain

$$A \otimes B |C\rangle\!\rangle = \sum_{i=1}^{d_{\rm A}} \sum_{j=1}^{d_{\rm B}} C_{j,i} A|f_j\rangle \otimes B|e_i\rangle.$$

Finally, applying $I_{\rm B} \otimes I_{\rm D}$ to the left on both sides, using the form of equation (10.30), we get

$$A \otimes B |C\rangle\!\rangle = \sum_{j'=1}^{d_{\rm C}} \sum_{i'=1}^{d_{\rm D}} \sum_{i=1}^{d_{\rm A}} \sum_{j=1}^{d_{\rm B}} C_{j,i} A_{j',j} B_{i',i} |f'_{j'}\rangle |e'_{i'}\rangle,$$

where $\{e'_i\}_{i=1}^{d_{\rm D}}$ and $\{f'_j\}_{j=1}^{d_{\rm C}}$ are orthonormal bases in $\mathcal{H}_{\rm D}$ and $\mathcal{H}_{\rm C}$, respectively, and

$$A_{j',j} := \langle f'_{j'}|A|f_j\rangle, \quad B_{i',i} := \langle e'_{i'}|B|e_i\rangle.$$

Finally, reminding that $(B^T)_{i,i'} = B_{i',i}$, we obtain

$$A \otimes B |C\rangle\!\rangle = \sum_{j'=1}^{d_{\rm C}} \sum_{i'=1}^{d_{\rm D}} (ACB^T)_{j',i'} |f'_{j'}\rangle |e'_{i'}\rangle,$$

namely equation (10.46). Let us then consider the vector $|I_{\rm A}\rangle\!\rangle \in \mathcal{H}_{\rm A} \otimes \mathcal{H}_{\rm A'}$, with $\mathcal{H}_{\rm A} \simeq \mathcal{H}_{\rm A'}$. Now, one simply needs to calculate $\mathrm{Tr}_{\rm A}[|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|]$, since using equation (10.36)

$$\begin{aligned} \mathrm{Tr}_{\rm A}[|A\rangle\!\rangle\langle\!\langle B|] &= \mathrm{Tr}_{\rm A}[(A \otimes I_{\rm A})|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|(B^\dagger \otimes I_{\rm A})] \\ &= A\, \mathrm{Tr}_{\rm A}[|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|] B^\dagger, \qquad (10.50) \\ \mathrm{Tr}_{\rm B}[|A\rangle\!\rangle\langle\!\langle B|] &= \mathrm{Tr}_{\rm B}[(I_{\rm B} \otimes A^T)|I_{\rm B}\rangle\!\rangle\langle\!\langle I_{\rm B}|(I_{\rm B} \otimes B^*)] \\ &= A^T\, \mathrm{Tr}_{\rm B}[|I_{\rm B}\rangle\!\rangle\langle\!\langle I_{\rm B}|] B^*. \qquad (10.51) \end{aligned}$$

Now, consider that

$$\begin{aligned} |I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}| &= \sum_{i,i'=1}^{d_{\rm A}} |e_i\rangle|e_i\rangle\langle e_{i'}|\langle e_{i'}| \\ &= \sum_{i,i'=1}^{d_{\rm A}} |e_i\rangle\langle e_{i'}|_{\rm A} \otimes |e_i\rangle\langle e_{i'}|_{\rm A'}, \end{aligned}$$

and taking $\mathrm{Tr_A}$ we obtain

$$\begin{aligned}\mathrm{Tr_A}[|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|] &= \sum_{i,i'=1}^{d_\mathrm{A}} \mathrm{Tr}[|e_i\rangle\langle e_{i'}|]|e_i\rangle\langle e_{i'}|_{\mathrm{A}'} \\ &= \sum_{i=1}^{d_\mathrm{A}} |e_{i'}\rangle\langle e_{i'}|_{\mathrm{A}'} \\ &= I_{\mathrm{A}'}.\end{aligned}$$

The same argument leads to $\mathrm{Tr_B}[|I_\mathrm{B}\rangle\!\rangle\langle\!\langle I_\mathrm{B}|] = I_{\mathrm{B}'}$. Thus, equations (10.50) and (10.51) lead to

$$\mathrm{Tr_A}[|A\rangle\!\rangle\langle\!\langle B|] = AB^\dagger, \quad \mathrm{Tr_B}[|A\rangle\!\rangle\langle\!\langle B|] = A^T B^*.$$

Finally, reminding lemma 10.36 we have

$$\begin{aligned}\mathrm{Tr}[|A\rangle\!\rangle\langle\!\langle B|] &= \mathrm{Tr_B}[\mathrm{Tr_A}[|A\rangle\!\rangle\langle\!\langle B|]] = \mathrm{Tr_A}[\mathrm{Tr_B}[|A\rangle\!\rangle\langle\!\langle B|]] \\ &= \mathrm{Tr}[AB^\dagger] = \mathrm{Tr}[A^T B^*].\end{aligned}$$ □

Using the double-ket notation it is easy to prove the so-called Schmidt decomposition of vectors in $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$.

Theorem 10.39 (Schmidt decomposition)**.** *Let $|\Psi\rangle\!\rangle \in \mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}$. Then one can find two orthonormal sets $\{\psi_i\}_{i=1}^{d_\mathrm{B}}$ and $\{\varphi_j\}_{j=1}^{d_\mathrm{A}}$ and a set of positive reals $\{\sigma_i\}_{i=1}^{s}$ with $s \leq \min(d_\mathrm{A}, d_\mathrm{B})$ such that*

$$|\Psi\rangle\!\rangle = \sum_{i=1}^{s} \sigma_i |\psi_i\rangle|\varphi_i\rangle. \tag{10.52}$$

The integer number s is called Schmidt rank *of $|\Psi\rangle\!\rangle$.*

Proof. We use the inverse of the Vec isomorphism to study the operator $\Psi : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$

$$\Psi = \mathsf{Vec}^{-1}|\Psi\rangle\!\rangle.$$

Now, by the singular value decomposition 10.25 and its corollary 10.26 we know that we can write

$$\Psi = V\Sigma U,$$

with $V : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ isometric on $\mathsf{Rng}(\Sigma)$ and $U : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ unitary. Thus we have

$$\begin{aligned}|\Psi\rangle\!\rangle &= V \otimes U^T |\Sigma\rangle\!\rangle \\ &= V \otimes U^T \sum_{i=1}^{s} \sigma_i |e_i\rangle|e_i\rangle \\ &= \sum_{i=1}^{s} \sigma_i |\psi_i\rangle|\varphi_i\rangle,\end{aligned}$$

where $|\psi_i\rangle = V|e_i\rangle$ and $|\varphi_j\rangle = U^T|e_j\rangle$. Since V is isometric on $\mathsf{Span}(\{e_i\}_{i=1}^s)$ and U^T is unitary, the sets $\{\psi_i\}_{i=1}^s$ and $\{\varphi_j\}_{j=1}^s$ are orthonormal, and they can be completed to orthonormal bases in the corresponding spaces. □

10.8 Linear maps on $\mathcal{L}(\mathcal{H})$

We remind that $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B})$ is a complex Hilbert space of dimension $d_\mathrm{A} d_\mathrm{B}$, as witnessed by the Vec isomorphism. The sesquilinear product is the Hilbert-Schmidt product $\langle\!\langle A|B \rangle\!\rangle = \mathrm{Tr}[A^\dagger B]$. We can thus define linear maps $\mathcal{L}(\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}) \to \mathcal{L}(\mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{B}))$. For simplicity, and for other reasons that will be clearer later, we will denote the set of such linear maps as $\mathrm{A} \to \mathrm{B}$. Moreover, since in $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ one has an ordering relation $A \geq B$ defined by

$$A \geq B \;\Leftrightarrow\; A - B \geq 0. \tag{10.53}$$

The space $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$ equipped with the ordering $\geq$ is an *ordered linear space*. Having a partial order, we can also define *order-preserving* linear maps.

Definition 10.40 (Linear map)**.** A map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is *linear* if for all $A, B : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ and $a, b \in \mathbb{C}$ one has

$$\mathscr{M}(aA + bB) = a\mathscr{M}(A) + b\mathscr{M}(B). \tag{10.54}$$

Being a set of linear maps on a complex vector space, also the set of maps $\mathrm{A} \to \mathrm{B}$ is a complex vector space.

Definition 10.41 (Identity map)**.** The identity map $\mathscr{I}_\mathrm{A} : \mathrm{A} \to \mathrm{A}$ is defined as

$$\mathscr{I}_\mathrm{A}(A) = A, \quad \forall A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}. \tag{10.55}$$

Since unit trace positive operators on $\mathcal{H}_\mathrm{A}$ are complete in $\mathcal{L}(\mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A})$, we have the following lemma.

Lemma 10.42. *Let $\mathscr{M}, \mathscr{N} : \mathrm{A} \to \mathrm{B}$. Then $\mathscr{M} = \mathscr{N}$ iff*

$$\mathscr{M}(\rho) = \mathscr{N}(\rho), \quad \forall\ \rho : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A},\ \rho \geq 0,\ \mathrm{Tr}[\rho] = 1. \tag{10.56}$$

Definition 10.43 (Adjoint preserving)**.** A map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is *adjoint-preserving* if

$$\mathscr{M}(A^\dagger) = \mathscr{M}(A)^\dagger, \quad , \forall A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}. \tag{10.57}$$

Lemma 10.44. *The map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is adjoint-preserving iff*

$$\mathscr{M}(A) = \mathscr{M}(A)^\dagger, \quad \forall A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A},\ A = A^\dagger. \tag{10.58}$$

Proof. Let us write $T = A + iB$. Then

$$\mathscr{M}(T) = \mathscr{M}(A) + i\mathscr{M}(B).$$

Now, if $\mathscr{M}(A) = \mathscr{M}(A)^\dagger$ for every selfadjoint A, we have

$$\begin{aligned} \mathscr{M}(T^\dagger) &= \mathscr{M}(A) - i\mathscr{M}(B) \\ &= \mathscr{M}(T)^\dagger, \end{aligned}$$

for every $T : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$. □

10.8.1 Positive maps

Order-preserving maps are called *positive*, because they map positive operators to positive operators. Here is the precise definition.

Definition 10.45 (Positive map)**.** A map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is *positive* if

$$\mathscr{M}(P) \geq 0, \quad \forall\ P \in \mathcal{P}(\mathrm{A}). \tag{10.59}$$

10.9 Maps on tensor products

The tensor product structure, that is lifted on linear operators, is also lifted to the level of linear maps on operators. In the following we will denote by $\mathscr{M} : \mathrm{AB} \to \mathrm{CD}$ a map from $\mathcal{L}(\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B})$ to $\mathcal{L}(\mathcal{H}_\mathrm{C} \otimes \mathcal{H}_\mathrm{D})$. In the set of maps $\{\mathscr{M} : \mathrm{AB} \to \mathrm{CD}\}$ there is a special class that we define here.

Definition 10.46 (Tensor-product maps)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{C}$ and $\mathscr{N} : \mathrm{B} \to \mathrm{D}$. Then $\mathscr{M} \otimes \mathscr{N} : \mathrm{AB} \to \mathrm{CD}$ is the map defined as

$$\mathscr{M} \otimes \mathscr{N}(A \otimes B) = \mathscr{M}(A) \otimes \mathscr{N}(B), \quad \forall A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A},\ B : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{B}. \tag{10.60}$$

Remark 11. Notice that since every operator $T : \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ can be expanded as in equation (10.32) $T = \sum_i A_i \otimes B_i$, tensor-product maps are well defined, indeed

$$\mathscr{M} \otimes \mathscr{N}(T) = \sum_i \mathscr{M}(A_i) \otimes \mathscr{N}(B_i). \tag{10.61}$$

The most important property of the tensor product of maps is its interplay with sequential composition.

Lemma 10.47. *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$, $\mathscr{N} : \mathrm{D} \to \mathrm{E}$, $\mathscr{M}' : \mathrm{B} \to \mathrm{C}$, $\mathscr{N}' : \mathrm{E} \to \mathrm{F}$. Then the following identity holds*

$$(\mathscr{M}' \otimes \mathscr{N}')(\mathscr{M} \otimes \mathscr{N}) = (\mathscr{M}'\mathscr{M}) \otimes (\mathscr{N}'\mathscr{N}). \tag{10.62}$$

Proof. Since tensor product operators are complete in $\mathcal{L}(\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B})$, we just need to prove that the identity holds when both members are applied to an operator of the form $A \otimes B$. In this case we have

$$\begin{aligned}
[(\mathscr{M}' \otimes \mathscr{N}')(\mathscr{M} \otimes \mathscr{N})](A \otimes B) &= (\mathscr{M}' \otimes \mathscr{N}')[\mathscr{M}(A) \otimes \mathscr{N}(B)] \\
&= [(\mathscr{M}'\mathscr{M})(A)] \otimes [\mathscr{N}'\mathscr{N}(B)] \\
&= [(\mathscr{M}'\mathscr{M}) \otimes (\mathscr{N}'\mathscr{N})](A \otimes B),
\end{aligned}$$

where in both equalities we used the defining equation (10.60). □

Corollary 10.48. *Let* $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ *and* $\mathscr{N} : \mathrm{C} \to \mathrm{D}$. *Then the following identity holds*

$$(\mathscr{M} \otimes \mathscr{I}_{\mathrm{D}})(\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}) = (\mathscr{I}_{\mathrm{B}} \otimes \mathscr{N})(\mathscr{M} \otimes \mathscr{I}_{\mathrm{C}}) = \mathscr{M} \otimes \mathscr{N}. \tag{10.63}$$

Remark 12. Notice that the trace is a special linear map from $\mathrm{A} \to \mathrm{I}$, and thus corollary 10.48 in the special case where $\mathscr{M} = \mathscr{N} = \mathrm{Tr}$ gives lemma 10.36 as a corollary.

Corollary 10.49. *Let* $\mathscr{N} : \mathrm{B} \to \mathrm{C}$. *Then for every* $T : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$ *one has*

$$\mathrm{Tr}_{\mathrm{A}}[(\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N})(T)] = \mathscr{N}(\mathrm{Tr}_{\mathrm{A}}[T]). \tag{10.64}$$

Proof. The thesis is just the statement of corollary 10.48 with $\mathscr{M} = \mathrm{Tr} : \mathrm{A} \to \mathrm{I}$. Indeed

$$(\mathrm{Tr}_{\mathrm{A}} \otimes \mathscr{I}_{\mathrm{C}})(\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}) = \mathscr{N}(\mathrm{Tr}_{\mathrm{A}} \otimes \mathscr{I}_{\mathrm{B}}). \qquad \square$$

Finally, we prove that the tensor product distributes over linear combinations.

Lemma 10.50. *Let* $\mathscr{M}, \mathscr{N} : \mathrm{A} \to \mathrm{B}$. *Then for every* $a, b \in \mathbb{C}$ *and every* $\mathcal{H}_{\mathrm{C}}$, *the following identities hold*

$$(a\mathscr{M} + b\mathscr{N}) \otimes \mathscr{I}_{\mathrm{C}} = a\mathscr{M} \otimes \mathscr{I}_{\mathrm{C}} + b\mathscr{N} \otimes \mathscr{I}_{\mathrm{C}}, \tag{10.65}$$
$$\mathscr{I}_{\mathrm{C}} \otimes (a\mathscr{M} + b\mathscr{N}) = a\mathscr{I}_{\mathrm{C}} \otimes \mathscr{M} + b\mathscr{I}_{\mathrm{C}} \otimes \mathscr{N}. \tag{10.66}$$

Proof. We prove the first identity, the proof for the second being straightforwardly analogous. It is sufficient to consider the action of the map on the l.h.s. on factorised operators $A \otimes B$. In this case

$$\begin{aligned}
[(a\mathscr{M} + b\mathscr{N}) \otimes \mathscr{I}_{\mathrm{C}}](A \otimes B) &= [(a\mathscr{M} + b\mathscr{N})(A)] \otimes B \\
&= [a\mathscr{M}(A) + b\mathscr{N}(A)] \otimes B \\
&= a(\mathscr{M} \otimes \mathscr{I}_{\mathrm{C}})(A \otimes B) + b(\mathscr{N} \otimes \mathscr{I}_{\mathrm{C}})(A \otimes B). \qquad \square
\end{aligned}$$

The desired result is obtained invoking corollary 10.48, in the form of the following corollary.

Corollary 10.51. *Let* $\mathscr{M}, \mathscr{N} : \mathrm{A} \to \mathrm{B}$ *and* $\mathscr{L} : \mathrm{C} \to \mathrm{D}$. *Then for every* $a, b \in \mathbb{C}$ *the following identity holds*

$$(a\mathscr{M} + b\mathscr{N}) \otimes \mathscr{L} = a\mathscr{M} \otimes \mathscr{L} + b\mathscr{N} \otimes \mathscr{L}.$$

10.10 Complete positivity

We now introduce a class of maps that have a crucial importance in quantum theory: completely positive maps. We need a preliminary definition.

Definition 10.52 (n-positive map)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. We say that $\mathscr{M}$ is *n-positive* if for $\mathcal{H}_\mathrm{C} \simeq \mathbb{C}^n$ the map $\mathscr{M} \otimes \mathscr{I}_\mathrm{C}$ is positive.

Definition 10.53 (Completely positive map)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. We say that $\mathscr{M}$ is *completely positive (CP)* if it is n positive for every $n \in \mathbb{N}$.

Remark 13. Notice that the notion of complete positivity is strictly stronger than simple positivity. There are indeed examples of maps that are positive but not completely positive. The typical example is the transpose $\mathscr{T}$. Indeed the *partial transpose* $\mathscr{I}_\mathrm{A} \otimes \mathscr{T}_\mathrm{B}$, defined for $T = \sum_i A_i \otimes B_i : \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{B}$ as

$$\mathscr{I}_\mathrm{A} \otimes \mathscr{T}_\mathrm{B}(T) = \sum_i A_i \otimes B_i^T, \tag{10.67}$$

is not positive (see example 11.9).

Lemma 10.54. *Let $\mathscr{M} : \mathrm{B} \to \mathrm{C}$ and $\mathscr{N} : \mathrm{A} \to \mathrm{B}$ be both CP. Then $\mathscr{M}\mathscr{N}$ is CP.*

Proof. Since it is

$$\begin{aligned}[\mathscr{M}\mathscr{N} \otimes \mathscr{I}_\mathrm{C}](P) &= (\mathscr{M} \otimes \mathscr{I}_\mathrm{C})(\mathscr{N} \otimes \mathscr{I}_\mathrm{C})(P) \\ &= (\mathscr{M} \otimes \mathscr{I}_\mathrm{C})[(\mathscr{N} \otimes \mathscr{I}_\mathrm{C})(P)],\end{aligned}$$

then clearly if $0 \leq P : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is positive then

$$[\mathscr{M}\mathscr{N} \otimes \mathscr{I}_\mathrm{C}](P) \geq 0. \qquad \square$$

Lemma 10.55. *Let $\mathscr{M}, \mathscr{N} : \mathrm{A} \to \mathrm{B}$ be CP, and $a, b \geq 0$. Then $a\mathscr{M} + b\mathscr{N}$ is CP.*

Proof. Let $0 \leq P : \mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{C}$. By lemma 10.50 one has

$$[(a\mathscr{M} + b\mathscr{N}) \otimes \mathscr{I}_\mathrm{C}](P) = a(\mathscr{M} \otimes \mathscr{I}_\mathrm{C})(P) + b(\mathscr{N} \otimes \mathscr{I}_\mathrm{C})(P).$$

Then, since $\mathscr{M}, \mathscr{N}$ are CP, $\mathscr{M} \otimes \mathscr{I}_\mathrm{C}(P) \geq 0$ and $\mathscr{N} \otimes \mathscr{I}_\mathrm{C}(P) \geq 0$, and finally

$$[(a\mathscr{M} + b\mathscr{N}) \otimes \mathscr{I}_\mathrm{C}](P) \geq 0. \qquad \square$$

Definition 10.56. Let $X : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$. We define $\mathscr{N}_X : \mathrm{A} \to \mathrm{B}$ as

$$\mathscr{N}_X(A) := XAX^\dagger. \tag{10.68}$$

Lemma 10.57. *For every $\mathscr{N}_X : \mathrm{A} \to \mathrm{B}$ and for every $\mathcal{H}_\mathrm{C}$ one has $\mathscr{N}_X \otimes \mathscr{I}_\mathrm{C} = \mathscr{N}_{X \otimes I_\mathrm{C}}$.*

Proof. It is sufficient to prove the statement for the action on factorised operators $A \otimes B : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}}$. In this case one has

$$\begin{aligned} \mathscr{N}_X \otimes \mathscr{I}_{\mathrm{C}}(A \otimes B) &= \mathscr{N}_X(A) \otimes B \\ &= XAX^\dagger \otimes B \\ &= (X \otimes I_{\mathrm{C}})(A \otimes B)(X^\dagger \otimes I_{\mathrm{C}}) \\ &= \mathscr{N}_{X \otimes I_{\mathrm{C}}}(A \otimes B). \end{aligned}$$

□

Lemma 10.58. *Every map $\mathscr{N}_X : \mathrm{A} \to \mathrm{B}$ is positive.*

Proof. Let $P \geq 0$. By corollary 10.16 one has $P = A^\dagger A$ for some $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{C}}$. Then

$$\begin{aligned} \mathscr{N}_X(P) &= XA^\dagger AX^\dagger \\ &= (AX^\dagger)^\dagger (AX^\dagger), \end{aligned}$$

which by corollary 10.16 is non-negative. □

Corollary 10.59. *Every map $\mathscr{N}_X : \mathrm{A} \to \mathrm{B}$ is CP.*

Proof. Indeed, for every $\mathcal{H}_{\mathrm{C}}$ the map $\mathscr{N}_X \otimes \mathscr{I}_{\mathrm{C}}$ is $\mathscr{N}_{X \otimes I_{\mathrm{C}}}$ by lemma 10.57, and thus by lemma 10.58 $\mathscr{N}_X \otimes \mathscr{I}_{\mathrm{C}}$ is positive. □

Corollary 10.60. *For every collection $\{\mathscr{N}_{X_i}\}_{i \in \mathrm{Y}}$ and for coefficients $a_i \geq 0$ the map $\sum_{i \in \mathrm{Y}} a_i \mathscr{N}_{X_i}$ is CP.*

Proof. This is a consequence of corollary 10.59 and lemma 10.55. □

10.11 Trace-preserving and unital maps

Definition 10.61 (Trace-preserving maps)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. We say that $\mathscr{M}$ is *trace-preserving* if for all $T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$

$$\mathrm{Tr}[\mathscr{M}(T)] = \mathrm{Tr}[T]. \tag{10.69}$$

Definition 10.62 (Trace non-increasing maps)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. We say that $\mathscr{M}$ is *trace non-increasing* if for all $0 \leq T : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$

$$\mathrm{Tr}[\mathscr{M}(T)] \leq \mathrm{Tr}[T]. \tag{10.70}$$

Definition 10.63 (Dual map)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. We define the *dual map* $\mathscr{M}^\dagger$ as the map satisfying the following identity for every $X : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ and $Y : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{B}}$

$$\mathrm{Tr}[Y\mathscr{M}(X)] = \mathrm{Tr}[\mathscr{M}^\dagger(Y)X]. \tag{10.71}$$

Lemma 10.64. *The dual map $\mathscr{M}^\dagger$ is well defined and linear.*

Proof. Let us consider the set of operators $|\varphi_i\rangle\langle\varphi_j|$, where $\{\varphi_i\}_{i=1}^{d_{\rm A}}$ is an orthonormal basis in $\mathcal{H}_{\rm A}$. Then for every $Y : \mathcal{H}_{\rm B} \to \mathcal{H}_{\rm B}$ we have

$$\begin{aligned}\mathscr{M}^\dagger(Y)_{j,i} &= \mathrm{Tr}[\mathscr{M}^\dagger(Y)|\varphi_i\rangle\langle\varphi_j|] \\ &= \mathrm{Tr}[Y\mathscr{M}(|\varphi_i\rangle\langle\varphi_j|)].\end{aligned}$$

Thus, if $Y = 0$ one has $\mathscr{M}^\dagger(Y) = 0$. Moreover the map $\mathscr{M}^\dagger$ is linear. Indeed

$$\begin{aligned}\mathrm{Tr}[\mathscr{M}^\dagger(aY_1 + bY_2)X] &= \mathrm{Tr}[(aY_1 + bY_2)\mathscr{M}(X)] \\ &= a\,\mathrm{Tr}[Y_1\mathscr{M}(X)] + b\,\mathrm{Tr}[Y_2\mathscr{M}(X)] \\ &= a\,\mathrm{Tr}[\mathscr{M}^\dagger(Y_1)X] + b\,\mathrm{Tr}[\mathscr{M}^\dagger(Y_2)X] \\ &= \mathrm{Tr}[(a\mathscr{M}^\dagger(Y_1) + b\mathscr{M}^\dagger(Y_2))X].\end{aligned}$$

Thus, $\mathscr{M}^\dagger(aY_1 + bY_2) = a\mathscr{M}^\dagger(Y_1) + b\mathscr{M}^\dagger(Y_2)$. □

The dual map is relevant for conditions of trace-preservation. Indeed, we have the following lemmas.

Definition 10.65 (Unital map)**.** We say that the map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is *unital* if $\mathscr{M}(I_{\rm A}) = I_{\rm B}$.

Lemma 10.66. *The map $\mathscr{M}$ is trace-preserving iff $\mathscr{M}^\dagger$ is unital.*

Proof. It is clear that if $I_{\rm A} = \mathscr{M}^\dagger(I_{\rm B})$ then for every X

$$\begin{aligned}\mathrm{Tr}[X] &= \mathrm{Tr}[I_{\rm A}X] \\ &= \mathrm{Tr}[\mathscr{M}^\dagger(I_{\rm B})X] \\ &= \mathrm{Tr}[I_{\rm B}\mathscr{M}(X)] \\ &= \mathrm{Tr}[\mathscr{M}(X)],\end{aligned}$$

namely $\mathscr{M}$ is trace-preserving. Viceversa, if $\mathrm{Tr}[X] = \mathrm{Tr}[\mathscr{M}(X)]$ for every X, then

$$\begin{aligned}\mathrm{Tr}[\mathscr{M}^\dagger(I_{\rm B})X] &= \mathrm{Tr}[\mathscr{M}(X)] \\ &= \mathrm{Tr}[X] \\ &= \mathrm{Tr}[I_{\rm A}X],\end{aligned}$$

namely $\mathscr{M}^\dagger(I_{\rm B}) = I_{\rm A}$ □

Remark 14. Notice that we used the property that for $A, B : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm A}$ one has $A = B$ iff $\mathrm{Tr}[AX] = \mathrm{Tr}[BX]$ for every $X : \mathcal{H}_{\rm A} \to \mathcal{H}_{\rm A}$. Actually, it is sufficient to consider the case of unit-trace positive X. Indeed, this case includes all operators of the form $X = |\varphi\rangle\langle\varphi|$ for every $\varphi \in \mathcal{H}_{\rm A}$ with $\|\varphi\| = 1$. Thus, the condition $\mathrm{Tr}[AX] = \mathrm{Tr}[BX]$ for $X = |\varphi\rangle\langle\varphi|$ implies

$$\begin{aligned}\langle\psi|A|\psi\rangle &= \|\psi\|^2\langle\varphi|A|\varphi\rangle \\ &= \|\psi\|^2\,\mathrm{Tr}[|\varphi\rangle\langle\varphi|A] \\ &= \|\psi\|^2\,\mathrm{Tr}[|\varphi\rangle\langle\varphi|B] \\ &= \|\psi\|^2\langle\varphi|B|\varphi\rangle \\ &= \langle\psi|B|\psi\rangle \quad \forall\ \psi \in \mathcal{H}_{\rm A},\end{aligned} \tag{10.72}$$

where $|\psi\rangle\langle\psi| = \|\psi\|^2|\varphi\rangle\langle\varphi|$ for $\|\varphi\| = 1$. Then, by the polarisation identity, equation 10.72 this implies $A = B$. Analogously, one can easily prove that if $\mathrm{Tr}[X\rho] \geq 0$ for every $\rho \geq 0$, then $X \geq 0$. Indeed, all the operators $\rho = |\varphi\rangle\langle\varphi|$ are positive semidefinite.

Finally, we have the following lemma

Lemma 10.67. *The map $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ is trace non-increasing iff $\mathscr{M}^\dagger(I_\mathrm{B}) \leq I_\mathrm{A}$.*

Proof. The map is trace non-increasing iff for every $P \in \mathcal{P}(\mathrm{A})$

$$\begin{aligned}\mathrm{Tr}[\mathscr{M}(P)] &= \mathrm{Tr}[I_\mathrm{B}\mathscr{M}(P)] \\ &= \mathrm{Tr}[\mathscr{M}^\dagger(I_\mathrm{B})P] \\ &\leq \mathrm{Tr}[P] \\ &= \mathrm{Tr}[I_\mathrm{A}P],\end{aligned}$$

which is equivalent to $\mathrm{Tr}[(I_\mathrm{A} - \mathscr{M}^\dagger(I_\mathrm{B}))P] \geq 0$ for all $P \geq 0$, and then to

$$I_\mathrm{A} - \mathscr{M}^\dagger(I_\mathrm{B}) \geq 0. \qquad \square$$

Chapter 11

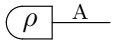
Lecture 11: Quantum Theory and the Choi isomorphism

11.1 Quantum theory

Analogously to classical information theory, quantum theory is a theory dealing with *quantum* information carrying systems, that can be thought of as quantum n-levels systems. A quantum system A is in correspondence with a complex Hilbert space $\mathcal{H}_{\mathrm{A}}$, having d_{A} different levels, but also allowing for the storage of superpositions, and for the measurement of complementary quantities. The *type* of a quantum system A is the dimension d_{A} of the corresponding space. Also in the quantum case, for the purpose of information encoding, all systems of the same type are equivalent.

11.2 States

The encoding of information in a system or an array of systems corresponds to the preparation of a special *state* of the system. For a given system A, states are represented by positive operators ρ with $\mathrm{Tr}[\rho] \leq 1$. The preparation may occur according to a probabilistic algorithm, and $\mathrm{Tr}[\rho]$ represents the preparation probability. The set of states of A is denoted by $\mathsf{St}(\mathrm{A})$. A state ρ is *deterministic* if and only if it is *normalised*, namely $\mathrm{Tr}[\rho] = 1$, and the set of deterministic states is denoted by $\mathsf{St}_1(\mathrm{A})$. In the general case, a preparation of system A corresponds to an *ensemble*, that is a collection of states $\{\rho_i\}_{i\in\mathrm{X}}$ such that $\sum_{i\in\mathrm{X}} \rho_i$ is deterministic. The ensemble represents a preparation algorithm where the occurrence of a particular classical outcome $i \in \mathrm{X}$—that can be a digit on a display, or any other classical signal from a device—heralds preparation of the specific state ρ_i in the ensemble. A state ρ of system A is represented by the following diagram



(11.1)

Composing two systems A of type d_{A} and B of type d_{B} *in parallel* allows one to encode states of a new system C whose type is $d_{\mathrm{A}}d_{\mathrm{B}}$ and $\mathcal{H}_{\mathrm{AB}} = \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$. The parallel run of preparation algorithms ρ and σ for A and B, respectively, is represented by the state $\rho \otimes \sigma$.

Notice that both the set of states $\mathsf{St}(\mathrm{A})$ and the set of deterministic states $\mathsf{St}_1(\mathrm{A})$ are convex. Let us consider in particular $\mathsf{St}_1(\mathrm{A})$. It is clear from the spectral theorem that every element $\rho \in \mathsf{St}_1(\mathrm{A})$ can be decomposed as a convex combination of elements of the form $\rho_i = |\psi_i\rangle\langle\psi_i|$, namely

$$\rho = \sum_{i=1}^{\mathsf{rank}(\rho)} p_i |\psi_i\rangle\langle\psi_i|.$$

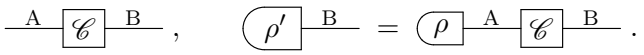
In mathematical terms, the elements $|\psi_i\rangle\langle\psi_i|$ correspond to *extremal elements* of the convex set $\mathsf{St}_1(\mathrm{A})$. We call these states *pure states* of system A. The reason is that they correspond to preparations where we have maximal control of the system. Every other preparation can be perfectly simulated via a random preparation algorithm.

What is the relation of the above formalism with the traditional representation of quantum mechanical systems? The connection is through pure states. In the traditional terminology, the vector $|\psi\rangle$ is used to denote the state $\rho_\psi = |\psi\rangle\langle\psi|$. Differently from ρ_ψ, however, $|\psi\rangle$ carries irrelevant information, e.g. any overall phase factor. This makes the correspondence $\psi \mapsto \rho_\psi$ not invertible. On the other hand, the full information about preparation of system A can be summarised in the specification of a state $\rho \in \mathsf{St}_1(\mathrm{A})$.

11.2.1 Quantum operations and channels

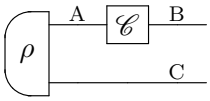
As in the classical case, also transformations have a type $\mathrm{A} \to \mathrm{B}$, that accounts for the type A of the input system and the type B of the output.

A transformation of type $\mathrm{A} \to \mathrm{B}$ maps states of system A into states of system B. The diagrammatic representation of a transformation $\mathscr{C}$ of type $\mathrm{A} \to \mathrm{B}$ is the following



Also in the quantum case, the transformation algorithm cannot behave differently depending on the probability distribution $\mathbb{P}_{\mathrm{X}}(i) := \mathrm{Tr}[\rho_i]$ of an input ensemble $\{\rho_i\}_{i\in\mathrm{X}}$. We already know from the classical case that the requirement on the transformation $\mathscr{C}$, of behaving in a fixed way independently of the probability distribution of the ensemble, is mathematically expressed through the property of *linearity*: The state ρ' of B is a linear function of the state ρ of A. Then, the transformation $\mathscr{C}$ is a linear map $\mathscr{C} : \mathrm{A} \to \mathrm{B}$.

In order to respect positivity, also when applied in parallel with the identical channel $\mathscr{I}_{\mathrm{C}}$ on an arbitrary independent system C, as in the following diagram

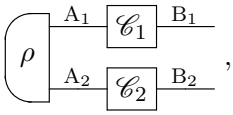


the linear transformation must preserve the positivity of joint states $\mathsf{St}(\mathrm{AC})$. Thus, a transformation $\mathscr{C} : \mathrm{A} \to \mathrm{B}$ must be a linear CP map.

Finally, in order to preserve (sub-)normalisation the map $\mathscr{C} : \mathrm{A} \to \mathrm{B}$ must be trace non-increasing. Linear CP trace non-increasing maps are called *quantum operations*, and their set will be denoted by $\mathsf{QO}(\mathrm{A} \to \mathrm{B})$. Deterministic quantum operations are called *channels*, and they are trace-preserving. The set of quantum channels is denoted by

$\mathsf{QC}(\mathrm{A} \to \mathrm{B})$. A test in quantum theory is a collection $\{\mathscr{A}_i\}_{i\in \mathrm{X}}$ of quantum operations summing to a channel. In quantum theory a test is called a *quantum instrument*, and represents a collection of possible events in a measurement, with outcome $i \in \mathrm{X}$ heralding the occurrence of the event corresponding to the quantum operation $\mathscr{A}_i$.

The parallel action of a channel (or operation) $\mathscr{C}_1$ from A_1 to B_1 and of a channel (or operation) $\mathscr{C}_2$ from A_2 to B_2, as in the following diagram



is represented by the tensor product channel (or operation)

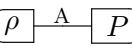
$$\mathscr{D} = \mathscr{C}_1 \otimes \mathscr{C}_2.$$

11.2.2 Effects

One can think of states $\mathsf{St}(\mathrm{A})$ as special quantum operations $\mathsf{QO}(\mathrm{I} \to \mathrm{A})$, where the system I has $d_{\mathrm{I}} = 1$. The system I is such that $\mathrm{AI} = \mathrm{IA} = \mathrm{A}$. Thanks to the system I we can define another special type of operations, $\mathsf{Eff}(\mathrm{A}) := \mathsf{QO}(\mathrm{A} \to \mathrm{I})$. These operations are called *effects*, and map states into probabilities given by the *Born rule*

$$p(P, \rho) = \mathrm{Tr}[\rho P]. \tag{11.2}$$

In circuit notation, we will write

$$p(P, \rho) = \mathrm{Tr}[\rho P] = \text{  } \tag{11.3}$$

There is clearly a unique deterministic effect P such that $\mathrm{Tr}[P\rho] = 1$ for all $\rho \in \mathsf{St}_1(\mathrm{A})$, namely I_{A}. A complete instrument in this case is thus represented by a *POVM*, i.e. a collection of positive operators $\{P_i\}$ such that $\sum_i P_i = I_{\mathrm{A}}$.

11.3 The qubit

A system type of particular relevance is the qubit, i.e. the type of systems whose dimension is $d = 2$. In this case $\mathcal{H} = \mathbb{C}^2$. By the Vec isomorphism, the space $\mathcal{L}(\mathcal{H} \to \mathcal{H})$ of linear operators on $\mathcal{H}$ is isomorphic to $\mathcal{H} \otimes \mathcal{H}$, and one can then expand such operators on any set that corresponds to an orthonormal basis in $\mathcal{H} \otimes \mathcal{H}$. A basis that is particularly suited for the expansion of operators on $\mathcal{H}$ is represented by the identity I along with the three Pauli sigma operators

$$\sigma_1 = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Indeed, defining $\sigma_0 \coloneqq I$, since, for $i = 1, 2, 3$, $\sigma_i^\dagger = \sigma_i$ with $\mathrm{Tr}[\sigma_i] = 0$ and $\sigma_j \sigma_k = \delta_{jk} I + i\varepsilon_{jkl}\sigma_l$, one has that $\{(1/\sqrt{2})|\sigma_i\rangle\rangle\}_{i=0}^3$ is an orthonormal basis in $\mathcal{H} \otimes \mathcal{H}$. Thus, for every operator T we can write

$$T = t_0 I + \mathbf{t} \cdot \boldsymbol{\sigma}. \tag{11.4}$$

Moreover, T is selfadjoint if and only if $t_0, t_1, t_2, t_3 \in \mathbb{R}$. One can easily prove that, for $t_k \in \mathbb{R}$, $k = 1, 2, 3$, the spectrum of $\mathbf{t} \cdot \boldsymbol{\sigma}$ is $\{-\|\mathbf{t}\|, \|\mathbf{t}\|\}$. As a consequence, the spectrum of a selfadjoint operator expressed as in Eq. (11.4) is $\{t_0 + \|\mathbf{t}\|, t_0 - \|\mathbf{t}\|\}$. It is now easy to conclude that a necessary and sufficient condition for T to be positive is $t_0 \geq \|\mathbf{t}\|$. Moreover, $\mathrm{Tr}[T] = 2t_0$. Thus, the set of deterministic states of a qubit system is of the form

$$T = \frac{1}{2}(I + \mathbf{n} \cdot \boldsymbol{\sigma}), \quad \|\mathbf{n}\| \leq 1. \tag{11.5}$$

In other words, the set of deterministic states is a 3-dimensional sphere, with pure states corresponding to vectors on the boundary ($\|\mathbf{n}\| = 1$), and the center represented by the maximally mixed state $I/2$. Such a representation of the set of deterministic states is known as the *Bloch sphere* representation.

11.4 Purification

The most important feature of quantum theory is a property of its states called *purification*. Indeed, purification can be used as an axiom to derive the Hilbert space structure, along with other five axioms that are shared by classical information theory. In this respect, one can say that purification is not only one of the properties of quantum theory, but it really captures its essence.

Definition 11.1 (Purification of a state)**.** Let $\rho \in \mathsf{St}(\mathrm{A})$. We say that $R = |\Psi\rangle\!\rangle\langle\!\langle\Psi| \in \mathsf{St}(\mathrm{AB})$ is a purification of ρ if ρ is the marginal of R on system A. In formula

$$\rho = \mathrm{Tr}_{\mathrm{B}}[R] \tag{11.6}$$

The property of purification is expressed by the following theorem.

Theorem 11.2 (Purification)**.** *Every state $\rho \in \mathsf{St}(\mathrm{A})$ has a purification $R \in \mathsf{St}(\mathrm{AB})$ with $d_{\mathrm{B}} = \mathsf{rank}(\rho)$. If $R' \in \mathsf{St}(\mathrm{AC})$ is another purification of ρ, then there is an isometry $V : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}}$ such that*

$$R' = (\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}_V)(R). \tag{11.7}$$

Proof. First of all, let $|\Psi\rangle\!\rangle\langle\!\langle\Psi| \in \mathsf{St}(\mathrm{AB})$ be a purification of ρ. By theorem 10.38, one has

$$\rho = \mathrm{Tr}_{\mathrm{B}}[|\Psi\rangle\!\rangle\langle\!\langle\Psi|] = \Psi\Psi^\dagger. \tag{11.8}$$

We observe that since $\mathsf{rank}(AB) \leq \min\{\mathsf{rank}(A), \mathsf{rank}(B)\}$, it must be $d_{\mathrm{B}} \geq \mathsf{rank}(\Psi) \geq \mathsf{rank}(\rho)$. Existence of a *minimal* purification—that is a purification $R \in \mathsf{St}(\mathrm{AB})$ with $d_{\mathrm{B}} = \mathsf{rank}(\rho)$—is proved by exhibiting R. For this purpose, let us diagonalise ρ as

$$\rho = \sum_{i=1}^{\mathsf{rank}(\rho)} p_i |\psi_i\rangle\langle\psi_i|.$$

By definition $p_i > 0$ and $\sum_{i=1}^{\mathsf{rank}(\rho)} p_i \leq 1$. with equality holding iff $\mathrm{Tr}[\rho] = 1$. We then define $R := |\Psi\rangle\!\rangle\langle\!\langle\Psi|$, with

$$|\Psi\rangle\!\rangle := \sum_{i=1}^{\mathsf{rank}(\rho)} \sqrt{p_i}|\psi_i\rangle|\varphi_i\rangle, \tag{11.9}$$

where $\{\varphi_i\}_{i=1}^{\mathsf{rank}(\rho)}$ is an orthonormal complete set in $\mathcal{H}_{\mathrm{B}} \simeq \mathbb{C}^{\mathsf{rank}(\rho)}$. It is easy to check that $\mathrm{Tr}_{\mathrm{B}}[R] = \rho$. Notice that by the Vec isomorphism (with respect to the bases $\{\psi_i\}_{i=1}^{d_{\mathrm{A}}}$ and $\{\varphi_i\}_{i=1}^{d_{\mathrm{B}}}$) it is

$$\Psi = \sum_{i=1}^{\mathsf{rank}(\rho)} \sqrt{p_i}|\psi_i\rangle\langle\varphi_i|,$$

and thus $\mathsf{rank}(\Psi) = \mathsf{rank}(\rho) = d_{\mathrm{B}}$. In order to prove the second statement, i.e. that there is an isometry $V : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}}$ such that $R' = (\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}_V)(R)$, we start observing that

$$|\Psi^\dagger| = \sum_{i=1}^{\mathsf{rank}(\rho)} \sqrt{p_i}|\psi_i\rangle\langle\psi_i|, \quad |\Psi| = \sum_{i=1}^{d_{\mathrm{B}}} \sqrt{p_i}|\varphi_i\rangle\langle\varphi_i|.$$

namely $|\Psi^\dagger|$ and $|\Psi|$ have the same spectrum (neglecting the eigenvalue 0), with the same degeneracy of non-null eigenvalues. Let us now suppose that $R' = |\Phi\rangle\!\rangle\langle\!\langle\Phi| \in \mathsf{St}(\mathrm{AC})$ is a purification of ρ. Then we have

$$|\Psi^\dagger|^2 = \Psi\Psi^\dagger = \rho = \mathrm{Tr}_{\mathrm{C}}[R'] = \Phi\Phi^\dagger = |\Phi^\dagger|^2.$$

Also $|\Phi^\dagger|$ has then the same spectrum as $|\Psi|$, and by the polar decomposition theorem 10.23 we can write

$$\Phi^\dagger = U|\Phi^\dagger| = \sum_{i=1}^{d_{\mathrm{B}}} \sqrt{p_i}|\eta_i\rangle\langle\psi_i|,$$

with the (co-)isometry U such that $U|\psi_i\rangle = |\eta_i\rangle$. Now, we have

$$\Phi = |\Phi^\dagger|U^\dagger = \sum_{i=1}^{d_{\mathrm{B}}} \sqrt{p_i}|\psi_i\rangle\langle\eta_i|. \tag{11.10}$$

If we define the isometry $V^* : \mathcal{H}_{\mathrm{B}} \to \mathcal{H}_{\mathrm{C}}$ such that $V^*|\varphi_i\rangle = |\eta_i\rangle$, we can write

$$|\Phi\rangle\!\rangle = |\Psi V^T\rangle\!\rangle = (I \otimes V)|\Psi\rangle\!\rangle.$$

Finally, we have

$$R' = (I \otimes V)R(I \otimes V^\dagger) = (\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}_V)(R). \qquad \square$$

Theorem 11.3 (Essential uniqueness of purification)**.** *Let $R_0, R_1 \in \mathsf{St}(\mathrm{AC})$ be two purifications of ρ. Then there is a unitary $U : \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{C}}$ such that*

$$R_1 = (\mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}_U)(R_0). \tag{11.11}$$

Proof. By the proof of theorem 11.3, we know that R_0 and R_1 must be of the form

$$R_i = |\Phi_i\rangle\!\rangle\langle\!\langle\Phi_i|, \quad , i = 0, 1,$$

with

$$\Phi_i = \sum_{j=1}^{\mathsf{rank}(\rho)} \sqrt{p_j}|\psi_j\rangle\langle\eta_j^{(i)}|.$$

One can define a unitary U^* such that $U^*|\eta_j^{(0)}\rangle = |\eta_j^{(1)}\rangle$, and then

$$|\Phi_1\rangle\!\rangle = I \otimes U|\Phi_0\rangle\!\rangle.$$

Finally, this implies that

$$R_1 = (\mathscr{I} \otimes \mathscr{N}_U)(R_0). \qquad \square$$

11.5 The Choi correspondence

We now introduce the Choi correspondence, which allows us to treat the complete positivity constraint in a very convenient way. The correspondence is actually a general correspondence between linear maps of type $\mathrm{A} \to \mathrm{B}$ (we remind that this means $\mathcal{L}(\mathcal{H}_\mathrm{A}) \to \mathcal{L}(\mathcal{H}_\mathrm{B})$) and linear operators on $\mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}$ (namely elements of $\mathcal{L}(\mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A})$).

Definition 11.4 (Choi correspondence)**.** Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ be a linear map. Its Choi representative (also called Choi state, matrix, operator) is

$$\mathsf{Ch}(\mathscr{M}) := \mathscr{M} \otimes \mathscr{I}_\mathrm{A}(|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|) \quad \in \mathcal{L}(\mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}). \tag{11.12}$$

Notice that according to the above definition one has

$$\mathsf{Ch}(\mathscr{N}_X) = |X\rangle\!\rangle\langle\!\langle X|, \tag{11.13}$$

for every X, as one can straightforwardly verify using the properties of the Vec map. The first property that we show is that the Choi correspondence is linear and invertible.

Theorem 11.5. *The correspondence* $\mathsf{Ch} : \mathscr{M} \mapsto \mathsf{Ch}(\mathscr{M})$ *is linear and invertible. The inverse map* $\mathsf{Ch}^{-1} : \mathcal{L}(\mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{A}) \to \mathcal{L}[\mathcal{L}(\mathcal{H}_\mathrm{A}) \to \mathcal{L}(\mathcal{H}_\mathrm{B})]$ *is defined through the identity*

$$\mathsf{Ch}^{-1}[R](X) = \mathrm{Tr}_\mathrm{A}[(I_\mathrm{B} \otimes X^T)R] \quad \in \mathcal{L}(\mathcal{H}_\mathrm{B}). \tag{11.14}$$

Proof. Let us first prove linearity. Consider the map $\alpha\mathscr{M} + \beta\mathscr{N}$ with $\alpha, \beta \in \mathbb{C}$ and $\mathscr{M}, \mathscr{N} : \mathrm{A} \to \mathrm{B}$. Then by definition

$$\begin{aligned}
\mathsf{Ch}(\alpha\mathscr{M} + \beta\mathscr{N}) &= (\alpha\mathscr{M} + \beta\mathscr{N}) \otimes \mathscr{I}_\mathrm{A}(|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|) \\
&= \alpha\mathscr{M} \otimes \mathscr{I}_\mathrm{A}(|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|) + \beta\mathscr{N} \otimes \mathscr{I}_\mathrm{A}(|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|) \\
&= \alpha\mathsf{Ch}(\mathscr{M}) + \beta\mathsf{Ch}(\mathscr{N}).
\end{aligned}$$

Let us now consider the expression in equation (11.14). First of all, this defines a linear correspondence, since

$$\begin{aligned}\mathsf{Ch}^{-1}[\alpha R+\beta S](X) &= \mathrm{Tr}_{\mathrm{A}}[(I_{\mathrm{B}}\otimes X^T)(\alpha R+\beta S)]\\ &= \alpha\,\mathrm{Tr}_{\mathrm{A}}[(I_{\mathrm{B}}\otimes X^T)R]+\beta\,\mathrm{Tr}_{\mathrm{A}}[(I_{\mathrm{B}}\otimes X^T)S]\\ &= \alpha\mathsf{Ch}^{-1}(R)(X)+\beta\mathsf{Ch}^{-1}(S)(X).\end{aligned}$$

Let us now calculate $\mathsf{Ch}^{-1}[\mathsf{Ch}(\mathscr{M})](X)$ explicitly.

$$\begin{aligned}\mathsf{Ch}^{-1}[\mathsf{Ch}(\mathscr{M})](X) &= \mathrm{Tr}_{\mathrm{A}}[(I_{\mathrm{B}}\otimes X^T)\mathsf{Ch}(\mathscr{M})]\\ &= \mathrm{Tr}_{\mathrm{A}}[(\mathscr{I}_{\mathrm{B}}\otimes\mathscr{L}_{X^T})\mathsf{Ch}(\mathscr{M})]\\ &= \mathrm{Tr}_{\mathrm{A}}[(\mathscr{M}\otimes\mathscr{I}_{\mathrm{A}})(\mathscr{I}_{\mathrm{A}}\otimes\mathscr{L}_{X^T})(|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|)]\\ &= \mathrm{Tr}_{\mathrm{A}}[(\mathscr{M}\otimes\mathscr{I}_{\mathrm{A}})(I\otimes X^T|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|)]\\ &= \mathscr{M}(\mathrm{Tr}_{\mathrm{A}}[|X\rangle\!\rangle\langle\!\langle I|])\\ &= \mathscr{M}(X),\end{aligned}$$

where for every $X:\mathcal{H}_{\mathrm{A}}\to\mathcal{H}_{\mathrm{A}}$ the linear map $\mathscr{L}_X$ acts as $\mathscr{L}_X(Y)\coloneqq XY$. Thus, we can conclude that $\mathsf{Ch}^{-1}[\mathsf{Ch}(\mathscr{M})](X)=\mathscr{M}(X)$ for every $X\in\mathcal{L}(\mathcal{H}_{\mathrm{A}})$, namely

$$\mathsf{Ch}^{-1}[\mathsf{Ch}(\mathscr{M})]=\mathscr{M}. \tag{11.15}$$

Moreover, by equation 11.14 if $\mathsf{Ch}^{-1}(R)=0$ one has $R=0$, which implies invertibility of the map Ch^{-1}. Thus, since by equation 11.15 one has

$$\mathsf{Ch}^{-1}\{\mathsf{Ch}[\mathsf{Ch}^{-1}(R)]\}=\mathsf{Ch}^{-1}(R),$$

we must conclude that $\mathsf{Ch}[\mathsf{Ch}^{-1}(R)]=R$. □

We can now prove the most important property of the Choi correspondence, which is given by the next theorem.

Theorem 11.6. *Let $\mathscr{M}:\mathrm{A}\to\mathrm{B}$ be a linear map. Then $\mathscr{M}$ is CP iff*

$$\mathsf{Ch}(\mathscr{M})\geq 0. \tag{11.16}$$

Proof. We first prove that equation (11.16) is necessary for $\mathscr{M}$ to be CP. Indeed, since the operator $|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|$ is non-negative definite, if $\mathscr{M}$ is CP one has

$$\mathsf{Ch}(\mathscr{M})=\mathscr{M}\otimes\mathscr{I}_{\mathrm{A}}(|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|)\geq 0.$$

On the other hand, the condition is sufficient. Indeed, let $\mathsf{Ch}(\mathscr{M})\geq 0$. Then for every $|\Psi\rangle\!\rangle\in\mathcal{H}_{\mathrm{A}}\otimes\mathcal{H}_{\mathrm{C}}$ we have

$$\begin{aligned}\mathscr{M}\otimes\mathscr{I}_{\mathrm{C}}(|\Psi\rangle\!\rangle\langle\!\langle\Psi|) &= (\mathscr{M}\otimes\mathscr{I}_{\mathrm{C}})(\mathscr{I}_{\mathrm{A}}\otimes\mathscr{N}_{\Psi^T})(|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|)\\ &= (\mathscr{I}_{\mathrm{B}}\otimes\mathscr{N}_{\Psi^T})(\mathscr{M}\otimes\mathscr{I}_{\mathrm{A}})(|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|)\\ &= (\mathscr{I}_{\mathrm{B}}\otimes\mathscr{N}_{\Psi^T})\mathsf{Ch}(\mathscr{M}).\end{aligned}$$

Finally, since $\mathscr{N}_X$ is CP for every X and $\mathsf{Ch}(\mathscr{M})$ is non-negative definite by hypothesis, we have

$$\mathscr{M} \otimes \mathscr{I}_{\mathrm{C}}(|\Psi\rangle\!\rangle\langle\!\langle\Psi|) \geq 0,$$

for any C and $|\Psi\rangle\!\rangle \in \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}}$. Finally, since every positive operator $\rho \in \mathcal{L}(\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_C)$ can be diagonalised and has positive eigenvalues, $\rho = \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{C}}} p_i |\Psi_i\rangle\!\rangle\langle\!\langle\Psi_i|$, the result can be straightforwardly extended from rank-one positive operators $|\Psi\rangle\!\rangle\langle\!\langle\Psi|$ to general positive operators $\rho \geq 0$. Thus, $\mathscr{M} \otimes \mathscr{I}_{\mathrm{C}}$ is positive for every C, namely $\mathscr{M}$ is CP. □

This result bears two important consequences, given by the following two corollaries.

Corollary 11.7. *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. Then $\mathscr{M}$ is CP iff it is d_{A}-positive.*

The second one is a celebrated theorem known as Kraus theorem, leading to the so-called *operator-sum* representation of CP maps.

Corollary 11.8 (Kraus' theorem)**.** *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$. Then $\mathscr{M}$ is CP iff there exists a set of operators $\{K_i\}_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}}$ such that*

$$\mathscr{M}(X) = \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}} K_i X K_i^{\dagger}. \tag{11.17}$$

The operators K_i are called Kraus operators *of $\mathscr{M}$, and equation 11.17 is called* Kraus form *or* operator-sum *representation of $\mathscr{M}$.*

Proof. A map $\mathscr{M}$ defined as in equation (11.17) is of the form

$$\mathscr{M} = \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}} \mathscr{N}_{K_i},$$

and since $\mathscr{N}_{K_i}$ is CP for any K_i, and $\alpha\mathscr{N}_1 + \beta\mathscr{N}_2$—with $\alpha, \beta \geq 0$ and $\mathscr{N}_i$ CP—is itself CP, we conclude that such a map $\mathscr{M}$ is CP. On the other hand, let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ be CP. By theorem 11.6 $\mathsf{Ch}(\mathscr{M}) \geq 0$, and then $\mathsf{Ch}(\mathscr{M}) \in \mathcal{L}(\mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{A}})$ is diagonalisable with non-negative eigenvalues. Then we have

$$\begin{aligned}\mathsf{Ch}(\mathscr{M}) &= \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}} q_i |H_i\rangle\!\rangle\langle\!\langle H_i| \\ &= \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}} |K_i\rangle\!\rangle\langle\!\langle K_i|,\end{aligned}$$

where $\langle\!\langle H_i|H_j\rangle\!\rangle = \delta_{i,j}$, while $K_i := \sqrt{q_i} H_i$. Thus,

$$\mathsf{Ch}(\mathscr{M}) = \sum_{i=1}^{d_{\mathrm{A}}d_{\mathrm{B}}} \mathsf{Ch}(\mathscr{N}_{K_i}),$$

and by linearity of Ch^{-1} we conclude

$$\mathscr{M} = \sum_{i=1}^{d_{\mathrm{A}} d_{\mathrm{B}}} \mathscr{N}_{K_i}. \qquad \square$$

As a consequence of theorem 11.6, one can easily prove that the transpose map, which is positive, is not completely positive.

Example 11.9. Let the transpose map $\mathscr{T} : \mathrm{A} \to \mathrm{A}$ be defined by $\mathscr{T}(X) = X^T$ for every $X \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})$. The positive map $\mathscr{T}$ is not CP. Indeed, its Choi representative is

$$\begin{aligned} \mathsf{Ch}(\mathscr{T}) &= \mathscr{T} \otimes \mathscr{I}_{\mathrm{A}}(|I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|) \\ &= \sum_{m,n=1}^{d_{\mathrm{A}}} |e_m\rangle\langle e_n| \otimes |e_n\rangle\langle e_m| = E, \end{aligned}$$

where E is the swap operator $E(|\varphi\rangle \otimes |\psi\rangle) = |\psi\rangle \otimes |\varphi\rangle$. Clearly, the swap operator E has negative eigenvalue -1, as one can immediately see applying E to an antisymmetric vector $|\Psi\rangle\!\rangle = (|\varphi\rangle \otimes |\psi\rangle - |\psi\rangle \otimes |\varphi\rangle)$. One indeed obtains

$$E|\Psi\rangle\!\rangle = -|\Psi\rangle\!\rangle.$$

This proves that $\mathsf{Ch}(\mathscr{T}) \ngeq 0$.

We can now move on to the study of the second relevant property of quantum channels, namely the property of being trace-preserving. For this purpose, it is useful to prove the following consequence of the Kraus theorem 11.8.

Corollary 11.10. *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ be CP, with Kraus form*

$$\mathscr{M}(X) = \sum_{i=1}^{d_{\mathrm{A}} d_{\mathrm{B}}} K_i X K_i^\dagger.$$

then the dual map $\mathscr{M}^\dagger$ is CP and has Kraus form

$$\mathscr{M}^\dagger(Y) = \sum_{i=1}^{d_{\mathrm{A}} d_{\mathrm{B}}} K_i^\dagger Y K_i. \tag{11.18}$$

We can now prove the following lemma.

Lemma 11.11. *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ be CP. Then $\mathscr{M}$ is trace-preserving iff*

$$\mathrm{Tr}_{\mathrm{B}}[\mathsf{Ch}(\mathscr{M})] = I_{\mathrm{A}}. \tag{11.19}$$

Proof. We proved that a necessary and sufficient condition for $\mathscr{M}$ to be TP is $\mathscr{M}^\dagger(I_{\mathrm{B}}) = I_{\mathrm{A}}$. Now, for a map $\mathscr{M}$ of the form of equation (11.17), by corollary 11.10 one has

$$\mathscr{M}^\dagger(Y) = \sum_i K_i^\dagger Y K_i.$$

This implies $\mathscr{M}^\dagger(I_\mathrm{B}) = \sum_i K_i^\dagger K_i$. In other words, one has

$$\mathscr{M}^\dagger(I_\mathrm{B}) = I_\mathrm{A} \;\Leftrightarrow\; \sum_i K_i^\dagger K_i = I_\mathrm{A},$$

and finally, considering that $\sum_i K_i^\dagger K_i = \sum_i \mathrm{Tr}_\mathrm{B}[\mathsf{Ch}(\mathscr{N}_{K_i})]^T = \mathrm{Tr}_\mathrm{B}[\mathsf{Ch}(\mathscr{M})]^T$, we conclude that

$$\mathscr{M}^\dagger(I_\mathrm{B}) = I_\mathrm{A} \;\Leftrightarrow\; \mathrm{Tr}_\mathrm{B}[\mathsf{Ch}(\mathscr{M})] = I_\mathrm{A}. \qquad \square$$

This result has as an immediate consequence a characterisation of Choi representatives of quantum operations, given by the following corollary.

Corollary 11.12. *Let $\mathscr{M} : \mathrm{A} \to \mathrm{B}$ be CP. Then $\mathscr{M} \in \mathsf{QO}(\mathrm{A} \to \mathrm{B})$ iff*

$$\mathrm{Tr}_\mathrm{B}[\mathsf{Ch}(\mathscr{M})] \leq I_\mathrm{A}. \tag{11.20}$$

Proof. As we showed in the proof of lemma 11.11, one has

$$\mathscr{M}^\dagger(I_\mathrm{B}) = \sum_i K_i^\dagger K_i.$$

Now, a CP map $\mathscr{M}$ is trace-non-increasing iff

$$\mathrm{Tr}[\mathscr{M}(\rho)] = \mathrm{Tr}[\mathscr{M}^\dagger(I_\mathrm{B})\rho] \leq \mathrm{Tr}[\rho], \quad \forall \rho \geq 0.$$

This is equivalent to

$$\mathrm{Tr}[(I_\mathrm{A} - \mathscr{M}^\dagger(I_\mathrm{B}))\rho] \geq 0 \quad \forall \rho \geq 0,$$

and this clearly implies

$$\begin{aligned} I_\mathrm{A} &\geq \mathscr{M}^\dagger(I_\mathrm{B}) \\ &= \sum_i K_i^\dagger K_i \\ &= \mathrm{Tr}_\mathrm{B}[\mathsf{Ch}(\mathscr{M})]^T. \end{aligned}$$ $\square$

Corollary 11.13. *Let $\mathscr{N}_V : \mathrm{A} \to \mathrm{B}$ be a quantum channel. Then $V : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ is an isometry.*

Proof. The map $\mathscr{N}_V$ is CP. Now, it is TP iff

$$\mathscr{N}_V^\dagger(I_\mathrm{B}) = \mathscr{N}_{V^\dagger}(I_\mathrm{B}) = V^\dagger V = I_\mathrm{A}. \qquad \square$$

11.5.1 Effects

We now consider the particular case of quantum operations $\mathsf{QO}(\mathrm{A} \to \mathrm{I})$, the effects. Since $\mathcal{H}_\mathrm{I} = \mathbb{C}$, we have that the Choi representative of an effect is a positive operator $\mathsf{Ch}(a) = P \geq 0$, with the property that

$$\mathrm{Tr}_\mathrm{I}[P] = P \leq I_\mathrm{A}.$$

Notice that application of the Choi map in the case of effects would require to define $\mathsf{Ch}(a) = P^T$, since

$$\begin{aligned}\mathrm{Tr}_\mathrm{A}[\rho^T \mathsf{Ch}(a)] &= \mathrm{Tr}[\rho^T P] \\ &= \mathrm{Tr}[\rho P^T].\end{aligned}$$

However, we will stick to the notation of the Born rule (11.2), and omit the transpose from the effect.

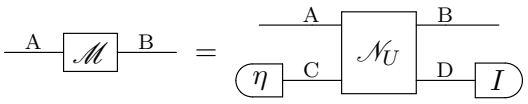
11.6 Unitary dilation of channels

We now prove one of the most important theorems of quantum information theory, that is a consequence of purification.

Theorem 11.14. *Let $\mathscr{M} \in \mathsf{QC}(\mathrm{A} \to \mathrm{B})$. Then there exist*

- *Two systems* C *and* D *with $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{C} \simeq \mathcal{H}_\mathrm{B} \otimes \mathcal{H}_\mathrm{D}$;*
- *A unitary operator $U \in \mathcal{L}(\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{C})$;*
- *A pure state $\eta \in \mathsf{St}_1(\mathrm{C})$;*

such that

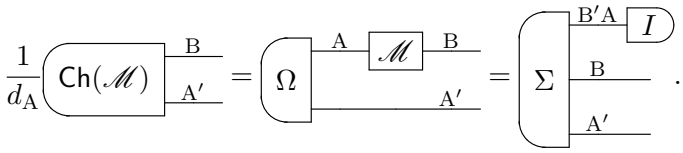


(11.21)

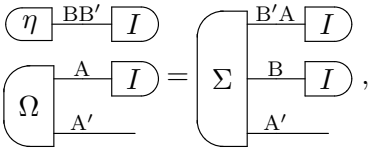
In formula

$$\mathscr{M}(\rho) = \mathrm{Tr}_\mathrm{D}[U(\rho_\mathrm{A} \otimes \eta_\mathrm{C})U^\dagger]. \tag{11.22}$$

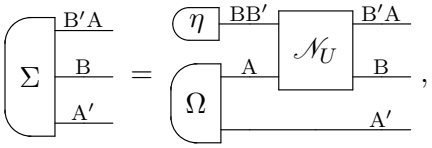
Proof. Let us denote by $\Omega \in \mathsf{St}_1(\mathrm{AA}')$ the pure state $\frac{1}{d_\mathrm{A}}|I_\mathrm{A}\rangle\!\rangle\langle\!\langle I_\mathrm{A}|$. Then we have $\frac{1}{d_\mathrm{A}}\mathsf{Ch}(\mathscr{M}) = (\mathscr{M} \otimes \mathscr{I}_{\mathrm{A}'})\Omega$. Let now $\Sigma \in \mathsf{St}_1(\mathrm{ABA'B'})$ be a purification of $\frac{1}{d_\mathrm{A}}\mathsf{Ch}(\mathscr{M})$. Then



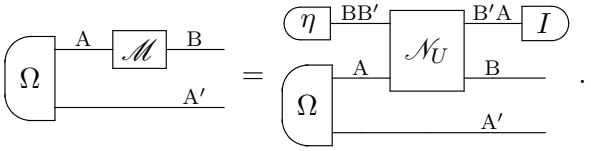
Since the channel $\mathscr{M}$ is normalised, we have $\mathrm{Tr}_{\mathrm{B}}[\mathsf{Ch}(\mathscr{M})] = I_{\mathrm{A}} = d_{\mathrm{A}}\,\mathrm{Tr}_{\mathrm{A}}[\Omega]$, thus, introducing $\mathrm{C} = \mathrm{BB}'$ and a pure state $\eta \in \mathsf{St}_1(\mathrm{C})$, we have



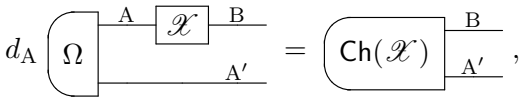
and by the essential uniqueness of purification 11.3 we obtain



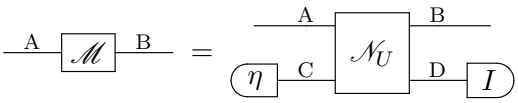
for some unitary operator $U : \mathcal{H}_{\mathrm{B}'\mathrm{AB}} \to \mathcal{H}_{\mathrm{B}'\mathrm{AB}}$. Now, applying the deterministic effect I_{D} with $\mathrm{D} = \mathrm{B}'\mathrm{A}$, we have



Finally, considering that



by the invertibility of the map Ch we conclude that



□

Chapter 12

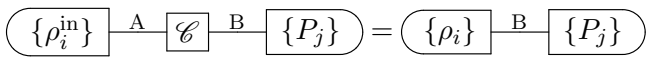
Lecture 13: von Neumann entropy

12.1 Quantum transmission of classical information

In a quantum transmission of classical information, such as in a low-signal communication along an optical fibre, the input character, say x_i, is transmitted by sending the quantum state $\rho_i^{\rm in}$. At the output of the channel, described by $\mathscr{C}$, the state is degraded to $\rho_i = \mathscr{C}(\rho_i^{\rm in})$. The receiver then performs a measurement for the purpose of discriminating the possible input states ρ_i. Such a measurement will be described in general by a POVM $\mathbf{P} = \{P_j\}$ where the element P_j corresponds to the effect giving the probability of detection of the output character y_j. Therefore, the joint probability of the channel will be given by the Born rule

$$\mathbb{P}_{\rm XY}(x_i, y_j) = \mathrm{Tr}[\rho_i P_j] = p_i \,\mathrm{Tr}[\tilde{\rho}_i P_j], \tag{12.1}$$

where $p_i := \mathrm{Tr}[\rho_i] = \mathrm{Tr}[\rho_i^{\rm in}]$ is the probability of preparing the state $\rho_i^{\rm in}$, namely the probability of sending the symbol x_i. This corresponds to the following scheme



(12.2)

In summary, the ingredients for the quantum communication scheme are:

1. an ensemble of states $\mathsf{E} = \{\rho_i\} \equiv \{p_i\tilde{\rho}_i\}$
2. a POVM $\mathbf{P} = \{P_j\}$.

where we remind that the quantum ensemble E could be equivalently described by a set of normalised positive operators $\tilde{\rho}_j$ along with their probabilities p_j, so that $\rho_j = p_j\tilde{\rho}_j$, and

$$\mathbb{P}_{\rm Y|X=x_i}[y_j|x_i] = \mathrm{Tr}[P_j\tilde{\rho}_i] = \frac{\mathrm{Tr}[P_j\rho_i]}{\mathrm{Tr}[\rho_i]}. \tag{12.3}$$

We can now regard the mutual information of a quantum communication scheme as a functional of the ensemble E and the POVM $\mathbf{P}$, and we will then write equivalently

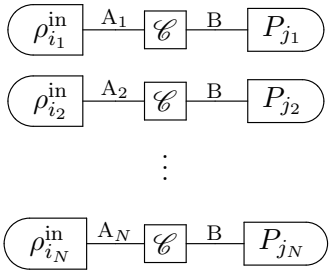
$$I(\mathsf{E}, \mathbf{P}) \equiv I(\mathrm{X} : \mathrm{Y}). \tag{12.4}$$

This scenario naturally leads to the following definition

Definition 12.1 (Accessible information of an ensemble)**.** Given an ensemble $\mathsf{E} = \{\rho_i\} \subseteq \mathsf{St}(\mathrm{A})$, the *accessible information* of E is

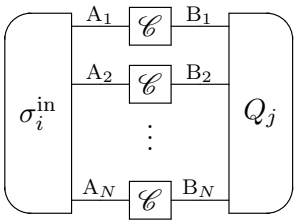
$$\mathsf{Acc}(\mathsf{E}) := \max_{\mathbf{P}} I(\mathrm{X} : \mathrm{Y}) = \max_{\mathbf{P}} I(\mathsf{E}, \mathbf{P}). \tag{12.5}$$

If we use the channel N times independently on N copies of the ensemble, as in the following scheme



, (12.6)

the accessible information clearly provides a bound on the classical capacity of $\mathscr{C}$. However, differently from the classical case, the existence of quantum entangled states (and non-local measurements) offers the possibility of achieving better performances. Indeed, the N uses of $\mathscr{C}$ can be exploited as follows



(12.7)

This possibility makes it harder to find a closed formula—what is usually called a *single letter formula*—for the classical capacity of a quantum channel. However, *in the case of separable input states* the Holevo-Schumacher-Westmoreland theorem characterises the classical capacity in terms of the so-called Holevo χ-quantity, which represents an upper bound on the accessible information of an ensemble.

The purpose of the next two chapters is precisely to prove the latter relevant bound, due to Holevo.

12.2 Von Neumann entropy

We will see that the role of the Shannon entropy here is played by the von Neumann entropy $S(\rho)$ of the quantum state ρ, defined as

$$S(\rho) := -\operatorname{Tr}[\rho \log_2 \rho]. \tag{12.8}$$

Indeed, the von Neumann entropy plays a dual role. Not only it quantifies the maximum amount of *classical* information that one can gain about a classical symbol encoded in the input state by performing the best possible measurement, but it also

quantifies the maximum amount of *quantum* information content per state of the ensemble E, i.e. the minimum number of *qubits* needed to reliably encode the quantum state.

Thus quantification of quantum information is largely concerned with von Neumann entropy, in the same way as classical information uses the Shannon entropy. And, indeed, much of the mathematical framework of quantum information theory resembles the classical one, even though the conceptual context is radically different. Clearly, the central issue in quantum information is the existence of entangled states, which bear features with no classical analogue.

Remark 15. Since every state $\rho \in \mathsf{St}_1(\mathrm{A})$ is positive by definition, the spectral theorem holds, and one can write

$$\rho = \sum_k p_k |\varphi_k\rangle\langle\varphi_k|,$$

and for every function $f : \mathbb{R} \to \mathbb{R}$ one has

$$f(\rho) := \sum_k f(p_k) |\varphi_k\rangle\langle\varphi_k|.$$

In particular, the function f defining the von Neumann entropy is

$$f(x) := \begin{cases} -x \log_2 x & x > 0, \\ 0 & x = 0. \end{cases}$$

Thus, if $\{p_k\}_{k=1}^{d_\mathrm{A}}$ is the spectrum of ρ, one has $p_k \geq 0$ and $\sum_k p_k = \mathrm{Tr}[\rho] = 1$, and then

$$S(\rho) = H(\mathbf{p}).$$

This allows us to observe that *the von Neumann entropy only depends on the spectrum of ρ, and every map $\mathscr{M}$ that does not change the spectrum leaves the von Neumann entropy invariant, $S(\rho) = S(\mathscr{M}(\rho))$.* In particular, we have

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|, \;\Rightarrow\; \rho^T = \sum_i p_i |\psi_i^*\rangle\langle\psi_i^*|,$$

and since $\langle\psi_i^*|\psi_j^*\rangle = \langle\psi_j|\psi_i\rangle = \delta_{i,j}$, the transpose leaves the spectrum of ρ invariant, then

$$S(\rho^T) = S(\rho). \tag{12.9}$$

Remark 16. As regards notation, we will often write $S(\mathrm{A})$ instead of $S(\rho)$ when system A is in state $\rho \in \mathsf{St}_1(\mathrm{A})$. In particular, when dealing with bipartite systems AB we will write $\rho_{\mathrm{AB}} \in \mathsf{St}_1(\mathrm{AB})$, and

$$\rho_\mathrm{A} := \mathrm{Tr}_\mathrm{B}[\rho_{\mathrm{AB}}], \quad \rho_\mathrm{B} := \mathrm{Tr}_\mathrm{A}[\rho_{\mathrm{AB}}],$$

thus

$$S(\mathrm{AB}) = S(\rho_{\mathrm{AB}}), \quad S(\mathrm{A}) = S(\rho_\mathrm{A}), \quad S(\mathrm{B}) = S(\rho_\mathrm{B}).$$

Remark 17. It is very natural to describe encoding of classical information on quantum states. In particular, given a classical random variable of type n, one has $\mathsf{St}_1(\mathrm{X}) = \{(p_1, p_2, \ldots, p_n) | p_i \geq 0, \sum_i p_i = 1\}$. Given a quantum system A of type n, we can fix an orthonormal basis $\{\varphi_i\}_{i=1}^n$, and given a state $(p_1, p_2, \ldots p_n)$ one can encode it in the quantum state $\rho \in \mathsf{St}_1(\mathrm{A})$ as follows

$$(p_1, p_2, \ldots p_n) \mapsto \rho = \sum_{i=1}^{n} p_i |\varphi_i\rangle\langle\varphi_i|.$$

In particular, for a bit X and a qubit A one has

$$\mathsf{St}_1(\mathrm{X}) = \{(p, 1-p) | 0 \leq p \leq 1\}, \quad (p, 1-p) \mapsto p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1| \in \mathsf{St}_1(\mathrm{A}).$$

However, the set of states of A contains also projections on arbitrary superpositions of $|0\rangle$ and $|1\rangle$—and mixtures thereof—which are not simply encodings of classical sources under the above encoding map. Conversely, one might in general encode classical information into a set of states that are not mutually orthogonal, and not even pure.

We now define the analogues of relative and conditional entropy, and mutual information.

Definition 12.2 (Quantum relative entropy)**.** We define the *quantum relative entropy* of ρ and σ in $\mathsf{St}(\mathrm{A})$ as follows

$$S(\rho\|\sigma) := \begin{cases} \mathrm{Tr}[\rho \log_2 \rho] - \mathrm{Tr}[\rho \log_2 \sigma] & \mathsf{Ker}(\sigma) \subseteq \mathsf{Ker}(\rho) \\ +\infty & \text{otherwise.} \end{cases} \tag{12.10}$$

Similarly to the Kullback-Leibler divergence, also the quantum relative entropy is a useful proving tool.

Definition 12.3 (Quantum conditional entropy)**.** We define the *quantum conditional entropy* of system A conditional on system B, with AB in state $\rho \in \mathsf{St}_1(\mathrm{AB})$, as follows

$$S(\mathrm{A}|\mathrm{B}) := S(\mathrm{AB}) - S(\mathrm{B}). \tag{12.11}$$

Definition 12.4 (Quantum mutual information)**.** We define the *quantum mutual information* of systems A and B, with AB in state $\rho \in \mathsf{St}_1(\mathrm{AB})$, as follows

$$\begin{aligned} S(\mathrm{A}:\mathrm{B}) :=& S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{AB}) \\ =& S(\mathrm{A}) - S(\mathrm{A}|\mathrm{B}) \\ =& S(\mathrm{B}) - S(\mathrm{B}|\mathrm{A}). \end{aligned} \tag{12.12}$$

Notice that, differently from the classical case, the joint von Neumann entropy $S(\mathrm{AB})$ is not necessarily bounded from below by the two marginals. Indeed, for a pure maximally entangled state $\rho_{\mathrm{AB}} = \frac{1}{d_{\mathrm{A}}} |I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|$ one has $\rho_{\mathrm{A}} = I_{\mathrm{A}}/d_{\mathrm{A}}$ and $\rho_{\mathrm{B}} = I_{\mathrm{B}}/d_{\mathrm{B}}$. Then, $S(\mathrm{AB}) = 0$ and $S(\mathrm{A}) = \log d_{\mathrm{A}}$, hence $S(\mathrm{B}|\mathrm{A}) < 0$. In particular, if ρ_{AB} is pure, then $S(\mathrm{B}|\mathrm{A})$ is negative if and only if ρ_{AB} is entangled.

For the special case of diagonal factorised states we have

$$\rho_{\mathrm{AB}} = \sum_{(x_i,y_j)\in \mathsf{Rng}(\mathrm{X})\times\mathsf{Rng}(\mathrm{Y})} p_{ij}^{\mathrm{AB}} |i\rangle\langle i|_{\mathrm{A}} \otimes |j\rangle\langle j|_{\mathrm{B}},$$

and

$$\rho_{\mathrm{A}} = \sum_{x_i\in\mathsf{Rng}(\mathrm{X})} p_i^{\mathrm{A}} |i\rangle\langle i|_{\mathrm{A}}, \quad p_i^{\mathrm{A}} = \sum_{y_j\in\mathsf{Rng}(\mathrm{Y})} p_{ij}^{\mathrm{AB}},$$
$$\rho_{\mathrm{B}} = \sum_{y_j\in\mathsf{Rng}(\mathrm{Y})} p_j^{\mathrm{B}} |j\rangle\langle j|_{\mathrm{B}}, \quad p_j^{\mathrm{B}} = \sum_{x_i\in\mathsf{Rng}(\mathrm{X})} p_{ij}^{\mathrm{AB}}.$$

In this case

$$S(\mathrm{A}:\mathrm{B}) = I(\mathrm{X}:\mathrm{Y}), \quad \mathbb{P}_{\mathrm{X},\mathrm{Y}}[x_i, y_j] = p_{ij}$$

In consideration of the relevant role played by the von Neumann entropy, we now devote a subsection to analyse its properties.

12.2.1 Klein's inequality

We first prove the main property of quantum relative entropy.

Theorem 12.5 (Klein's inequality)**.** *Let the states $\rho, \sigma \in \mathsf{St}(\mathrm{A})$ have the same probability* $\mathrm{Tr}[\rho] = \mathrm{Tr}[\sigma]$*. The quantum relative entropy of ρ with respect to σ is non-negative*

$$S(\rho\|\sigma) \geq 0, \qquad \textit{with equality iff } \rho = \sigma. \tag{12.13}$$

Proof. The thesis is straightforward if $\mathsf{Ker}(\sigma) \not\subseteq \mathsf{Ker}(\rho)$. Let us focus then on the complementary case. One can diagonalise both ρ and σ, as

$$\rho = \sum_i p_i P_i, \quad \sigma = \sum_j q_j Q_j,$$

where P_i and Q_j are the projections on the i-th and j-th eigenspace of ρ and σ, respectively, and $p_i \neq p_{i'}$ for $i \neq i'$, as well as $q_j \neq q_{j'}$ for $j \neq j'$. Let $\mathrm{Tr}[P_i] = n_i$ and $\mathrm{Tr}[Q_j] = m_j$. We then write the relative entropy as follows

$$\begin{aligned} S(\rho\|\sigma) &= \sum_i n_i p_i \log_2 p_i - \sum_{i,j} p_i \log_2 q_j \,\mathrm{Tr}[P_i Q_j] \\ &= \sum_i n_i p_i \left(\log_2 p_i - \sum_j S_{ji} \log_2 q_j \right), \end{aligned}$$

where the matrix with elements $S_{ji} := \mathrm{Tr}[P_i Q_j]/n_i$ is stochastic, namely $S_{ji} \geq 0$ for every i, j, and $\sum_j S_{ji} = \mathrm{Tr}[P_i]/n_i = 1$, for every i. Since the logarithm is a strictly concave function, by Jensen's inequality we then have

$$\sum_{j=1}^{d_{\mathrm{A}}} S_{ji} \log_2 q_j \leq \log_2 \left(\sum_{j=1}^{d_{\mathrm{A}}} S_{ji} q_j \right) = \log_2 r_i, \qquad r_i := \sum_{j=1}^{d_{\mathrm{A}}} S_{ji} q_j. \tag{12.14}$$

Notice that

$$\sum_i n_i r_i = \sum_{i,j} n_i S_{ji} q_j = \sum_{i,j} \mathrm{Tr}[P_i Q_j] q_j = \mathrm{Tr}[\sigma] = \mathrm{Tr}[\rho],$$

thus the vector $\mathbf{r}$ where r_i has the same multiplicity n_i as p_i, has the same sum as $\mathbf{p}$. For deterministic states ($\mathrm{Tr}[\rho] = \mathrm{Tr}[\sigma] = 1$), we can then write

$$S(\rho\|\sigma) \geq \sum_i n_i p_i \log_2 \frac{p_i}{r_i} \equiv H(\mathbf{p}\|\mathbf{r}),$$

while for sub-normalised states one can define $\mathbf{p}' = \mathbf{p}/\mathrm{Tr}[\rho]$ and $\mathbf{r}' = \mathbf{r}/\mathrm{Tr}[\sigma]$, and

$$S(\rho\|\sigma) \geq \frac{1}{\mathrm{Tr}[\rho]} \sum_i n_i p_i \log_2 \frac{p_i}{r_i} \equiv \frac{1}{\mathrm{Tr}[\rho]} H(\mathbf{p}'\|\mathbf{r}'),$$

From the Gibbs' inequality for $H(\mathbf{p}\|\mathbf{r})$ we thus get

$$S(\rho\|\sigma) \geq 0.$$

Equality holds upon two conditions. The first one is saturation of the bound in equation (12.14), which is Jensen's inequality for the $\log_2$ function, and by strict concavity of $\log_2$ one must have $S_{ji} = \delta_{j,\varphi(i)}$ for every i. By making this condition explicit, we have

$$\mathrm{Tr}[P_i Q_{\varphi(i)}] = n_i, \quad \mathrm{Tr}[P_i Q_j] = 0, \ j \neq \varphi(i).$$

Then, since $\mathrm{Tr}[P_i Q_j] = \mathrm{Tr}[P_i Q_j P_i]$ is the trace of the non-negative definite operator $P_i Q_j P_i$, the trace is null iff $P_i Q_j P_i = 0$. Now, since $P_i Q_j P_i = (Q_j P_i)^\dagger (Q_j P_i)$, one has $Q_j P_i = 0$ for $j \neq \varphi(i)$. Also, $P_i Q_j = (Q_j P_i)^\dagger = 0$ for $j \neq \varphi(i)$. Then

$$\begin{aligned} P_i &= \sum_j Q_j P_i = Q_{\varphi(i)} P_i \\ &= P_i^\dagger = P_i Q_{\varphi(i)}. \end{aligned}$$

Thus, $[P_i, Q_j] = 0$ for every i, j and $P_i Q_{\varphi(i)} = Q_{\varphi(i)} P_i = P_i$. Moreover, since $S_{ji} = \delta_{j,\varphi(i)}$, one has $r_i = q_{\varphi(i)}$. Suppose now that for some j there is no i such that $j = \varphi(i)$. This implies that

$$Q_j = Q_j \sum_i P_i = \sum_i (Q_j P_i) = 0,$$

namely $Q_j = 0$. This implies that there cannot be any such j, and φ is thus surjective on the set of values of j, namely the spectrum of σ is the collection of the r_i's.

The second condition is saturation of Gibbs' inequality for $H(\mathbf{p}\|\mathbf{r})$ [or $H(\mathbf{p}'\|\mathbf{r}')$], namely $\mathbf{p} = \mathbf{r}$. Then it must be $p_i = r_i = q_{\varphi(i)}$, and thus, ρ and σ must have the same spectrum. Moreover, for $i \neq i'$ one has $q_{\varphi(i)} = r_i = p_i \neq p_{i'} = r_{i'} = q_{\varphi(i')}$, which implies $\varphi(i) \neq \varphi(i')$, thus φ is an injective function. Then, one can write

$$\begin{aligned} Q_{\varphi(j)} &= \sum_i Q_{\varphi(j)} P_i = Q_{\varphi(j)} P_j \\ &= \sum_l Q_l P_j = P_j. \end{aligned}$$

In conclusion, ρ and σ have the same eigenspaces and the same spectrum, which implies

$$S(\rho\|\sigma) = 0 \;\Leftrightarrow\; \rho = \sigma. \qquad \square$$

12.2.2 Preliminary lemmas

We now need two lemmas that we use in the following.

Lemma 12.6. *Let* $0 \leq X, Y \in \mathcal{L}(\mathcal{H}_\mathrm{A})$ *with* $[X, Y] = 0$. *Then*

$$\log_2 XY = \log_2 X + \log_2 Y. \tag{12.15}$$

Proof. The operators X and Y are simultaneously diagonalisable, with

$$X = \sum_{i=1}^{d_\mathrm{A}} x_i |\psi_i\rangle\langle\psi_i|, \quad Y = \sum_{i=1}^{d_\mathrm{A}} y_i |\psi_i\rangle\langle\psi_i|.$$

Then $XY = \sum_{i=1}^{d_\mathrm{A}} x_i y_j |\psi_i\rangle\langle\psi_i|$, and

$$\begin{aligned}
\log_2 XY &= \sum_{i=1}^{d_\mathrm{A}} \log_2(x_i y_i) |\psi_i\rangle\langle\psi_i| \\
&= \sum_{i=1}^{d_\mathrm{A}} (\log_2 x_i + \log_2 y_i) |\psi_i\rangle\langle\psi_i| \\
&= \log_2 X + \log_2 Y. \qquad \square
\end{aligned}$$

Lemma 12.7. *Let* $0 \leq X \in \mathcal{L}(\mathcal{H}_\mathrm{A})$. *Then one has*

$$\log_2(X \otimes I_\mathrm{B}) = (\log_2 X) \otimes I_\mathrm{B} \tag{12.16}$$

Proof. One has

$$X = \sum_{i=1}^{d_\mathrm{A}} x_i |\psi_i\rangle\langle\psi_i| \;\Rightarrow X \otimes I_\mathrm{B} = \sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} x_i |\psi_i\rangle\langle\psi_i| \otimes |\varphi_j\rangle\langle\varphi_j|.$$

Then

$$\begin{aligned}
\log_2(X \otimes I_\mathrm{B}) &= \log_2 \left(\sum_{i=1}^{d_\mathrm{A}} \sum_{j=1}^{d_\mathrm{B}} x_i |\psi_i\rangle\langle\psi_i| \otimes |\varphi_j\rangle\langle\varphi_j| \right) \\
&= \sum_{i=1}^{d_\mathrm{A}} \log_2(x_i) |\psi_i\rangle\langle\psi_i| \otimes \sum_{j=1}^{d_\mathrm{B}} |\varphi_j\rangle\langle\varphi_j| \\
&= (\log_2 X) \otimes I_\mathrm{B}. \qquad \square
\end{aligned}$$

12.3 Mathematical properties of the von Neumann entropy

1. **Entropy of pure states**. The entropy of a pure state $\rho = |\psi\rangle\langle\psi|$ is $S(\rho) = 0$. We notice that in the definition of the von Neumann entropy we follow the same rule as for the Shannon entropy, namely we put $0 \log 0 := 0$. Indeed, the von Neumann entropy is equal to the Shannon entropy $S(\rho) = H(\mathbf{p})$ of the eigenvalues $\{p_i\}_{i=1}^{d_\mathrm{A}}$ of the state $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, and for a pure state the eigenvalues are all zero, apart from a single one which is unit.

2. **Maximum value**. The entropy is upper bounded as $S(\rho) \leq \log_2 d_\mathrm{A}$, where $\rho \in \mathsf{St}_1(\mathrm{A})$. This clearly follows from the fact that $S(\rho) = H(\mathbf{p})$, where $\{p_i\}_{i=1}^{d_\mathrm{A}}$ is the spectrum of ρ, and the maximum value of the Shannon entropy is achieved when the probability distribution is uniform: $p_i = 1/d_\mathrm{A}$, i.e. for $\rho = I_\mathrm{A}/d_\mathrm{A}$.

3. **Isometric invariance**. The entropy is left unchanged by an isometric transformation, i.e.

$$S(V\rho V^\dagger) = S(\rho), \tag{12.17}$$

with $V : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{B}$ is an isometry, namely $V^\dagger V = I_\mathrm{A}$. This is just a consequence of the fact that $S(\rho)$ depends only on the spectrum of ρ, and the isometric transformation preserves the diagonal form of the state:

$$\rho = \sum_{i=1}^{d_\mathrm{A}} p_i |\psi_i\rangle\langle\psi_i| \;\Rightarrow\; V\rho V^\dagger = \sum_{i=1}^{d_\mathrm{A}} p_i V|\psi_i\rangle\langle\psi_i|V^\dagger, \tag{12.18}$$

and by the properties of isometries $\langle\psi_i|V^\dagger V|\psi_j\rangle = \langle\psi_i|\psi_j\rangle = \delta_{i,j}$.

4. **Marginal entropies**. For a composite system AB in a pure state $|\Psi\rangle\!\rangle\langle\!\langle\Psi| \in \mathsf{St}_1(\mathrm{AB})$, the marginal entropies are equal, i.e. $S(\mathrm{A}) = S(\mathrm{B})$.

Proof. Let us fix $d_\mathrm{A} \leq d_\mathrm{B}$, which is clearly not restrictive, as the role of A and B can be interchanged. For a pure joint state $|\Psi\rangle\!\rangle$ with $\Psi = V|\Psi|$ we have

$$\begin{aligned}
\rho_\mathrm{A} &= \mathrm{Tr}_\mathrm{B}[|\Psi\rangle\!\rangle\langle\!\langle\Psi|] = V|\Psi|^2 V^\dagger, \\
\rho_\mathrm{B} &= \mathrm{Tr}_\mathrm{A}[|\Psi\rangle\!\rangle\langle\!\langle\Psi|] = |\Psi|^T V^T V^* |\Psi|^T = (|\Psi|^2)^T
\end{aligned}$$

namely $\rho_\mathrm{A} = V\rho_\mathrm{B}^T V^\dagger$. Since the transpose as well as isometries do not change the spectrum, by equations (12.9) and (12.17), the two states have the same von Neumann entropy. □

5. **Entropy of statistically independent systems**.

$$S(\rho \otimes \sigma) = S(\rho) + S(\sigma). \tag{12.19}$$

Proof. The statement is an immediate consequence of the identity

$$\rho \otimes \sigma = (\rho \otimes I_{\mathrm{B}})(I_{\mathrm{A}} \otimes \sigma),$$

which by lemmas 12.6 and 12.7 implies

$$\log_2(\rho \otimes \sigma) = \log_2 \rho \otimes I_{\mathrm{B}} + I_{\mathrm{A}} \otimes \log_2 \sigma. \tag{12.20}$$

More explicitly, the latter follows from $\log_2(XY) = \log_2(X) + \log_2(Y)$ for $[X, Y] = 0$, and from $\log_2(X \otimes I_{\mathrm{B}}) = \log_2(X) \otimes I_{\mathrm{B}}$ [and $\log_2(I_{\mathrm{A}} \otimes Y) = I_{\mathrm{A}} \otimes \log_2(Y)$]. □

6. **Subadditivity**. For any state $\rho_{\mathrm{AB}} \in \mathsf{St}_1(\mathrm{AB})$ one has

$$S(\mathrm{AB}) \leq S(\mathrm{A}) + S(\mathrm{B}), \qquad \text{with equality iff A and B are factorised.} \tag{12.21}$$

Proof. Using the Klein's inequality for $S(\rho_{\mathrm{AB}} \| \rho_{\mathrm{A}} \otimes \rho_{\mathrm{B}})$ and identity (12.20), one has

$$\begin{aligned} S(\rho_{\mathrm{AB}}) \leq & -\operatorname{Tr}[\rho_{\mathrm{AB}} \log(\rho_{\mathrm{A}} \otimes \rho_{\mathrm{B}})] \qquad (\text{with equality iff } \rho_{\mathrm{AB}} = \rho_{\mathrm{A}} \otimes \rho_{\mathrm{B}}) \\ = & -\operatorname{Tr}[\rho_{\mathrm{AB}}(\log \rho_{\mathrm{A}} \otimes I_{\mathrm{B}})] - \operatorname{Tr}[\rho_{\mathrm{AB}}(I_{\mathrm{A}} \otimes \log \rho_{\mathrm{B}})] \\ = & S(\mathrm{A}) + S(\mathrm{B}). \end{aligned}$$ □

7. **States with orthogonal supports**. For a set of states $\{\rho_i\}$ having orthogonal supports, i.e. $\rho_i \rho_j = 0$ for $i \neq j$

$$S\left(\sum_i p_i \tilde{\rho}_i\right) = H(\mathbf{p}) + \sum_i p_i S(\tilde{\rho}_i). \tag{12.22}$$

Proof. Since the states have orthogonal supports, we can write the joint spectral decomposition of the normalised states $\tilde{\rho}_i$ as

$$\tilde{\rho}_i |e_j^{(i)}\rangle = \lambda_j^{(i)} |e_j^{(i)}\rangle.$$

Consequently, we have

$$\sum_i p_i \tilde{\rho}_i |e_j^{(k)}\rangle = p_k \lambda_j^{(k)} |e_j^{(k)}\rangle, \quad \sum_j \lambda_j^{(k)} = 1 \, \forall k.$$

We therefore have

$$\begin{aligned} S\left(\sum_i p_i \tilde{\rho}_i\right) = & -\sum_{jk} p_k \lambda_j^{(k)} \log p_k \lambda_j^{(k)} \\ = & -\sum_k p_k \log p_k - \sum_k p_k \sum_j \lambda_j^{(k)} \log \lambda_j^{(k)} \\ = & H(\mathbf{p}) + \sum_i p_i S(\tilde{\rho}_i). \end{aligned}$$ □

As a corollary of this case we have that the quantum von Neumann entropy of a joint classical-quantum system AX in the *quantum-classical state*

$$\rho_{\mathrm{AX}} = \sum_j p_j \rho_{\mathrm{A}}^j \otimes |j\rangle\langle j|_{\mathrm{X}},$$

associated with the ensemble of states $\mathsf{E}_{\mathrm{A}} = \{p_j \rho_{\mathrm{A}}^j\}$, is given by

$$S(\mathrm{AX}) = \sum_j p_j S(\rho_{\mathrm{A}}^j) + H(\mathbf{p}), \tag{12.23}$$

and the quantum-classical mutual information is rewritten as

$$S(\mathrm{A}:\mathrm{X}) = S\left(\sum_j p_j \rho_{\mathrm{A}}^j\right) - \sum_j p_j S(\rho_{\mathrm{A}}^j). \tag{12.24}$$

8. **Concavity**. For $\mathbf{p} := (p_1, p_2, \dots p_k)$ a general probability vector and $\tilde{\rho}_1, \tilde{\rho}_2, \dots \tilde{\rho}_k \in \mathsf{St}_1(\mathrm{A})$, one has

$$S\left(\sum_{i=1}^k p_i \tilde{\rho}_i\right) \geq \sum_{i=1}^k p_i S(\tilde{\rho}_i). \tag{12.25}$$

Proof. Introduce an auxiliary system B such that the state of AB is

$$\rho_{\mathrm{AB}} = \sum_{n=1}^k p_n \tilde{\rho}_n \otimes |n\rangle\langle n|.$$

Clearly we have $S(\mathrm{A}) = S\left(\sum_{n=1}^k p_n \tilde{\rho}_n\right)$, $S(\mathrm{B}) = H(\mathbf{p})$, hence, using subadditivity and identity (12.23), we obtain

$$H(\mathbf{p}) + \sum_{i=1}^k p_i S(\tilde{\rho}_i) = S(\mathrm{AB}) \leq S(\mathrm{A}) + S(\mathrm{B}) = S\left(\sum_{n=1}^k p_n \tilde{\rho}_n\right) + H(\mathbf{p}),$$

namely the statement. □

9. **Non-selective projective measurements**. Let $\mathscr{P} : \mathrm{A} \to \mathrm{A}$ be a projective measurement, i.e.

$$\mathscr{P}(\rho) = \sum_{i=1}^k P_i \rho P_i, \qquad P_i P_j = \delta_{ij} P_i, \quad \sum_{i=1}^k P_i = I_{\mathrm{A}}. \tag{12.26}$$

Then one has

$$S(\mathscr{P}(\rho)) \geq S(\rho), \tag{12.27}$$

with equality holding iff $\mathscr{P}(\rho) = \rho$.

Proof. Notice that $\mathscr{P}$ is an idempotent projective map, i.e. $\mathscr{P}^2 = \mathscr{P}$, and that $\mathscr{P}(\rho)$ commutes with P_i for every i. Now, reminding that $\sum_{i=1}^k P_i = I_{\rm A}$, one has

$$\begin{aligned}\mathrm{Tr}[\rho \log_2 \mathscr{P}(\rho)] &= \mathrm{Tr}[\sum_{i=1}^k P_i \rho \log_2 \mathscr{P}(\rho)] \\ &= \mathrm{Tr}[\sum_{i=1}^k P_i \rho \log_2 \mathscr{P}(\rho) P_i] = \mathrm{Tr}[\sum_i P_i \rho P_i \log \mathscr{P}(\rho)]\end{aligned}$$

namely

$$\mathrm{Tr}[\rho \log \mathscr{P}(\rho)] = \mathrm{Tr}[\mathscr{P}(\rho) \log \mathscr{P}(\rho)].$$

This implies that

$$\begin{aligned}S(\rho \| \mathscr{P}(\rho)) &= -S(\rho) - \mathrm{Tr}[\rho \log \mathscr{P}(\rho)] \\ &= -S(\rho) + S(\mathscr{P}(\rho)),\end{aligned}$$

and the statement then follows from Klein's inequality. Consequently, equality holds iff $\mathscr{P}(\rho) = \rho$. □

10. **Upper bound**. The von Neumann entropy is bounded as

$$S\left(\sum_{n=1}^k p_n \tilde{\rho}_n\right) \leq \sum_{n=1}^k p_n S(\tilde{\rho}_n) + H(\mathbf{p}), \quad \text{(with equality iff } \tilde{\rho}_i \tilde{\rho}_j = 0 \text{ for } i \neq j\text{)}. \tag{12.28}$$

Proof. Let us first consider the pure case $\tilde{\rho}_n = |\psi_n\rangle\langle\psi_n| \in \mathsf{St}_1(\mathrm{A})$. Take the following purification of $\rho := \sum_{n=1}^k p_n \rho_n$ on $\mathcal{H}_{\rm A} \otimes \mathcal{H}_{\rm B}$

$$|\Psi\rangle\!\rangle = \sum_{n=1}^k \sqrt{p_n} |\psi_n\rangle \otimes |n\rangle.$$

Since the system AB is in a pure state, we have

$$S(\rho) = S(\mathrm{A}) = S(\mathrm{B}).$$

Suppose now we perform a projective measurement over B in the orthonormal basis $\{|n\rangle\}$. After the measurement we have

$$\rho_{\mathrm{AB}'} = \sum_n p_n |\psi_n\rangle\langle\psi_n| \otimes |n\rangle\langle n|, \quad \rho_{\mathrm{B}'} = \sum_n p_n |n\rangle\langle n|.$$

Now, since non-selective projective measurements never decrease the entropy [see identity (12.27)], we have $S(\mathrm{AB}') \geq S(\mathrm{AB}) = 0$, and at the same time $S(\mathrm{B}') \geq S(\mathrm{B}) = S(\rho)$. Therefore,

$$S(\rho) \leq S(\mathrm{B}') = H(\mathbf{p}), \quad \text{with equality iff } \{|\psi_n\rangle\} \text{ are orthogonal},$$

and since $S(\tilde{\rho}_n) = 0$, we have trivially

$$S(\rho) \leq \sum_{n=1}^{k} p_n S(\tilde{\rho}_n) + H(\mathbf{p}).$$

The case of mixed ρ_n is proved as follows. Consider the orthonormal decomposition

$$\tilde{\rho}_n = \sum_{j=1}^{d_{\mathrm{A}}} p_j^{(n)} |e_j^{(n)}\rangle\langle e_j^{(n)}|.$$

This provides a pure ensemble $\{p_n p_j^{(n)} |e_j^{(n)}\rangle\langle e_j^{(n)}|\}$ for ρ, since

$$\rho = \sum_{nj} p_n p_j^{(n)} |e_j^{(n)}\rangle\langle e_j^{(n)}|,$$

and using our theorem for the pure case, we have

$$\begin{aligned} S(\rho) \leq H(\{p_n p_j^{(n)}\}) &= -\sum_{jn} p_n p_j^{(n)} \log p_n p_j^{(n)} \\ &= -\sum_{n} p_n \log p_n - \sum_{n} p_n \sum_{j} p_j^{(n)} \log p_j^{(n)} \\ &= H(\mathbf{p}) + \sum_{n} p_n S(\tilde{\rho}_n), \end{aligned}$$

which is the statement. Notice that the bound is achieved iff all vectors $|e_j^{(n)}\rangle$ are mutually orthogonal, i.e. the states $\tilde{\rho}_n$ have orthogonal supports. □

11. **Entropy of state preparation**. If a pure state is drawn randomly from the ensemble $\mathsf{E} = \{p_n, |\psi_n\rangle\langle\psi_n|\}$, one has

$$S\left(\sum_n p_n |\psi_n\rangle\langle\psi_n|\right) \leq H(\mathbf{p}), \tag{12.29}$$

with equality iff the states are mutually orthogonal.

Proof. Indeed, the statement is just the upper bound (12.28) of item 10 for an ensemble of pure states. □

Notice that this result implies that the von Neumann entropy could be defined as

$$S(\rho) := \min_{\mathsf{E}_\rho^P} H(\mathbf{p}_\mathsf{E}),$$

where E_ρ^P is the set of pure ensembles $\mathsf{E} = \{p_i |\psi_i\rangle\langle\psi_i|\}$ such that $\sum_i p_i |\psi\rangle\langle\psi|_i = \rho$, and $\mathbf{p}_\mathsf{E} := \{p_i\}$.

12. **Entropy of measurement**. Let $\rho \in \mathsf{St}_1(\mathrm{A})$, and let $\mathbf{P}$ be an *atomic* POVM $\{|\psi_i\rangle\langle\psi_i|\}$. Then

$$S(\rho) \leq H(\mathrm{X}), \tag{12.30}$$

where X is the random variable

$$\mathrm{X} = \begin{cases} \mathsf{Rng}(\mathrm{X}) & = \{x_n\}, \\ \mathbb{P}_{\mathrm{X}}[\mathrm{X} = x_n] & = \{\langle\psi_n|\rho|\psi_n\rangle\}. \end{cases} \tag{12.31}$$

Proof. Let us define the following isometric operator

$$V := \sum_{i=1}^{|\mathsf{Rng}(\mathrm{X})|} |i\rangle\langle\psi_i| : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{B}} = \mathbb{C}^{|\mathsf{Rng}(\mathrm{X})|},$$

where $\langle i|j\rangle = \delta_{i,j}$. Indeed, it is easy to check that $V^\dagger V = I_{\mathrm{A}}$. Now, let us define $\mathscr{P} : \mathrm{B} \to \mathrm{B}$ as

$$\mathscr{P}(\sigma) := \sum_{i=1}^{d_{\mathrm{B}}} |i\rangle\langle i|\sigma|i\rangle\langle i|.$$

By items 9 and 3, we now have

$$S(\mathscr{P}(V\rho V^\dagger)) \geq S(V\rho V^\dagger) = S(\rho).$$

Observing that

$$\begin{aligned} \mathscr{P}(V\rho V^\dagger) &= \sum_{i=1}^{d_{\mathrm{B}}} |i\rangle\langle\psi_i|\rho|\psi_i\rangle\langle i| \\ &= \sum_{i=1}^{d_{\mathrm{B}}} |i\rangle\langle i|\mathbb{P}_{\mathrm{X}}[\mathrm{X} = x_i], \end{aligned}$$

we conclude that $S(\mathscr{P}(V\rho V^\dagger)) = H(\mathrm{X})$. Notice that, having used item 9, equality holds if and only if $\mathscr{P}(V\rho V^\dagger) = V\rho V^\dagger$, i.e.

$$\begin{aligned} V\rho V^\dagger &= \sum_{i,i'=1}^{d_{\mathrm{B}}} |i\rangle\langle\psi_i|\rho|\psi_{i'}\rangle\langle i'| \\ &= \sum_{i=1}^{d_{\mathrm{B}}} |i\rangle\langle\psi_i|\rho|\psi_i\rangle\langle i|, \\ &\Leftrightarrow\ \langle\psi_i|\rho|\psi_{i'}\rangle = \delta_{i,i'}\langle\psi_i|\rho|\psi_i\rangle, \end{aligned}$$

namely if and only if $|\psi_i\rangle\langle\psi_i|$ are projections on the eigenvectors of ρ. □

Notice that also this result provides an alternate definition of von Neumann entropy:

$$S(\rho) := \min_{\mathbf{P}\in P_A} H(\mathbf{p}_{\mathbf{P},\rho}),$$

where P_A is the set of *atomic* POVMs, i.e. POVMs made of rank one operators $\mathbf{P} = \{|\psi_i\rangle\langle\psi_i|\}$, and $\mathbf{p}_{\mathbf{P},\rho}$ is the probability vector $\mathbf{p} = (\langle\psi_i|\rho|\psi_i\rangle)$.

13. **Araki-Lieb triangle inequality**. The following bound holds

$$S(\mathrm{AB}) \geq |S(\mathrm{A}) - S(\mathrm{B})|. \tag{12.32}$$

Proof. Let AB be prepared in the state ρ_{AB}. We introduce a system R which purifies ρ_{AB}. From subadditivity, we have

$$S(\mathrm{AR}) \leq S(\mathrm{A}) + S(\mathrm{R}). \tag{12.33}$$

However, since ABR is pure, we have

$$S(\mathrm{AR}) = S(\mathrm{B}), \quad S(\mathrm{AB}) = S(\mathrm{R}),$$

namely, using inequality (12.33), we obtain

$$S(\mathrm{B}) = S(\mathrm{AR}) \leq S(\mathrm{A}) + S(\mathrm{R}) = S(\mathrm{A}) + S(\mathrm{AB}),$$

and finally

$$S(\mathrm{AB}) \geq S(\mathrm{B}) - S(\mathrm{A}).$$

Due to symmetry between A and B in the same way we find also

$$S(\mathrm{AB}) \geq S(\mathrm{A}) - S(\mathrm{B}),$$

and the two bounds together are equivalent to the statement. □

Chapter 13

Lecture 15: Properties of quantum relative entropy and Lieb's theorem

13.1 Mathematical properties of the quantum relative entropy

1. **Isometric invariance**. The quantum relative entropy is invariant under isometric transformations, namely

$$S(V\rho V^\dagger \| V\sigma V^\dagger) = S(\rho\|\sigma), \quad V^\dagger V = I. \tag{13.1}$$

Proof. In the case where $\mathsf{Ker}(\sigma) \not\subseteq \mathsf{Ker}(\rho)$, clearly $\mathsf{Ker}(V\sigma V^\dagger) \not\subseteq \mathsf{Ker}(V\rho V^\dagger)$. Let indeed $|\varphi\rangle \in \mathsf{Ker}(\sigma)$ while $|\varphi\rangle \notin \mathsf{Ker}(\rho)$. Then $V|\varphi\rangle \in \mathsf{Ker}(V\sigma V^\dagger)$ while $V|\varphi\rangle \notin \mathsf{Ker}(V\rho V^\dagger)$. In the complementary case the thesis is an immediate consequence of the identity $f(VAV^\dagger) = Vf(A)V^\dagger$ for any selfadjoint operator A, and the invariance of the trace under cyclic permutations of its argument. □

2. **Invariance under extension**. The quantum relative entropy is invariant under extension with a fixed state, namely

$$S(\rho\otimes\nu\|\sigma\otimes\nu) = S(\rho\|\sigma). \tag{13.2}$$

Proof. If $\mathsf{Ker}(\sigma) \not\subseteq \mathsf{Ker}(\rho)$, then also $\mathsf{Ker}(\sigma\otimes\nu) \not\subseteq \mathsf{Ker}(\rho\otimes\nu)$. When $\mathsf{Ker}(\sigma) \subseteq \mathsf{Ker}(\rho)$, the thesis is an easy consequence of the definition. One has

$$\begin{aligned} S(\rho\otimes\nu\|\sigma\otimes\nu) =& \operatorname{Tr}[(\rho\otimes\nu)\log_2(\rho\otimes\nu)] - \operatorname{Tr}[(\rho\otimes\nu)\log_2(\sigma\otimes\nu)] \\ =& \operatorname{Tr}[(\rho\otimes\nu)(\log_2\rho\otimes I_{\mathrm{B}} + I_{\mathrm{A}}\otimes\log_2\nu)] \\ &- \operatorname{Tr}[(\rho\otimes\nu)(\log_2\sigma\otimes I_{\mathrm{B}} + I_{\mathrm{A}}\otimes\log_2\nu)] \\ =& \operatorname{Tr}[\rho\log_2\rho] + \operatorname{Tr}[\nu\log_2\nu] - \operatorname{Tr}[\rho\log_2\sigma] - \operatorname{Tr}[\nu\log_2\nu] \\ =& S(\rho\|\sigma). \end{aligned}$$

□

3. **Monotonicity under marginalisation**. Let ρ and σ be two states of a composite system AB. The quantum relative entropy is monotonically decreasing when discarding one system, namely

$$S(\operatorname{Tr}_2(\rho)\|\operatorname{Tr}_2(\sigma)) \leq S(\rho\|\sigma). \tag{13.3}$$

This result requires joint convexity of relative entropy, which we prove later, using Lieb's theorem.

4. **Uhlmann monotonicity theorem**. The quantum relative entropy is monotonically non-increasing under the action of a channel, namely one has

$$S(\mathscr{E}(\rho)\|\mathscr{E}(\sigma)) \leq S(\rho\|\sigma). \tag{13.4}$$

Proof. This result easily follows from items 1, 2, and 3. □

13.2 Preliminary results for Lieb's theorem

The aim of this chapter is to provide a proof of Lieb's theorem, a technical result whose proof was provided by E. Lieb, and used for strong subadditivity by Lieb and Ruskai. The theorem is preliminary to many crucial results about properties of von Neumann entropy. The proof in this chapter is based on J. Watrous's lecture notes [3]. We start by introducing some definitions and lemmas that will be used to prove the main result.

Definition 13.1 (Spectral radius)**.** Given an operator $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$, we define its spectral radius $\rho(A)$ as the maximum among absolute values of its eigenvalues. In formula

$$\rho(A) := \max_{\lambda \in \mathsf{Spec}(A)} |\lambda|, \tag{13.5}$$

$\mathsf{Spec}(A)$ denoting the spectrum of A.

Remark 18. In the infinite-dimensional case the spectral radius $\rho(A)$ is defined as the supremum of $|\lambda|$ for $\lambda \in \mathsf{Spec}(A)$.

The spectral radius is related to the *uniform norm*, also called *sup-norm*, inducing the *uniform operator topology* in $\mathcal{L}(\mathcal{H}_{\mathrm{A}})$.

Definition 13.2 (Uniform norm)**.** Given an operator $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$, we define its uniform norm (sup-norm) $\|A\|_\infty$ as follows

$$\|A\|_\infty := \sup_{\psi \in \mathcal{H}_{\mathrm{A}}} \frac{\|A\psi\|_{\mathrm{A}}}{\|\psi\|_{\mathrm{A}}} = \sup_{\substack{\psi \in \mathcal{H}_{\mathrm{A}} \\ \|\psi\|_{\mathrm{A}}=1}} \|A\psi\|_{\mathrm{A}} \tag{13.6}$$

where $\|\cdot\|_{\mathrm{A}}$ denotes the norm induced by the inner product in $\mathcal{H}_{\mathrm{A}}$.

The following property (the proof is left as an exercise) holds for the uniform norm

$$\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty.$$

Now, the spectral radius is bounded by the sup-norm as follows.

Lemma 13.3. *For an operator $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$, the spectral radius is upper-bounded by the uniform norm.*

Proof. Let ψ be any normalised eigenvector of A. Then, by definition, the sup norm satisfies

$$\|A\|_\infty \geq \|A\psi\|_{\mathrm{A}} = |\lambda|.$$

This implies that the maximum (supremum) of $|\lambda|$ over $\mathsf{Spec}(A)$ is also bounded from above by $\|A\|_\infty$. □

For normal operators $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ a stronger result holds.

Lemma 13.4. *Let $A : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ be normal. Then $\|A\|_\infty = \rho(A)$.*

Proof. If A is normal, it is diagonalizable, and can then be written as

$$A = \sum_{\lambda_i \in \mathsf{Spec}(A)} \lambda_i \Pi_i, \quad \Pi_i \Pi_j = \delta_{ij} \Pi_i, \quad \sum_{\lambda_i \in \mathsf{Spec}(A)} \Pi_i = I_{\mathrm{A}} \tag{13.7}$$

where Π_i denotes the projection on the eigenspace corresponding to the eigenvalue λ_i. For $\|\psi\|_{\mathrm{A}} = 1$ one then has

$$\begin{aligned} \|A\psi\|_{\mathrm{A}} &= \left(\sum_{\lambda_i \in \mathsf{Spec}(A)} |\lambda_i|^2 \langle \psi | \Pi_i | \psi \rangle \right)^{\frac{1}{2}} \\ &\leq \rho(A) \left(\sum_{\lambda_i \in \mathsf{Spec}(A)} \langle \psi | \Pi_i | \psi \rangle \right)^{\frac{1}{2}} \\ &= \rho(A) \|\psi\|_{\mathrm{A}}. \end{aligned}$$

Taking the sup on l.h.s. we have $\|A\|_\infty \leq \rho(A)$. Along with lemma 13.3, this implies $\|A\|_\infty = \rho(A)$. □

In the case of selfadjoint operators, we have also the following useful result.

Lemma 13.5. *Let $A = A^\dagger : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ be a selfadjoint operator. Then*

$$-\|A\|_\infty I_{\mathrm{A}} \leq A \leq \|A\|_\infty I_{\mathrm{A}}. \tag{13.8}$$

Proof. Since A is normal, by lemma 13.4 we have $\rho(A) = \|A\|_\infty$. Moreover, the eigenvalues of A are real: $\mathsf{Spec}(A) \subseteq \mathbb{R}$. Then we have $-\rho(A) \leq -|\lambda_i| \leq \lambda_i \leq |\lambda_i| \leq \rho(A)$ for all $\lambda_i \in \mathsf{Spec}(A)$. Upon diagonalizing A as $A = \sum_{\lambda_i \in \mathsf{Spec}(A)} \lambda_i \Pi_i$, this finally implies

$$\begin{aligned} -\|A\|_\infty I_{\mathrm{A}} &= \sum_{\lambda_i \in \mathsf{Spec}(A)} [-\rho(A)] \Pi_i \\ &\leq \sum_{\lambda_i \in \mathsf{Spec}(A)} \lambda_i \Pi_i \\ &\leq \sum_{\lambda_i \in \mathsf{Spec}(A)} \rho(A) \Pi_i \\ &= \|A\|_\infty I_{\mathrm{A}}. \end{aligned}$$

□

The last result has a relevant consequence on the limit of Cauchy sequences of operators in the uniform norm.

Definition 13.6. Let $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{L}(\mathcal{H}_\mathrm{A})$ be a sequence of operators. We say that the sequence uniformly converges to A, in formula $\lim_{n\to\infty} A_n = A$, if, $\forall \varepsilon > 0$, $\exists n_0$ such that $\forall n > n_0$

$$\|A_n - A\|_\infty < \varepsilon. \tag{13.9}$$

We now prove that if a sequence $\{A_n\}_{n\in\mathbb{N}}$ of selfadjoint operators uniformly converges to A, then also A is selfadjoint.

Lemma 13.7. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of selfadjoint operators $A_n^\dagger = A_n$, that uniformly converges to A. Then $A = A^\dagger$.*

Proof. Let us consider the polar decomposition $B = U|B|$ of a general operator B, with $U : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ unitary. Then one has $\|B^\dagger\|_\infty = \|U^\dagger B U^\dagger\|_\infty = \|B\|_\infty$. Thus,

$$\begin{aligned} \|A_n - A^\dagger\|_\infty &= \|A_n^\dagger - A^\dagger\|_\infty \\ &= \|A_n - A\|_\infty, \end{aligned}$$

and thus A_n converges to $A^\dagger$. By uniqueness of the limit we finally have $A = A^\dagger$. □

Lemma 13.5 allows us now to prove the following theorem.

Theorem 13.8. *If the sequence $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{L}(\mathcal{H}_\mathrm{A})$ of selfadjoint operators uniformly converges to A, and for all $n > \bar{n}$ we have $A_n \geq 0$, then also $A \geq 0$.*

Proof. By definition of uniform convergence, for every $\varepsilon > 0$ there exists n_0 such that for $n > n_0$ we have $\|A_n - A\|_\infty < \varepsilon$. This implies that

$$-\varepsilon I_\mathrm{A} \leq A_n - A \leq \varepsilon I_\mathrm{A},$$

and finally, taking the expectation value of all members on the arbitrary vector $\psi \in \mathcal{H}_\mathrm{A}$, we have

$$|\langle\psi|A_n|\psi\rangle - \langle\psi|A|\psi\rangle| \leq \varepsilon.$$

Finally, since a sequence of positive real numbers $\{a_n\}_{n\in\mathbb{N}}$ such that for $n > \bar{n}$ $a_n \geq 0$ converges to a positive $a \geq 0$, we conclude that also $\langle\psi|A|\psi\rangle \geq 0$. □

13.3 Lieb's theorem

In this lecture we prove Lieb's theorem and subsequently we will use it to prove some key properties of von Neumann entropy. For this purpose, we need some preliminary lemmas. The next few results regard partitioned matrices. Let the Hilbert space $\mathcal{H}_\mathrm{A}$ be partitioned into orthogonal subspaces $\mathcal{H}_\mathrm{A} = \bigoplus_{r=1}^b \mathcal{H}_r$. If we define the projections $\{P_r\}_{r=1}^b$ on the b subspaces, then any operator $A \in \mathcal{L}(\mathcal{H}_\mathrm{A})$ can be partitioned into blocks $A^{(p,q)} := P_p A P_q$. The matrix representing A in a basis that is a collection made of of one orthonormal basis for each subspace $\mathcal{H}_r$ is then a partitioned matrix, with blocks $A^{(p,q)}$. In the following, when we write *partitioned matrix* the partitioning of the Hilbert space $\mathcal{H}_\mathrm{A}$ will be implicit.

Lemma 13.9. *Let $A, B : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be partitioned matrices. Then the following identities hold*

$$(AB)^{(p,q)} = \sum_{r=1}^{b} A^{(p,r)} B^{(r,q)}, \quad (A^\dagger)^{(p,q)} = A^{(q,p)\dagger}. \tag{13.10}$$

Proof. We can write $(AB)^{(p,q)} = P_p ABP_q$ as

$$P_p A I_\mathrm{A} B P_q = \sum_{r=1}^{b} P_p A P_r B P_q = \sum_{r=1}^{b} A^{(p,r)} B^{(r,q)}.$$

Moreover, $(A^\dagger)^{(p,q)} = P_p A^\dagger P_q = (P_q A P_p)^\dagger = A^{(q,p)\dagger}$. □

For positive partitioned matrices the following result holds.

Lemma 13.10. *Let $A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ be a partitioned matrix. Then A is positive if and only if there exists a Hilbert space $\mathcal{H}_\mathrm{B}$ and a set $\{T^{(r)}\}_{r=1}^{b} \subseteq \mathcal{L}(\mathcal{H}_\mathrm{A}, \mathcal{H}_\mathrm{B})$ such that $\mathsf{Supp}(T^{(r)}) \subseteq \mathcal{H}_r$ and $A^{(p,q)} = T^{(p)\dagger} T^{(q)}$.*

Proof. First suppose that $A^{(p,q)} = T^{(p)\dagger} T^{(q)}$. Define $X := \sum_{p=1}^{b} T^{(p)}$. It is immediate to verify that $A = \sum_{p,q=1}^{b} A^{(p,q)}$ can be written as $A = X^\dagger X \geq 0$. Conversely, if $A \geq 0$ then $A = X^\dagger X$ for some $X : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$, and by lemma 13.9 we have

$$A^{(p,q)} = \sum_{r=1}^{b} X^{(r,p)\dagger} X^{(r,q)}.$$

Notice that $X^{(r,p)\dagger} X^{(s,q)} = (X^\dagger)^{(p,r)} X^{(s,q)} = \delta_{rs} X^{(r,p)\dagger} X^{(s,q)}$. If we now define $T^{(p)} := \sum_{r=1}^{b} X^{(r,p)}$, we then have $A^{(p,q)} = T^{(p)\dagger} T^{(q)}$. □

Lemma 13.11. *For a partitioned matrix $A : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$, A is positive iff $A^{(p,p)} \geq 0$ and $A^{(p,q)} = (A^{(p,p)})^{\frac{1}{2}} U_p^\dagger U_q (A^{(q,q)})^{\frac{1}{2}}$, where $U_p : \mathcal{H}_\mathrm{A} \to \mathcal{H}_\mathrm{A}$ is unitary for every p.*

Proof. By lemma 13.10, if $A \geq 0$ one has $A^{(p,p)} = T^{(p)\dagger} T^{(p)}$ with $T^{(p)} \in \mathcal{L}(\mathcal{H}_\mathrm{A})$, and then by the polar decomposition $T^{(p)} = U_p (A^{(p,p)})^{\frac{1}{2}}$, where U_p is unitary. Then $A^{(p,q)} = T^{(p)\dagger} T^{(q)} = (A^{(p,p)})^{\frac{1}{2}} U_p^\dagger U_q (A^{(q,q)})^{\frac{1}{2}}$. On the other hand, if $A^{(p,q)} = (A^{(p,p)})^{\frac{1}{2}} U_p^\dagger U_q (A^{(q,q)})^{\frac{1}{2}}$, one can define $T^{(p)} := U_p (A^{(p,p)})^{\frac{1}{2}}$, and clearly

$$A = \sum_{p,q} A^{(p,q)} = \sum_{p,q} T^{(p)\dagger} T^{(q)}.$$ □

Lemma 13.12. *Let $M : \mathcal{H} \to \mathcal{H}$ with $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ and $\mathcal{H}_0 \sim \mathcal{H}_1 \sim \mathcal{K}$. Let M correspond to a matrix of the form*

$$M = \left(\begin{array}{c|c} X & H \\ \hline H & Y \end{array} \right), \tag{13.11}$$

with $H, X, Y \in \mathcal{L}(\mathcal{K})$. Let $H = H^\dagger$ and $[X, Y] = 0$. If $M \geq 0$ then $X, Y \geq 0$ and $H \leq X^{\frac{1}{2}} Y^{\frac{1}{2}}$.

Proof. Let $W_i : \mathcal{K} \to \mathcal{H}$ be the isometry such that $W_i\mathcal{K} = \mathcal{H}_i$. Thus, $W_0 X W_0^\dagger = M^{(0,0)}$, $W_1 Y W_1^\dagger = M^{(1,1)}$, $W_1 H W_0^\dagger = M^{(1,0)}$, and $W_0 H W_1^\dagger = M^{(0,1)}$. By lemma 13.11, M is positive if and only if $X, Y \geq 0$ and

$$\begin{aligned} H &= W_0^\dagger M^{(0,1)} W_1 \\ &= W_0^\dagger M^{(0,0)\frac{1}{2}} U_0^\dagger U_1 M^{(1,1)\frac{1}{2}} W_1 \\ &= X^{\frac{1}{2}} V_0^\dagger V_1 Y^{\frac{1}{2}}, \end{aligned}$$

where $V_i \coloneqq U_i W_i$, and $\{U_i\}$ are the unitary operators of lemma 13.11. Moreover, since $H = H^\dagger$, we have $H = X^{\frac{1}{2}} V_0^\dagger V_1 Y^{\frac{1}{2}} = Y^{\frac{1}{2}} V_1^\dagger V_0 X^{\frac{1}{2}}$. Now, denoting by Z^{-1} the generalized inverse of Z, and defining $Q := X^{-\frac{1}{2}} H Y^{-\frac{1}{2}}$ we have

$$\rho(Q) \leq \|Q\|_\infty = \|P_X V_0^\dagger V_1 P_Y\|_\infty \leq \|P_X\|_\infty \|V_0^\dagger\|_\infty \|V_1\|_\infty \|P_Y\|_\infty = 1,$$

where P_Z denotes the projection on the support $\mathsf{Supp}(Z)$ of Z. Notice that $P_X H P_Y = P_Y H P_X = H$, and then $P_H \leq P_X$ and $P_H \leq P_Y$. This implies that both X and Y are invertible on $\mathsf{Supp}(H)$. As a consequence, the selfadjoint operator $R \coloneqq X^{-\frac{1}{4}} Y^{-\frac{1}{4}} H Y^{-\frac{1}{4}} X^{-\frac{1}{4}}$ can be written as $R = Y^{-\frac{1}{4}} X^{\frac{1}{4}} Q X^{-\frac{1}{4}} Y^{\frac{1}{4}} = TQT^{-1}$, where

$$T \coloneqq Y^{-\frac{1}{4}} X^{\frac{1}{4}} + W(I - P_{Y^{-\frac{1}{4}} X^{\frac{1}{4}}}),$$

and $W\mathsf{Ker}(Y^{-\frac{1}{4}} X^{\frac{1}{4}}) = \mathsf{Co}-\mathsf{Ker}(Y^{-\frac{1}{4}} X^{\frac{1}{4}})$. The operator T is invertible. Since $R = TQT^{-1}$ for an invertible T, it holds that $\mathsf{Spec}(R) = \mathsf{Spec}(Q)$, and then $\rho(R) = \rho(Q)$. Finally, since R is selfadjoint, we have

$$\|R\|_\infty = \rho(R) \leq 1.$$

Now, by lemma 13.5 we conclude $R \leq I_{\mathrm{A}}$, and finally

$$H = X^{\frac{1}{4}} Y^{\frac{1}{4}} R Y^{\frac{1}{4}} X^{\frac{1}{4}} \leq X^{\frac{1}{4}} Y^{\frac{1}{2}} X^{\frac{1}{4}} = X^{\frac{1}{2}} Y^{\frac{1}{2}}. \qquad \square$$

Lemma 13.13. *Let $\mathcal{L}(\mathcal{H}_{\mathrm{A}}) \ni A_0, A_1 \geq 0$, $\mathcal{L}(\mathcal{H}_{\mathrm{B}}) \ni B_0, B_1 \geq 0$. Then for every $p \in [0,1]$ we have*

$$(A_0 + A_1)^p \otimes (B_0 + B_1)^{1-p} \geq A_0^p \otimes B_0^{1-p} + A_1^p \otimes B_1^{1-p}. \tag{13.12}$$

Proof. Let $X_0(p) := A_0^p \otimes B_0^{1-p}$, $X_1(p) := A_1^p \otimes B_1^{1-p}$, and $X_2(p) := (A_0 + A_1)^p \otimes (B_0 + B_1)^{1-p}$. It is clear that

$$[X_i(p), X_i(q)] = 0, \quad \forall i = 0,1,2, \quad \forall p, q \in [0,1].$$

Moreover,

$$(X_i(p) X_i(q))^{\frac{1}{2}} = X_i(\tfrac{p+q}{2}), \quad \forall i = 0,1,2.$$

Let us now consider the partitioned matrices $Y_0(p,q)$ and $Y_1(p,q)$ defined as

$$Y_i(p,q) := \left(\begin{array}{c|c} X_i(p) & X_i(\frac{p+q}{2}) \\ \hline X_i(\frac{p+q}{2}) & X_i(q) \end{array}\right) = \left(\begin{array}{c|c} X_i(p)^{\frac{1}{2}} & 0 \\ \hline X_i(q)^{\frac{1}{2}} & 0 \end{array}\right) \left(\begin{array}{c|c} X_i(p)^{\frac{1}{2}} & X_i(q)^{\frac{1}{2}} \\ \hline 0 & 0 \end{array}\right).$$

It is clear that $Y_i(p,q) \geq 0$ for all $i = 0,1$ and for all $p,q \in [0,1]$. Moreover, if $X_2(p) \geq X_0(p) + X_1(p)$ and $X_2(q) \geq X_0(q) + X_1(q)$, then

$$\left(\begin{array}{c|c} X_2(p) - X_0(p) - X_1(p) & 0 \\ \hline 0 & X_2(q) - X_0(q) - X_1(q) \end{array}\right) + Y_0(p,q) + Y_1(p,q) =$$
$$\left(\begin{array}{c|c} X_2(p) & X_0(\frac{p+q}{2}) + X_1(\frac{p+q}{2}) \\ \hline X_0(\frac{p+q}{2}) + X_1(\frac{p+q}{2}) & X_2(q) \end{array}\right) \geq 0.$$

By lemma 13.12 we then have $X_0(\frac{p+q}{2}) + X_1(\frac{p+q}{2}) \leq (X_2(p)X_2(q))^{\frac{1}{2}} = X_2(\frac{p+q}{2})$. It is easy to check that for $p = 0$ and $p = 1$ we have $X_2(p) \geq X_0(p) + X_1(p)$, since $A_i^0 = P_{A_i} \leq P_{A_0+A_1} = (A_0 + A_1)^0$, and $B_i^0 = P_{B_i} \leq P_{B_0+B_1} = (B_0 + B_1)^0$. Iterating this argument, we find that the thesis is true for all $p \in J := \{\frac{n}{2^k} \,|\, k \in \mathbb{N},\ 0 \leq n \leq 2^k\}$, which is dense in $[0,1]$. In order to complete the proof, it is then sufficient to prove continuity of the function $p \mapsto \Delta(p) := X_2(p) - X_0(p) - X_1(p)$ in the uniform topology. This can be proved considering that for every $\varepsilon > 0$ there exists δ such that if $|p - q| < \delta$ we have

$$\begin{aligned} \|\Delta(p) - \Delta(q)\|_\infty &\leq \sum_{i=0,1,2} \|X_i(p) - X_i(q)\|_\infty \\ &\leq \sum_{i=0,1,2} \|X_i(p)\|_\infty \|I - A_i^{q-p} \otimes B_i^{p-q}\|_\infty \\ &\leq \sum_{i=0,1,2} \rho(X_i(p))|1 - k_i^{q-p}| < \varepsilon, \end{aligned}$$

where we defined $A_2 := A_0 + A_1$ and $B_2 := B_0 + B_1$. Now, since for any sequence $\{p_n\}_{n\in\mathbb{N}} \subseteq J$ converging to $p \in [0,1]$ we have $\Delta(p_n) \geq 0$, by theorem 13.8 we also have $\Delta(p) \geq 0$ for all $p \in [0,1]$. □

Using the last result, we can prove Lieb's theorem. We introduce the set $\mathcal{P}(\mathrm{A}) := \{X \in \mathcal{L}(\mathrm{A}) |\, X \geq 0\} = \{\lambda\rho |\, 0 \leq \lambda \in \mathbb{R},\, \rho \in \mathsf{St}_1(\mathrm{A})\}$

Theorem 13.14 (Lieb's theorem)**.** *Let $F_p^{(K)} : \mathcal{P}(\mathrm{A}) \times \mathcal{P}(\mathrm{B}) \to \mathbb{R}$ be defined as $F_p^{(K)}(A,B) := \mathrm{Tr}[A^p K B^{1-p} K^\dagger]$ for an arbitrary $K : \mathcal{H}_\mathrm{B} \to \mathcal{H}_\mathrm{A}$. The function F_p is jointly concave in A, B for any K and any $p \in [0,1]$, i.e.*

$$F_p^{(K)}(\lambda A_1 + (1-\lambda)A_2, \lambda B_1 + (1-\lambda)B_2) \geq \lambda F_p^{(K)}(A_1, B_1) + (1-\lambda) F_p^{(K)}(A_2, B_2), \quad (13.13)$$

for every $0 \leq \lambda \leq 1$.

Proof. It is sufficient to notice that

$$F_p^{(K)}(A,B) = \langle\!\langle K | A^p \otimes (B^T)^{1-p} | K \rangle\!\rangle. \quad (13.14)$$

If we then apply Eq. (13.12) with λA_1 and $(1-\lambda)A_2$ instead of A_1 and A_2, and λB_1^T and $(1-\lambda)B_2^T$ instead of B_1 and B_2, by lemma 13.13 we have

$$(\lambda A_1 + (1-\lambda)A_2)^p \otimes (\lambda B_1^T + (1-\lambda)B_2^T)^{1-p} \geq \lambda A_1^p \otimes (B_1^T)^{1-p} + (1-\lambda) A_2^p \otimes (B_2^T)^{1-p}.$$

The thesis then follows reminding Eq. (13.14). □

Chapter 14

Lecture 16: Monotonicity and Holevo bound

In the following we use Lieb's theorem to prove four crucial results: Joint convexity of quantum relative entropy, Uhlmann's monotonicity, concavity of quantum conditional entropy and strong subadditivity. The logical scheme shown in fig. 14.1 is useful to clarify the importance of Lieb's theorem, as well as the relation between the above mentioned theorems.

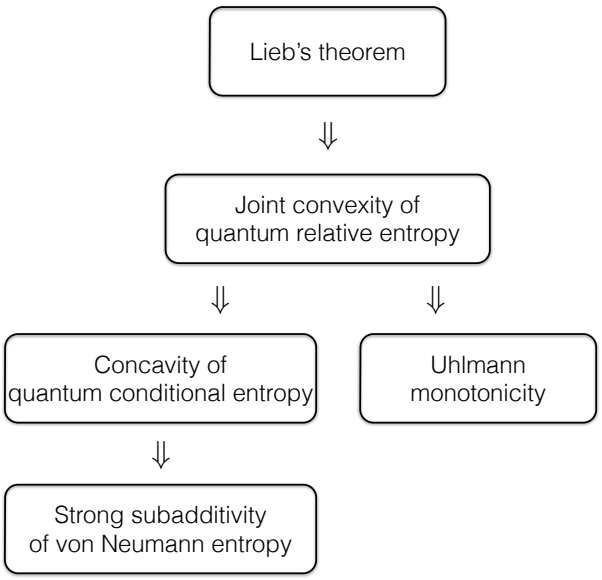


Figure 14.1 Logical dependence of the main results in the chapter

14.1 Joint convexity of quantum relative entropy

Now that we have proved the Lieb's theorem, we are in position to prove the joint convexity of relative entropy. First, we prove an equivalent way of expressing $S(\rho\|\sigma)$.

Definition 14.1. Let us define the function $G_t : \mathcal{P}(\mathrm{A}) \times \mathcal{P}(\mathrm{B}) \to \mathbb{R}$

$$G_t(A, B) := \mathrm{Tr}[A^{1-t}B^t] - \mathrm{Tr}[A].$$

Let us also define

$$G(A, B) := \lim_{t\to 0^+} \frac{G_t(A, B)}{t},$$

where we recall that $\lim_{t\to 0^+} A^t := P_{\mathsf{Supp}(A)}$ for $A \geq 0$.

Lemma 14.2. *The quantum relative entropy $S(\rho\|\sigma)$ can be expressed as*

$$S(\rho\|\sigma) = -\frac{1}{\ln 2} G(\rho, \sigma). \tag{14.1}$$

Proof. Let us remind definition 12.2 of $S(\rho\|\sigma)$, in particular equation (12.10):

$$S(\rho\|\sigma) := \begin{cases} \mathrm{Tr}[\rho \log_2 \rho] - \mathrm{Tr}[\rho \log_2 \sigma] & \mathsf{Ker}(\sigma) \subseteq \mathsf{Ker}(\rho) \\ \infty & \text{otherwise.} \end{cases}$$

Let us first consider the case $\mathsf{Ker}(\sigma) \not\subseteq \mathsf{Ker}(\rho)$. Since by definition one has $\lim_{t\to 0^+} A^t := P_{\mathsf{Supp}(A)}$, we have

$$\lim_{t\to 0^+} G_t(\rho, \sigma) = -\mathrm{Tr}[\rho(I - P_{\mathsf{Supp}(\sigma)})] = -\mathrm{Tr}[\rho P_{\mathsf{Ker}(\sigma)}] < 0.$$

Thus,

$$-\frac{1}{\ln 2} \lim_{t\to 0^+} \frac{G_t(\rho, \sigma)}{t} = +\infty.$$

Let us now consider the case $\mathsf{Ker}(\sigma) \subseteq \mathsf{Ker}(\rho)$. In this case we have

$$\begin{aligned} G_0(\rho, \sigma) &= -\mathrm{Tr}[\rho(I - P_{\mathsf{Supp}(\sigma)})] \\ &= -\mathrm{Tr}[\rho P_{\mathsf{Ker}(\sigma)}] \\ &= 0, \end{aligned}$$

and thus

$$\begin{aligned} \lim_{t\to 0^+} \frac{G_t(\rho, \sigma)}{t} &= \lim_{t\to 0^+} \frac{G_t(\rho, \sigma) - G_0(\rho, \sigma)}{t} \\ &= \lim_{t\to 0^+} \frac{d}{dt} \mathrm{Tr}[\rho^{1-t}\sigma^t] \\ &= -\lim_{t\to 0^+} \{\mathrm{Tr}[\ln \rho \rho^{1-t}\sigma^t] - \mathrm{Tr}[\rho^{1-t} \ln \sigma \sigma^t]\}. \end{aligned}$$

Finally, this implies

$$-\frac{1}{\ln 2} G(\rho, \sigma) = \mathrm{Tr}[\rho \log_2 \rho] - \mathrm{Tr}[\rho \log_2 \sigma].$$

□

Theorem 14.3 (Joint convexity of quantum relative entropy)**.** *The relative entropy $S(\rho\|\sigma)$ is jointly convex in ρ and σ, that is*

$$S(p\rho_1 + (1-p)\rho_2\|p\sigma_1 + (1-p)\sigma_2) \leq pS(\rho_1\|\sigma_1) + (1-p)S(\rho_2\|\sigma_2), \tag{14.2}$$

for all $p \in [0,1]$.

Proof. If we use the expression 14.1, we have

$$S(\rho\|\sigma) = -\frac{1}{\ln 2}\lim_{t\to 0^+}\frac{\mathrm{Tr}[\rho^{1-t}\sigma^t] - \mathrm{Tr}[\rho]}{t}.$$

We can write

$$G_t(\rho,\sigma) = \mathrm{Tr}[\rho^{1-t}\sigma^t] - \mathrm{Tr}[\rho] = \langle\!\langle I|\rho^{1-t}\otimes(\sigma^T)^t|I\rangle\!\rangle - \mathrm{Tr}[\rho],$$

and using now Lieb's theorem 13.14, we have

$$G_t(p\rho_1 + (1-p)\rho_2, p\sigma_1 + (1-p)\sigma_2) \geq pG_t(\rho_1,\sigma_1) + (1-p)G_t(\rho_2,\sigma_2).$$

Then, we have

$$\begin{aligned}
S(p\rho_1 + (1-p)\rho_2\|p\sigma_1 + (1-p)\sigma_2) &= -\frac{1}{\ln 2}\lim_{t\to 0^+}\frac{G_t(p\rho_1 + (1-p)\rho_2, p\sigma_1 + (1-p)\sigma_2)}{t} \\
&\leq -\frac{1}{\ln 2}\lim_{t\to 0^+}\frac{pG_t(\rho_1,\sigma_1) + (1-p)G_t(\rho_2,\sigma_2)}{t} \\
&= p\left\{-\frac{1}{\ln 2}\lim_{t\to 0^+}\frac{G_t(\rho_1,\sigma_1)}{t}\right\} + (1-p)\left\{-\frac{1}{\ln 2}\lim_{t\to 0^+}\frac{G_t(\rho_2,\sigma_2)}{t}\right\} \\
&= pS(\rho_1\|\sigma_1) + (1-p)S(\rho_2\|\sigma_2).
\end{aligned}$$

□

14.2 Concavity of quantum conditional entropy

From joint convexity of relative entropy, concavity of the conditional entropy follows.

Theorem 14.4 (Concavity of quantum conditional entropy)**.** *The quantum conditional entropy is concave versus the joint state ρ_{AB}, that is*

$$S_{p\rho^{(1)}+(1-p)\rho^{(2)}}(\mathrm{A}|\mathrm{B}) \geq pS_{\rho^{(1)}}(\mathrm{A}|\mathrm{B}) + (1-p)S_{\rho^{(2)}}(\mathrm{A}|\mathrm{B}). \tag{14.3}$$

Proof. Let us use the following identity

$$\begin{aligned}
S(\rho_{\mathrm{AB}}\|\tfrac{I_{\mathrm{A}}}{d_{\mathrm{A}}}\otimes\rho_{\mathrm{B}}) &= -S(\mathrm{AB}) - \mathrm{Tr}[\rho_{\mathrm{AB}}\log_2(\tfrac{I_{\mathrm{A}}}{d_{\mathrm{A}}}\otimes\rho_{\mathrm{B}})] \\
&= -S(\mathrm{AB}) - \mathrm{Tr}[\rho_{\mathrm{AB}}\{\log_2(\tfrac{I_{\mathrm{A}}}{d_{\mathrm{A}}})\otimes I_{\mathrm{B}}\}] - \mathrm{Tr}[\rho_{\mathrm{AB}}\{I_{\mathrm{A}}\otimes\log_2(\rho_{\mathrm{B}})\}] \\
&= -S(\mathrm{AB}) + S(\mathrm{B}) + \log_2 d_{\mathrm{A}} \\
&= -S(\mathrm{A}|\mathrm{B}) + \log_2 d_{\mathrm{A}}.
\end{aligned}$$

Since the partial trace Tr_A is linear, we have $\mathrm{Tr}_\mathrm{B}[p\rho^{(1)}_\mathrm{AB} + (1-p)\rho^{(2)}_\mathrm{AB}] = p\rho^{(1)}_\mathrm{B} + (1-p)\rho^{(2)}_\mathrm{B}$. Then, using joint convexity of quantum relative entropy, one has

$$\begin{aligned}
&S_{p\rho^{(1)}+(1-p)\rho^{(2)}}(\mathrm{A}|\mathrm{B}) = \log_2 d_\mathrm{A} - S(p\rho^{(1)}_\mathrm{AB} + (1-p)\rho^{(2)}_\mathrm{AB} \| p(\tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(1)}_\mathrm{B}) + (1-p)(\tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(2)}_\mathrm{B})) \\
&\geq \log_2 d_\mathrm{A} - pS(\rho^{(1)}_\mathrm{AB} \| \tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(1)}_\mathrm{B}) - (1-p)S(\rho^{(2)}_\mathrm{AB} \| \tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(2)}_\mathrm{B}) \\
&= p(\log_2 d_\mathrm{A} - S(\rho^{(1)}_\mathrm{AB} \| \tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(1)}_\mathrm{B})) + (1-p)(\log_2 d_\mathrm{A} - S(\rho^{(2)}_\mathrm{AB} \| \tfrac{I_\mathrm{A}}{d_\mathrm{A}} \otimes \rho^{(2)}_\mathrm{B})) \\
&= pS_{\rho^{(1)}}(\mathrm{A}|\mathrm{B}) + (1-p)S_{\rho^{(2)}}(\mathrm{A}|\mathrm{B})
\end{aligned}$$

□

14.3 Strong subadditivity of von Neumann entropy

Using concavity of conditional entropy we can now prove the strong subadditivity property of the von Neumann entropy.

Theorem 14.5 (Strong subadditivity of von Neumann entropy)**.** *For any three quantum systems* A*,* B*, and* C*, one has*

$$S(\mathrm{A}) + S(\mathrm{B}) \leq S(\mathrm{AC}) + S(\mathrm{BC}), \tag{14.4}$$

or equivalently

$$S(\mathrm{ABC}) + S(\mathrm{B}) \leq S(\mathrm{AB}) + S(\mathrm{BC}). \tag{14.5}$$

Proof. The two inequalities are equivalent, as we will see. Define the function

$$T(\rho_\mathrm{ABC}) := -S(\mathrm{C}|\mathrm{A}) - S(\mathrm{C}|\mathrm{B}) = S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{AC}) - S(\mathrm{BC}),$$

Since conditional entropies are concave versus their joint state, the function T is convex versus ρ_ABC. Then, it follows that for $\rho_\mathrm{ABC} = \sum_i p_i |i\rangle\langle i|$ one has

$$T(\rho_\mathrm{ABC}) \leq \sum_i p_i T(|i\rangle\langle i|).$$

Now, T is zero on pure states, since in this case $S(\mathrm{AC}) = S(\mathrm{B})$ and $S(\mathrm{BC}) = S(\mathrm{A})$. Thus, one has $T(\rho_\mathrm{ABC}) \leq 0$, namely

$$S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{AC}) - S(\mathrm{BC}) \leq 0, \tag{14.6}$$

which is the first inequality of the statement. To obtain the second inequality, let us purify ρ_ABC by adding a reference system R. Inequality (14.6) for systems R, B, C is

$$S(\mathrm{R}) + S(\mathrm{B}) \leq S(\mathrm{RC}) + S(\mathrm{BC}),$$

but since ρ_RABC is now pure, one has $S(\mathrm{R}) = S(\mathrm{ABC})$ and $S(\mathrm{RC}) = S(\mathrm{AB})$, and the last inequality becomes

$$S(\mathrm{ABC}) + S(\mathrm{B}) \leq S(\mathrm{AB}) + S(\mathrm{BC}),$$

which is the second inequality of the statement. □

Theorem 14.6 (Conditioning reduces the quantum entropy)**.**

$$S(\mathrm{A}|\mathrm{BC}) \leq S(\mathrm{A}|\mathrm{B}). \tag{14.7}$$

Proof. Using the definition of conditional entropy, the statement of the theorem is equivalent to

$$S(\mathrm{ABC}) - S(\mathrm{BC}) \leq S(\mathrm{AB}) - S(\mathrm{B}),$$

which is just strong subadditivity. □

Theorem 14.7 (Discarding systems and quantum mutual information)**.** *Discarding quantum systems never increases the quantum mutual information, namely*

$$S(\mathrm{A} : \mathrm{B}) \leq S(\mathrm{A} : \mathrm{BC}). \tag{14.8}$$

Proof. Using the definition of mutual information, we see that the statement is equivalent to

$$S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{AB}) \leq S(\mathrm{A}) + S(\mathrm{BC}) - S(\mathrm{ABC}),$$

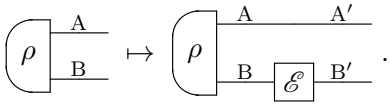
namely

$$S(\mathrm{ABC}) + S(\mathrm{B}) \leq S(\mathrm{AB}) + S(\mathrm{BC}),$$

which is strong subadditivity. □

Theorem 14.8 (Local channels and quantum mutual information)**.** *Local quantum channels never increase the quantum mutual information.*

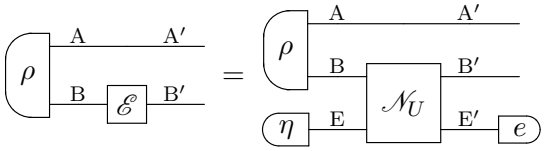
Proof. Denote by A' and B' the two quantum systems A and B after the action of the local channel $\mathscr{E}$ on B, as in the following diagram



We want to prove that

$$S(\mathrm{A}' : \mathrm{B}') \leq S(\mathrm{A} : \mathrm{B}).$$

By the unitary dilation theorem, the action of $\mathscr{E}$ on B is equivalent to a unitary interaction $\mathscr{N}_U$ with another system E which is then discarded, as represented in the following diagram



Before the action of $\mathscr{N}_U$, B and E are uncorrelated, then

$$\begin{aligned} S(\mathrm{A} : \mathrm{BE}) =& S(\mathrm{A}) + S(\mathrm{BE}) - S(\mathrm{ABE}) \\ =& S(\mathrm{A}) + S(\mathrm{B}) + S(\mathrm{E}) - S(\mathrm{AB}) - S(\mathrm{E}) = S(\mathrm{A} : \mathrm{B}). \end{aligned}$$

Invariance of von Neumann entropy under unitary transformations leads to

$$S(\mathrm{A}' : \mathrm{B}'\mathrm{E}') = S(\mathrm{A} : \mathrm{BE}).$$

Finally, discarding a quantum system cannot increase the mutual information, thus

$$S(\mathrm{A}' : \mathrm{B}') \leq S(\mathrm{A}' : \mathrm{B}'\mathrm{E}') = S(\mathrm{A} : \mathrm{BE}) = S(\mathrm{A} : \mathrm{B}).$$ □

14.4 Uhlmann monotonicity theorem

In this section we derive the last of the main consequences of Lieb's theorem: Uhlmann monotonicity theorem for joint entropy. In particular, the proof uses joint convexity of quantum relative entropy.

The first lemma that we need involves the so-called *shift-and-multiply group* for a system A. This is a group of d_{A}^2 unitary operators $\{U_{p,q}\}_{p,q=0}^{d_{\mathrm{A}}-1}$ that are given in the canonical basis $\{\varphi_i\}_{i=0}^{d_{\mathrm{A}}-1}$ as $U_{p,q} := Z^p W^q$ in terms of the phase operator W and shift operator Z, defined as

$$\begin{aligned} W &:= \sum_{i=0}^{d_{\mathrm{A}}-1} \omega_{\mathrm{A}}^i |\varphi_i\rangle\langle\varphi_i|, & \omega_{\mathrm{A}} &:= \exp(-i\tfrac{2\pi}{d_{\mathrm{A}}}), \\ Z &:= \sum_{i=0}^{d_{\mathrm{A}}-1} |\varphi_{i\oplus 1}\rangle\langle\varphi_i|, & a\oplus b &:= a+b \pmod{d_{\mathrm{A}}}. \end{aligned} \tag{14.9}$$

The set $U_{p,q}$ provides a *projective unitary representation* of the group $\mathbb{Z}_{d_{\mathrm{A}}} \times \mathbb{Z}_{d_{\mathrm{A}}}$. The property we are interested in is given by the following lemma.

Lemma 14.9. *Let $\{U_{p,q}\}_{p,q=0}^{d_{\mathrm{A}}-1}$ be the shift-and-multiply group for a system* A *with dimension d_{A}. then the following identity holds*

$$\frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} \otimes U_{p,q}^* = \frac{1}{d_{\mathrm{A}}} |I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|. \tag{14.10}$$

Proof. By straightforward algebra we have

$$\begin{aligned} \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} \otimes U_{p,q}^* &= \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} Z^p W^q \otimes Z^p W^{-q} \\ &= \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} \sum_{j,k=0}^{d_{\mathrm{A}}-1} \omega^{(j-k)q} |\varphi_{j\oplus p}\rangle\langle\varphi_j| \otimes |\varphi_{k\oplus p}\rangle\langle\varphi_k|. \end{aligned}$$

Now, since

$$\frac{1}{d_{\mathrm{A}}} \sum_{q=0}^{d_{\mathrm{A}}-1} \omega_{\mathrm{A}}^{(j-k)q} = \begin{cases} 1 & j = k, \\ 0 & j \neq k, \end{cases}$$

we conclude

$$\begin{aligned} \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} \otimes U_{p,q}^* &= \frac{1}{d_{\mathrm{A}}} \sum_{p=0}^{d_{\mathrm{A}}-1} \sum_{j=0}^{d_{\mathrm{A}}-1} |\varphi_{j\oplus p}\rangle\langle\varphi_j| \otimes |\varphi_{j\oplus p}\rangle\langle\varphi_j| \\ &= \frac{1}{d_{\mathrm{A}}} \sum_{p=0}^{d_{\mathrm{A}}-1} |\varphi_p\rangle|\varphi_p\rangle \sum_{j=0}^{d_{\mathrm{A}}-1} \langle\varphi_j|\langle\varphi_j| \\ &= \frac{1}{d_{\mathrm{A}}} |I_{\mathrm{A}}\rangle\!\rangle\langle\!\langle I_{\mathrm{A}}|. \end{aligned}$$

□

Corollary 14.10. *For any system* A *and every operator* $X : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$ *the following identity holds*

$$\frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} X U_{p,q}^{\dagger} = \mathrm{Tr}[X] \frac{I_{\mathrm{A}}}{d_{\mathrm{A}}} \tag{14.11}$$

Proof. Let us use the Vec isomorphism. We have

$$\begin{aligned} \left| \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} X U_{p,q}^{\dagger} \right\rangle\!\!\rangle &= \frac{1}{d_{\mathrm{A}}^2} \sum_{p,q=0}^{d_{\mathrm{A}}-1} U_{p,q} \otimes U_{p,q}^{*} |X\rangle\!\rangle \\ &= \frac{1}{d_{\mathrm{A}}} |I_{\mathrm{A}}\rangle\!\rangle \langle\!\langle I_{\mathrm{A}} | X \rangle\!\rangle \\ &= \frac{\mathrm{Tr}[X]}{d_{\mathrm{A}}} |I_{\mathrm{A}}\rangle\!\rangle. \end{aligned}$$

□

We can now prove the following result.

Lemma 14.11 (Monotonicity of $S(\rho\|\sigma)$ under marginalisation)**.** *Let* ρ *and* σ *be states of a composite system* AB*. The quantum relative entropy is monotonically decreasing when ignoring one system, namely*

$$S(\rho_{\mathrm{A}}\|\sigma_{\mathrm{A}}) \leq S(\rho_{\mathrm{AB}}\|\sigma_{\mathrm{AB}}). \tag{14.12}$$

Proof. Using the result (14.11) of corollary 14.10, we can write the partial trace Tr_{B} as follows

$$\tau_{\mathrm{A}} \otimes \frac{I}{d_{\mathrm{B}}} = \frac{1}{d_{\mathrm{B}}^2} \sum_{p,q=0}^{d_{\mathrm{B}}-1} (I_{\mathrm{A}} \otimes U_{p,q}) \tau_{\mathrm{AB}} (I_{\mathrm{A}} \otimes U_{p,q}^{\dagger}).$$

Using properties (13.1) and (13.2), along with the convexity of the relative entropy we have

$$\begin{aligned} S(\rho_{\mathrm{A}}\|\sigma_{\mathrm{A}}) &= S\left(\rho_{\mathrm{A}} \otimes \tfrac{I_{\mathrm{B}}}{d_{\mathrm{B}}} \| \sigma_{\mathrm{A}} \otimes \tfrac{I_{\mathrm{B}}}{d_{\mathrm{B}}}\right) \\ &\leq \frac{1}{d_{\mathrm{B}}^2} \sum_{p,q=0}^{d_{\mathrm{B}}-1} S((I_{\mathrm{A}} \otimes U_{p,q}) \rho_{\mathrm{AB}} (I_{\mathrm{A}} \otimes U_{p,q}^{\dagger}) \| (I_{\mathrm{A}} \otimes U_{p,q}) \sigma_{\mathrm{AB}} (I_{\mathrm{A}} \otimes U_{p,q}^{\dagger})) \\ &= S(\rho_{\mathrm{AB}}\|\sigma_{\mathrm{AB}}). \end{aligned}$$

□

We can finally prove Uhlmann's monotonicity theorem.

Theorem 14.12 (Uhlmann monotonicity of relative entropy)**.** *The quantum relative entropy is monotonically non-increasing under the action of a channel, namely for every* $\mathscr{E} : \mathrm{A} \to \mathrm{B}$ *one has*

$$S(\mathscr{E}(\rho)\|\mathscr{E}(\sigma)) \leq S(\rho\|\sigma). \tag{14.13}$$

Proof. The inequality in equation 14.13 can be simply proved using properties (13.1), (13.2), and (14.12). From the unitary dilation theorem for channels we know indeed that there exist systems C and D such that $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}} \simeq \mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{D}}$, a unitary operator $U : \mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{B}} \otimes \mathcal{H}_{\mathrm{D}}$, and a (pure) state $\nu \in \mathsf{St}_1(\mathrm{C})$, such that the channel $\mathscr{E}$ can be written as $\mathscr{E}(\rho) = \mathrm{Tr}_{\mathrm{D}}[U(\rho \otimes \nu)U^\dagger]$. Therefore, one has

$$\begin{aligned} S(\mathscr{E}(\rho)\|\mathscr{E}(\sigma)) &= S(\mathrm{Tr}_{\mathrm{D}}[U(\rho \otimes \nu)U^\dagger]\|\,\mathrm{Tr}_{\mathrm{D}}[U(\sigma \otimes \nu)U^\dagger]) \\ &\leq S(U(\rho \otimes \nu)U^\dagger\|U(\sigma \otimes \nu)U^\dagger) \\ &= S(\rho \otimes \nu\|\sigma \otimes \nu) \\ &= S(\rho\|\sigma). \end{aligned}$$

□

14.5 The Holevo bound

We now use Uhlmann's monotonicity theorem to prove the Holevo bound. This proof exploits the channel associated with a POVM. This notion is a particular case of channel associated with a quantum instrument. This kind of channel describes a transformation

$$\text{quantum} \to \text{quantum} + \text{classical},$$

namely a transformation that consists in a measurement with corresponding state reduction along with a classical record of the outcome. The precise mathematical definition is the following.

Definition 14.13 (Channel associated with an instrument)**.** Let $\{\mathscr{A}_i\}_{i\in\mathrm{X}} \subseteq \mathsf{QO}(\mathrm{A} \to \mathrm{B})$ be a quantum instrument. Let C be a system with $d_{\mathrm{C}} = |\mathrm{X}|$, and $\{\varphi\}_{i\in\mathrm{X}}$ an orthonormal basis for $\mathcal{H}_{\mathrm{C}} \simeq \mathbb{C}^{|\mathrm{X}|}$. The channel associated with the instrument $\{\mathscr{A}_i\}_{i\in\mathrm{X}}$ is the channel $\mathscr{C}_{\mathscr{A}} : \mathrm{A} \to \mathrm{BC}$

$$\mathscr{C}_{\mathscr{A}}(\rho) := \sum_{i\in\mathrm{X}} \mathscr{A}_i(\rho) \otimes |\varphi_i\rangle\langle\varphi_i|. \tag{14.14}$$

When we discard the reduced state of system B, and only consider the outcomes, the instrument becomes a POVM, and the associated channel becomes

$$\text{quantum} \to \text{classical}.$$

The precise definition of a channel associated with a POVM is the following.

Definition 14.14 (Channel associated with a POVM)**.** Let $\mathbf{P} := \{P_i\}_{i\in\mathrm{X}} \subseteq \mathsf{Eff}(\mathrm{A})$ be a POVM. Let C be a system with $d_{\mathrm{C}} = |\mathrm{X}|$, and $\{\varphi\}_{i\in\mathrm{X}}$ an orthonormal basis for $\mathcal{H}_{\mathrm{C}} \simeq \mathbb{C}^{|\mathrm{X}|}$. The channel associated with the POVM $\mathbf{P}$ is the channel $\mathscr{C}_{\mathbf{P}} : \mathrm{A} \to \mathrm{C}$

$$\mathscr{C}_{\mathbf{P}}(\rho) := \sum_{i\in\mathrm{X}} \mathrm{Tr}[\rho P_i]|\varphi_i\rangle\langle\varphi_i|. \tag{14.15}$$

Remark 19. As one can write the channel $\mathscr{C}_{\mathbf{P}}$ as

$$\mathscr{C}_{\mathbf{P}}(\rho) = \mathrm{Tr}_{\mathrm{A}}[(I_{\mathrm{C}} \otimes \rho^T) \sum_{i\in\mathrm{X}} |\varphi_i\rangle\langle\varphi_i| \otimes P_i^T],$$

it is immediate to deduce the Choi representative of $\mathscr{C}_P$ as

$$\mathsf{Ch}(\mathscr{C}_{\mathbf{P}}) = \sum_{i\in \mathrm{X}} |\varphi_i\rangle\langle\varphi_i| \otimes P_i^T.$$

Let now $\mathsf{E} = \{\rho_i\}_{i\in\mathrm{X}}$ be an ensemble. We define $\rho^{\mathsf{E}} := \sum_{i\in\mathrm{X}} \rho_i = \sum_{i\in\mathrm{X}} p_i\tilde{\rho}_i$, with $p_i := \mathrm{Tr}[\rho_i]$. The *Holevo χ-quantity* of the ensemble E is

$$\chi(\mathsf{E}) := \sum_{j\in\mathrm{X}} p_j S(\tilde{\rho}_j \| \rho^{\mathsf{E}}) = S(\rho^{\mathsf{E}}) - \sum_{j\in\mathrm{X}} p_j S(\tilde{\rho}_j). \tag{14.16}$$

Notice that the equality is obtained from the definition of the quantum relative entropy, considering that since $\rho^{\mathsf{E}} = \sum_{i\in\mathrm{X}} p_i\tilde{\rho}_i$ one has $\mathsf{Ker}(\rho^{\mathsf{E}}) \subseteq \mathsf{Ker}(\rho_i)$ for every $i \in \mathrm{X}$. The Holevo bound sets an upper bound on the accessible information $\mathsf{Acc}(\mathsf{E})$ in terms of the χ-quantity.

Theorem 14.15 (Holevo bound)**.** *For every POVM $\mathbf{P} = \{P_k\}_{k\in\mathrm{Y}}$ and any ensemble of states $\mathsf{E} = \{\rho_j\}_{j\in\mathrm{X}}$ the mutual information is bounded as follows*

$$I(\mathsf{E}, \mathbf{P}) \leq \chi(\mathsf{E}). \tag{14.17}$$

Proof. By applying Uhlmann's monotonicity Theorem 14.12 with the channel $\mathscr{C}_{\mathbf{P}}$ associated with the POVM $\mathbf{P}$ of Eq. (14.15) and the states $\tilde{\rho}_j$ and $\rho^{\mathsf{E}} = \sum_{j\in\mathrm{X}} p_j\tilde{\rho}_j$, and averaging on j, one obtains

$$\sum_j p_j S(\mathscr{C}_{\mathbf{P}}(\tilde{\rho}_j) \| \mathscr{C}_{\mathbf{P}}(\rho^{\mathsf{E}})) \leq \sum_{j\in\mathrm{X}} p_j S(\tilde{\rho}_j \| \rho^{\mathsf{E}}) = \chi(\mathsf{E}).$$

We now prove that the left hand side of the bound is just the mutual information $I(\mathsf{E}, \mathbf{P})$. Indeed, defining $p_{k|j} := \mathrm{Tr}[P_k\tilde{\rho}_j]$, $p_{k,j} := \mathrm{Tr}[\rho_j P_k] = p_j p_{k|j}$ and $p_k = \mathrm{Tr}[P_k\rho^{\mathsf{E}}]$ one has

$$S(\mathscr{C}_{\mathbf{P}}(\tilde{\rho}_j) \| \mathscr{C}_{\mathbf{P}}(\rho)) = S\left(\sum_{k\in\mathrm{Y}} p_{k|j} |\varphi_k\rangle\langle\varphi_k| \,\middle\|\, \sum_{k\in\mathrm{Y}} p_k |\varphi_k\rangle\langle\varphi_k|\right) = \sum_{k\in\mathrm{Y}} p_{k|j} \log_2 \frac{p_{k|j}}{p_k},$$

and finally, averaging over X this implies

$$\sum_{j\in\mathrm{X}} p_j S(\mathscr{C}_{\mathbf{P}}(\rho_j) \| \mathscr{C}_{\mathbf{P}}(\rho)) = \sum_{j\in\mathrm{X}} \sum_{k\in\mathrm{Y}} p_{k,j} \log_2 \frac{p_{k,j}}{p_j p_k} = I(\mathsf{E}, \mathbf{P}). \qquad \square$$

It should be emphasised that the Holevo bound is not generally saturated, namely there are ensembles of states for which the bound is not achieved for any POVM. The bound can indeed be achieved if and only if the states of the ensemble commute, namely

$$[\rho_i, \rho_j] = 0. \tag{14.18}$$

In particular, for orthogonal states ρ_j one has $I(\mathrm{X} : \mathrm{Y}) = H(\mathrm{X})$ and the POVM achieving the bound is $P_i = |\psi_i\rangle\langle\psi_i|$, where $\{\psi_i\}_{i=1}^{d_{\mathrm{A}}}$ is the orthonormal basis of common eigenvectors of the states ρ_j. Generally, one has the following chain of bounds

$$I(\mathsf{E}, \mathbf{P}) \leq \mathsf{Acc}(\mathsf{E}) \leq \chi(\mathsf{E}) \leq S(\rho^{\mathsf{E}}) \leq \log_2 d_{\mathrm{A}}. \tag{14.19}$$

The third bound is achieved with equality for ensembles made of pure states.

The Holevo theorem is one of the most relevant results in quantum information theory. Indeed, it establishes the maximal amount of information that can be extracted by a measurement on a quantum system. The last bound in Eq. (14.19) gives the maximum accessible information from a quantum system in any state and for any measurement, which is just the $\log_2$ of the Hilbert space dimension.

Notice that the rightmost bound tells us that the maximum amount of classical information that can be encoded in a quantum system A coincides with its dimension (measured in qubits): $\log_2 d_{\rm A}$. From this point of view, quantum systems do not provide any advantage versus classical systems.

Exercise 14.1

Show that the Holevo theorem is saturated for Abelian ensembles, and that for orthogonal states ones one has $I(X:Y) = H(X)$. Show that the optimal POVM is the von Neumann measurement on the orthonormal basis on which the states ρ_j are jointly diagonal. [It can be shown that condition (14.18) is also necessary for achieving the Holevo bound].

Answer of exercise 14.1

Consider the POVM $P_n = |n\rangle\langle n|$ where $\{|n\rangle\}$ denote the orthonormal basis jointly diagonalizing the states ρ_j, namely

$$\tilde{\rho}_j = \sum_n \lambda_n^{(j)} |n\rangle\langle n| \tag{14.20}$$

One has

$$\mathscr{C}_{\mathbf{P}}(\tilde{\rho}_j) = \sum_n \mathrm{Tr}[P_n \tilde{\rho}_j] |n\rangle\langle n| = \sum_n \lambda_n^{(j)} |n\rangle\langle n| = \tilde{\rho}_j, \tag{14.21}$$

which implies also $\mathscr{C}_{\mathbf{P}}(\rho) = \rho$, and then

$$S(\mathscr{C}_{\mathbf{P}}(\tilde{\rho}_j) \| \mathscr{C}_{\mathbf{P}}(\rho)) = S(\tilde{\rho}_j \| \rho). \tag{14.22}$$

Finally, this implies

$$I(\mathsf{E}, \mathbf{P}) = \sum_j p_j S(\mathscr{C}_{\mathbf{P}}(\tilde{\rho}_j) \| \mathscr{C}_{\mathbf{P}}(\rho)) = \sum_j p_j S(\tilde{\rho}_j \| \rho) = \chi(\mathsf{E}). \tag{14.23}$$

If the states are orthogonal, one has $\mathrm{Tr}[P_n \tilde{\rho}_j] = \lambda_n^{(j)}$ if $|n\rangle \in \mathsf{Supp}(\rho_j)$, and $\mathrm{Tr}[P_n \tilde{\rho}_j] = 0$ otherwise. Then, the random variable X with values j is a deterministic function of Y with values n, and thus $H(\mathrm{X}|\mathrm{Y}) = 0$. Finally, this implies that$I(\mathrm{X}:\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X})$.

Example 14.16 (Two equiprobable nonorthogonal pure states)**.** To get an idea of the bound at work, it is interesting to consider the mutual information when the states are non-orthogonal. For simplicity, let's consider just the case of two equiprobable pure states: $|\psi_0\rangle = |0\rangle$ and $|\psi_1\rangle = \cos\theta|0\rangle + \sin\theta|1\rangle$. The *a priori* density matrix of the ensemble is

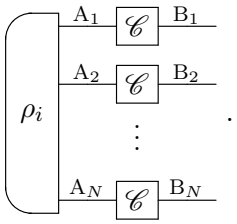
$$\rho^{\mathsf{E}} = \frac{1}{2} \begin{pmatrix} 1 + \cos^2\theta & \cos\theta \sin\theta \\ \cos\theta \sin\theta & \sin^2\theta \end{pmatrix}, \tag{14.24}$$

with eigenvalues $\lambda_\pm = \frac{1}{2}(1 \pm \cos\theta)$. Since the two states are pure, the Holevo theorem gives $\chi(\mathsf{E}) = S(\rho) = H_2(\lambda_+)$. The bound is maximized for $\theta = \frac{\pi}{2}$, namely when the states are orthogonal.

14.6 Holevo-Schumacher-Westmoreland theorem

Asymptotically for large n the Holevo bound for accessible information on output states of a channel $\mathscr{C}$ can be achieved as the classical transmission rate of $\mathscr{C}$. In particular, this is the case if one uses separable encodings, i.e. without exploiting entanglement. This is the content of Holevo-Schumacher-Westmoreland theorem. The theorem thus provides a lower bound on the classical capacity of $\mathscr{C}$, which could be in principle overcome by using entangled encodings. Let us see this result in more detail.

We are considering the following scenario



Let us consider the ensemble $\mathsf{E}_o^N := \{\mathscr{C}^{\otimes N}(\rho_i)\}_{i\in\mathrm{X}}$ of output states provided an input ensemble $\mathsf{E}^N := \{\rho_i\}_{i\in\mathrm{X}}$. The classical capacity of the quantum channel $\mathscr{C}$ can be defined as

$$C_C(\mathscr{C}) := \lim_{N\to\infty} \frac{\max_{\mathsf{E}^N,\mathbf{P}^N} I(\mathsf{E}_o^N, \mathbf{P}^N)}{N}.$$

This quantity is generally difficult to calculate. One can try to make the problem simpler by considering separable encodings. The Holevo-Schumacher-Westmoreland theorem indeed provides a nice expression for the classical capacity

$$C_C^P(\mathscr{C}) := \lim_{N\to\infty} \frac{\max_{\mathsf{E}^{\otimes N},\mathbf{P}^N} I(\mathsf{E}_o^{\otimes N}, \mathbf{P}^N)}{N} \tag{14.25}$$

under the restrictive hypothesis of product encoding, in terms of the single-use Holevo quantity $\max_{\mathsf{E}} \chi(\mathscr{C}(\mathsf{E}))$. This result then shifts the question about classical capacity to whether the assumption of separable encodings is really restrictive—which means that an entangled encoding provides better performances—, or the use of entanglement does not provide an advantage in communication of classical information.

In order to prove that $C_C^P(\mathscr{C}) \leq \max_{\mathsf{E}} \chi(\mathscr{C}(\mathsf{E}))$, let us define the ensemble $\mathsf{E}^{\otimes N}$ as follows

$$\mathsf{E}^{\otimes N} := \{p_{i_1,i_2,\ldots,i_N} \tilde{\rho}_{i_1} \otimes \tilde{\rho}_{i_2} \otimes \ldots \otimes \tilde{\rho}_{i_N}\}, \tag{14.26}$$

for arbitrary probability distribution $p_{i_1,i_2,\ldots,i_N}$. We define

$$\tilde{\rho}_{\mathbf{i}} := \tilde{\rho}_{i_1} \otimes \tilde{\rho}_{i_2} \otimes \ldots \otimes \tilde{\rho}_{i_N},$$
$$p_{\mathbf{i}} := p_{i_1,i_2,\ldots,i_N}.$$

If Alice is allowed to encode over tensor product states, by choosing any prior probability distribution, then she can use any ensemble of the form $\mathsf{E}^{\otimes N}$ defined in equation (14.26), obtaining an output ensemble of the same form

$$\begin{aligned}\mathsf{E}_o^{\otimes N} &= \{p_{i_1,i_2,\ldots,i_N}\mathscr{C}(\tilde{\rho}_{i_1})\otimes\mathscr{C}(\tilde{\rho}_{i_2})\otimes\ldots\otimes\mathscr{C}(\tilde{\rho}_{i_N})\}\\ &= \{p_{i_1,i_2,\ldots,i_N}\tilde{\rho}'_{i_1}\otimes\tilde{\rho}'_{i_2}\otimes\ldots\otimes\tilde{\rho}'_{i_N}\},\end{aligned}$$

where $\tilde{\rho}'_i := \mathscr{C}(\tilde{\rho}_i)$. We will also denote by E_l the l-th marginal ensemble of $\mathsf{E}_o^{\otimes N}$, namely

$$\mathsf{E}_l := \{p_{i_l}^{(l)}\tilde{\rho}'_{i_l}\}, \quad p_{i_l}^{(l)} := \sum_{i_1,i_2,\ldots,i_{l-1},i_{l+1},\ldots,i_N} p_{i_1,i_2,\ldots,i_N}.$$

Notice that

$$\begin{aligned}&\sum_{i_1,i_2,\ldots,i_N} p_{i_1,i_2,\ldots,i_n}S(\tilde{\rho}'_{i_1}\otimes\tilde{\rho}'_{i_2}\ldots\otimes\tilde{\rho}'_{i_N})\\ &= \sum_{i_1} p_{i_1}^{(1)}S(\tilde{\rho}'_{i_1}) + \sum_{i_2} p_{i_2}^{(2)}S(\tilde{\rho}'_{i_2}) + \ldots + \sum_{i_N} p_{i_N}^{(N)}S(\tilde{\rho}'_{i_N}),\end{aligned}$$

or, more compactly

$$\sum_{\mathbf{i}} p_{\mathbf{i}}S(\tilde{\rho}'_{\mathbf{i}}) = \sum_{l=1}^{N}\sum_{i_l} p_{i_l}^{(l)}S(\tilde{\rho}'_{i_l}).$$

The last identity, along with subadditivity of the von Neumann entropy

$$S(\rho^{\mathsf{E}_o^{\otimes N}}) \leq \sum_{l=1}^{N} S(\rho'^{\mathsf{E}_l}),$$

gives the bound

$$\chi(\mathsf{E}_o^{\otimes N}) \leq \sum_{l=1}^{N}\chi(\mathsf{E}_l) \leq N\max_{\mathsf{E}}\chi(\mathscr{C}(\mathsf{E})). \tag{14.27}$$

If we now define single-shot channel capacity as

$$C_C^{(1)}(\mathscr{C}) := \max_{\mathsf{E}}\chi(\mathscr{C}(\mathsf{E})), \tag{14.28}$$

it is clear that from the Holevo bound and from equation 14.27 one has

$$\begin{aligned}C_C^P(\mathscr{C}) &\leq \lim_{N\to\infty}\frac{1}{N}C_C^{(1)}(\mathscr{C}^{\otimes N})\\ &\leq C_C^{(1)}(\mathscr{C}).\end{aligned}$$

The Holevo-Schumacher-Westmoreland theorem assures that the capacity in terms of transmission rate is actually given by the Holevo quantity, and that the last bound is achievable asymptotically for large N, namely

$$\lim_{N\to\infty}\frac{1}{N}\chi(\mathsf{E}_o^{\otimes N}) = C_C^{(1)}(\mathscr{C}) = \max_{\mathsf{E}}\chi(\mathscr{C}(\mathsf{E})), \tag{14.29}$$

for suitable encoding and decoding on ensembles of factorised states.

The question about classical capacity without constraints then boils down to whether $\frac{1}{k}C_C^{(1)}(\mathscr{C}^{\otimes k}) > C_C^{(1)}(\mathscr{C})$. A weaker version of the question—i.e. whether $C_C^{(1)}(\mathscr{C}_1 \otimes \mathscr{C}_2) > C_C^{(1)}(\mathscr{C}_1) + C_C^{(1)}(\mathscr{C}_2)$—is known as the *superadditivity question*, and was solved for the positive by an example provided in 2009 by Hastings. This implies that entanglement can make the classical capacity better than the Holevo-Schumacher-Westmoreland rate. However, this also implies that in general one has no simple formula to express the classical capacity.

Chapter 15

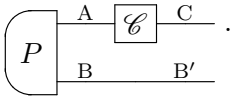
Lecture 17: Quantum information, compression, and Uhlmann fidelity

What is quantum information? In order to answer this question, let us analyse the task of transmitting an unknown quantum state ρ over a quantum channel $\mathscr{C}$, as in the following diagram

Let us consider, for example, the channel $\mathscr{C} : \mathrm{A} \to \mathrm{A}$ given by the following Kraus form, in terms of an orthonormal basis $\{\varphi_i\}_{i=1}^{d_\mathrm{A}}$

$$\mathscr{C}(\rho) = \sum_{i=1}^{d_\mathrm{A}} |\varphi_i\rangle\langle\varphi_i|\rho|\varphi_i\rangle\langle\varphi_i|.$$

One can easily verify that this channel has a very large classical capacity, essentially behaving as an ideal classical channel. However, it is also clear that $\mathscr{C}$ completely destroys superpositions of the states φ_i. Let us also consider the following scenario, in which the systems A and B of two agents Alice and Bob are entangled, and Alice sends her quantum system to Charlie C:



The question is: are the systems C and B′ still entangled? As we will see in the following, the answer to this question is strictly related to the possibility of sending superpositions. In other words, the ability to send quantum superpositions is equivalent to the ability to send entanglement: quantum information is equivalent to entanglement.

In quantitative terms, the amount of quantum information sent by a quantum channel represents the dimension of the largest Hilbert space whose superpositions can be reliably transmitted, or equivalently the amount of entanglement that can be reliably transmitted. This amount of quantum information is normally evaluated in *qubits*—the units of quantum information—and it corresponds then to the base two logarithm of the dimension.

The task that we consider in this chapter is *compression*.

We start from an example. Let $\mathsf{E} = \{p_i|\varphi_i\rangle\langle\varphi_i|\} \subseteq \mathsf{St}(\mathrm{A})$. Let us consider the state

$$(\rho^{\mathsf{E}})^{\otimes N} = \rho^{\mathsf{E}^{\otimes N}}.$$

The encoding $\mathscr{E}(\rho^{\mathsf{E}^{\otimes N}})$ generally provides a compressed state with smaller support than $(\rho^{\mathsf{E}})^{\otimes N}$—let us say $\dim \mathsf{Supp}(\mathscr{E}(\rho^{\mathsf{E}^{\otimes N}})) = d_{\mathrm{E}}(N)$. This will introduce some error probability in the decoding, but the optimal encoding will make the error arbitrarily small in the asymptotic limit $N \to \infty$. The compression ratio is given by

$$R = \lim_{N\to\infty} \frac{\log_2 d_{\mathrm{E}}(N)}{N}$$

We will prove Schumacher's theorem, that provides the optimal compression ratio for a reliable compression scheme as $d_{\mathrm{E}}(N) = 2^{NS(\rho^{\mathsf{E}})}$, i.e.

$$R = S(\rho^{\mathsf{E}}).$$

Therefore, the von Neumann entropy can be interpreted as the number of qubits of quantum information per use that are needed in average to store or transmit a quantum message described by an ensemble with average state ρ^{E}.

15.1 Quantum compression

As an example, suppose that the ensemble is binary, corresponding to the two pure non-orthogonal states

$$|\psi_0\rangle = \begin{pmatrix}1\\0\end{pmatrix},\ p_0 = \frac{1}{2}, \qquad |\psi_1\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},\ p_1 = \frac{1}{2},$$

so that the *a priori* state is given by

$$\rho = \frac{1}{4}\begin{pmatrix}3 & 1\\ 1 & 1\end{pmatrix}.$$

The eigenstates of ρ are given by

$$|0'\rangle = \begin{pmatrix}\cos\frac{\pi}{8}\\ \sin\frac{\pi}{8}\end{pmatrix}, \qquad |1'\rangle = \begin{pmatrix}\sin\frac{\pi}{8}\\ -\cos\frac{\pi}{8}\end{pmatrix},$$

corresponding to the eigenvalues $\lambda_0 = \cos^2\frac{\pi}{8}$ and $\lambda_1 = \sin^2\frac{\pi}{8}$. The eigenstate $|0'\rangle$ has large overlap with both input states

$$|\langle 0'|\psi_0\rangle|^2 = |\langle 0'|\psi_1\rangle|^2 = \cos^2\frac{\pi}{8} = .8535,$$

whereas $|1'\rangle$ have small overlap with both

$$|\langle 1'|\psi_0\rangle|^2 = |\langle 1'|\psi_1\rangle|^2 = \sin^2\frac{\pi}{8} = .1465.$$

Therefore, if Alice compresses to $\mathcal{H} = \mathbb{C}$ (0 qubits), Bob can only guess the input state. The best guess that Bob can make is $|\psi\rangle = |0'\rangle$. This observation is based on the overlap with the input states.

Notice that just guessing the state would have given an average overlap

$$\frac{1}{2}|\langle\psi_0|\psi_0\rangle|^2 + \frac{1}{2}|\langle\psi_0|\psi_1\rangle|^2 = 3/4. \tag{15.1}$$

Now, suppose that Alice needs to send a message of 3 letters to Bob, but she can afford using only two qubits. Can she have Bob reconstruct her state, and with which overlap? If she sends to Bob only two of the three letters and asks Bob to guess the third one, then Bob will receive two letters with overlap 1, and one letter with overlap .8535, hence with overall overlap .8535.

Is there a better procedure? Indeed, there is. By diagonalising ρ we decomposed the Hilbert space into a "high overlap subspace" $\mathsf{Span}\{|0'\rangle\}$ and an "low overlap subspace" $\mathsf{Span}\{|1'\rangle\}$. Similarly we can decompose the Hilbert space of three qubits. Indeed, for any input string of states $|\Psi\rangle = |\psi_{i_1}\rangle|\psi_{i_2}\rangle|\psi_{i_3}\rangle$ we have

$$\begin{aligned}
|\langle 0'0'0'|\Psi\rangle|^2 &= \cos^6\frac{\pi}{8} = .6219,\\
|\langle 1'0'0'|\Psi\rangle|^2 &= |\langle 0'1'0'|\Psi\rangle|^2 = |\langle 0'0'1'|\Psi\rangle|^2 = \cos^4\frac{\pi}{8}\sin^2\frac{\pi}{8} = .1067,\\
|\langle 1'1'0'|\Psi\rangle|^2 &= |\langle 0'1'1'|\Psi\rangle|^2 = |\langle 1'0'1'|\Psi\rangle|^2 = \cos^2\frac{\pi}{8}\sin^4\frac{\pi}{8} = .0183,\\
|\langle 1'1'1'|\Psi\rangle|^2 &= \sin^6\frac{\pi}{8} = .0031.
\end{aligned}$$

Thus, the "high overlap subspace" L will be the span of the four states

$$\mathsf{L} = \mathsf{Span}\{|0'0'0'\rangle, |1'0'0'\rangle, |0'1'0'\rangle, |0'0'1'\rangle\},$$

and the "low overlap subspace" will be the orthogonal complement $\mathsf{L}^\perp$. Now, if we want to project the state $|\Psi\rangle$ over L, the probability of success is

$$p = \mathrm{Tr}[P_{\mathsf{L}}|\Psi\rangle\langle\Psi|] = .6219 + 3(.1067) = .9419,$$

whereas the probability of projecting over the low overlap space is $1 - p = .0581$. The compression technique of Alice is precisely based on such projection. This can be achieved by the following measuring procedure. First she makes a unitary transformation that maps L and $\mathsf{L}^\perp$ into two-qubit spaces as follows

$$\begin{array}{ll}
U|0'0'0'\rangle = |0\rangle|0\rangle|0\rangle & U|1'1'1'\rangle = |0\rangle|0\rangle|1\rangle\\
U|1'0'0'\rangle = |1\rangle|0\rangle|0\rangle & U|1'1'0'\rangle = |1\rangle|0\rangle|1\rangle\\
U|0'1'0'\rangle = |0\rangle|1\rangle|0\rangle\text{'} & U|0'1'1'\rangle = |0\rangle|1\rangle|1\rangle\\
U|0'0'1'\rangle = |1\rangle|1\rangle|0\rangle & U|1'0'1'\rangle = |1\rangle|1\rangle|1\rangle
\end{array}$$

that is to say

$$|\Psi\rangle \in \mathsf{L},\ U|\Psi\rangle = |\cdot\rangle|\cdot\rangle|0\rangle, \quad |\Psi\rangle \in \mathsf{L}^\perp,\ U|\Psi\rangle = |\cdot\rangle|\cdot\rangle|1\rangle.$$

Then she measures the third qubit, and if she finds 0 (i.e. the input state has been successfully projected) she sends the remaining two qubits to Bob. Let's call the state of the two qubits that she sends $|\Psi_{\rm comp}\rangle$. When Bob receives the two qubits in the state $|\Psi_{\rm comp}\rangle$, he decompresses them by appending $|0\rangle$ and performing the inverse unitary transformation U^{-1}, obtaining

$$|\Psi'\rangle = U^{-1}|\Psi_{\rm comp}\rangle|0\rangle.$$

If Alice's measurement of the third qubit yields 1, this means that the state has been projected to $\mathsf{L}^\perp$, in which case, the best thing she can do is to send the state that Bob will decompress to the highest overlap state $|0'0'0'\rangle$, namely she sends the state $|\Psi_{\rm comp}\rangle = |0\rangle|0\rangle$ such that

$$|\Psi'\rangle = U^{-1}|\Psi_{\rm comp}\rangle|0\rangle = |0'0'0'\rangle.$$

The overall compression channel from Alice to Bob is then given by

$$\mathscr{C}(\rho) = P_{\mathsf{L}}\rho P_{\mathsf{L}} + {\rm Tr}[(I - P_{\mathsf{L}})\rho]|0'0'0'\rangle\langle 0'0'0'|,$$

with ρ being any 3-qubit state, and with $P_{\mathsf{L}} = U^\dagger(I \otimes I \otimes |0\rangle\langle 0|)U$. Thus one has

$$\mathscr{C}(|\Psi\rangle\langle\Psi|) = P_{\mathsf{L}}|\Psi\rangle\langle\Psi|P_{\mathsf{L}} + |0'0'0'\rangle\langle\Psi|(I - P_{\mathsf{L}})|\Psi\rangle\langle 0'0'0'|,$$

with overlap

$$\begin{aligned}\langle\Psi|\mathscr{C}(|\Psi\rangle\langle\Psi|)|\Psi\rangle &= \langle\Psi|P_{\mathsf{L}}|\Psi\rangle^2 + \langle\Psi|(I - P_{\mathsf{L}})|\Psi\rangle|\langle 0'0'0'|\Psi\rangle|^2 \\ &= .9419^2 + .0581 \times .6219 \\ &= .9234,\end{aligned}$$

which is much better than the value .8535 corresponding to the naive strategy of sending only two qubits and guessing the third one. Clearly, as we make the string of qubits longer, the overlap can be improved. According to the Schumacher compression theorem, we can shorten a long message by a factor

$$S(\rho) = H_2(\cos^2 \pi/8) = .6009$$

and asymptotically get the decoding close to the input state at will. Up to now, we considered the overlap as an intuitive measure of quality of the compression. However, we now need to provide a justification for this, through the introduction of *fidelity* as a figure of merit.

15.2 Fidelity

We start with the definition if fidelity, due to Uhlmann.

Definition 15.1 (Uhlmann's fidelity)**.** For any two normalized states ρ and σ in $\mathsf{St}_1(\mathrm{A})$ the Uhlmann's fidelity is defined as

$$\begin{aligned}F(\rho,\sigma) &:= {\rm Tr}[(\sigma^{\frac{1}{2}}\rho\sigma^{\frac{1}{2}})^{\frac{1}{2}}] \\ &= {\rm Tr}[|\sqrt{\rho}\sqrt{\sigma}|] \\ &= \|\sqrt{\rho}\sqrt{\sigma}\|_1. \end{aligned} \tag{15.2}$$

We remind that the trace-norm $\|X\|_1$ is defined as follows.

Definition 15.2 (Trace-norm)**.** Let $X : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$. The trace-norm $\|X\|_1$ of X is defined as

$$\|X\|_1 := \mathrm{Tr}[(X^\dagger X)^{\frac{1}{2}}]. \tag{15.3}$$

Lemma 15.3. *Let $A, U \in \mathcal{L}(\mathcal{H}_{\mathrm{A}})$, and U be unitary. The following bound holds*

$$|\,\mathrm{Tr}[AU]| \leq \mathrm{Tr}[|A|] = \|A\|_1, \tag{15.4}$$

with equality iff $A = U^\dagger |A|$.

Proof. Using the polar decomposition $A = V|A|$, one can write

$$|\,\mathrm{Tr}[AU]| = |\,\mathrm{Tr}[V|A|U]| = |\,\mathrm{Tr}[|A|^{\frac{1}{2}}|A|^{\frac{1}{2}}UV]|, \tag{15.5}$$

and using the Cauchy-Schwarz inequality for the Frobenius product $|\,\mathrm{Tr}[A^\dagger B]|^2 \leq \mathrm{Tr}[A^\dagger A]\,\mathrm{Tr}[B^\dagger B]$ one has

$$|\,\mathrm{Tr}[AU]| \leq \sqrt{\mathrm{Tr}[|A|]\,\mathrm{Tr}[V^\dagger U^\dagger |A| UV]} = \mathrm{Tr}[|A|], \tag{15.6}$$

with equality attained iff $V^\dagger U^\dagger |A| = |A|$, i.e. $U^\dagger |A| = V|A| = A$. □

We can now prove the following important theorems, providing two useful ways of calculating the fidelity

Theorem 15.4 (Uhlmann's theorem)**.** *For any two normalised states ρ and σ in $\mathsf{St}_1(\mathrm{A})$ and any fixed system* B*, the Uhlmann's fidelity is given by*

$$F(\rho, \sigma) = \max_{\Psi_\rho, \Psi_\sigma} |\langle\!\langle \Psi_\rho | \Psi_\sigma \rangle\!\rangle|, \tag{15.7}$$

where $\Psi_\rho, \Psi_\sigma \in \mathsf{St}_1(\mathrm{AB})$ denote any two purifications of ρ and σ, respectively.

Proof. First of all, in order to have a purification of σ and ρ in $\mathcal{H}_{\mathrm{A}} \otimes \mathcal{H}_{\mathrm{B}}$, it must be $d_B \geq \max\{\mathsf{rank}(\rho), \mathsf{rank}(\sigma)\}$. By the purification theorem, a purification $|\Psi_\rho\rangle\!\rangle\langle\!\langle\Psi_\rho|$ of ρ can be obtained by $|\Psi_\rho\rangle\!\rangle = |\rho^{\frac{1}{2}} U\rangle\!\rangle$, where $UU^\dagger \geq P_{\mathsf{Supp}(\rho)}$. Analogously for σ. Having fixed the system B, the purification theorem 11.2 then states that every other purification of ρ can be obtained from a given one $|\Psi_\rho\rangle\!\rangle = |\rho^{\frac{1}{2}} U\rangle\!\rangle$ by a unitary map W on system B, i.e. $|\Psi'_\rho\rangle\!\rangle = |\rho^{\frac{1}{2}} UW\rangle\!\rangle$ (the same for σ, with V and Z in place of U and W). Let us then consider the expression

$$|\langle\!\langle \Psi'_\rho | \Psi'_\sigma \rangle\!\rangle| = |\langle\!\langle \rho^{\frac{1}{2}} UW | \sigma^{\frac{1}{2}} VZ \rangle\!\rangle| = |\,\mathrm{Tr}[W^\dagger U^\dagger \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} VZ]| = |\,\mathrm{Tr}[ZW^\dagger U^\dagger \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} V]|.$$

By lemma 15.3, considering that $ZW^\dagger$ is unitary, one has

$$|\langle\!\langle \Psi'_\rho | \Psi'_\sigma \rangle\!\rangle| \leq \|U^\dagger \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} V\|_1 = \mathrm{Tr}[(V^\dagger \sigma^{\frac{1}{2}} \rho^{\frac{1}{2}} UU^\dagger \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} V)^{\frac{1}{2}}].$$

Now, since $UU^\dagger \geq P_{\mathsf{Supp}(\rho)}$, the above inequality becomes

$$|\langle\langle \Psi'_\rho | \Psi'_\sigma \rangle\rangle| \leq \mathrm{Tr}[(V^\dagger \sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}} V)^{\frac{1}{2}}].$$

Moreover, since $VV^\dagger \geq P_{\mathsf{Supp}(\sigma)}$, and $\mathsf{Supp}[(\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}}] \subseteq \mathsf{Supp}(\sigma)$, one has

$$V^\dagger (\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}} V V^\dagger (\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}} V = V^\dagger \sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}} V,$$

namely $(V^\dagger \sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}} V)^{\frac{1}{2}} = V^\dagger (\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}} V$. Finally, this gives

$$|\langle\langle \Psi'_\rho | \Psi'_\sigma \rangle\rangle| \leq \mathrm{Tr}[V V^\dagger (\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}}] = \mathrm{Tr}[(\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}})^{\frac{1}{2}}] = F(\rho, \sigma).$$

By lemma 15.3, the bound can be achieved by a suitable choice of $ZW^\dagger$. □

Theorem 15.5. *The Uhlmann's fidelity is given by the minimum over all possible measurements of the classical fidelity between the probability distributions of the two states, namely*

$$F(\rho, \sigma) = \min_{\mathbf{P}} F_C(\mathbf{p}, \mathbf{q}), \quad \mathbf{p} := \mathrm{Tr}[\mathbf{P}\rho], \ \mathbf{q} := \mathrm{Tr}[\mathbf{P}\sigma], \tag{15.8}$$

where

$$F_C(\mathbf{p}, \mathbf{q}) := \sum_n \sqrt{p_n q_n}. \tag{15.9}$$

Proof. From lemma 15.3 we learned that there exists a unitary U on $\mathcal{H}_\mathrm{A}$ such that $F(\rho, \sigma) = \mathrm{Tr}(\rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} U)$. Then

$$\begin{aligned} F(\rho, \sigma) &= \mathrm{Tr}(U \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}}) = \sum_n \mathrm{Tr}(U \rho^{\frac{1}{2}} \sqrt{P_n} \sqrt{P_n} \sigma^{\frac{1}{2}}) \\ &\leq \sum_n \sqrt{\mathrm{Tr}(\rho P_n) \, \mathrm{Tr}(\sigma P_n)} =: F_C(\mathbf{p}, \mathbf{q}) \end{aligned} \tag{15.10}$$

where we used the Cauchy-Schwarz inequality, along with the normalisation of the arbitrary POVM $\mathbf{P}$, $\sum_n P_n = I_\mathrm{A}$. The Cauchy-Schwarz inequality is saturated when, for every n, either

$$U \rho^{\frac{1}{2}} \sqrt{P_n} - \alpha_n \sigma^{\frac{1}{2}} \sqrt{P_n} = 0, \tag{15.11}$$

or $\sigma^{\frac{1}{2}} \sqrt{P_n} = 0$. Multiplying by $\sqrt{P_n}$ on the right hand side and by $P_{\mathsf{Supp}(\sigma)}$ on the left, one obtains the following necessary and sufficient condition

$$(P_{\mathsf{Supp}(\sigma)} U \rho^{\frac{1}{2}} - \alpha_n \sigma^{\frac{1}{2}}) P_n = 0. \tag{15.12}$$

Now, reminding that U is such that $U \rho^{\frac{1}{2}} \sigma^{\frac{1}{2}} = \sqrt{\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}}}$, we have

$$\sigma^{-\frac{1}{2}} U \rho^{\frac{1}{2}} P_{\mathsf{Supp}(\sigma)} = \sigma^{-\frac{1}{2}} \sqrt{\sigma^{\frac{1}{2}} \rho \sigma^{\frac{1}{2}}} \sigma^{-\frac{1}{2}} =: M. \tag{15.13}$$

The condition of Eq. (15.12) can be satisfied by taking the POVM $P_n = |n\rangle\langle n|$, with $M|n\rangle = \beta_n |n\rangle$, since by definition of M it is $\mathsf{Supp}(M) \subseteq \mathsf{Supp}(\sigma)$. Thus one obtains $\sigma^{\frac{1}{2}} M P_n = P_{\mathsf{Supp}(\sigma)} U \rho^{\frac{1}{2}} P_n = \sigma^{\frac{1}{2}} \beta_n P_n$, which means that condition (15.12) is satisfied taking $\alpha_n = \beta_n$. For $|n\rangle \in \mathsf{Ker}(\sigma)$, clearly $\mathrm{Tr}[\sigma P_n] = 0$, then the n-th term does not contribute to the sum in Eq. (15.10). □

Lemma 15.6. *The Uhlmann's fidelity satisfies the following properties:*

1. $0 \leq F(\rho,\sigma) \leq 1$, *and* $F(\rho,\sigma) = 1$ *iff* $\rho = \sigma$, *and* $F(\rho,\sigma) = 0$ *iff* $\mathsf{Supp}(\rho) \perp \mathsf{Supp}(\sigma)$;
2. $F(\rho,\sigma) = F(\sigma,\rho)$;
3. $F(\rho, |\psi\rangle\langle\psi|) = \langle\psi|\rho|\psi\rangle^{\frac{1}{2}}$.

Proof. Item 1 can be proved considering that $0 \leq (\rho^{\frac{1}{2}}\sigma\rho^{\frac{1}{2}})^{\frac{1}{2}}$, and thus $F(\rho,\sigma) \geq 0$. Moreover, by theorem 15.4 one clearly has $F(\rho,\sigma) \leq 1$. Now, the case $F(\rho,\sigma) = 0$ corresponds to the case $\|\rho^{\frac{1}{2}}\sigma^{\frac{1}{2}}\|_1 = 0$, i.e. $\rho^{\frac{1}{2}}\sigma^{\frac{1}{2}} = 0$. It is then easy to obtain $\rho\sigma = \sigma\rho = 0$. Moreover, again by theorem 15.4, if $F(\rho,\sigma) = 1$ then there exist purifications Ψ_ρ, Ψ_σ such that $\langle\!\langle \Psi_\rho | \Psi_\sigma \rangle\!\rangle = 1$, and thus $\Psi_\rho = \Psi_\sigma$, which in turn implies $\rho = \sigma$. Item 2 is an immediate consequence of theorem 15.4. Finally, for item 3 it is sufficient to observe that $(|\psi\rangle\langle\psi|)^{\frac{1}{2}} = |\psi\rangle\langle\psi|$ for every $|\psi\rangle\langle\psi| \in \mathsf{St}_1(\mathrm{A})$. □

Theorem 15.7 (Monotonicity of Uhlmann's fidelity)**.** *For any quantum channel* $\mathscr{E}$*, one has*

$$F[\mathscr{E}(\rho), \mathscr{E}(\sigma)] \geq F(\rho,\sigma). \tag{15.14}$$

Proof. Denote by $|\Psi_\rho\rangle\!\rangle$ and $|\Psi_\sigma\rangle\!\rangle$ any two purifications achieving the fidelity $F(\rho,\sigma)$. The reversible dilation theorem 11.14 then provides a purification for the output of the channel, and one has

$$F[\mathscr{E}(\rho), \mathscr{E}(\sigma)] \geq |\langle\!\langle \Psi_\rho | \langle 0 | (I_\mathrm{A} \otimes U^\dagger U) | \Psi_\sigma \rangle\!\rangle | 0 \rangle| = |\langle\!\langle \Psi_\rho | \Psi_\sigma \rangle\!\rangle| = F(\rho,\sigma). \qquad \square$$

Since the (partial) trace is a special channel, fidelity is monotonic under partial trace.

Corollary 15.8 (Monotonicity of Uhlmann's fidelity under partial trace)**.** *For any two bipartite states* $\rho_{\mathrm{AB}}, \sigma_{\mathrm{AB}}$ *one has*

$$F(\rho_\mathrm{A}, \sigma_\mathrm{A}) \geq F(\rho_{\mathrm{AB}}, \sigma_{\mathrm{AB}}). \tag{15.15}$$

Lemma 15.9 (Strong concavity of Uhlmann's fidelity)**.** *For any two ensembles of states* $\{r_i\tilde{\rho}_i\}$ *and* $\{s_i\tilde{\sigma}_i\}$*, one has*

$$F\left(\sum_i r_i\tilde{\rho}_i, \sum_i s_i\tilde{\sigma}_i\right) \geq \sum_i \sqrt{r_i s_i} F(\tilde{\rho}_i, \tilde{\sigma}_i). \tag{15.16}$$

Proof. Denote by $|\psi_i\rangle$ and $|\phi_i\rangle$ the vectors on $\mathcal{H}_\mathrm{A} \otimes \mathcal{H}_\mathrm{A}$ corresponding to purifications of ρ_i and σ_i such that $\langle\psi_i|\phi_i\rangle = F(\tilde{\rho}_i, \tilde{\sigma}_i)$. Then $|\Psi\rangle = \sum_i \sqrt{r_i}|\psi_i\rangle \otimes |i\rangle$ and $|\Phi\rangle = \sum_i \sqrt{s_i}|\phi_i\rangle \otimes |i\rangle$ are purifications of $\sum_i r_i\tilde{\rho}_i$ and $\sum_i s_i\tilde{\sigma}_i$, respectively. By Uhlmann's theorem one has

$$\begin{aligned} F\left(\sum_i r_i\tilde{\rho}_i, \sum_i s_i\tilde{\sigma}_i\right) \geq |\langle\!\langle \Psi | \Phi \rangle\!\rangle| &= \sum_i \sqrt{r_i s_i}\langle\psi_i|\phi_i\rangle \\ &= \sum_i \sqrt{r_i s_i} F(\tilde{\rho}_i, \tilde{\sigma}_i). \end{aligned} \qquad \square$$

Corollary 15.10. *The Uhlmann's fidelity is jointly concave in both entries.*

Proof. It is sufficient to take $r_i = s_i$ in equation (15.16), obtaining

$$F\left(\sum_i r_i\tilde{\rho}_i, \sum_i r_i\tilde{\sigma}_i\right) \geq \sum_i r_i F(\tilde{\rho}_i, \tilde{\sigma}_i). \qquad \square$$

The last result that we prove relates the Uhlmann fidelity to the *trace-norm distance* defined as

Definition 15.11 (Trace-norm distance)**.** Let $X, Y : \mathcal{H}_{\mathrm{A}} \to \mathcal{H}_{\mathrm{A}}$. The trace-norm distance $D(X, Y)$ is defined as

$$D(X, Y) := \frac{1}{2}\|X - Y\|_1. \tag{15.17}$$

We now show that the trace norm distance for states has an immediate operational interpretation in terms of discrimination probability. The scenario that we consider is the following. An agent is provided with a system A prepared in a state, which is either ρ_0 or ρ_1, both cases having prior probability $1/2$. The agent then performs a binary measurement $\{P_0, P_1\}$ with outcomes 0 and 1, summarising the discrimination strategy: upon reading 0 the agent will declare "ρ_0", and upon reading 1 they declare "ρ_1". The total error probability for the above strategy is thus given by

$$\begin{aligned}
p_{\mathrm{err}} &= \frac{1}{2}(\mathrm{Tr}[\rho_0 P_1] + \mathrm{Tr}[\rho_1 P_0]) \\
&= \frac{1}{2}(\mathrm{Tr}[\rho_0 P_1] + \mathrm{Tr}[\rho_1 (I_{\mathrm{A}} - P_1)]) \\
&= \frac{1}{2}(1 - \mathrm{Tr}[(\rho_1 - \rho_0) P_1]) \\
&= \frac{1}{2}(\mathrm{Tr}[\rho_0 (I_{\mathrm{A}} - P_0)] + \mathrm{Tr}[\rho_1 P_0]) \\
&= \frac{1}{2}(1 + \mathrm{Tr}[(\rho_1 - \rho_0) P_0]) \\
&= \frac{1}{4}(2 - \mathrm{Tr}[(\rho_1 - \rho_0)(P_1 - P_0)]).
\end{aligned}$$

Notice that $\mathrm{Tr}[(\rho_1 - \rho_0)(P_1 + P_0)] = \mathrm{Tr}[\rho_1 - \rho_0] = 0$, and thus $\mathrm{Tr}[(\rho_1 - \rho_0)(P_1 - P_0)] = 2\,\mathrm{Tr}[(\rho_1 - \rho_0)P_1]$. If we minimise the error probability for the above discrimination strategy, we obtain

$$p_{\mathrm{err}} = \frac{1}{2}(1 - \max_{0 \leq P \leq I_{\mathrm{A}}} \mathrm{Tr}[(\rho_1 - \rho_0) P]). \tag{15.18}$$

The next lemma shows that the above expression is related to the trace-norm distance as follows

$$p_{\mathrm{err}} = \frac{1}{4}(2 - \|\rho_1 - \rho_0\|_1]).$$

Lemma 15.12. *Let* $\rho_0, \rho_1 \in \mathsf{St}_1(\mathrm{A})$. *Then*

$$\frac{1}{2}\|\rho_0 - \rho_1\|_1 = \max_{0 \le P \le I_\mathrm{A}} \mathrm{Tr}[(\rho_0 - \rho_1)P]. \tag{15.19}$$

Proof. First, we remark that $\rho_0 - \rho_1$ is selfadjoint, and thus it can be expressed as $N_+ - N_-$, with $N_+ N_- = N_- N_+ = 0$, and $N_\pm \ge 0$. Now, clearly $|\rho_0 - \rho_1| = \{(N_+ - N_-)^2\}^{\frac{1}{2}} = (N_+^2 + N_-^2)^{\frac{1}{2}} = N_+ + N_-$, since $(N_+ + N_-)^2 = N_+^2 + N_-^2$. Moreover, since $\mathrm{Tr}[N_+ - N_-] = \mathrm{Tr}[\rho_0 - \rho_1] = 0$, we have $\mathrm{Tr}[N_+] = \mathrm{Tr}[N_-] = \frac{1}{2}\mathrm{Tr}[N_+ + N_-] = \frac{1}{2}\mathrm{Tr}[|\rho_0 - \rho_1|] = \frac{1}{2}\|\rho_0 - \rho_1\|_1$. Then, one has

$$\begin{aligned}
\mathrm{Tr}[(\rho_0 - \rho_1)P] &= \mathrm{Tr}[N_+ P] - \mathrm{Tr}[N_- P] \\
&\le \mathrm{Tr}[N_+ P] \\
&\le \mathrm{Tr}[N_+] \\
&= \frac{1}{2}\|\rho_0 - \rho_1\|_1.
\end{aligned}$$

Moreover, one can clearly achieve the upper bound by taking $P = P_+$, where P_+ is the projection on the support of N_+, i.e. the projection on the direct sum of all the eigenspaces of $\rho_0 - \rho_1$ with non-negative eigenvalues. Thus,

$$\max_{0 \le P \le I_\mathrm{A}} \mathrm{Tr}[(\rho_0 - \rho_1)P] = \frac{1}{2}\|\rho_0 - \rho_1\|_1. \qquad \square$$

The above result can be generalised as follows.

Theorem 15.13. *The trace-norm distance of two states* $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$ *is equal to*

$$\|\rho - \sigma\|_1 = \max_{\mathbf{P}} \sum_{i=1}^{l(\mathbf{P})} |\,\mathrm{Tr}[P_i(\rho - \sigma)]|, \tag{15.20}$$

where $\mathbf{P}$ *is a generic POVM with finitely many elements, and* $l(\mathbf{P})$ *is the number of elements of* $\mathbf{P}$.

Proof. Let us evaluate the argument of the maximum on r.h.s. in Eq. (15.20) separating the indices i of terms for which $\mathrm{Tr}[P_i(\rho - \sigma)] \ge 0$ from those for which $\mathrm{Tr}[P_i(\rho - \sigma)] < 0$, thus forming the sets S_+ and S_-:

$$\begin{aligned}
\sum_{i=1}^{l(\mathbf{P})} |\,\mathrm{Tr}[P_i(\rho - \sigma)]| &= \sum_{i \in \mathsf{S}_+} \mathrm{Tr}[P_i(\rho - \sigma)] - \sum_{i \in \mathsf{S}_-} \mathrm{Tr}[P_i(\rho - \sigma)] \\
&= \mathrm{Tr}[P_+(\rho - \sigma)] - \mathrm{Tr}[P_-(\rho - \sigma)],
\end{aligned}$$

where $P_\pm = \sum_{i \in \mathsf{S}_\pm} P_i$. Thus, it is not restrictive to evaluate the maximum on the set of binary POVMs. As a result, one obtains the thesis by virtue of lemma 15.12. $\square$

For generally mixed states, we have the following relevant theorem

Theorem 15.14 (Equivalence of trace-norm and Uhlmann's distances)**.** *The following double bound holds*

$$1 - F(\rho,\sigma) \leq \frac{1}{2}\|\rho-\sigma\|_1 \leq \sqrt{1-F(\rho,\sigma)^2}, \tag{15.21}$$

with both equalities iff $\rho = \sigma$*.*

Proof. By theorem 15.5, one has

$$1 - F(\rho,\sigma) = 1 - \min_{\mathbf{P}} F_C(\mathbf{p},\mathbf{q}) = 1 - F_C(\mathrm{Tr}[\rho\mathbf{P}], \mathrm{Tr}[\sigma\mathbf{P}]),$$

where $\mathbf{P}$ is the POVM minimising the classical fidelity $F_C(\mathrm{Tr}[\rho\mathbf{P}], \mathrm{Tr}[\sigma\mathbf{P}])$. We then have

$$\begin{aligned}
1 - F(\rho,\sigma) &= 1 - \sum_{i=1}^{l(\mathbf{P})} \sqrt{\mathrm{Tr}[\rho P_i]\,\mathrm{Tr}[\sigma P_i]} \\
&= \frac{1}{2}\sum_{i=1}^{l(\mathbf{P})} (\sqrt{\mathrm{Tr}[\rho P_i]} - \sqrt{\mathrm{Tr}[\sigma P_i]})^2 \\
&\leq \frac{1}{2}\sum_{i=1}^{l(\mathbf{P})} \{|(\sqrt{\mathrm{Tr}[\rho P_i]} - \sqrt{\mathrm{Tr}[\sigma P_i]})|(\sqrt{\mathrm{Tr}[\rho P_i]} + \sqrt{\mathrm{Tr}[\sigma P_i]})\} \\
&= \frac{1}{2}\sum_{i=1}^{l(\mathbf{P})} |(\mathrm{Tr}[\rho P_i] - \mathrm{Tr}[\sigma P_i])| \\
&\leq \frac{1}{2}\|\rho-\sigma\|_1,
\end{aligned}$$

where I^+ and I^- in the fifth line are the sets of values of i such that $(\mathrm{Tr}[\rho P_i] - \mathrm{Tr}[\sigma P_i]) \geq 0$ and $(\mathrm{Tr}[\rho P_i] - \mathrm{Tr}[\sigma P_i]) < 0$, respectively, and $P_0' = \sum_{i\in I^+} P_i$. In the second to last step, we used the observation that $\mathrm{Tr}[(\rho-\sigma)(P_0' - P_1')] = \mathrm{Tr}[(\rho-\sigma)(2P_0' - I)] = 2\,\mathrm{Tr}[(\rho-\sigma)P_0']$ Now, by theorem 15.13 we have

$$\begin{aligned}
\frac{1}{2}\|\rho-\sigma\|_1 =& \frac{1}{2}\max_{\mathbf{P}} \sum_{i=1}^{l(\mathbf{P})} |\,\mathrm{Tr}[P_i\rho] - \mathrm{Tr}[P_i\sigma]| \\
\leq& \frac{1}{2}\max_{\mathbf{P}} \sum_{i=1}^{l(\mathbf{P})} \{|\sqrt{\mathrm{Tr}[P_i\rho]} - \sqrt{\mathrm{Tr}[P_i\sigma]}||\sqrt{\mathrm{Tr}[P_i\rho]} + \sqrt{\mathrm{Tr}[P_i\sigma]}|\} \\
\leq& \frac{1}{2}\max_{\mathbf{P}} \left\{\sum_{i=1}^{l(\mathbf{P})} |\sqrt{\mathrm{Tr}[P_i\rho]} - \sqrt{\mathrm{Tr}[P_i\sigma]}|^2\right\}^{\frac{1}{2}} \left\{\sum_{i'=1}^{l(\mathbf{P})} |\sqrt{\mathrm{Tr}[P_{i'}\rho]} + \sqrt{\mathrm{Tr}[P_{i'}\sigma]}|^2\right\}^{\frac{1}{2}} \\
=& \frac{1}{2}\max_{\mathbf{P}} \{[2 - 2F_C(\mathbf{p},\mathbf{q})][2 + 2F_C(\mathbf{p},\mathbf{q})]\}^{\frac{1}{2}} \\
=& \max_{\mathbf{P}} [1 - F_C(\mathbf{p},\mathbf{q})^2]^{\frac{1}{2}} \\
=& \{1 - [\min_{\mathbf{P}} F_C(\mathbf{p},\mathbf{q})]^2\}^{\frac{1}{2}} \\
=& \sqrt{1 - F(\rho,\sigma)^2}.
\end{aligned}$$

□

Chapter 16

Lecture 19: Schumacher quantum source coding theorem

16.1 Entanglement fidelity

The square of the fidelity $\langle\!\langle\Phi|\mathscr{E}\otimes\mathscr{I}_{\rm B}(|\Phi\rangle\!\rangle\langle\!\langle\Phi|)|\Phi\rangle\!\rangle$ between input and output of a channel acting locally on a pure bipartite state $|\Phi\rangle\!\rangle\langle\!\langle\Phi|\in\mathsf{St}_1(\mathrm{AB})$ only depends on the channel $\mathscr{E}:\mathrm{A}\to\mathrm{A}$ and on the local state $\rho_{\rm A}:=\mathrm{Tr}_{\rm B}[|\Phi\rangle\!\rangle\langle\!\langle\Phi|]$. Indeed, for every purification $|\Phi'\rangle\!\rangle\langle\!\langle\Phi'|\in\mathsf{St}_1(\mathrm{AC})$ of the same state $\rho_{\rm A}$ one has

$$\begin{aligned}\rho_{\rm A}&=\mathrm{Tr}_{\rm C}[|\Phi'\rangle\!\rangle\langle\!\langle\Phi'|]\\&=\Phi'\Phi'^{\dagger}\\&=\Phi\Phi^{\dagger},\end{aligned}$$

and using the identity

$$\begin{aligned}|\Phi'\rangle\!\rangle\langle\!\langle\Phi'|&=(I_{\rm A}\otimes\Phi'^{T})|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|(I_{\rm A}\otimes\Phi'^{*})\\&=\mathscr{I}_{\rm A}\otimes\mathscr{N}_{\Phi'^{T}}(|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|),\end{aligned}$$

one can easily verify that

$$\begin{aligned}\langle\!\langle\Phi'|\mathscr{E}\otimes\mathscr{I}_{\rm C}(|\Phi'\rangle\!\rangle\langle\!\langle\Phi'|)|\Phi'\rangle\!\rangle&=\langle\!\langle\Phi'|(\mathscr{E}\otimes\mathscr{I}_{\rm C})(\mathscr{I}_{\rm A}\otimes\mathscr{N}_{\Phi'^{T}})(|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|)|\Phi'\rangle\!\rangle\\&=\langle\!\langle\Phi'|(\mathscr{I}_{\rm A}\otimes\mathscr{N}_{\Phi'^{T}})(\mathscr{E}\otimes\mathscr{I}_{\rm A'})(|I_{\rm A}\rangle\!\rangle\langle\!\langle I_{\rm A}|)|\Phi'\rangle\!\rangle\\&=\langle\!\langle\Phi'|(I_{\rm A}\otimes\Phi'^{T})\mathsf{Ch}(\mathscr{E})(I\otimes\Phi'^{*})|\Phi'\rangle\!\rangle\\&=\langle\!\langle\Phi'\Phi'^{\dagger}|\mathsf{Ch}(\mathscr{E})|\Phi'\Phi'^{\dagger}\rangle\!\rangle\\&=\langle\!\langle\rho|\mathsf{Ch}(\mathscr{E})|\rho\rangle\!\rangle.\end{aligned}$$

This observation motivates the following definition

Definition 16.1 (Entanglement fidelity of a channel)**.** The entanglement fidelity of a channel $\mathscr{E}$ for a state ρ is defined as the square of the fidelity between input and output of a channel acting locally on any purification $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$ of ρ, i.e. $\rho=\Phi_\rho\Phi_\rho^{\dagger}$

$$\begin{aligned}\mathcal{F}(\rho,\mathscr{E})&:=\langle\!\langle\Phi_\rho|\mathscr{E}\otimes\mathscr{I}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)|\Phi_\rho\rangle\!\rangle\\&=\langle\!\langle\rho|\mathsf{Ch}(\mathscr{E})|\rho\rangle\!\rangle.\end{aligned}\tag{16.1}$$

By the above observation, the entanglement fidelity $\mathcal{F}(\rho, \mathscr{E})$ is well defined, since it does not depend on the particular purification $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$ of ρ.

Lemma 16.2. *The following identity holds*

$$\mathcal{F}(\rho, \mathscr{E}) = \sum_i |\operatorname{Tr}[\rho E_i]|^2, \tag{16.2}$$

for an arbitrary Kraus decomposition $\mathscr{E}(\rho) = \sum_i E_i \rho E_i^\dagger$ *of* $\mathscr{E}$.

Proof. We remind that if $\mathscr{E}(\rho) = \sum_i E_i \rho E_i^\dagger$ of $\mathscr{E}$ is a Kraus decomposition of $\mathscr{E}$ then one has

$$\mathsf{Ch}(\mathscr{E}) = \sum_i |E_i\rangle\!\rangle\langle\!\langle E_i|,$$

and by the second line of the definition (16.1), one has

$$\mathcal{F}(\rho, \mathscr{E}) = \sum_i \langle\!\langle \rho | E_i \rangle\!\rangle \langle\!\langle E_i | \rho \rangle\!\rangle = \sum_i |\operatorname{Tr}[\rho E_i]|^2. \qquad \square$$

Lemma 16.3. *The entanglement fidelity satisfies the following properties*

1. $0 \leq \mathcal{F}(\rho, \mathscr{E}) \leq 1$;
2. $\mathcal{F}(\rho, \mathscr{E})$ *is linear in* $\mathscr{E}$;
3. *For pure states one has*

$$\mathcal{F}(|\psi\rangle\langle\psi|, \mathscr{E}) = F(|\psi\rangle\langle\psi|, \mathscr{E}(|\psi\rangle\langle\psi|))^2. \tag{16.3}$$

Proof. Item 1 can be proved reminding that $\mathcal{F}(\rho, \mathscr{E})$ is the square of the fidelity of two states, and the Uhlmann fidelity is positive and bounded by one. As to item 2, it is easy to observe that the second line in the definition (16.1) is linear in $\mathsf{Ch}(\mathscr{E})$, and the Choi map Ch is linear, too. Finally, let us consider item 3. An arbitrary purification of a pure state $|\psi\rangle\langle\psi|$ is a factorised state

$$|\Phi_\psi\rangle\!\rangle\langle\!\langle\Phi_\psi| = |\psi\rangle\langle\psi| \otimes |\eta\rangle\langle\eta|,$$

for some pure state $|\eta\rangle\langle\eta| \in \mathsf{St}_1(\mathrm{C})$. Thus,

$$\begin{aligned}
\mathcal{F}(|\psi\rangle\langle\psi|, \mathscr{E}) &= \langle\psi|\langle\eta|(\mathscr{E} \otimes \mathscr{I}_\mathrm{C})[|\psi\rangle\langle\psi| \otimes |\eta\rangle\langle\eta|]|\psi\rangle|\eta\rangle \\
&= \langle\psi|\mathscr{E}(|\psi\rangle\langle\psi|)|\psi\rangle|\langle\eta|\eta\rangle|^2 \\
&= \langle\psi|\mathscr{E}(|\psi\rangle\langle\psi|)|\psi\rangle \\
&= F(|\psi\rangle\langle\psi|, \mathscr{E}(|\psi\rangle\langle\psi|))^2. \qquad \square
\end{aligned}$$

Theorem 16.4. *The entanglement fidelity $\mathcal{F}(\rho, \mathscr{E})$ is convex in ρ and for every ensemble $\{p_i\tilde{\rho}_i\}$ satisfies the following bound*

$$\begin{aligned}\mathcal{F}\left(\sum_j p_j\tilde{\rho}_j, \mathscr{E}\right) &\leq \sum_j p_j \mathcal{F}(\tilde{\rho}_j, \mathscr{E}) \\ &\leq \sum_j p_j F(\tilde{\rho}_j, \mathscr{E}(\rho_j))^2. \end{aligned} \tag{16.4}$$

Proof. From the definition of entanglement fidelity and the monotonicity of the Uhlmann fidelity under partial trace, one has

$$\mathcal{F}(\rho, \mathscr{E}) = F(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|, \mathscr{E} \otimes \mathscr{J}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|))^2 \leq F(\rho, \mathscr{E}(\rho))^2.$$

Moreover, the entanglement fidelity is a convex function of ρ, since one has

$$\begin{aligned}\mathcal{F}(p\rho_1 + (1-p)\rho_2, \mathscr{E}) = {} & p^2 \langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_1 \rangle\!\rangle + (1-p)^2 \langle\!\langle \rho_2 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle \\ & + p(1-p)\langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle + p(1-p)\langle\!\langle \rho_2 | \mathsf{Ch}(\mathscr{E}) | \rho_1 \rangle\!\rangle\end{aligned}$$

for all $0 \leq p \leq 1$. By the Cauchy-Schwartz inequality, one has $|\langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle| \leq \sqrt{\langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_1 \rangle\!\rangle \langle\!\langle \rho_2 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle}$, and finally

$$\begin{aligned}\mathcal{F}(p\rho_1 + (1-p)\rho_2, \mathscr{E}) &\leq (p\sqrt{\langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_1 \rangle\!\rangle} + (1-p)\sqrt{\langle\!\langle \rho_2 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle})^2 \\ &\leq p\langle\!\langle \rho_1 | \mathsf{Ch}(\mathscr{E}) | \rho_1 \rangle\!\rangle + (1-p)\langle\!\langle \rho_2 | \mathsf{Ch}(\mathscr{E}) | \rho_2 \rangle\!\rangle \\ &= p\mathcal{F}(\rho_1, \mathscr{E}) + (1-p)\mathcal{F}(\rho_2, \mathscr{E}),\end{aligned}$$

where the second inequality is a straightforward consequence of convexity of $f(x) = x^2$ in the variable x. Therefore, one has

$$F\left(\sum_j p_j\rho_j, \mathscr{E}\right) \leq \sum_j p_j \mathcal{F}(\rho_j, \mathscr{E}) \leq \sum_j p_j F(\rho_j, \mathscr{E}(\rho_j))^2. \qquad \square$$

16.2 Refinement set

Let us say that the state σ *refines* ρ if the preparation of the state ρ can be achieved by an algorithm that involves preparation in state σ with some non-null probability. The remainder of the section is dedicated to the relation between entanglement fidelity of a channel $\mathscr{E}$ for state ρ and the behaviour of $\mathscr{E}$ on the set of every possible state σ that refines ρ. This set is called the *refinement set* of ρ.

Definition 16.5 (Refinement set)**.** The *refinement set* of a state $\rho \in \mathsf{St}_1(\mathrm{A})$ is the set $\mathsf{Ref}_1(\rho)$ of all the deterministic states σ that refine ρ, namely those states $\sigma \in \mathsf{St}_1(\mathrm{A})$ for which there exist $0 < p \leq 1$ and $\tau \in \mathsf{St}_1(\mathrm{A})$ such that

$$\rho = p\sigma + (1-p)\tau. \tag{16.5}$$

We now prove a theorem that regards the structure of the refinement set of a state ρ, and will be useful for the following results.

Theorem 16.6. *Let $\rho, \sigma \in \mathsf{St}_1(\mathrm{A})$ Then one has $\mathsf{Supp}(\sigma) \subseteq \mathsf{Supp}(\rho)$ iff $\sigma \in \mathsf{Ref}_1(\rho)$.*

Proof. Let us prove that the condition is necessary. Indeed, if $\rho = \lambda\sigma + (1-\lambda)\tau$ with $0 < \lambda \leq 1$ then $0 \leq \lambda\sigma \leq \rho$, and consequently

$$\langle\psi|\rho|\psi\rangle = 0 \;\Rightarrow \langle\psi|\sigma|\psi\rangle = 0, \quad \forall\psi \in \mathcal{H}_{\mathrm{A}}.$$

This implies that $\mathsf{Ker}(\rho) \subseteq \mathsf{Ker}(\sigma)$, i.e. $\mathsf{Supp}(\sigma) \subseteq \mathsf{Supp}(\rho)$. On the other hand, the condition is sufficient. Indeed, if $\mathsf{Ker}(\rho) \subseteq \mathsf{Ker}(\sigma)$, let ρ^{-1} denote the Moore-Penrose generalised inverse of ρ, namely $\rho\rho^{-1} = \rho^{-1}\rho = \Pi_{\mathsf{Supp}(\rho)}$, and $\mathsf{Supp}(\rho^{-1}) = \mathsf{Supp}(\rho)$. In this case, let us define

$$\lambda_M := \rho(\rho^{-\frac{1}{2}}\sigma\rho^{-\frac{1}{2}}),$$

where $\rho(X)$ denotes the spectral radius of X. Since $\rho^{-\frac{1}{2}}\sigma\rho^{-\frac{1}{2}}$ is selfadjoint, by lemma 13.4 it is $\lambda_M = \|\rho^{-\frac{1}{2}}\sigma\rho^{-\frac{1}{2}}\|_\infty$. Then by lemma 13.5 we have

$$\rho^{-\frac{1}{2}}\sigma\rho^{-\frac{1}{2}} \leq \lambda_M I_{\mathrm{A}}. \tag{16.6}$$

Now, multiplying on both sides by $\rho^{\frac{1}{2}}$, and reminding that $\mathsf{Supp}(\sigma) \subseteq \mathsf{Supp}(\rho)$, we have

$$\sigma \leq \lambda_M \rho.$$

There must then exist a positive operator $\tilde{\tau}$ such that

$$\rho = \frac{1}{\lambda_M}\sigma + \tilde{\tau},$$

and clearly $0 \leq \mathrm{Tr}[\tilde{\tau}] = 1 - \frac{1}{\lambda_M}$, thus $\lambda_M \geq 1$. Finally

$$\rho = \mu\sigma + (1-\mu)\tau,$$

with $\mu = \frac{1}{\lambda_M} \leq 1$ and $(1-\mu)\tau = \tilde{\tau}$. □

Then, the above theorem states that σ refines ρ if and only if its support is contained in the support of ρ. In other words,

$$\mathsf{Ref}_1(\rho) = \{\sigma \in \mathsf{St}_1(\mathrm{A}) | \mathsf{Supp}(\sigma) \subseteq \mathsf{Supp}(\rho)\}. \tag{16.7}$$

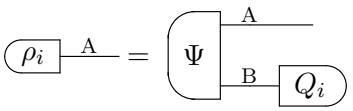
We now introduce the notion of a *refinement* of the state ρ, as follows

Definition 16.7. Let $\rho \in \mathsf{St}_1(\mathrm{A})$. The ensemble $\{\rho_i\}_{i\in\mathrm{I}}$ is a *refinement* of ρ if

$$\sum_{i\in\mathrm{I}} \rho_i = \rho. \tag{16.8}$$

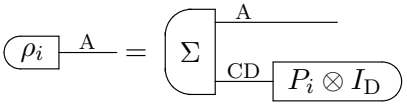
Clearly, for every ρ_i in a refinement of ρ, the deterministic state $\tilde{\rho}_i \coloneqq \rho_i / \operatorname{Tr}[\rho_i]$ refines ρ. The next result deals with refinements of a given state ρ.

Theorem 16.8. *Let $\rho \in \mathsf{St}_1(\mathrm{A})$. Then $\{\rho_i\}_{i\in\mathrm{I}}$ is a refinement of ρ iff for every purification $\Psi \in \mathsf{St}_1(\mathrm{AB})$ of ρ, there exists a POVM $\{Q_i\}_{i\in\mathrm{I}}$ such that*

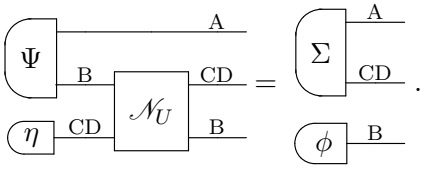


(16.9)

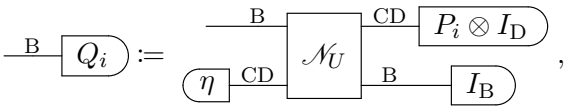
Proof. Eq. (16.9) is clearly sufficient for $\{\rho_i\}_{i\in\mathrm{I}}$ to be an ensemble. Let us now prove that it is also necessary. First, let us construct the state $P = \sum_{i\in\mathrm{I}} \rho_i \otimes |\varphi_i\rangle\langle\varphi_i| \in \mathsf{St}_1(\mathrm{AC})$, then consider its purification $\Sigma \in \mathsf{St}_1(\mathrm{ACD})$. Clearly, the POVM $\{P_i \otimes I_\mathrm{D}\}_{i\in\mathrm{I}}$ with $P_i \coloneqq |\varphi_i\rangle\langle\varphi_i|$ is such that



Now, let $\Psi \in \mathsf{St}_1(\mathrm{AB})$ be any purification of ρ. Clearly, given two pure states $\eta \in \mathsf{St}_1(\mathrm{CD})$ and $\phi \in \mathsf{St}_1(\mathrm{B})$, the pure states $\Sigma \otimes \phi$ and $\Psi \otimes \eta$ are purifications of ρ with the same purifying system BCD, thus by theorem 11.3 there exists a unitary operator U on $\mathcal{H}_\mathrm{BCD}$ such that

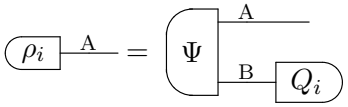


Thus, upon defining the POVM $\{Q_i\}_{i\in\mathrm{I}}$ as



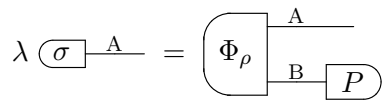
,

one obtains



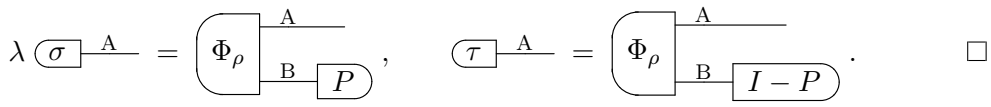
. □

Corollary 16.9 (Conditional marginal)**.** *Let $\sigma, \rho \in \mathsf{St}_1(\mathrm{A})$. Then $\sigma \in \mathsf{Ref}_1(\rho)$ iff there exists $0 < \lambda \leq 1$, and for any purification $\Phi_\rho \in \mathsf{St}_1(\mathrm{AB})$ there exists an effect $P \in \mathsf{Eff}(\mathrm{B})$, such that*



(16.10)

Proof. By definition $\sigma \in \mathsf{Ref}_1(\rho)$ iff there exists $0 < \lambda \leq 1$ and $\tau \in \mathsf{St}_1(\mathrm{A})$ such that the ensemble $\{\lambda\sigma, (1-\lambda)\tau\}$ refines ρ. By theorem 16.8, this is true iff for every purification $\Phi_\rho \in \mathsf{St}_1(\mathrm{AB})$ there exists a binary POVM $\{P, I-P\}$ such that one has



□

16.3 Equality upon input

The notion that we introduce now is that of *equality upon input*, and determines when two channels behave in the same way on some subspace, that is typically identified with the support of a state ρ.

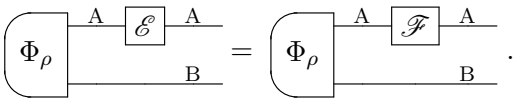
Definition 16.10 (Equality upon input)**.** Let $\mathscr{E}, \mathscr{F}$ be two quantum operations in $\mathsf{QO}(\mathrm{A} \to \mathrm{B})$. We say that $\mathscr{E}$ is equal to $\mathscr{F}$ upon input of ρ, and write

$$\mathscr{E} =_\rho \mathscr{F}, \tag{16.11}$$

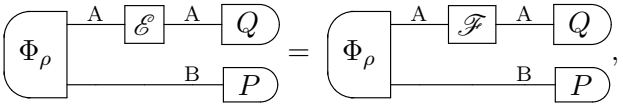
if $\mathscr{E}(\sigma) = \mathscr{F}(\sigma)$ for all $\sigma \in \mathsf{Ref}_1(\rho)$.

Lemma 16.11. *Let $\mathscr{E}, \mathscr{F} \in \mathsf{QO}(\mathrm{A} \to \mathrm{B})$. Then one has $\mathscr{E} =_\rho \mathscr{F}$ if and only if $\mathscr{E} \otimes \mathscr{I}_{\mathrm{C}}(\Phi_\rho) = \mathscr{F} \otimes \mathscr{I}_{\mathrm{C}}(\Phi_\rho)$ for any purification $\Phi_\rho \in \mathsf{St}_1(\mathrm{AC})$ of ρ.*

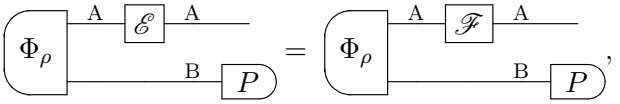
Proof. Indeed, $\mathscr{E} \otimes \mathscr{I}_{\mathrm{B}}(\Phi_\rho) = \mathscr{F} \otimes \mathscr{I}_{\mathrm{B}}(\Phi_\rho)$ if and only if



Then, equivalently we can write



for every $P \in \mathsf{Eff}(\mathrm{A})$ and $Q \in \mathsf{Eff}(\mathrm{B})$, which in turn is equivalent to



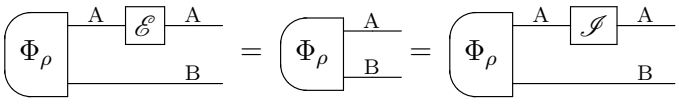
for every $P \in \mathsf{Eff}(\mathrm{A})$. Finally, by corollary 16.9 the latter is equivalent to $\lambda\mathscr{E}(\sigma) = \lambda\mathscr{F}(\sigma)$ for every $\sigma \in \mathsf{Ref}_1(\rho)$, and some $0 < \lambda \leq 1$ depending on σ, i.e. $\mathscr{E}(\sigma) = \mathscr{F}(\sigma)$ for every $\sigma \in \mathsf{Ref}_1(\rho)$. □

Now, the following result holds

Theorem 16.12. *Let $\mathscr{E} \in \mathsf{QC}(\mathrm{A} \to \mathrm{A})$ be a quantum channel. Then*

$$\mathcal{F}(\rho, \mathscr{E}) = 1 \quad \Leftrightarrow \quad \mathscr{E} =_\rho \mathscr{I}_{\mathrm{A}} \tag{16.12}$$

Proof. By definition, $\mathcal{F}(\rho, \mathscr{E}) = F[\Phi_\rho, (\mathscr{E} \otimes \mathscr{I}_{\mathrm{B}})(\Phi_\rho)]^2$, and thus $\mathcal{F}(\rho, \mathscr{E}) = 1$ if and only if



(16.13)

for any purification Φ_ρ of ρ. Finally, by lemma 16.11, this is equivalent to $\mathscr{E} =_\rho \mathscr{I}_{\mathrm{A}}$. □

What can we say when the entanglement fidelity is *approximately* unit? One expects that, in some sense, in that case $\mathscr{E}$ is approximately equal to $\mathscr{I}$ upon input of ρ. Is that true? The answer to this question is provided by the following result.

Theorem 16.13. *Approximate equality upon input For every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{F}(\rho, \mathscr{E}) \geq 1 - \delta$, then for every refinement $\{\rho_i\}_{i \in \mathrm{I}} = \{p_i \tilde{\rho}_i\}$ of ρ one has*

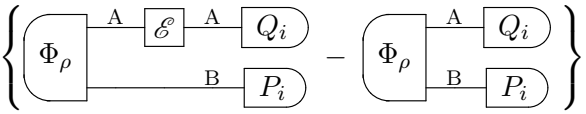
$$\frac{1}{2} \sum_{i \in \mathrm{I}} p_i \| \mathscr{E}(\tilde{\rho}_i) - \rho_i \|_1 \leq \varepsilon. \tag{16.14}$$

Proof. Let Φ_ρ be any purification of ρ. Then

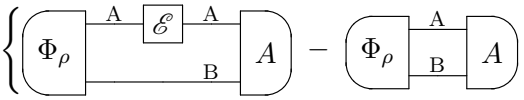
$$\frac{1}{2} \sum_{i \in \mathrm{I}} p_i \| \mathscr{E}(\tilde{\rho}_i) - \rho_i \|_1 = \sum_{i \in \mathrm{I}} p_i \operatorname{Tr}[\{\mathscr{E}(\tilde{\rho}_i) - \rho_i\} Q_i],$$

for suitably chosen effects Q_i. By theorem 16.8, this is equivalent to

$$\frac{1}{2} \sum_{i \in \mathrm{I}} p_i \| \mathscr{E}(\tilde{\rho}_i) - \rho_i \|_1 = \sum_{i \in \mathrm{I}}$$



$$= \left\{ \right.$$



$$\left. \right\},$$

where $A = \sum_{i \in \mathrm{I}} P_i \otimes Q_i$ is an effect, corresponding to a coarse graining of the POVM $\{P_i \otimes Q_i, P_i \otimes (I - Q_i)\}_{i \in \mathrm{I}}$. Then, by equation (15.19), this implies that

$$\frac{1}{2} \sum_{i \in \mathrm{I}} p_i \| \mathscr{E}(\tilde{\rho}_i) - \rho_i \|_1 \leq \frac{1}{2} \| \mathscr{E} \otimes \mathscr{I}(\Phi_\rho) - \Phi_\rho \|_1.$$

Now, by the Fuchs-van de Graaf inequality we have

$$\begin{aligned} \frac{1}{2} \sum_{i \in \mathrm{I}} p_i \| \mathscr{E}(\tilde{\rho}_i) - \rho_i \|_1 &\leq \sqrt{1 - F^2[\mathscr{E} \otimes \mathscr{I}(\Phi_\rho), \Phi_\rho]} \\ &= \sqrt{1 - \mathcal{F}(\rho, \mathscr{E})} \leq \sqrt{\delta}. \end{aligned}$$

□

16.4 Schumacher's quantum source coding theorem

The main idea at the basis of the quantum source coding theorem is a quantum version of the notions of typical sequence and typical set.

Definition 16.14 (Typical subspace)**.** Given a quantum state ρ with orthonormal decomposition $\rho = \sum_{x_i \in \mathsf{Rng}(\mathrm{X})} p_i |x_i\rangle\langle x_i|$, the ε-*typical* subspace $\mathsf{H}_{N,\varepsilon}(\rho)$ of $\mathcal{H}^{\otimes N}$ is defined as

$$\mathsf{H}_{N,\varepsilon}(\rho) := \mathsf{Span}\{|x_{\mathbf{i}}\rangle \mid x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})\}. \tag{16.15}$$

where $|x_{\mathbf{i}}\rangle := |x_{i_1}\rangle|x_{i_2}\rangle \ldots |x_{i_N}\rangle$, and X is the random variable with $\mathsf{Rng}(\mathrm{X}) = \{x_i\}$ and $\mathbb{P}_{\mathrm{X}}(x_i) := p_i$.

We remind that N i.i.d. copies of the state ρ have the following form

$$\rho^{\otimes N} = \sum_{x_{\mathbf{i}} \in \mathsf{Rng}(\mathrm{X})^N} p_{\mathbf{i}} |x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|,$$

where $p_{\mathbf{i}} = p_{i_1} p_{i_2} \ldots p_{i_N}$.

It is an immediate consequence of the definition of typical subspace that

$$\mathsf{H}_{N,\varepsilon}(\rho) := \mathsf{Span}\left\{ |x_{\mathbf{i}}\rangle \mid \left| \frac{1}{N} \log_2 \frac{1}{\mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}})} - S(\rho) \right| \leq \varepsilon \right\}.$$

We will also denote the projector on the typical subspace as

$$\begin{aligned} P_{N,\varepsilon}(\rho) &:= \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} |x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}| \\ &= \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} |x_{i_1}\rangle\langle x_{i_1}| \otimes |x_{i_2}\rangle\langle x_{i_2}| \otimes \ldots |x_{i_N}\rangle\langle x_{i_N}|, \end{aligned} \tag{16.16}$$

and we have that

$$\dim(\mathsf{H}_{N,\varepsilon}(\rho)) = \mathrm{Tr}[P_{N,\varepsilon}(\rho)] = |\mathsf{T}_{N,\varepsilon}(\mathrm{X})| \tag{16.17}$$

It is immediate to see that

$$\mathrm{Tr}[P_{N,\varepsilon}(\rho)\rho^{\otimes N}] = \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) = \mathbb{P}_{\mathrm{X}^N}[x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})]. \tag{16.18}$$

Theorem 16.15 (Typical subspace)**.** *The following statements hold:*

1. *For every $\varepsilon > 0$ and $\delta > 0$ there exists N_0 such that for every $N \geq N_0$*

$$\mathrm{Tr}[P_{N,\varepsilon}(\rho)\rho^{\otimes N}] \geq 1 - \delta. \tag{16.19}$$

2. *For every $\epsilon > 0$ and $\delta > 0$ there exists N_0 such that for every $N \geq N_0$ the dimension of the typical subspace $\mathsf{H}_{N,\varepsilon}(\rho)$ is bounded as*

$$(1-\delta)2^{N(S(\rho)-\varepsilon)} \leq \dim(\mathsf{H}_{N,\varepsilon}(\rho)) \leq 2^{N(S(\rho)+\varepsilon)} \tag{16.20}$$

3. *Let S_N denote for every N an arbitrary orthogonal projection on a subspace of $\mathcal{H}^{\otimes N}$ with dimension $\mathrm{Tr}(S_N) < 2^{NR}$, with $R < S(\rho)$ fixed. Then for every $\delta > 0$ there exists N_0 such that for every $N \geq N_0$ and every choice of S_N*

$$\mathrm{Tr}[S_N \rho^{\otimes N}] \leq \delta. \tag{16.21}$$

Proof. 1. This is an immediate consequence of identity (16.18) and of item 2 of the equipartition theorem 4.15.

2. Using identity (16.17), this is just item 3 of the equipartition theorem.

3. This is the only item that is not an immediate consequence of the asymptotic equipartition theorem: let us choose ε such that $R < S(\rho) - \varepsilon$, and write

$$\mathrm{Tr}[S_N \rho^{\otimes N}] = \mathrm{Tr}[S_N \rho^{\otimes N} P_{N,\varepsilon}(\rho)] + \mathrm{Tr}[S_N \rho^{\otimes N}(I - P_{N,\varepsilon}(\rho))].$$

Reminding that $\rho^{\otimes N}$ commutes with $P_{N,\varepsilon}(\rho)$, we have

$$\mathrm{Tr}[S_N \rho^{\otimes N} P_{N,\varepsilon}(\rho)] = \mathrm{Tr}[S_N P_{N,\varepsilon}(\rho) \rho^{\otimes N} P_{N,\varepsilon}(\rho)],$$

and since $P_{N,\varepsilon}(\rho)\rho^{\otimes N} P_{N,\varepsilon}(\rho) = \sum_{x_{\mathbf{i}} \in \mathsf{T}_{N,\varepsilon}(\mathrm{X})} \mathbb{P}_{\mathrm{X}^N}(x_{\mathbf{i}}) |x_{\mathbf{i}}\rangle\langle x_{\mathbf{i}}|$, using item 2 of the asymptotic equipartition theorem one has

$$P_{N,\varepsilon}(\rho)\rho^{\otimes N} P_{N,\varepsilon}(\rho) \leq 2^{-N(S(\rho)-\varepsilon)} I,$$

which implies

$$\begin{aligned} \mathrm{Tr}[S_N \rho^{\otimes N} P_{N,\varepsilon}(\rho)] &\leq \mathrm{Tr}[S_N] 2^{-N(S(\rho)-\epsilon)} \\ &\leq 2^{N[R-S(\rho)+\varepsilon]} \\ &\leq \frac{\delta}{2}, \end{aligned}$$

since the dimension of the subspace on which S_N projects is smaller than 2^{NR}. Notice that the bound above holds with the same δ for every S_N satisfying condition (16.21). On the other hand, using item 2 of this theorem, for every $\alpha > 0$ there is N_0 such that for $N \geq N_0$ we can bound the second term as

$$\mathrm{Tr}[S_N \rho^{\otimes N}(I - P_{N,\varepsilon})] \leq \mathrm{Tr}[\rho^{\otimes N}(I - P_{N,\varepsilon})] \leq 1 - (1 - \alpha) = \alpha.$$

Notice that also the second bound above holds with the same α independently of S_N. Therefore, in conclusion for sufficiently large N we have

$$\mathrm{Tr}[S_N \rho^{\otimes N}] \leq \alpha + 2^{N(R-S(\rho)+\varepsilon)}. \tag{16.22}$$

Since $R < S(\rho)$, for every $\delta > 0$ one can find N_0 such that for every $N \geq N_0$ the last bound can be made smaller than δ, by making $\alpha < \delta/2$ and $2^{N(R-S(\rho)+\varepsilon)} \leq \delta/2$.

□

Theorem 16.16 (Schumacher)**.** *For every i.i.d. quantum source with Hilbert space* $\mathcal{H}_{\mathrm{A}}$ *and prior state* $\rho \in \mathsf{St}_1(\mathrm{A})$*, for every* $\delta > 0$ *and* $R > S(\rho)$ *there exists* N_0 *such that for every* $N \geq N_0$ *one has a compression scheme* $\{\mathscr{E}^N, \mathscr{D}^N\}$ *with rate* R*, and* $\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N\mathscr{E}^N) \geq 1 - \delta$*. Conversely, for every* $R < S(\rho)$ *there is* $\delta \geq 0$ *such that for every compression scheme* $\{\mathscr{E}^N, \mathscr{D}^N\}$ *with rate* R *one has* $\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N\mathscr{E}^N) \leq \delta$*.*

Proof. Let $R > S(\rho)$. Take $\varepsilon > 0$ such that $R \geq S(\rho) + \varepsilon$. Then, by the typical subspace theorem one has that for every $\delta > 0$ there exists N_0 such that for every $N \geq N_0$

$$\mathrm{Tr}[P_{N,\varepsilon}(\rho)\rho^{\otimes N}] \geq 1 - \delta, \quad \dim(\mathsf{H}_{N,\varepsilon}(\rho)) \leq 2^{N(S(\rho)+\varepsilon)} \leq 2^{NR}.$$

Now, consider a subspace $\mathcal{H}^N_{\mathrm{comp}}$ of $\mathcal{H}_{\mathrm{A}}^{\otimes N}$ with dimension 2^{NR}, containing $\mathsf{H}_{N,\varepsilon}$, and use the following compression scheme.

1. Perform the von Neumann-Lüders measurement $\{P_{N,\varepsilon}(\rho), I - P_{N,\varepsilon}(\rho)\}$. If the outcome corresponding to $P_{N,\varepsilon}(\rho)$ occurs, then leave the state unchanged. Otherwise, if the outcome corresponding to $I - P_{N,\varepsilon}(\rho)$ occurs, replace the state by a standard state $|S\rangle\langle S|$, with $|S\rangle \in \mathsf{H}_{N,\varepsilon}(\rho)$. Finally, isometrically embed $\mathcal{H}^N_{\mathrm{comp}}$ into $\mathcal{H}_{\mathrm{B}^N} \simeq \mathcal{H}^N_{\mathrm{comp}}$ by the co-isometry $V^\dagger$ [with $VV^\dagger = P_{\mathcal{H}^N_{\mathrm{comp}}}$], complemented by an arbitrary channel from $(\mathcal{H}^N_{\mathrm{comp}})^\perp$ to $\mathcal{H}_{\mathrm{B}^N}$, e.g. $\mathscr{A}(\sigma) = |\psi_0\rangle\langle\psi_0|$. Such an encoding is described by the quantum channel

$$\mathscr{E}^N : \mathrm{A}^{\otimes N} \to \mathrm{B}^N, \quad \mathscr{E}^N(\sigma) := V^\dagger P_{N,\varepsilon}(\rho)\sigma P_{N,\varepsilon}(\rho)V + \sum_i A_i \sigma A_i^\dagger,$$

   where $A_i := V^\dagger|S\rangle\langle i|$, $\{|i\rangle\}$ denoting an orthonormal basis for $\mathsf{H}_{N,\varepsilon}(\rho)^\perp$.

2. The decoding is the channel $\mathscr{D}^N : \mathrm{B}^N \to \mathrm{A}^{\otimes N}$ which is just the isometry V inverting the embedding $V^\dagger$.

With the above compression scheme, we have according to the typical subspace theorem

$$\begin{aligned}
\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N\mathscr{E}^N) &= |\,\mathrm{Tr}[\rho^{\otimes N} P_{N,\varepsilon}(\rho)]|^2 + \sum_i |\,\mathrm{Tr}(\rho^{\otimes N} V A_i)|^2 \\
&\geq |\,\mathrm{Tr}[\rho^{\otimes N} P_{N,\varepsilon}(\rho)]|^2 \\
&\geq |1 - \delta|^2 \\
&\geq 1 - 2\delta,
\end{aligned}$$

where we used identity (16.2) in the first equality. Hence there exists a reliable compression scheme with rate R for $R > S(\rho)$.

To prove the converse, suppose now that $R < S(\rho)$, and consider any compression scheme $\{\mathscr{E}^N, \mathscr{D}^N\}$ which compresses states on $\mathcal{H}_{\mathrm{A}}^{\otimes N}$ to states on a system B^N with $\dim(\mathcal{H}_{\mathrm{B}^N}) = 2^{NR}$. In terms of the Kraus decompositions $\mathscr{E}^N(\sigma) = \sum_k C_k \sigma C_k^\dagger$ and $\mathscr{D}^N(\tau) = \sum_k D_k \tau D_k^\dagger$ we have

$$\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N\mathscr{E}^N) = \sum_{ij} |\,\mathrm{Tr}(D_i C_j \rho^{\otimes N})|^2.$$

In order to compress to $\mathcal{H}_{\mathrm{B}}^N$, the map $\mathscr{E}$ clearly has Kraus operators with $\mathsf{Rng}(C_j) \subseteq \mathcal{H}_{\mathrm{B}^N}$, and the map $\mathscr{D}$ has Kraus operators with support $\mathsf{Supp}(D_i) \subseteq \mathcal{H}_{\mathrm{B}^N}$. Denote by P_i the projection on $\mathsf{Rng}(D_i) \simeq \mathsf{Supp}(D_i)$, namely the space to which the compressed state is mapped by D_i. Thus $D_i C_j = P_i D_i C_j$, and

$$\begin{aligned}\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N \mathscr{E}^N) &= \sum_{ij} |\operatorname{Tr}[D_i C_j \rho^{\otimes N} P_i]|^2 \\ &\leq \sum_{ij} \operatorname{Tr}[D_i C_j \rho^{\otimes N} C_j^\dagger D_i^\dagger] \operatorname{Tr}[P_i \rho^{\otimes N}],\end{aligned}$$

where we applied the Cauchy-Schwarz inequality

$$|\operatorname{Tr}[A^\dagger B]|^2 \leq \operatorname{Tr}[A^\dagger A] \operatorname{Tr}[B^\dagger B],$$

with $A = \sqrt{\rho}^{\otimes N} C_j^\dagger D_i^\dagger$, and $B = \sqrt{\rho}^{\otimes N} P_i$. Since for every value of the index i one has $\operatorname{Tr}[P_i] = \dim(\mathsf{Supp}(P_i)) \leq \dim(\mathcal{H}_{\mathrm{B}^N}) = 2^{NR}$, applying item 3 of the typical subspace theorem we have that for every i and for any $\delta > 0$ there is N_0 such that for $N \geq N_0$ one has $\operatorname{Tr}[P_i \rho^{\otimes N}] \leq \delta$. Moreover, N_0 does not depend on the particular projections P_i as they all project on subspaces whose dimension is $\operatorname{Tr}[P_i] \leq 2^{NR}$. Thus the following bound holds *for all* schemes $(\mathscr{E}^N, \mathscr{D}^N)$ with rate R and $N \geq N_0$

$$\mathcal{F}(\rho^{\otimes N}, \mathscr{D}^N \mathscr{E}^N) \leq \delta \sum_{ij} \operatorname{Tr}(D_i C_j \rho^{\otimes N} C_j^\dagger D_i^\dagger) = \delta,$$

since both maps $\mathscr{D}^N, \mathscr{E}^N$ are trace-preserving. □

Chapter 17

Lecture 20: Entropy exchange and coherent information

17.1 Entropy exchange

17.1.1 Quantum data-processing theorem

Is there a quantum version of the data-processing theorem? If so, what is the quantity that plays the role of mutual information? Indeed, as we will see in this chapter, there exists a quantum version of the data-processing theorem, in which the role of mutual information is played by the *coherent information*. Before introducing the coherent information, we need to introduce the notion of *entropy exchange*.

Definition 17.1 (Entropy exchange)**.** Let $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}'$ be a quantum channel and $\rho \in \mathsf{St}_1(\mathrm{Q})$ be a state of the input system. We define the entropy exchange of $\mathscr{E}$ upon input of ρ as the following quantity

$$S(\rho, \mathscr{E}) := S(\mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)), \tag{17.1}$$

where $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho| \in \mathsf{St}_1(\mathrm{QR})$ is an arbitrary purification of ρ.

We remind that any purification $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$ of ρ can be written as

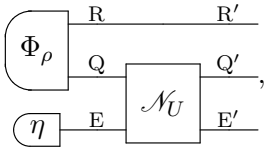
$$\begin{aligned} &|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho| = \mathscr{I}_{\mathrm{Q}} \otimes \mathscr{N}_V(|\Phi\rangle\!\rangle\langle\!\langle\Phi|), \\ &|\Phi\rangle\!\rangle = \sum_{j=1}^{r} \sqrt{p_i}|\psi_i\rangle|i\rangle, \\ &r := \mathsf{rank}(\rho), \quad \{|i\rangle\}_{i=1}^{r} \text{ o.n.b. in } \mathbb{C}^r =: \mathcal{H}_{\mathrm{C}}, \end{aligned} \tag{17.2}$$

where $|\Phi\rangle\!\rangle\langle\!\langle\Phi| \in \mathsf{St}_1(\mathrm{QC})$ is the *minimal purification* of ρ, and $V : \mathcal{H}_{\mathrm{C}} \to \mathcal{H}_{\mathrm{R}}$ is an isometry. Then, due to isometric invariance of von Neumann entropy, the entropy exchange is invariant under local isometries on the purifying system, namely

$$\begin{aligned} S(\mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)) &= S(\mathscr{E} \otimes \mathscr{N}_V(|\Phi\rangle\!\rangle\langle\!\langle\Phi|)) \\ &= S[\mathscr{I}_{\mathrm{Q}'} \otimes \mathscr{N}_V(\mathscr{E} \otimes \mathscr{I}_{\mathrm{C}}(|\Phi\rangle\!\rangle\langle\!\langle\Phi|))] \\ &= S(\mathscr{E} \otimes \mathscr{I}_{\mathrm{C}}(|\Phi\rangle\!\rangle\langle\!\langle\Phi|)). \end{aligned} \tag{17.3}$$

The definition of entropy exchange is then independent of the particular purification $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$.

Why the name entropy exchange? Because, if one considers that the channel $\mathscr{E}$ has a unitary dilation in terms of an interaction between the quantum system Q and an environment E in a pure state η, and one purifies the initial state by the reference system R in the pure state $\Phi_\rho := |\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$ as in the following diagram



(17.4)

then the state of $\mathrm{Q}'\mathrm{E}'\mathrm{R}'$ after the interaction is still pure, and one has

$$S(\mathrm{E}') = S(\mathrm{Q}'\mathrm{R}') = S(\rho, \mathscr{E}). \tag{17.5}$$

In other words, the entropy exchange is exactly the amount of entropy introduced ("exchanged") in the environment. Due to the Holevo bound, the classical mutual information between Alice preparing the system Q and any user Eve controlling the environment E is bounded by

$$I(\mathsf{E}, \mathbf{P}) \leq S(\rho^{\mathrm{E}'}) - \sum_{i \in \mathrm{X}} p_i S(\rho_i^{\mathrm{E}'}) \leq S(\rho^{\mathrm{E}'}) = S(\rho, \mathscr{E}).$$

A small entropy exchange is thus not only a signature of a reliable communication, but also of a reliably *private* one. This feature is distinctive of quantum communication theory compared to the classical one.

Notice that this amount of entropy depends neither on the particular unitary interaction with the environment used to model the channel $\mathscr{E}$, nor on the particular purification Φ_ρ of ρ, but only on the channel itself and the initial state ρ of the system Q.

It is interesting to write down explicitly the state of the environment after the application of the unitary transformation which dilates the channel on the system Q. Let $\{E_n\}_{n \in \mathrm{X}} \subseteq \mathcal{L}(\mathcal{H}_\mathrm{Q} \to \mathcal{H}_{\mathrm{Q}'})$ be the canonical Kraus of $\mathscr{E}$. Then the operator $V : \mathcal{H}_\mathrm{Q} \to \mathcal{H}_{\mathrm{Q}'\mathrm{E}'}$ defined as

$$V := \sum_{n \in \mathrm{X}} E_n \otimes |n\rangle,$$

for an orthonormal set $\{|n\rangle\}_{n=1}^{d'_\mathrm{E}}$ on $\mathcal{H}_{\mathrm{E}'}$ with $d_{\mathrm{E}'} \geq |\mathrm{X}|$ is isometric, since

$$V^\dagger V = \sum_{n \in \mathrm{X}} E_n^\dagger E_n = I_\mathrm{Q}.$$

The isometry V can be extended to a unitary transformation over system Q and environment E provided that $d_\mathrm{E} d_\mathrm{Q} = d_{\mathrm{E}'} d_{\mathrm{Q}'}$. Such a unitary must then satisfy

$$U|\psi\rangle \otimes |0\rangle = \sum_{n \in \mathrm{X}} E_n |\psi\rangle \otimes |n\rangle, \tag{17.6}$$

where $|\psi\rangle$ is an arbitrary vector in $\mathcal{H}_{\mathrm{Q}}$. For mixed input state we have

$$U(\rho \otimes |0\rangle\langle 0|)U^\dagger = \sum_{nm} E_n \rho E_m^\dagger \otimes |n\rangle\langle m|,$$

and the state of the environment E' after the interaction is given by

$$\sigma := \rho^{E'} = \sum_{nm} \sigma_{nm} |n\rangle\langle m|, \quad \sigma_{nm} := \mathrm{Tr}[E_n \rho E_m^\dagger].$$

Therefore, according to Eq. (17.5) the entropy exchange can be written as

$$S(\rho, \mathscr{E}) = S(\sigma) = -\mathrm{Tr}(\sigma \log_2 \sigma). \tag{17.7}$$

Exercise 17.1

Prove that the entropy exchange has the property that $S(d_{\mathrm{Q}}^{-1} I_{\mathrm{Q}}, \mathscr{E}) = 0$ if and only if the channel $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}'$ is an isometry.

Answer of exercise 17.1

The state of the environment E' after the interaction is given by

$$\sigma_{m,n} = \mathrm{Tr}[E_m \rho E_n^\dagger], \tag{17.8}$$

and if we diagonalise it we obtain

$$\delta_{i,j} q_i = \sum_{m,n=1}^{d_{\mathrm{E}'}} W_{i,m} \sigma_{m,n} W_{n,j}^\dagger = \mathrm{Tr}[F_i \rho F_j^\dagger],$$

where $F_i := \sum_{m=1}^{d_{\mathrm{E}'}} W_{i,m} E_m$ is an alternate Kraus for $\mathscr{E}$. The entropy exchange in terms of the canonical environment density matrix is given by

$$-\mathrm{Tr}(\sigma \log_2 \sigma) = -\sum_i q_i \log_2 q_i, \tag{17.9}$$

where $\{q_i\} := \mathsf{Spec}(\sigma)$, and

$$\mathrm{Tr}[F_i \rho F_i^\dagger] = q_i, \quad \mathrm{Tr}[F_i \rho F_j^\dagger] = 0 \quad i \neq j. \tag{17.10}$$

Now, $S(d_{\mathrm{Q}}^{-1} I_{\mathrm{Q}}, \mathscr{E}) = 0$ if and only if $q_i = \delta_{i i_0}$, namely

$$\frac{1}{d_{\mathrm{Q}}} \mathrm{Tr}[F_i F_i^\dagger] = \delta_{i,i_0}, \quad \mathrm{Tr}[F_i F_j^\dagger] = 0 \quad i \neq j. \tag{17.11}$$

Since $0 \leq F_{i_0}^\dagger F_{i_0} \leq I_{\mathrm{Q}}$, the only way to satisfy the first identity in 17.11 is to have $F_i^\dagger F_i = \delta_{i,i_0} I_{\mathrm{Q}}$, which implies that F_{i_0} is isometric and, since the map $\mathscr{E}$ is trace preserving, all the remaining Kraus operators are equal to zero.

Exercise 17.2

Prove concavity of $S(\rho, \mathscr{E})$ versus the channel $\mathscr{E}$.

Answer of exercise 17.2

The property follows immediately from concavity of the von Neumann entropy. Intuitively this means that mixing two different maps produces to an increase of the noise, and consequently to a larger inflation of entropy into the environment. Indeed, one has

$$\begin{aligned} S(\rho, p\mathscr{E}_1 + (1-p)\mathscr{E}_2) &= S[(p\mathscr{E}_1 + (1-p)\mathscr{E}_2) \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)] \\ &\leq pS(\mathscr{E}_1 \otimes \mathscr{I}_{\mathrm{Q}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)) + (1-p)S(\mathscr{E}_2 \otimes \mathscr{I}_{\mathrm{Q}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)) \\ &= pS(\rho, \mathscr{E}_1) + (1-p)S(\rho, \mathscr{E}_2). \end{aligned}$$

17.1.2 Quantum Fano inequality

A very interesting property of the entropy exchange is the quantum version of Fano's inequality

Theorem 17.2 (Quantum Fano's inequality)**.** *The following inequality holds*

$$S(\rho, \mathscr{E}) \leq H_2(F(\rho, \mathscr{E})) + (1 - F(\rho, \mathscr{E})) \log_2(d_{\mathrm{Q}} r - 1), \tag{17.12}$$

for any quantum channel $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}$ *and any state* $\rho \in \mathsf{St}_1(\mathrm{Q})$ *with rank* $r := \mathsf{rank}(\rho)$.

Proof. Consider an orthonormal basis $\{|B_i\rangle\!\rangle\}_{i=1}^{d_{\mathrm{Q}} r}$ for system and a reference R with $d_{\mathrm{R}} = r := \mathsf{rank}(\rho)$. Let $B_1 = \Phi_\rho$, and consider the probability distribution

$$p_i = \langle\!\langle B_i | \mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|) | B_i \rangle\!\rangle.$$

Using bound (12.27) we have

$$\begin{aligned} S(\rho, \mathscr{E}) &= S(\mathscr{E} \otimes \mathscr{I}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)) \\ &\leq S\left(\sum_{i=1}^{d_{\mathrm{Q}} r} |B_i\rangle\!\rangle\langle\!\langle B_i| \mathscr{E} \otimes \mathscr{I}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|) |B_i\rangle\!\rangle\langle\!\langle B_i|\right) \\ &= S\left(\sum_{i=1}^{d_{\mathrm{Q}} r} p_i |B_i\rangle\!\rangle\langle\!\langle B_i|\right) \\ &= H(\mathbf{p}). \end{aligned}$$

Then, it is easy to see that

$$\begin{aligned} H(\mathbf{p}) &= H_2(p_1) + (1-p_1) H(\mathbf{q}) \\ &\leq H_2(p_1) + (1-p_1) \log_2(d_{\mathrm{Q}} r - 1), \\ q_j &:= \frac{p_j}{1-p_1}, \quad 2 \leq j \leq d_{\mathrm{Q}} r \end{aligned}$$

On the other hand, since we have chosen $B_1 = \Phi_\rho$, we have

$$p_1 = \langle\!\langle \Phi_\rho | \mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|)|\Phi_\rho\rangle\!\rangle = F(\rho, \mathscr{E}),$$

Finally, by substitution we get the statement. □

The quantum Fano inequality allows us to conclude that a necessary condition for a reliable quantum communication, corresponding to $F(\rho, \mathscr{E}) > 1 - \delta$, is granted by a small entropy exchange, namely a small leakage of information in the environment.

17.2 Reversibility upon input

In this section we introduce the notion of reversibility upon input of a channel, which lies at the basis of the theory of quantum error correction.

Definition 17.3 (Channel reversible upon input)**.** A channel $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}'$ is reversible upon input of $\rho \in \mathsf{St}_1(\mathrm{Q})$ if there exists another channel $\mathscr{R} : \mathrm{Q}' \to \mathrm{Q}$ such that $\mathscr{R}\mathscr{E} =_\rho \mathscr{I}_{\mathrm{Q}}$, or equivalently $F(\rho, \mathscr{R}\mathscr{E}) = 1$.

According to corollary 16.11, the above definition is equivalent to the condition

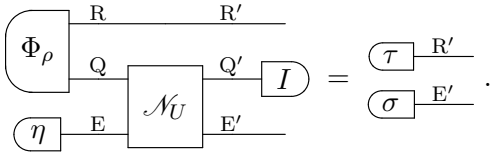
$$\mathscr{R}\mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|) = |\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|, \tag{17.13}$$

for any purification $|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho| \in \mathsf{St}_1(\mathrm{QR})$ of ρ.

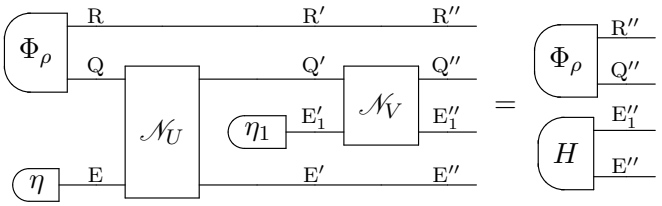
Theorem 17.4. *The channel $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}'$ is reversible upon input of $\rho \in \mathsf{St}_1(\mathrm{Q})$ if and only if the reference and the environment remain uncorrelated after the action of the channel. In formula, if we consider a general purification $\Phi_\rho := |\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho| \in \mathsf{St}_1(\mathrm{QR})$ of ρ, and a general unitary dilation (E, η, U) of $\mathscr{E}$, the equivalent condition for reversibility of $\mathscr{E}$ upon input of ρ is*

$$\rho_{\mathrm{R}'\mathrm{E}'} = \mathrm{Tr}_{\mathrm{Q}'}[(I_{\mathrm{R}} \otimes U)(\Phi_\rho \otimes \eta)(I_{\mathrm{R}} \otimes U^\dagger)] = \rho_{\mathrm{R}'} \otimes \rho_{\mathrm{E}'} = \tau \otimes \sigma. \tag{17.14}$$

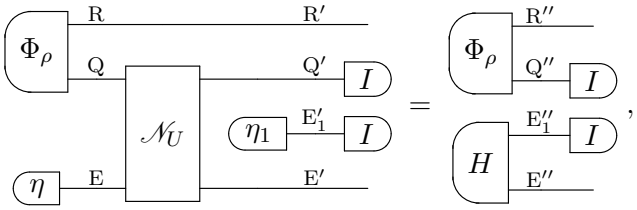
In diagrams



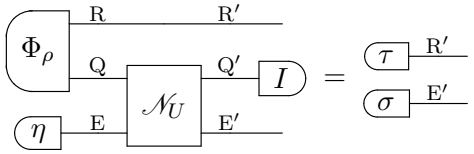
Proof. The converse statement can be proved by considering a reversible dilation $(\mathrm{E}_1', \eta_1, V)$ of the correcting channel $\mathscr{R}$. In this case, we have that $\mathscr{I}_{\mathrm{R}} \otimes \mathscr{R}\mathscr{E}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|) = |\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$, and then



for some pure state $H \in \mathsf{St}_1(\mathrm{E}_1''\mathrm{E}'')$, being the state of $R''Q''E_1''E''$ a purification of Φ_ρ. It is now easy to check that, applying the deterministic effect on $\mathrm{Q}''\mathrm{E}_1''$ we obtain

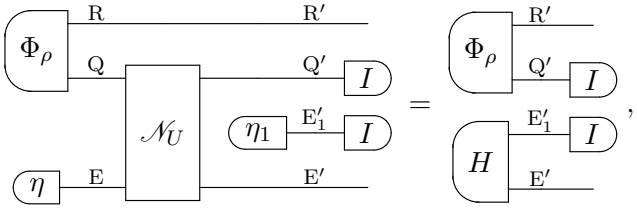


namely

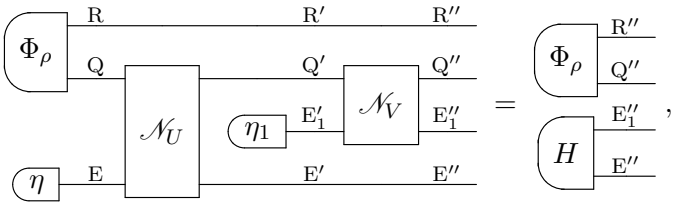


(17.15)

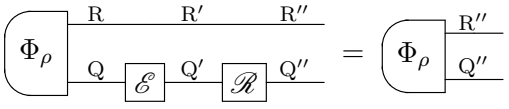
Starting from equation (17.15) one can now prove sufficiency. Indeed, if equation (17.15) holds, then we can introduce a pure state $\eta_1 \in \mathsf{St}_1(\mathrm{E}_1')$ such that



where $H \in \mathsf{St}_1(\mathrm{E}'\mathrm{E}_1')$ is a purification of σ. Thus, by uniqueness of purification there must exist a unitary $V \in \mathcal{L}(\mathcal{H}_{\mathrm{Q}'\mathrm{E}_1'})$ such that



and then discarding $\mathrm{E}''\mathrm{E}_1''$ we obtain



namely $\mathscr{R}\mathscr{E} \otimes \mathscr{I}_{\mathrm{R}}(|\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|) = |\Phi_\rho\rangle\!\rangle\langle\!\langle\Phi_\rho|$. □

We now provide a lemma that justifies the subject of the next section.

Lemma 17.5. *The condition* $\rho_{\mathrm{R}'\mathrm{E}'} = \rho_{\mathrm{R}'}\rho_{\mathrm{E}'} = \tau \otimes \sigma$ *is equivalent to*

$$S(\mathscr{E}(\rho)) = S(\rho) + S(\rho, \mathscr{E}). \tag{17.16}$$

Proof. Notice that since the state $\rho_{\mathrm{R}'\mathrm{Q}'\mathrm{E}'}$ is pure one has

$$S(\mathscr{E}(\rho)) = S(\mathrm{Q}') = S(\mathrm{R}'\mathrm{E}').$$

Moreover, by equation 17.5 one has

$$S(\rho, \mathscr{E}) = S(\mathrm{E}').$$

Finally,

$$S(\rho) = S(\mathrm{Q}) = S(\mathrm{R}) = S(\mathrm{R}'),$$

since QR is in a pure state, and $\rho_{\mathrm{R}'} = \rho_{\mathrm{R}}$. Now, identity (17.16) is equivalent to

$$S(\mathrm{R}'\mathrm{E}') = S(\mathrm{R}') + S(\mathrm{E}'),$$

which is satisfied if and only if R' and E' after the interaction remain in an uncorrelated state. □

17.3 Coherent information and quantum data processing theorem

Let us consider again the identity (17.16), which is equivalent to $S(\mathrm{R}'\mathrm{E}') = S(\mathrm{R}') + S(\mathrm{E}')$. In general, one has $S(\mathrm{R}'\mathrm{E}') \leq S(\mathrm{R}') + S(\mathrm{E}')$, namely $S(\mathscr{E}(\rho)) \leq S(\rho) + S(\rho, \mathscr{E})$. This observation suggests the next definition.

Definition 17.6 (Coherent information)**.** Let $\mathscr{E} : \mathrm{Q} \to \mathrm{Q}'$ be a quantum channel, and $\rho \in \mathsf{St}_1(\mathrm{Q})$ a state of the input system. The coherent information of $\mathscr{E}$ for state ρ is defined as

$$I(\rho, \mathscr{E}) := S(\mathscr{E}(\rho)) - S(\rho, \mathscr{E}). \tag{17.17}$$

Remark 20. Notice that in the quantum Fano inequality, the role of $H(\mathrm{X}|\mathrm{Y})$ is played by $S(\rho, \mathscr{E})$. In a similar way, we will now see that there is a quantum version of the data processing theorem, and the role of $I(\mathrm{X} : \mathrm{Y})$ in this case is played by the coherent information $I(\rho, \mathscr{E})$.

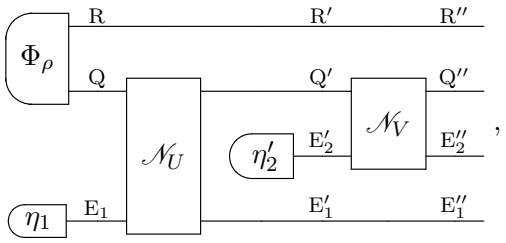
Theorem 17.7 (Quantum data-processing theorem)**.** *Let $\rho \in \mathsf{St}_1(\mathrm{A})$, $\mathscr{E}_1 : \mathrm{A} \to \mathrm{B}$ and $\mathscr{E}_2 : \mathrm{B} \to \mathrm{C}$ two quantum channels. The following chain of inequalities holds.*

$$I(\rho, \mathscr{E}_2\mathscr{E}_1) \leq I(\rho, \mathscr{E}_1) \leq S(\rho). \tag{17.18}$$

Moreover, $I(\rho, \mathscr{E}_1) = S(\rho)$ if and only if $\mathscr{E}_1$ is reversible upon input of ρ.

Proof. The second bound is just a special case of the first one for $\mathscr{E}_1 = \mathscr{I}_{\mathrm{A}}$. The fact that the last inequality is an identity if and only if $\mathscr{E}$ is reversible upon input of ρ is already proved in lemma 17.5. Therefore, it is now sufficient to prove the first inequality. We

consider a dilation $(\mathrm{E}_1, \eta_1, U)$ and $(\eta_2', \mathrm{E}_2', V)$ for $\mathscr{E}_1$ and $\mathscr{E}_2$, respectively. Including also the reference system R purifying ρ, we have



after the application of both channels. Using the strong subadditivity inequality for the von Neumann entropy

$$S(\mathrm{R}''\mathrm{E}_1''\mathrm{E}_2'') + S(\mathrm{E}_1'') \leq S(\mathrm{R}''\mathrm{E}_1'') + S(\mathrm{E}_1''\mathrm{E}_2''), \tag{17.19}$$

and since the total state of the four systems is pure, one has

$$S(\mathrm{R}''\mathrm{E}_1''\mathrm{E}_2'') = S(\mathrm{Q}'') = S(\mathscr{E}_2\mathscr{E}_1(\rho)).$$

Since R and E_1 are not involved in the second map, one has

$$S(\mathrm{R}''\mathrm{E}_1'') = S(\mathrm{R}'\mathrm{E}_1') = S(\mathrm{Q}') = S(\mathscr{E}_1(\rho)),$$

where we used also the fact that the state of $\mathrm{R}'\mathrm{E}_1'\mathrm{Q}'$ is pure. The environmental entropies are, by definition, entropy exchanges, namely

$$S(E_1'') = S(E_1') = S(\rho, \mathscr{E}_1), \quad S(E_1''E_2'') = S(\rho, \mathscr{E}_2\mathscr{E}_1),$$

and substitution into (17.19) gives

$$S(\mathscr{E}_2\mathscr{E}_1(\rho)) + S(\rho, \mathscr{E}_1) \leq S(\mathscr{E}_1(\rho)) + S(\rho, \mathscr{E}_2\mathscr{E}_1),$$

which is the inequality

$$I(\rho, \mathscr{E}_2\mathscr{E}_1) \leq I(\rho, \mathscr{E}_1). \qquad \square$$

Exercise 17.3

Show that the quantum data-processing theorem reduces to the classical one for classical states and channels.

Answer of exercise 17.3

Corresponding to a random variable X with probability distribution $p(x)$ we have the quantum state $\rho_X = \sum_x p(x)|x\rangle\langle x|$, with $\{|x\rangle\}$ orthonormal set, and with the identification of entropies

$$S(\rho_X) = H(X).$$

Corresponding to the classical channel with conditional probabilities $p(y|x)$ one has a quantum channel that prepares the state $|y\rangle\langle y|$ with probability $p(y|x)$ for input state $|x\rangle\langle x|$, namely

$$\mathscr{E}_1(\rho) = \sum_{xy} K_{xy}\rho K_{xy}^\dagger, \quad K_{xy} = \sqrt{p(y|x)}|y\rangle\langle x|.$$

The output state is then

$$\mathscr{E}_1(\rho) = \sum_{xy} p(xy)|y\rangle\langle y| \tag{17.20}$$

$$= \sum_{y} p(y)|y\rangle\langle y|, \tag{17.21}$$

and

$$S(\mathscr{E}_1(\rho)) = H(\mathrm{Y}).$$

The corresponding entropy exchange is given by

$$S(\rho, \mathscr{E}_1) = -\operatorname{Tr}[\sigma \log_2 \sigma],$$

where

$$\sigma_{(xy),(x'y')} = \operatorname{Tr}[K_{xy}\rho_X K_{x'y'}^\dagger] = \delta_{yy'}\delta_{xx'}p(x,y),$$

where $p(x,y) = p(y|x)p(x)$ is the joint probability of X, Y. Therefore, one obtains

$$S(\rho, \mathscr{E}_1) = H(X,Y).$$

Let now $\mathrm{X} \leftrightarrow \mathrm{Y} \leftrightarrow \mathrm{Z}$ be a Markov chain, and let us define the channel $\mathscr{E}_2$ as

$$\mathscr{E}_2(\sigma) = \sum_{yz} K_{yz}\rho K_{yz}^\dagger, \quad K_{yz} = \sqrt{p(z|y)}|z\rangle\langle y|.$$

We then have

$$\mathscr{E}_2\mathscr{E}_1(\tau) = \sum_{xyz} p(z|y)p(y|x)|z\rangle\langle x|\tau|x\rangle\langle z|.$$

By the Markov property, we have $p(z|y) = p(z|xy)$, then $p(z|y)p(y|x) = p(z|xy)p(y|x) = p(yz|x)$. This implies that

$$\mathscr{E}_2\mathscr{E}_1(\tau) = \sum_{xyz} p(yz|x)|z\rangle\langle x|\tau|x\rangle\langle z|$$

$$= \sum_{xz} p(z|x)|z\rangle\langle x|\tau|x\rangle\langle z|.$$

A Kraus form for $\mathscr{E}_2\mathscr{E}_1$ is then given by $H_{zx} = \sqrt{p(z|x)}|z\rangle\langle x|$, and

$$\mathscr{E}_2\mathscr{E}_1(\rho) = \sum_{z} p(z)|z\rangle\langle z|.$$

Repeating the argument that we used for $\mathscr{E}_1$, we obtain $S(\mathscr{E}_2\mathscr{E}_1(\rho)) = H(\mathrm{Z})$ and $S(\rho, \mathscr{E}_2\mathscr{E}_1) = H(\mathrm{X},\mathrm{Z})$. Upon substituting the above identifications in the quantum data-processing theorem

$$S(\mathscr{E}_2\mathscr{E}_1(\rho)) - S(\rho, \mathscr{E}_2\mathscr{E}_1) \leq S(\mathscr{E}_1(\rho)) - S(\rho, \mathscr{E}_1).$$

one obtains

$$H(Z) - H(X,Z) \leq H(Y) - H(X,Y),$$

and by adding $H(X)$ to both members we have

$$I(X:Z) \leq I(X:Y),$$

which is the classical data-processing theorem.

Chapter 18

Lecture 21: The theory of entanglement

18.1 Entanglement

We defined entanglement as the property of a state $\rho \in \mathsf{St}_1(\mathrm{AB})$ of being not separable.

Definition 18.1 (Separable state)**.** Let $\rho \in \mathsf{St}_1(\mathrm{AB})$. We say that ρ is *separable* if there exist $\{p_i\}_{i\in \mathrm{I}}$ $0 \leq p_i \ \forall \ i$, $\sum_{i\in \mathrm{I}} p_i = 1$, and $\{\sigma_i\}_{i\in \mathrm{I}} \subseteq \mathsf{St}_1(\mathrm{A})$, $\{\tau_i\}_{i\in \mathrm{I}} \subseteq \mathsf{St}_1(\mathrm{B})$, such that

$$\rho = \sum_{i\in \mathrm{I}} p_i(\sigma_i \otimes \tau_i).$$

Definition 18.2 (Entangled state)**.** Let $\rho \in \mathsf{St}_1(\mathrm{AB})$. We say that ρ is *entangled* if ρ is not separable.

The questions that we consider in these lectures is then: what do the properties of separability and entanglement mean operationally? Can we evaluate how much entanglement is there in a given state, and how? Let us start from the first question. Suppose Alice and Bob want to prepare a separable state $\rho = \sum_{i\in \mathrm{I}} p_i(\sigma_i \otimes \tau_i)$. Then Alice can use a random source to sample i with probability distribution $\{p_i\}_{i\in \mathrm{I}}$, and prepare σ_{i_0} upon reading $i = i_0$. Alice can then send a classical message to Bob, saying "I read $i = i_0$". Bob then prepares τ_{i_0}. At the end of the protocol, they prepared $\sigma_{i_0} \otimes \tau_{i_0}$, and this happens with probability p_{i_0}. Overall, they then prepare

$$\rho = \sum_{i\in \mathrm{I}} p_i(\sigma_i \otimes \tau_i)$$

Separable states are precisely those states that can be prepared by Local Operations and Classical Communication (LOCC). Entanglement is the main resource in quantum information—though not the only one. It enables:

- Teleportation
- Dense coding
- EPR - QKD

– Most quantum algorithms

Thus, we are not satisfied with a definition, we would also like to assess the *amount* of entanglement in a given state $\rho \in \mathsf{St}_1(\mathrm{AB})$. For this purpose, measures of entanglement are introduced. Every reasonable entanglement measure must obey two rules:

1. separable states have 0 entanglement
2. LOCC cannot increase entanglement

18.2 What does LOCC mean, precisely?

An instrument can be achieved by a LOCC (bipartite) protocol if it corresponds to a coarse graining of an instrument corresponding to:

– An instrument $\{\mathscr{A}_{i_1}\}_{i_1\in \mathrm{I}_1}$ of Alice, who keeps the output and sends the outcome i_1 (classical information) to Bob;

– An instrument $\{\mathscr{B}^{(i_1)}_{j_1}\}_{j_1\in \mathrm{J}_1}$ of Bob that is chosen depending on the information i_1. Bob keeps the output and sends the outcome j_1 to Alice;

– An instrument $\{\mathscr{A}^{(i_1,j_1)}_{i_2}\}_{i_2\in \mathrm{I}_2}$ of Alice chosen depending on i_1, j_1. Alice keeps the output and sends i_2 to Bob; ...

The protocol can last n rounds, with $n \in \mathbb{N}$. Notice that n can be arbitrarily large.

Definition 18.3. An instrument $(\mathscr{M}_i)_{i\in I}$ is *one-way* local with respect to Alice [Bob] if

$$\mathscr{M}_i = \mathscr{A}_i \otimes \mathscr{B}^{(i)} \qquad \Big[\mathscr{M}_i = \mathscr{A}^{(i)} \otimes \mathscr{B}_i\Big]$$

where $(\mathscr{A}_i)_{i\in \mathrm{I}}$ is an instrument $\big[(\mathscr{B}_i)_{i\in \mathrm{I}}\big]$ while $\forall\ i \in \mathrm{I}$, $\mathscr{B}^{(i)}$ is a channel $\big[\mathscr{A}^{(i)}\big]$

The interpretation in this case is easy: Alice [Bob] does the measurement described by the instrument $(\mathscr{A}_i)_{i\in \mathrm{I}}$ $\big[(\mathscr{B}_i)_{i\in \mathrm{I}}\big]$, sends i to Bob [Alice], who performs the channel $\mathscr{B}^{(i)}$ $[\mathscr{A}^{(i)}]$.

Definition 18.4. (LOCC-linked instrument). Let $(\mathscr{M}_i)_{i\in \mathrm{I}}$ be an instrument $\mathrm{AB} \to \mathrm{AB}$. We say that $(\mathscr{M}'_j)_{j\in \mathrm{J}}$ is LOCC-linked to $(\mathscr{M}_i)_{i\in \mathrm{I}}$ if there exists a collection of one-way local instruments $(\mathscr{F}^{(i)}_k)_{k\in \mathrm{K}}$ such that $(\mathscr{M}'_j)_{j\in \mathrm{J}}$ is a coarse-graining of $(\mathscr{F}^{(i)}_k \circ \mathscr{M}_i)_{i\in \mathrm{I}, k\in \mathrm{K}}$.

We can now give the following definitions:

– $(\mathscr{M}_i)_{i\in \mathrm{I}}$ is in LOCC_1 if $(\mathscr{M}_i)_{i\in \mathrm{I}}$ is one-way local followed by coarse-graining;

– $(\mathscr{M}_i)_{i\in \mathrm{I}}$ is in LOCC_n $(n \geq 2)$ if it is LOCC-linked to $(\mathscr{F}^{(i)}_j)_{j\in \mathrm{J}} \in \mathsf{LOCC}_{n-1}$ (it is implicitly assumed that if LOCC_1 is one-way local with respect to Alice [Bob], then the link with LOCC_2 is via a one-way local instrument with respect to Bob [Alice], etc.);

– $(\mathscr{M}_i)_{i\in \mathrm{I}}$ is in $\mathsf{LOCC}_{\mathbb{N}}$ if it belongs to LOCC_n for some $n \in \mathbb{N}$.

We then introduce a topology on instruments and close $\mathsf{LOCC}_{\mathbb{N}}$ with respect to such topology.

Definition 18.5 (Diamond norm)**.** We define the *diamond norm* of a linear map $\mathscr{M} : \mathrm{A} \to \mathrm{A}$ as follows

$$\|\mathscr{M}\|_\diamond := \max_{\mathrm{B};\rho\in\mathsf{St}_1(\mathrm{AB})} \|(\mathscr{M}\otimes\mathscr{I}_\mathrm{B})(\rho)\|_1$$

Definition 18.6 (Diamond norm distance)**.** We define the diamond norm distance of two instruments $(\mathscr{M}_i)_{i\in \mathrm{I}}$ and $(\mathscr{F}_i)_{i\in \mathrm{I}}$ as

$$D_\diamond(\underline{\mathscr{M}},\underline{\mathscr{F}}) := \max_{\mathrm{B};\rho\in\mathsf{St}_1(\mathrm{AB})} \sum_{i\in I} \|(\mathscr{M}_i\otimes\mathscr{I}_\mathrm{B} - \mathscr{F}_i\otimes\mathscr{I}_\mathrm{B})(\rho)\|_1$$

We can then first close $\mathsf{LOCC}_{\mathbb{N}}$ including protocols with infinitely many rounds, as follows:

– $(\mathscr{M}_i)_{i\in \mathrm{I}} \in \mathsf{LOCC}$ if there exists a sequence $(\mathscr{F}_i^{(n)})_{i\in \mathrm{I}_n} \subseteq \mathsf{LOCC}_{\mathbb{N}}$ such that

1. $(\mathscr{F}_i^{(n)})_{i\in I_n}$ is LOCC-linked to $(\mathscr{F}_j^{(n-1)})_{j\in I_{n-1}}$
2. $\forall n \in \mathbb{N}$ there exists a coarse-graining $(\mathscr{M}_i^{(n)})_{i\in I}$ of $(\mathscr{F}_i^{(n)})_{i\in I_n}$ such that $\forall\varepsilon > 0 \;\; \exists\, n_0 : \forall n \geq n_0 \; D_\diamond(\underline{\mathscr{M}}^{(n)},\underline{\mathscr{M}}) < \varepsilon$.

Finally, we can topologically close LOCC as follows:

– $(\mathscr{M}_i)_{i\in \mathrm{I}} \in \overline{\mathsf{LOCC}}$ if there exists a sequence $(\mathscr{M}_i^{(n)})_{i\in \mathrm{I}}$ in LOCC such that $\underline{\mathscr{M}}^{(n)}$ converges to $\underline{\mathscr{M}}$.

It holds that

$$\mathsf{LOCC}_1 \subset \mathsf{LOCC}_r \subset \mathsf{LOCC}_{r+1} \subset \mathsf{LOCC}_{\mathbb{N}} \subset \mathsf{LOCC} \subset \overline{\mathsf{LOCC}} \subset \mathsf{SEP},$$

where SEP is the set of all separable operations, that we define below.

Definition 18.7. An instrument $\underline{\mathscr{M}}$ belongs to SEP if $\underline{\mathscr{M}}$ is a coarse graining of $\underline{\mathscr{F}}$ with $\mathscr{F}_i = \mathscr{A}_i \otimes \mathscr{B}_i$ for a set of quantum operations $\mathscr{A}_i : \mathrm{A} \to \mathrm{A}$ and $\mathscr{B}_i : \mathrm{B} \to \mathrm{B}$.

While in the case of stases $\mathsf{LOCC}_1 \equiv \mathsf{SEP}$, for channels and instruments the hierarchy is non-trivial. More precisely, every inclusion is proved to be strict.

An interesting example of a SEP instrument that is not $\overline{\mathsf{LOCC}}$ is provided by the famous result "Quantum nonlocality without entanglement" [1]. The result states that in

$\mathbb{C}^3 \otimes \mathbb{C}^3$, the orthonormal basis:

$$
\begin{aligned}
|\psi_{11}\rangle &= |1\rangle \otimes |1\rangle \\
\left|\psi_{01}^{\pm}\right\rangle &= |0\rangle \otimes \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle) \\
\left|\phi_{01}^{\pm}\right\rangle &= \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle) \otimes |2\rangle \\
\left|\psi_{12}^{\pm}\right\rangle &= |2\rangle \otimes \frac{1}{\sqrt{2}}(|1\rangle \pm |2\rangle) \\
\left|\phi_{12}^{\pm}\right\rangle &= \frac{1}{\sqrt{2}}(|1\rangle \pm |2\rangle) \otimes |0\rangle
\end{aligned}
$$

cannot be discriminated by $\overline{\mathsf{LOCC}}$ protocols [1].

We omit the proof. However, notice that the result implies that the instrument $\{\mathscr{P}_i\}_{i\in \mathrm{I}}$ with $\mathscr{P}_i(\tau) = P_i \tau P_i$, where $\mathrm{I} = \{\phi_{01}^{\pm}, \psi_{01}^{\pm}, \phi_{12}^{\pm}, \psi_{12}^{\pm}, \psi_{11}\}$ and $P_i = |i\rangle\langle i|$, is not in $\overline{\mathsf{LOCC}}$, while it is in SEP.

For example:

$$P_{\phi_{01}^{\pm}} = P_{01}^{\pm} \otimes P_2, \text{ where } P_{ij}^{\pm} := |\eta_{ij}^{\pm}\rangle\langle\eta_{ij}^{\pm}|, |\eta_{ij}^{\pm}\rangle = \tfrac{1}{\sqrt{2}}(|i\rangle \pm |j\rangle) \text{ and } P_k = |k\rangle\langle k|.$$

If $\underline{\mathscr{P}}$ were in $\overline{\mathsf{LOCC}}$, this would indeed contradict the mentioned result of Ref. [1].

18.2.1 Entanglement and LOCC

Since $\overline{\mathsf{LOCC}} \subset \mathsf{SEP}$, it is clear that any $\overline{\mathsf{LOCC}}$ protocol will map separable states to separable states. Indeed, let $\rho = \sum_{i\in \mathrm{I}} p_i(\sigma_i \otimes \tau_i)$ and $\mathscr{C} = \sum_{j\in \mathrm{J}} \mathscr{A}_j \otimes \mathscr{B}_j$. Then

$$\mathscr{C}(\rho) = \sum_{i\in \mathrm{I}, j\in \mathrm{J}} p_i[\mathscr{A}_j(\sigma_i) \otimes \mathscr{B}_j(\tau_i)] = \sum_{i,j} \tilde{p}_{ij}(\tilde{\sigma}_{ij} \otimes \tilde{\tau}_{ij}),$$

where $\tilde{\sigma}_{ij} = \frac{\mathscr{A}_j(\sigma_i)}{\mathrm{Tr}[\mathscr{A}_j(\sigma_i)]}$, $\tilde{\tau}_{ij} = \frac{\mathscr{B}_j(\tau_i)}{\mathrm{Tr}[\mathscr{B}_j(\tau_i)]}$ and $\tilde{p}_{ij} = p_i \,\mathrm{Tr}[\mathscr{A}_j(\sigma_i)]\,\mathrm{Tr}[\mathscr{B}_j(\tau_i)]$.

Moreover, we expect that $\overline{\mathsf{LOCC}}$ protocols cannot increase the amount of entanglement of a given state. However, the last statement is not provable. Actually, the situation is exactly the opposite. We *define* measures of entanglement as to be non increasing under $\overline{\mathsf{LOCC}}$ maps. This is the paradigm of *resource theories*.

18.2.2 Resource theories

In any resource theory one starts defining *free operations*. Then set of free states must be invariant under free operations.

Normally, free operations form an OPT (operational probabilistic theory), as they have the following properties:

1. the sequence $\mathcal{A} \circ \mathcal{B}$ of any two free operations is free;

2. the parallel compositions $\mathcal{A} \otimes \mathcal{B}$ of free operations is free;

3. the identity map $\mathcal{I}_{\mathrm{A}}$ is considered as free for all systems A

One can define a relation between states as follows:

Definition 18.8. We say that $\rho \succ \sigma$ if $\exists \mathcal{E}$ free such that $\mathcal{E}(\rho) = \sigma$.

Since $\mathcal{I}_{\mathrm{A}}$ is free $\forall$A, clearly, the binary relation denoted as "$\succ$" is *reflexive*. Moreover, since $\forall \mathscr{A}, \mathscr{B}$ free, $\mathscr{B} \circ \mathscr{A}$ is free, $\succ$ is *transitive*. These two properties classify $\succ$ as a pseudo-order (or pre-order) relation[1]. One can then introduce the relation

Definition 18.9. $\rho \sim \sigma$ if $\rho \succ \sigma$ and $\sigma \succ \rho$.

$\rho \sim \sigma$ is an equivalence relation. One can then order equivalence classes ρ by a (partial) order: $[\rho] \succeq [\sigma]$ if $\rho \succeq \sigma$. One can easily check that $\succeq$ is well defined: the condition $[\rho] \succeq [\sigma]$ if $\rho \succeq \sigma$ holds independently of the elements $\rho \in [\rho]$ and $\sigma \in [\sigma]$. It is also easy to check that $\succeq$ is a *partial ordering* relation[2]. Thus, resource theories allow one to introduce an order relation on the set of states according to the "amount" of resource that the states contain. In the resource theory of entanglement, free operations are LOCC operations. In this way, entanglement is essentially defined as that resource that consists in those quantum correlations that cannot be produced by just exchanging classical information.

Any measure of the resource of interest must be monotonic under the above defined ordering.

18.2.3 Maximally entangled states

The ordering of states under the relation $\succeq$ naturally poses a question: are there states that represent "maximal" resources? Let us start considering two-qubit states. In this case we can suspect that maximally entangled states are maximal resources. Maximally entangled states are all pure states of the form

$$\rho = \frac{1}{2}|U\rangle\!\rangle\langle\!\langle U|$$
$$U^\dagger U = UU^\dagger = I_{\mathrm{A}}$$

Notice that given $\frac{1}{\sqrt{2}}|U\rangle\!\rangle$, $\frac{1}{\sqrt{2}}|W\rangle\!\rangle$, one can always find V such that $V^\dagger V = VV^\dagger = I_{\mathrm{A}}$, and

$$(V \otimes I_{\mathrm{A}'})\frac{1}{\sqrt{2}}|U\rangle\!\rangle = \frac{1}{\sqrt{2}}|W\rangle\!\rangle$$

Indeed, the choice of $V \coloneqq WU^\dagger$ will do the job. This implies that there exist LOCC maps that convert any two-qubit maximally entangled state into any other. In the following we will consider the singlet state $\frac{1}{\sqrt{2}}|i\sigma_y\rangle\!\rangle$ as our reference maximally entangled state.

[1] We remind that an ordering relation has the properties of being reflexive, antisymmetric and transitive.

[2] Reflexive, antisymmetric, transitive; however, differently from total ordering, given two states ρ, ρ' it can happen that neither $[\rho]_\sim \succeq [\rho']_\sim$, nor $[\rho']_\sim \succeq [\rho]_\sim$.

However, we now need to show that every two qubit state can be obtained via LOCC from a maximally entangled one. First, let $\rho = |\psi\rangle\langle\psi| \otimes |\phi\rangle\langle\phi|$. The LOCC map $\mathcal{Q}_\psi \circ \mathcal{Q}_\phi$, with

$$\mathcal{Q}_\eta(\sigma) \coloneqq \mathrm{Tr}[\sigma]\,|\eta\rangle\langle\eta|$$

will clearly turn $\frac{1}{\sqrt{2}}|i\sigma_y\rangle\!\rangle$ into $|\psi\rangle\langle\psi| \otimes |\phi\rangle\langle\phi|$. Let now $\rho = |F\rangle\!\rangle\langle\!\langle F|$ with $|F\rangle\!\rangle$ pure, entangled but not maximally entangled, i.e.

$$\mathsf{rank}(FF^\dagger) = 2$$
$$FF^\dagger \neq \frac{I_A}{2}$$

In general one can reduce $|F\rangle\!\rangle$ in the Schmidt form

$$|F\rangle\!\rangle = \sum_{i=0}^{1} f_i\,|\eta_i\rangle\,|\theta_i\rangle = f_0\,|\eta_0\rangle\,|\theta_0\rangle + f_1\,|\eta_1\rangle\,|\theta_1\rangle\,.$$

Via local unitary maps (that are LOCC) one can then turn $|F\rangle\!\rangle$ into $|\tilde{F}\rangle\!\rangle$, where

$$|\tilde{F}\rangle\!\rangle = f_0\,|0\rangle\,|1\rangle + f_1\,|1\rangle\,|0\rangle$$

and vice-versa. We then just need to prove that $\rho = \frac{1}{2}|U\rangle\!\rangle\langle\!\langle U| \succ |\tilde{F}\rangle\!\rangle\langle\!\langle\tilde{F}|$. We remind that $\frac{1}{\sqrt{2}}|i\sigma_y\rangle\!\rangle = \frac{1}{\sqrt{2}}(|0\rangle\,|1\rangle - |1\rangle\,|0\rangle)$. Let us then consider the following protocol: Alice prepares an ancillary system E in the pure state $|0\rangle\langle 0|$. The state $\rho = \frac{1}{2}|i\sigma_y\rangle\!\rangle\langle\!\langle i\sigma_y|$ is then mapped to $|0\rangle\langle 0| \otimes \frac{1}{2}|i\sigma_y\rangle\!\rangle\langle\!\langle i\sigma_y| = |\phi\rangle\langle\phi|$, where

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|0\rangle_E\,|0\rangle_A\,|1\rangle_B - |0\rangle_E\,|1\rangle_A\,|0\rangle_B)$$

Alice controls the system EA. She now performs the unitary map $\mathscr{N}_U$ defined by the following identities

$$\begin{aligned}
U|00\rangle &= f_0|00\rangle + f_1|11\rangle,\\
U|01\rangle &= -f_1|01\rangle - f_0|10\rangle,\\
U|10\rangle &= f_0^*|01\rangle - f_1^*|10\rangle,\\
U|11\rangle &= -f_1^*|00\rangle + f_0^*|11\rangle.
\end{aligned}$$

The matrix for U in the basis $\{|ij\rangle\}_{i,j=0}^{1}$ is

$$U = \begin{pmatrix} f_0 & 0 & 0 & -f_1^* \\ 0 & -f_1 & f_0^* & 0 \\ 0 & -f_0 & -f_1^* & 0 \\ f_1 & 0 & 0 & f_0^* \end{pmatrix}.$$

One has $\mathscr{N}_U \otimes \mathscr{I}_B\,(|\phi\rangle\langle\phi|) = |\phi'\rangle\langle\phi'|$, with

$$\begin{aligned}
|\phi'\rangle &= \frac{1}{\sqrt{2}}\,(f_0|0\rangle_E|0\rangle_A|1\rangle_B + f_1|1\rangle_E|1\rangle_A|1\rangle_B + f_1|0\rangle_E|1\rangle_A|0\rangle_B + f_0|1\rangle_E|0\rangle_A|0\rangle_B) =\\
&= \frac{1}{\sqrt{2}}\,[|0\rangle_E\,(f_0|0\rangle_A|1\rangle_B + f_1|1\rangle_A|0\rangle_B) + |1\rangle_E\,(f_0|0\rangle_A|0\rangle_B + f_1|1\rangle_A|1\rangle_B)]\,.
\end{aligned}$$

Now Alice can measure the ancillary system E with the POVM $\{|0\rangle\langle 0|, |1\rangle\langle 1|\}$, obtaining for the system AB the following conditional states:

$$\rho'_x \coloneqq \frac{\mathrm{Tr}_E[(P_x \otimes I_{\mathrm{AB}})\,|\phi'\rangle\langle\phi'|]}{\mathrm{Tr}[(P_x \otimes I_{\mathrm{AB}})\,|\phi'\rangle\langle\phi'|]}, \quad x = 0, 1.$$

Notice that

$$(P_0 \otimes I_{\mathrm{AB}})|\phi'\rangle = |0\rangle \otimes \frac{1}{\sqrt{2}}\,(f_0|0\rangle|1\rangle + f_1|1\rangle|0\rangle)\,,$$
$$(P_1 \otimes I_{\mathrm{AB}})|\phi'\rangle = |1\rangle \otimes \frac{1}{\sqrt{2}}\,(f_0|0\rangle|0\rangle + f_1|1\rangle|1\rangle)\,.$$

Thus $\rho_0 = |\tilde{F}\rangle\!\rangle\langle\!\langle\tilde{F}|$, $\rho_1 = |G\rangle\!\rangle\langle\!\langle G|$, where $|G\rangle\!\rangle \coloneqq \frac{1}{\sqrt{2}}\,(f_0|0\rangle|0\rangle + f_1|1\rangle|1\rangle)$. If Alice sends the outcome x to Bob, then Bob does:

- Nothing if $x = 0 \Longrightarrow \rho''_{\mathrm{AB}} = |\tilde{F}\rangle\!\rangle\langle\!\langle\tilde{F}|$
- Applies σ_x if $x = 1 \Longrightarrow \rho''_{\mathrm{AB}} = \mathscr{I}_{\mathrm{A}} \otimes \mathscr{N}_{\sigma_x}\,(|G\rangle\!\rangle\langle\!\langle G|) = |\tilde{F}\rangle\!\rangle\langle\!\langle\tilde{F}|$.
  Indeed, $I_{\mathrm{A}} \otimes \sigma_x|G\rangle\!\rangle = \frac{1}{\sqrt{2}}\,(f_0|0\rangle|1\rangle + f_1|1\rangle|0\rangle) = |\tilde{F}\rangle\!\rangle$.

Every maximally entangled two-qubit state can be mapped via LOCC to every pure two-qubit state. In order to produce mixed states

$$\rho = \sum_i p_i |\phi_i\rangle\!\rangle\langle\!\langle\phi_i|,$$

it is then sufficient that Alice produces a classical variable x_i with $p(x_i) = p_i$, then send x_i to Bob, and finally they start the LOCC protocol to turn $\frac{1}{2}|i\sigma_y\rangle\!\rangle\langle\!\langle i\sigma_y|$ into $|\phi_i\rangle\!\rangle\langle\!\langle\phi_i|$.

However, in the case of d-dimensional systems, there are *incomparable pairs* of pure maximal states ρ, σ. This means that, despite the fact that it both holds that $\tau \succ \rho$ only if $\tau \sim \rho$, and $\tau \succ \sigma$ only if $\tau \sim \sigma$, neither $\rho \succ \sigma$ nor $\sigma \succ \rho$. It is then impossible to establish a notion of *maximally entangled state* in the same way as we did for the singlet in the case of qubits.

Moreover, it can happen that $\rho \not\succ \sigma$, but $\rho \succ \tau_\varepsilon$ for a family of states $\{\tau_\varepsilon\}$ with $\|\sigma - \tau_\varepsilon\|_1 < \varepsilon$. The way in which equivalence classes are ordered in our resource theory of entanglement seems then to be too restrictive: if it is possible to produce an arbitrarily good approximation of σ via LOCC from ρ, it seems that our order relation should be extended so that $\rho \succ \sigma$.

To overcome these issues, one can introduce a scenario where, instead of changing the order relation between equivalence classes, one uses it in order to introduce a way of assessing an "exchange rate" between states, that allows one to compare also states of different systems. For example, let $\rho \in \mathsf{St}_1(\mathrm{A})$ and $\sigma \in \mathsf{St}_1(\mathrm{B})$. First of all, $\forall \varepsilon > 0$ we seek an LOCC protocol and a state τ_ε of suitably many copies of system B (say m), such that $\exists n$: $\rho^{\otimes n} \succ \tau_\varepsilon$ with $\|\sigma^{\otimes m} - \tau_\varepsilon\|_1 < \varepsilon$. The largest possible ratio $\frac{m}{n}$ for which the above LOCC transformation can be achieved is an "exchange rate" between ρ and σ, i.e. the relative entanglement content of σ with respect to ρ.

18.3 Entanglement cost and distillable entanglement

With the above observations in mind, we can now keep the singlet as our *standard unit* of entanglement, and define an entanglement measure for $\rho \in \mathsf{St}_1(\mathrm{AB})$ that accounts for the entanglement needed to produce ρ via LOCC. This measure is the *entanglement cost* of ρ. In the following, the symbol Q will denote a qubit system.

Definition 18.10 (Feasible production ratio)**.** Let $\rho \in \mathsf{St}_1(\mathrm{AB})$. We will say that the triple (p, q, ε), is *feasible* if there exists a protocol $\mathscr{L} \in \overline{\mathsf{LOCC}}$, with $\mathscr{L} : \mathrm{Q}^{\otimes 2p} \to (\mathrm{AB})^{\otimes q}$, such that

$$\frac{1}{2}\|\mathscr{L}(\Sigma^{\otimes p}) - \rho^{\otimes q}\|_1 < \varepsilon,$$

where $\Sigma := 1/2|i\sigma_y\rangle\!\rangle\langle\!\langle i\sigma_y|$ is the singlet state. The number p/q, in particular, is a *feasible production ratio* to tolerance ε. The set of feasible triples (p, q, ε) is denoted as F_C.

Let F_ε be the set of values of (m, n) such that $(m, n, \varepsilon) \in \mathsf{F}_C$ is feasible. The number m/n evaluates the number of singlets per copy of ρ needed to produce n copies of ρ via LOCC starting from m singlets. Now, the value $r(\varepsilon) = \inf_{(m,n)\in\mathsf{F}_\varepsilon} m/n$ denotes the ratio that can be achieved if the degree of approximation ε is tolerated. Notice that $r(\varepsilon)$ is a monotonically increasing function, since the ratios $r = m/n$ achievable for approximation ε_1 are straightforwardly achievable for every $\varepsilon_2 > \varepsilon_1$. The achievable ratio such that the approximation ε can be made arbitrarily small is then the limit $r(\varepsilon)$ for $\varepsilon \to 0$. Indeed, considering larger values of $r = m/n$ would mean overestimating the amount of singlets required to produce each copy of ρ. The definition of entanglement cost is then the following.

Definition 18.11 (Entanglement cost)**.** The entanglement cost of $\rho \in \mathsf{St}_1(\mathrm{AB})$ is

$$E_C(\rho) := \inf\{r \in \mathbb{R} \mid \forall \varepsilon > 0, \delta > 0, \exists (m, n, \varepsilon) \in \mathsf{F}_C, |r - m/n| < \delta\}.$$

$E_C(\rho)$ is then the optimal ratio $r = \frac{m}{n}$ of the number m of starting copies of a singlet state Σ and the number n of copies of ρ that can be obtained via LOCC, asymptotically, and with arbitrarily small approximation.

A second measure of entanglement accounts, on the contrary, for amount of entanglement that can be extracted from ρ vial LOCC. This quantity is called *distillable entanglement* of ρ, and is defined as follows.

Definition 18.12 (Feasible extraction ratio)**.** Let $\rho \in \mathsf{St}_1(\mathrm{AB})$. We will say that the triple (q, p, ε), is *feasible* if there exists a protocol $\mathscr{L} \in \overline{\mathsf{LOCC}}$, with $\mathscr{L} : (\mathrm{AB})^{\otimes q} \to \mathrm{Q}^{\otimes 2p}$, such that

$$\frac{1}{2}\|\mathscr{L}(\rho^{\otimes q}) - \Sigma^{\otimes p}\|_1 \leq \varepsilon.$$

The number p/q, in particular, is a *feasible extraction ratio* to tolerance ε. The set of feasible triples (p, q, ε) is denoted as F_D.

Let now G_ε be the set of values of (m,n) such that $(m,n,\varepsilon) \in \mathsf{F}_D$ is feasible. The number n/m evaluates the number of singlets that can be produced out of each copy of ρ, starting from m copies of ρ, via LOCC and producing n singlets. Similarly to the case of the entanglement cost, the value $r(\varepsilon) = \sup_{(m,n)\in\mathsf{G}_\varepsilon} n/m$ denotes the ratio that can be achieved if the degree of approximation ε is tolerated. Notice that $r(\varepsilon)$ is a monotonically increasing function, since the ratios $r = n/m$ achievable for approximation ε_1 are straightforwardly achievable for every $\varepsilon_2 > \varepsilon_1$. The achievable ratio such that the approximation ε can be made arbitrarily small is then the limit $r(\varepsilon)$ for $\varepsilon \to 0$. Indeed, considering smaller values of $r = n/m$ would mean underestimating the number of singlets that can be produced from each copy of ρ. The definition of distillable entanglement is then the following.

Definition 18.13. (Distillable entanglement) The distillable entanglement $E_D(\rho)$ of a state $\rho \in \mathsf{St}_1(AB)$ is

$$E_D(\rho) := \sup\{r \in \mathbb{R} \mid \forall \varepsilon > 0, \delta > 0, \exists (m,n,\varepsilon) \in \mathsf{F}_D, |r - n/m| < \delta\}.$$

$E_C(\rho)$ and $E_D(\rho)$ then measure the amount of entanglement measured in *singlets*, that is needed to produce one copy of ρ or that can be produced out of one copy of ρ, respectively, in the asymptotic scenario and with arbitrary accuracy. As one might expect, $E_C(\rho) \geq E_D(\rho)$. One might think that actually $E_C(\rho) = E_D(\rho)$, always. This is not the case, however. While $E_C(\rho) > 0$ for every entangled state, it can happen that $E_C(\rho) = 0$ even if ρ is entangled. In this case we speak of "bound entanglement", that can be interpreted as a number of singlets that must be spent to produce ρ but cannot be extracted anymore out of ρ.

18.4 Criteria for entanglement

We thus gave a thorough definition of entanglement through two measures with a clear operational interpretation. However, calculating $E_D(\rho)$ or $E_C(\rho)$ is hard. It is even hard just to establish whether a given ρ_{AB} is separable or not. For this reason it is useful to introduce criteria for entanglement that are easy to calculate, even though they are just sufficient, and not quantitative.

18.4.1 The PPT criterion - Positive partial transpose

Also known as Peres-Horodecki criterion, this criterion is a sufficient condition for ρ_{AB} to be entangled. It states that:

Lemma 18.14. *If $(\mathscr{I}_A \otimes \Theta_B)(\rho_{AB}) \ngeq 0$, then ρ_{AB} is entangled. (Alternatively if ρ_{AB} is separable, then $(\mathscr{I}_A \otimes \Theta_B)(\rho_{AB}) \geq 0$.)*

Proof. The proof is very simple. Let

$$\rho_{AB} = \sum_i p_i \sigma_i \otimes \tau_i \quad \text{with } p_i \geq 0, \sigma_i \geq 0, \tau_i \geq 0.$$

Then

$$(\mathscr{I}_A \otimes \Theta_B)(\rho_{AB}) = \sum_i p_i \sigma_i \otimes \tau_i^T \geq 0. \qquad \square$$

Remark 21. In the case where $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^2 \otimes \mathbb{C}^2$, $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^3 \otimes \mathbb{C}^2$, or $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^2 \otimes \mathbb{C}^3$, the PPT criterion becomes necessary and sufficient.

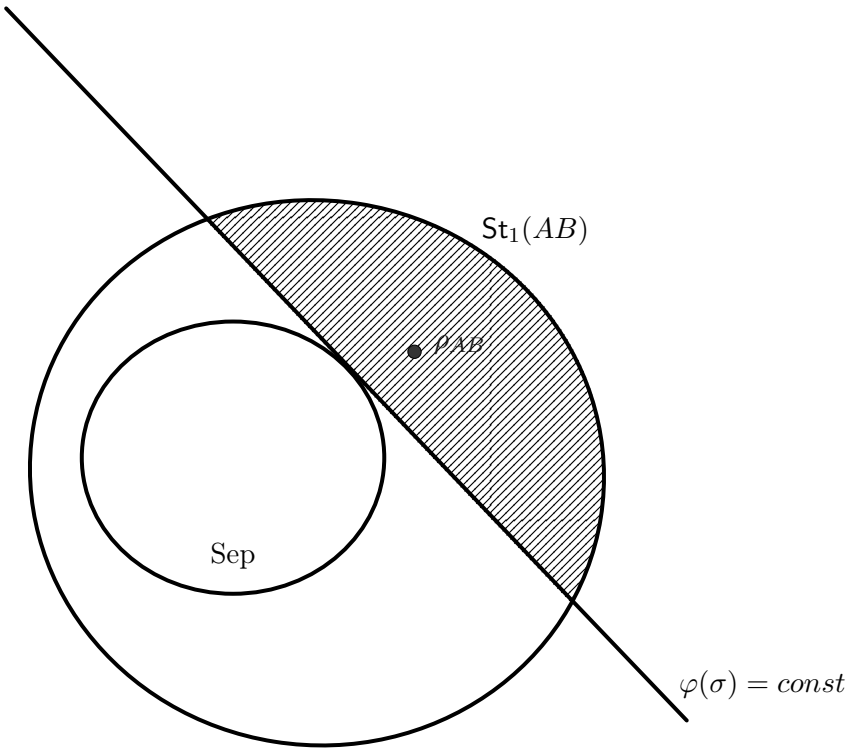
Remark 22. For every positive non-CP map $\mathscr{L}$, a similar criterion can be given: if ρ_{AB} is separable, then $(\mathscr{I}_A \otimes \mathscr{L}_B)(\rho_{AB}) \geq 0$.

18.4.2 Entanglement witnesses

Since separable states form a convex set, the separating hyperplane theorem ensures that given ρ_{AB} entangled, there is a functional $\phi : \mathsf{St}_1(AB) \to \mathbb{R}$ such that

$$\begin{aligned} &\phi(\sigma) \geq 0 \qquad \forall\ \sigma \text{ separable;} \\ &\phi(\rho_{AB}) < 0. \end{aligned}$$

A famous example is given by Bell's inequalities (or CHSH inequality).



Remark 23. The PPT criterion is sufficient for *bound entanglement*. Precisely, if ρ_{AB} is entangled but $(\mathscr{I}_{\mathrm{A}} \otimes \Theta_{\mathrm{B}})(\rho_{\mathrm{AB}}) \geq 0$, then $E_D(\rho) = 0$ while $(E_C(\rho) > 0)$.

18.4.3 Bound entanglement

Indeed, let $E_D(\rho) > 0$. Then $\exists n$, $\exists \mathscr{E} \in \overline{\mathsf{LOCC}}$, and r s.t. $|r - E_D(\rho)| < \delta$ and

$$\frac{1}{2}\|\mathscr{E}(\rho^{\otimes n}) - \Sigma^{\otimes rn}\|_1 \ < \varepsilon. \tag{18.1}$$

Now, since $\frac{1}{2}\|\tau - \eta\|_1 = \max_{0\leq P\leq I} \mathrm{Tr}\left[(\tau-\eta)P\right]$,

$$\frac{1}{2}\|\mathscr{C}(\tau) - \mathscr{C}(\eta)\|_1 = \max_{0\leq P\leq I} \mathrm{Tr}\left[(\tau-\eta)\mathscr{C}^\dagger(P)\right] \leq \max_{0\leq Q\leq I} \mathrm{Tr}\left[(\tau-\eta)Q\right] \leq \frac{1}{2}\|\tau-\eta\|_1.$$

Thus, discarding all the rn systems but one in Eq. (18.1), one has

$$\frac{1}{2}\|\mathrm{Tr}_{rn-1}\left[\mathscr{E}(\rho^{\otimes n}) - \Sigma^{\otimes rn}\right]\|_1 = \frac{1}{2}\|\mathscr{E}'(\rho^{\otimes n}) - \Sigma\|_1 \ < \varepsilon,$$

where we considered the channel $\mathscr{C} \coloneqq \mathrm{Tr}_{rn-1} : (\mathrm{AB})^{\otimes rn} \to \mathrm{AB}$. Thus, $\mathscr{E}'(\rho^{\otimes n})$ must be entangled. Indeed $F(\Sigma,\tau) = \frac{1}{2}\langle\!\langle i\sigma_y|\tau|i\sigma_y\rangle\!\rangle$ and for separable $\tau = \sum_i \ p_i \ \theta_i \otimes \gamma_i$

$$F(\Sigma,\tau) = \frac{1}{2}\sum_i \ p_i \,\mathrm{Tr}[\theta_i \sigma_y \gamma_i^T \sigma_y] \leq \frac{1}{2}\sum_i \ p_i \,\mathrm{Tr}[\theta_i] = \frac{1}{2}.$$

By the Fuchs-Van der Graaf inequalities one has

$$\frac{1}{2}\|\Sigma - \tau\|_1 \geq 1 - F(\Sigma,\tau) \geq \frac{1}{2}$$

for separable τ. Thus, for sufficiently small ε,

$$\frac{1}{2}\|\mathscr{E}'(\rho^{\otimes n}) - \Sigma\|_1 \ < \varepsilon \quad \Rightarrow \quad \mathscr{E}'(\rho^{\otimes n}) \text{ entangled.}$$

The map $\mathscr{E}'$ is in $\overline{\mathsf{LOCC}}$, thus it is in SEP:

$$\mathscr{E}'(\rho^{\otimes n}) = \sum_i (A_i \otimes B_i)\rho^{\otimes n}(A_i^\dagger \otimes B_i^\dagger).$$

There must then exists i_0 such that $(A_{i_0} \otimes B_{i_0})\rho^{\otimes n}(A_{i_0}^\dagger \otimes B_{i_0}^\dagger)$ is entangled. Since $A_{i_0} : \mathcal{H}_{\mathrm{A}}^{\otimes n} \to \mathbb{C}^2$, and similarly for $B_{i_0} : \mathcal{H}_{\mathrm{B}}^{\otimes n} \to \mathbb{C}^2$, by the singular values decomposition we have

$$\begin{aligned} A_{i_0} &= |0\rangle\langle\psi_0| + |1\rangle\langle\psi_1|, \\ B_{i_0} &= |0\rangle\langle\varphi_0| + |1\rangle\langle\varphi_1|. \end{aligned}$$

Defining P_{A} and P_{B} as the projections on $\mathsf{Span}(|\psi_0\rangle, |\psi_1\rangle)$ and $\mathsf{Span}(|\varphi_0\rangle, |\varphi_1\rangle)$ respectively, we have

$$(A_{i_0} \otimes B_{i_0})\rho^{\otimes n}(A_{i_0}^\dagger \otimes B_{i_0}^\dagger) = (A_{i_0} \otimes B_{i_0})(P_{\mathrm{A}} \otimes P_{\mathrm{B}})\rho^{\otimes n}(P_{\mathrm{A}} \otimes P_{\mathrm{B}})(A_{i_0}^\dagger \otimes B_{i_0}^\dagger),$$

and since LOCC maps cannot create entanglement, $\rho' = (P_{\mathrm{A}} \otimes P_{\mathrm{B}})\rho^{\otimes n}(P_{\mathrm{A}} \otimes P_{\mathrm{B}})$ must be entangled. Finally since ρ' has support in a $\mathbb{C}^2 \otimes \mathbb{C}^2$ space, ρ' has a non-positive partial transpose.

$$\rho'^{\tau_{\mathrm{B}}} = (I_{\mathrm{A}} \otimes \Theta_{\mathrm{B}})[(P_{\mathrm{A}} \otimes P_{\mathrm{B}})\rho^{\otimes n}(P_{\mathrm{A}} \otimes P_{\mathrm{B}})] = (P_{\mathrm{A}} \otimes P_{\mathrm{B}}^T)(\rho^{\tau_{\mathrm{B}}})^{\otimes n}(P_{\mathrm{A}} \otimes P_{\mathrm{B}}^T),$$

where we used the shorthand notation $X^{\tau_{\mathrm{B}}} \coloneqq \mathscr{I}_{\mathrm{A}} \otimes \Theta_{\mathrm{B}}(X)$. Since $\rho'^{\tau_{\mathrm{B}}}$ has a negative eigenvalue, it must be:

$$\langle\!\langle \Phi | (P_{\mathrm{A}} \otimes P_{\mathrm{B}}^T)(\rho^{\tau_{\mathrm{B}}})^{\otimes n}(P_{\mathrm{A}} \otimes P_{\mathrm{B}}^T) | \Phi \rangle\!\rangle < 0,$$

for some $|\Phi\rangle\!\rangle$—e.g. by taking $|\Phi\rangle\!\rangle$ as the corresponding eigenvector—i.e.

$$\langle\!\langle \Psi | (\rho^{\tau_{\mathrm{B}}})^{\otimes n} | \Psi \rangle\!\rangle < 0, \quad |\Psi\rangle\!\rangle \coloneqq P_{\mathrm{A}} \otimes P_{\mathrm{B}} |\Phi\rangle\!\rangle.$$

Finally, if $\rho^{\tau_{\mathrm{B}}} \geq 0$ also $(\rho^{\tau_{\mathrm{B}}})^{\otimes n} \geq 0$, thus it must be $\rho^{\tau_{\mathrm{B}}} \not\geq 0$. In other words we proved that if $E_D(\rho) > 0$ it must be $\rho^{\tau_{\mathrm{B}}} \not\geq 0$, i.e. if $\rho^{\tau_{\mathrm{B}}} \geq 0$, then $E_D(\rho) = 0$.

As a final remark on this point, we observe that it can be proved that entangled states with positive partial transpose exist, thus there exist states exhibiting bound entanglement.

18.4.4 Entanglement of formation

For pure states $\rho_{\mathrm{AB}} = |\Psi\rangle\!\rangle\langle\!\langle\Psi|$, a measure of entanglement is the *entropy of entanglement*:

$$S(\rho_{\mathrm{A}}) = S(\rho_{\mathrm{B}}) = S(\Psi\Psi^\dagger) =: E(\rho)$$

Every entanglement measure must reduce to $E(\rho)$ for pure states ρ.

Definition 18.15 (Entanglement of formation)**.**

$$E_F(\rho) \coloneqq \inf \left\{ \sum_i p_i E(|\psi_i\rangle\!\rangle\langle\!\langle\psi_i|) \,\middle|\, \rho = \sum_i p_i |\psi_i\rangle\!\rangle\langle\!\langle\psi_i| \right\}$$

The *regularized* entanglement of formation is:

$$E_F^\infty(\rho) \coloneqq \lim_{n\to\infty} \frac{E_F(\rho^{\otimes n})}{n}$$

It holds that $E_F^\infty(\rho) = E_C(\rho)$.

Acknowledgements

These notes are largely based on the lecture notes of G. M. D'Ariano who was formerly the teacher of the present course. Some parts are inspired by J. Watrous's lecture notes. I wish to express my gratitude to Maddalena Ghio, Francesco Tacchino, Andrea Olivo and Francesco Zagaria for their careful reading, and for pointing out various errors in earlier drafts of these lecture notes. A special thank to Gianmarco Ricciardi, Edoardo Centofanti, Daniele Gilio, Federico Quetti, Lia Rapella, Lorenzo Rossi and Alessandro Summer for turning my (awfully) handwritten notes for the last lecture into a LaTeX file. Finally, I acknowledge the thoughtful and detailed review of the above notes made by Leonardo Vaglini.

Bibliography

[1] Charles H. Bennett, David P. DiVincenzo, Christopher A. Fuchs, Tal Mor, Eric Rains, Peter W. Shor, John A. Smolin, and William K. Wootters. Quantum nonlocality without entanglement. *Phys. Rev. A*, 59:1070–1091, Feb 1999.

[2] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[3] J. Watrous. *Quantum entropy and source coding.* lecture notes available at https://cs.uwaterloo.ca/ watrous/CS766/DraftChapters/5.QuantumEntropy.pdf.