Lecture Notes

Quantum Information Theory

Matthias Christandl
based on lecture notes by Renato Renner

January 19, 2012

1 Introduction

The very process of doing *physics* is to acquire *information* about the world around us. At the same time, the storage and processing of information is necessarily a physical process. It is thus not surprising that physics and the theory of information are inherently connected.[1] *Quantum information theory* is an interdisciplinary research area whose goal is to explore this connection.

As the name indicates, the information carriers in *quantum information theory* are quantum-mechanical systems (e.g., the spin of a single electron). This is in contrast to *classical information theory* where information is assumed to be represented by systems that are accurately characterized by the laws of classical mechanics and electrodynamics (e.g., a classical computer, or simply a piece of paper). Because any such classical system can in principle be described in the language of quantum mechanics, classical information theory is a (practically significant) special case of quantum information theory.

The course starts with a quick introduction to classical probability and information theory. Many of the relevant concepts, e.g., the notion of *entropy* as a measure of uncertainty, can already be defined in the purely classical case. I thus consider this classical part as a good preparation as well as a source of intuition for the more general quantum-mechanical treatment.

We will then move on to the quantum setting, where we will spend a considerable amount of time to introduce a convenient framework for representing and manipulating quantum states and quantum operations. This framework will be the prerequisite for formulating and studying typical information-theoretic problems such as information storage and transmission (with possibly noisy devices). Furthermore, we will learn in what sense information represented by quantum systems is different from information that is represented classically. Finally, we will have a look at applications such as *quantum key distribution*.

I would like to emphasize that it is not an intention of this course to give a complete treatment of quantum information theory. Instead, the goal is to focus on certain key concepts and to study them in more detail. For further reading, I recommend the standard textbook by Nielsen and Chuang [10]. Also, I would like to mention the course on quantum computation [16] by Stefan Wolf (Computer Science Department). Wolf's course is somewhat complementary in the sense that it focuses on quantum *computation*, while this course is on quantum *information*.

[1] This connection has been noticed by numerous famous scientists over the past fifty years, among them Rolf Landauer with his claim "information is physical."

2 Probability Theory

Information theory is largely based on probability theory. Therefore, before introducing information-theoretic concepts, we need to recall some key notions of probability theory. The following section is, however, not thought as an introduction to probability theory. Rather, its main purpose is to summarize some basic facts as well as the notation we are going to use in this course.

2.1 What is probability?

This is actually a rather philosophical question and it is not the topic of this course to answer it.[1] Nevertheless, it might be useful to spend some thoughts about how probabilities are related to actual physical quantities.

For the purpose of this course, it might make sense to take a *Bayesian* point of view, meaning that probability distributions are generally interpreted as a *state of knowledge.* To illustrate the Bayesian approach, consider a game where a quizmaster hides a prize behind one of three doors, and where the task of a candidate is to find the prize. Let X be the number of the door (1, 2, or 3) which hides the prize. Obviously, as long as the candidate does not get any additional information, each of the doors is equally likely to hide the prize. Hence, the probability distribution P_X^{cand} that the candidate would assign to X is *uniform,*

$$P_X^{\mathrm{cand}}(1) = P_X^{\mathrm{cand}}(2) = P_X^{\mathrm{cand}}(3) = 1/3 \ .$$

On the other hand, the quizmaster knows where he has hidden the prize, so he would assign a *deterministic value* to X. For example, if the prize is behind door 1, the probability distribution P^{mast} the quizmaster would assign to X has the form

$$P_X^{\mathrm{mast}}(1) = 1 \quad \text{and} \quad P_X^{\mathrm{mast}}(2) = P_X^{\mathrm{mast}}(3) = 0 \ .$$

The crucial thing to note here is that, although the distributions P_X^{cand} and P_X^{mast} are referring to the same physical value X, they are different because they correspond to different states of knowledge.

We could extend this example arbitrarily. For instance, the quizmaster could open one of the doors, say 3, to reveal that the prize is *not* behind it. This additional information, of course, changes the state of knowledge of the candidate, resulting in yet another probability distribution $P_X^{\mathrm{cand}'}$ associated with X,[2]

[1] For a nice introduction to the philosophy of probability theory, I recommend the book [9].

[2] The situation becomes more intriguing if the quizmaster opens a door after the candidate has already made a guess. The problem of determining the probability distribution that the candidate assigns to X in this case is known as the *Monty Hall problem.* For further reading, I refer to [15].

$$P_X^{\text{cand}'}(1) = P_X^{\text{cand}'}(2) = 1/2 \quad \text{and} \quad P_X^{\text{cand}'}(3) = 0 \ .$$

When interpreting a probability distribution as a *state of knowledge* and, hence, as *subjective* quantity, we need to carefully specify *whose* state of knowledge we are referring to. This is particularly relevant for the analysis of information-theoretic settings, which usually involve more than one party. For example, in a communication scenario, we might have a *sender* who intends to transmit a message M to a *receiver*. Clearly, before M is sent, the sender and the receiver have different knowledge about M and, consequently, would assign different probability distributions to M. In the following, when describing such settings, we will typically understand all distributions as states of knowledge of an *outside observer*.

2.2 Definition of probability spaces and random variables

The concept of *random variables* is important in both physics and information theory. Roughly speaking, one can think of a random variable as the state of a classical probabilistic system. Hence, in classical information theory, it is natural to think of data as being represented by random variables.

In this section, we define random variables and explain a few related concepts. For completeness, we first give the general mathematical definition based on probability spaces. Later, we will restrict to *discrete* random variables (i.e., random variables that only take countably many values). These are easier to handle than general random variables but still sufficient for our information-theoretic considerations.

2.2.1 Probability space

A *probability space* is a triple $(\Omega, \mathcal{E}, P)$, where $(\Omega, \mathcal{E})$ is a measurable space, called *sample space*, and P is a probability measure. The *measurable space* consists of a set Ω and a σ-algebra $\mathcal{E}$ of subsets of Ω, called *events*.

By definition, the *σ-algebra* $\mathcal{E}$ must contain at least one event, and be closed under complements and countable unions. That is, (i) $\mathcal{E} \neq \emptyset$, (ii) if E is an event then so is its complement $E^c := \Omega \backslash E$, and (iii) if $(E_i)_{i\in\mathbb{N}}$ is a family of events then $\bigcup_{i\in\mathbb{N}} E_i$ is an event. In particular, Ω and $\emptyset$ are events, called the *certain event* and the *impossible event*.

The *probability measure* P on $(\Omega, \mathcal{E})$ is a function

$$P : \quad \mathcal{E} \to \mathbb{R}^+$$

that assigns to each event $E \in \mathcal{E}$ a nonnegative real number $P[E]$, called the *probability of* E. It must satisfy the probability axioms $P[\Omega] = 1$ and $P[\bigcup_{i\in\mathbb{N}} E_i] = \sum_{i\in\mathbb{N}} P[E_i]$ for any family $(E_i)_{in\in\mathbb{N}}$ of pairwise disjoint events.

2.2.2 Random variables

Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $(\mathcal{X}, \mathcal{F})$ be a measurable space. A *random variable* X is a function from Ω to $\mathcal{X}$ which is *measurable* with respect to the σ-algebras

$\mathcal{E}$ and $\mathcal{F}$. This means that the preimage of any $F \in \mathcal{F}$ is an event, i.e., $X^{-1}(F) \in \mathcal{E}$. The probability measure P on $(\Omega, \mathcal{E})$ induces a probability measure P_X on the measurable space $(\mathcal{X}, \mathcal{F})$, which is also called *range of* X,

$$P_X[F] := P[X^{-1}(F)] \quad \forall F \in \mathcal{F} \ . \tag{2.1}$$

A pair (X, Y) of random variables can obviously be seen as a new random variable. More precisely, if X and Y are random variables with range $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$, respectively, then (X, Y) is the random variable with range $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \times \mathcal{G})$ defined by[3]

$$(X, Y) : \quad \omega \mapsto X(\omega) \times Y(\omega) \ .$$

We will typically write P_{XY} to denote the *joint probability measure* $P_{(X,Y)}$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \times \mathcal{G})$ induced by (X, Y). This convention can, of course, be extended to more than two random variables in a straightforward way. For example, we will write $P_{X_1 \cdots X_n}$ for the probability measure induced by an n-tuple of random variables $(X_1, \ldots, X_n)$.

In a context involving only finitely many random variables $X_1, \ldots, X_n$, it is usually sufficient to specify the joint probability measure $P_{X_1 \cdots X_n}$, while the underlying probability space $(\Omega, \mathcal{E}, P)$ is irrelevant. In fact, as long as we are only interested in events defined in terms of the random variables $X_1, \ldots, X_n$ (see Section 2.2.3 below), we can without loss of generality identify the sample space $(\Omega, \mathcal{E})$ with the range of the tuple $(X_1, \ldots, X_n)$ and define the probability measure P to be equal to $P_{X_1 \cdots X_n}$.

2.2.3 Notation for events

Events are often defined in terms of random variables. For example, if the range of X is (a subset of) the set of real numbers $\mathbb{R}$ then $E := \{\omega \in \Omega : X(\omega) > x_0\}$ is the event that X takes a value larger than x_0. To denote such events, we will usually drop ω, i.e., we simply write $E = \{X > x_0\}$. If the event is given as an argument to a function, we also omit the curly brackets. For instance, we write $P[X > x_0]$ instead of $P[\{X > x_0\}]$ to denote the probability of the event $\{X > x_0\}$.

2.2.4 Conditioning on events

Let $(\Omega, \mathcal{E}, P)$ be a probability space. Any event $E' \in \mathcal{E}$ such that $P(E') > 0$ gives rise to a new probability measure $P[\cdot|E']$ on $(\Omega, \mathcal{E})$ defined by

$$P[E|E'] := \frac{P[E \cap E']}{P[E']} \quad \forall E \in \mathcal{E} \ .$$

$P[E|E']$ is called the *probability of E conditioned on E'* and can be interpreted as the probability that the event E occurs if we already know that the event E' has occurred. In particular, if E and E' are *mutually independent*, i.e., $P[E \cap E'] = P[E]P[E']$, then $P[E|E'] = P[E]$.

[3] $\mathcal{F} \times \mathcal{G}$ denotes the set $\{F \times G : F \in \mathcal{F}, G \in \mathcal{G}\}$. It is easy to see that $\mathcal{F} \times \mathcal{G}$ is a σ-algebra over $\mathcal{X} \times \mathcal{Y}$.

Similarly, we can define $P_{X|E'}$ as the *probability measure of a random variable X conditioned on E'*. Analogously to (2.1), it is the probability measure induced by $P[\cdot|E']$, i.e.,

$$P_{X|E'}[F] := P[X^{-1}(F)|E'] \quad \forall F \in \mathcal{F} \ .$$

2.3 Probability theory with discrete random variables

2.3.1 Discrete random variables

In the remainder of this script, if not stated otherwise, all random variables are assumed to be *discrete*. This means that their range $(\mathcal{X}, \mathcal{F})$ consists of a countably infinite or even finite set $\mathcal{X}$. In addition, we will assume that the σ-algebra $\mathcal{F}$ is the power set of $\mathcal{X}$, i.e., $\mathcal{F} := \{F \subseteq \mathcal{X}\}$.[4] Furthermore, we call $\mathcal{X}$ the *alphabet of X*. The probability measure P_X is then defined for any singleton set $\{x\}$. Setting $P_X(x) := P_X[\{x\}]$, we can interpret P_X as a *probability mass function*, i.e., a positive function

$$P_X : \quad \mathcal{X} \to \mathbb{R}^+$$

that satisfies the *normalization condition*

$$\sum_{x \in \mathcal{X}} P_X(x) = 1 \ . \tag{2.2}$$

More generally, for an event E' with $P[E'] > 0$, the *probability mass function of X conditioned on E'* is given by $P_{X|E'}(x) := P_{X|E'}[\{x\}]$, and also satisfies the normalization condition (2.2).

2.3.2 Marginals and conditional distributions

Although the following definitions and statements apply to arbitrary n-tuples of random variables, we will formulate them only for *pairs* (X, Y) in order to keep the notation simple. In particular, it suffices to specify a bipartite probability distribution P_{XY}, i.e., a positive function on $\mathcal{X} \times \mathcal{Y}$ satisfying the normalization condition (2.2), where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets of X and Y, respectively. The extension to arbitrary n-tuples is straightforward.[5]

Given P_{XY}, we call P_X and P_Y the *marginal distributions.* It is easy to verify that

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \quad \forall x \in \mathcal{X} \ , \tag{2.3}$$

and likewise for P_Y. Furthermore, for any $y \in \mathcal{Y}$ with $P_Y(y) > 0$, the *distribution $P_{X|Y=y}$ of X conditioned on the event $Y = y$* obeys

$$P_{X|Y=y}(x) = \frac{P_{XY}(x, y)}{P_Y(y)} \quad \forall x \in \mathcal{X} \ . \tag{2.4}$$

[4] It is easy to see that the power set of $\mathcal{X}$ is indeed a σ-algebra over $\mathcal{X}$.

[5] Note that X and Y can themselves be tuples of random variables.

2.3.3 Special distributions

Certain distributions are important enough to be given a name. We call P_X *flat* if all non-zero probabilities are equal, i.e.,

$$P_X(x) \in \{0, q\} \quad \forall x \in \mathcal{X}$$

for some $q \in [0,1]$. Because of the normalization condition (2.2), we have $q = \frac{1}{|\mathrm{supp} P_X|}$, where $\mathrm{supp} P_X := \{x \in \mathcal{X} : P_X(x) > 0\}$ is the *support* of the function P_X. Furthermore, if P_X is flat and has no zero probabilities, i.e.,

$$P_X(x) = \frac{1}{|\mathcal{X}|} \quad \forall x \in \mathcal{X} \ ,$$

we call it *uniform.*

2.3.4 Independence and Markov chains

Two discrete random variables X and Y are said to be *mutually independent* if the events $\{X = x\}$ and $\{Y = y\}$ are mutually independent for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Their joint probability mass function then satisfies $P_{XY} = P_X \times P_Y$.[6]

Related to this is the notion of *Markov chains.* A sequence of random variables $X_1, X_2, \ldots$ is said to have the *Markov property*, denoted $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_n$, if for all $i \in \{1, \ldots, n-1\}$

$$P_{X_{i+1}|X_1=x_1,\ldots,X_i=x_i} = P_{X_{i+1}|X_i=x_i} \quad \forall x_1, \ldots, x_i \ .$$

This expresses the fact that, given any fixed value of X_i, the random variable X_{i+1} is completely independent of all previous random variables $X_1, \ldots, X_{i-1}$. In particular, X_{i+1} can be computed given only X_i.

2.3.5 Functions of random variables, expectation values, and Jensen's inequality

Let X be a random variable with alphabet $\mathcal{X}$ and let f be a function from $\mathcal{X}$ to $\mathcal{Y}$. We denote by $f(X)$ the random variable defined by the concatenation $f \circ X$. Obviously, $f(X)$ has alphabet $\mathcal{Y}$ and, in the discrete case we consider here, the corresponding probability mass function $P_{f(X)}$ is given by

$$P_{f(X)}(y) = \sum_{x \in f^{-1}(\{y\})} P_X(x) \ .$$

For a random variable X whose alphabet $\mathcal{X}$ is a module over the reals $\mathbb{R}$ (i.e., there is a notion of addition and multiplication with reals), we define the *expectation value* of X by

$$\langle X \rangle_{P_X} := \sum_{x \in X} P_X(x) x \ .$$

[6] $P_X \times P_Y$ denotes the function $(x, y) \mapsto P_X(x) P_Y(y)$.

If the distribution P_X is clear from the context, we sometimes omit the subscript.

For a convex real function f on a convex set $\mathcal{X}$, the expectation values of X and $f(X)$ are related by *Jensen's inequality*

$$\langle f(X) \rangle \geq f(\langle X \rangle) \ .$$

The inequality is essentially a direct consequence of the definition of convexity (see Fig. 2.1).

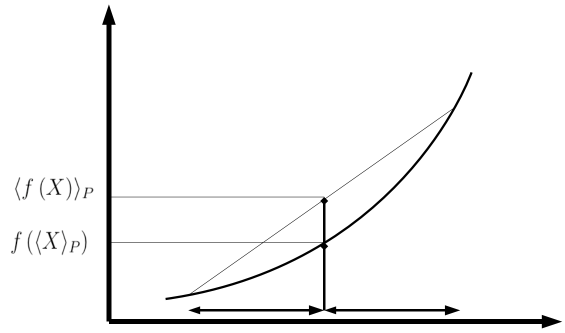


Figure 2.1: Jensen's inequality for a convex function

2.3.6 Trace distance

Let P and Q be two probability mass functions[7] on an alphabet $\mathcal{X}$. The *trace distance* δ between P and Q is defined by

$$\delta(P,Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$$

In the literature, the trace distance is also called *statistical distance*, *variational distance*, or *Kolmogorov distance*.[8] It is easy to verify that δ is indeed a *metric*, that is, it is symmetric, nonnegative, zero if and only if $P = Q$, and it satisfies the triangle inequality. Furthermore, $\delta(P,Q) \leq 1$ with equality if and only if P and Q have distinct support.

Because P and Q satisfy the normalization condition (2.2), the trace distance can equivalently be written as

$$\delta(P,Q) = 1 - \sum_{x \in \mathcal{X}} \min[P(x), Q(x)] \ . \tag{2.5}$$

The trace distance between the probability mass functions Q_X and $Q_{X'}$ of two random variables X and X' has a simple interpretation. It can be seen as the minimum probability that X and X' take different values.

[7] The definition can easily be generalized to probability measures.

[8] We use the term *trace distance* because, as we shall see, it is a special case of the trace distance for density operators.

Lemma 2.3.1. *Let Q_X and $Q_{X'}$ be probability mass functions on $\mathcal{X}$. Then*

$$\delta(Q_X, Q_{X'}) = \min_{P_{XX'}} P_{XX'}[X \neq X']$$

where the minimum ranges over all joint probability mass functions $P_{XX'}$ with marginals $P_X = Q_X$ and $P_{X'} = Q_{X'}$.

Proof. To prove the inequality $\delta(Q_X, Q_{X'}) \leq \min_{P_{XX'}} P_{XX'}[X \neq X']$, we use (2.5) and the fact that, for any joint probability mass function $P_{XX'}$, $\min[P_X(x), P_{X'}(x)] \geq P_{XX'}(x, x)$, which gives

$$\delta(P_X, P_{X'}) = 1 - \sum_{x \in \mathcal{X}} \min[P_X(x), P_{X'}(x)] \leq 1 - \sum_{x \in \mathcal{X}} P_{XX'}(x, x) = P_{XX'}[X \neq X'] \ .$$

We thus have $\delta(P_X, P_{X'}) \leq P_{XX'}[X \neq X']$, for any probability mass function $P_{XX'}$. Taking the minimum over all $P_{XX'}$ with $P_X = Q_X$ and $P_{X'} = Q_{X'}$ gives the desired inequality.

The proof of the opposite inequality is given in the exercises. □

An important property of the trace distance is that it can only decrease under the operation of taking marginals.

Lemma 2.3.2. *For any two density mass functions P_{XY} and Q_{XY},*

$$\delta(P_{XY}, Q_{XY}) \geq \delta(P_X, Q_X) \ .$$

Proof. Applying the triangle inequality for the absolute value, we find

$$\begin{aligned} \frac{1}{2} \sum_{x,y} |P_{XY}(x, y) - Q_{XY}(x, y)| &\geq \frac{1}{2} \sum_{x} |\sum_{y} P_{XY}(x, y) - Q_{XY}(x, y)| \\ &= \frac{1}{2} \sum_{x} |P_X(x) - Q_X(x)| \ , \end{aligned}$$

where the second equality is (2.3). The assertion then follows from the definition of the trace distance. □

2.3.7 I.i.d. distributions and the law of large numbers

An n-tuple of random variables $X_1, \ldots, X_n$ with alphabet $\mathcal{X}$ is said to be *independent and identically distributed (i.i.d.)* if their joint probability mass function has the form

$$P_{X_1 \cdots X_n} = P_X^{\times n} := P_X \times \cdots \times P_X \ .$$

The i.i.d. property thus characterizes situations where a certain process is repeated n times independently. In the context of information theory, the i.i.d. property is often used to describe the statistics of noise, e.g., in repeated uses of a communication channel (see Section 3.2).

The *law of large numbers* characterizes the "typical behavior" of real-valued i.i.d. random variables $X_1, \ldots, X_n$ in the limit of large n. It usually comes in two versions, called the *weak* and the *strong* law of large numbers. As the name suggests, the latter implies the first.

Let $\mu = \langle X_i \rangle$ be the expectation value of X_i (which, by the i.i.d. assumption, is the same for all $X_1, \ldots, X_n$), and let

$$Z_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

be the *sample mean.* Then, according to the *weak law of large numbers*, the probability that Z_n is ε-close to μ for any positive ε converges to one, i.e.,

$$\lim_{n \to \infty} P\big[|Z_n - \mu| < \varepsilon\big] = 1 \quad \forall \varepsilon > 0 \ . \tag{2.6}$$

The weak law of large numbers will be sufficient for our purposes. However, for completeness, we mention the *strong law of large numbers* which says that Z_n converges to μ with probability 1,

$$P\big[\lim_{n \to \infty} Z_n = \mu\big] = 1 \ .$$

2.3.8 Channels

A *channel* $\mathbf{p}$ is a probabilistic mapping that assigns to each value of an *input alphabet* $\mathcal{X}$ a value of the *output alphabet.* Formally, $\mathbf{p}$ is a function

$$\begin{aligned} \mathbf{p} : \quad & \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \\ & (x, y) \mapsto \mathbf{p}(y|x) \end{aligned}$$

such that $\mathbf{p}(\cdot|x)$ is a probability mass function for any $x \in \mathcal{X}$.

Given a random variable X with alphabet $\mathcal{X}$, a channel $\mathbf{p}$ from $\mathcal{X}$ to $\mathcal{Y}$ naturally defines a new random variable Y via the joint probability mass function P_{XY} given by[9]

$$P_{XY}(x, y) := P_X(x)\mathbf{p}(y|x) \ . \tag{2.7}$$

Note also that channels can be seen as generalizations of functions. Indeed, if f is a function from $\mathcal{X}$ to $\mathcal{Y}$, its description as a channel $\mathbf{p}$ is given by

$$\mathbf{p}(y|x) = \delta_{y, f(x)} \ .$$

Channels can be seen as abstractions of any (classical) physical device that takes an input X and outputs Y. A typical example for such a device is, of course, a *communication channel*, e.g., an optical fiber, where X is the input provided by a *sender* and where Y is the (possibly noisy) version of X delivered to a *receiver.* A practically relevant question

[9] It is easy to verify that P_{XY} is indeed a probability mass function.

then is how much information one can transmit *reliably* over such a channel, using an appropriate encoding.

But channels do not only carry information over space, but also over time. Typical examples are memory devices, e.g., a hard drive or a CD (where one wants to model the errors introduced between storage and reading out of data). Here, the question is how much redundancy we need to introduce in the stored data in order to correct these errors.

The notion of channels is illustrated by the following two examples.

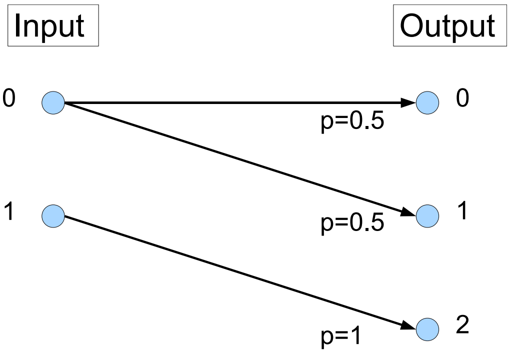


Figure 2.2: Example 1. A reliable channel

Example 2.3.3. *The channel depicted in Fig. 2.2 maps the input* 0 *with equal probability to either* 0 *or* 1*; the input* 1 *is always mapped to* 2*. The channel has the property that its input is uniquely determined by its output. As we shall see later, such a channel would allow to reliably transmit one classical bit of information.*

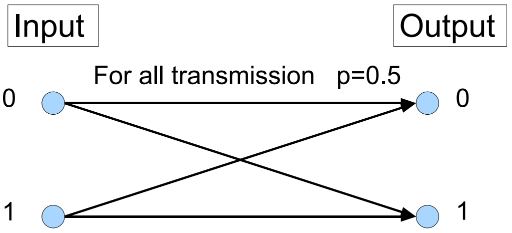


Figure 2.3: Example 2. An unreliable channel

Example 2.3.4. *The channel shown in Fig. 2.3 maps each possible input with equal probability to either* 0 *or* 1*. The output is thus completely independent of the input. Such a channel is obviously not useful to transmit information.*

The notion of i.i.d. random variables naturally translates to channels. A channel $\mathbf{p}_n$ from $\mathcal{X}\times\cdots\times\mathcal{X}$ to $\mathcal{Y}\times\cdots\times\mathcal{Y}$ is said to be *i.i.d.* if it can be written as $\mathbf{p}_n = \mathbf{p}^{\times n} := \mathbf{p}\times\cdots\times\mathbf{p}$.

3 Information Theory

3.1 Quantifying information

The main object of interest in information theory, of course, is information and the way it is processed. The quantification of information thus plays a central role. The aim of this section is to introduce some notions and techniques that are needed for the quantitative study of *classical* information, i.e., information that can be represented by the state of a classical (in contrast to *quantum*) system.

3.1.1 Approaches to define information and entropy

Measures of *information* and measures of *uncertainty*, also called *entropy measures*, are closely related. In fact, the information contained in a message X can be seen as the amount by which our uncertainty (measured in terms of entropy) decreases when we learn X.

There are, however, a variety of approaches to defining entropy measures. The decision what approach to take mainly depends on the type of questions we would like to answer. Let us thus consider a few examples.

Example 3.1.1 (Data transmission)**.** *Given a (possibly noisy) communication channel connecting a sender and a receiver (e.g., an optical fiber), we are interested in the time it takes to reliably transmit a certain document (e.g., the content of a textbook).*

Example 3.1.2 (Data storage)**.** *Given certain data (e.g., a movie), we want to determine the minimum space (e.g., on a hard drive) needed to store it.*

The latter question is related to *data compression*, where the task is to find a space-saving representation Z of given data X. In some sense, this corresponds to finding the shortest possible description of X. An elegant way to make this more precise is to view the description of X as an *algorithm* that generates X. Applied to the problem of data storage, this would mean that, instead of storing data X directly, one would store an (as small as possible) algorithm Z which can reproduce X.

The definition of *algorithmic entropy*, also known as *Kolmogorov complexity*, is exactly based on this idea. The *algorithmic entropy* of X is defined as the minimum length of an algorithm that generates X. For example, a bitstring $X = 00\cdots 0$ consisting of $n \gg 1$ zeros has small algorithmic entropy because it can be generated by a short program (the program that simply outputs a sequence of zeros). The same is true if X consists of the first n digits of π, because there is a short algorithm that computes the circular constant π. In contrast, if X is a sequence of n bits chosen at random, its algorithmic entropy will, with high probability, be roughly equal to n. This is because the shortest program generating the exact sequence of bits X is, most likely, simply the program that has the whole sequence already stored.[1]

Despite the elegance of its definition, the algorithmic entropy has a fundamental disadvantage when being used as a measure for uncertainty: it is *not computable*. This means that there cannot exist a method (e.g., a computer program) that estimates the algorithmic complexity of a given string X. This deficiency as well as its implications[2] render the algorithmic complexity unsuitable as a measure of entropy for most practical applications.

In this course, we will consider a different approach which is based on ideas developed in thermodynamics. The approach has been proposed in 1948 by Shannon [13] and, since then, has proved highly successful, with numerous applications in various scientific disciplines (including, of course, physics). It can also be seen as the theoretical foundation of modern information and communication technology. Today, Shannon's theory is viewed as *the* standard approach to information theory.

In contrast to the algorithmic approach described above, where the entropy is defined as a function of the actual data X, the information measures used in Shannon's theory depend on the probability distribution of the data. More precisely, the entropy of a value X is a measure for the likelihood that a particular value occurs. Applied to the above compression problem, this means that one needs to assign a probability mass function to the data to be compressed. The method used for compression might then be optimized for the particular probability mass function assigned to the data.

3.1.2 Entropy of events

We take an axiomatic approach to motivate the definition of the Shannon entropy and related quantities. In a first step, we will think of the entropy as a property of events E. More precisely, given a probability space $(\Omega, \mathcal{E}, P)$, we consider a function H that assigns to each event E a real value $H(E)$,

$$\begin{array}{rccc} H: & \mathcal{E} & \to & \mathbb{R} \cup \{\infty\} \\ & E & \mapsto & H(E) \ . \end{array}$$

For the following, we assume that the events are defined on a probability space with probability measure P. The function H should then satisfy the following properties.

1. *Independence of the representation:* $H(E)$ only depends on the probability $P[E]$ of the event E.

2. *Continuity:* H is continuous in the probability measure P (relative to the topology induced by the trace distance).

3. *Additivity:* $H(E \cap E') = H(E) + H(E')$ for two independent events E and E'.

4. *Normalization:* $H(E) = 1$ for E with $P[E] = \frac{1}{2}$.

[1] In fact, a (deterministic) computer can only generate *pseudo-random* numbers, i.e., numbers that cannot be distinguished (using any efficient method) from true random numbers.

[2] An immediate implication is that there cannot exist a compression method that takes as input data X and outputs a short algorithm that generates X.

The axioms appear natural if we think of H as a measure of uncertainty. Indeed, Axiom 3 reflects the idea that our total uncertainty about two independent events is simply the sum of the uncertainty about the individual events. We also note that the normalization imposed by Axiom 4 can be chosen arbitrarily; the convention, however, is to assign entropy 1 to the event corresponding to the outcome of a fair coin flip.

The axioms uniquely define the function H.

Lemma 3.1.3. *The function H satisfies the above axioms if and only if it has the form*

$$H : \quad E \longmapsto -\log_2 P[E] \ .$$

Proof. It is straightforward that H as defined in the lemma satisfies all the axioms. It thus remains to show that the definition is unique. For this, we make the ansatz

$$H(E) = f(-\log_2 P[E])$$

where f is an arbitrary function from $\mathbb{R}^+ \cup \{\infty\}$ to $\mathbb{R} \cup \{\infty\}$. We note that, apart from taking into account the first axiom, this is no restriction of generality, because any possible function of $P[E]$ can be written in this form.

From the continuity axiom, it follows that f must be continuous. Furthermore, inserting the additivity axiom for events E and E' with probabilities p and p', respectively, gives

$$f(-\log_2 p) + f(-\log_2 p') = f(-\log_2 pp') \ .$$

Setting $a := -\log_2 p$ and $a' := -\log_2 p'$, this can be rewritten as

$$f(a) + f(a') = f(a + a') \ .$$

Together with the continuity axiom, we conclude that f is linear, i.e., $f(x) = \gamma x$ for some $\gamma \in \mathbb{R}$. The normalization axiom then implies that $\gamma = 1$. □

3.1.3 Entropy of random variables

We are now ready to define entropy measures for random variables. Analogously to the entropy of an event E, which only depends on the probability $P[E]$ of the event, the entropy of a random variable X only depends on the probability mass function P_X.

We start with the most standard measure in classical information theory, the *Shannon entropy*, in the following denoted by H. Let X be a random variable with alphabet $\mathcal{X}$ and let $h(x)$ be the entropy of the event $E_x := \{X = x\}$, for any $x \in \mathcal{X}$, that is,

$$h(x) := H(E_x) = -\log_2 P_X(x) \ . \tag{3.1}$$

Then the *Shannon entropy* is defined as the *expectation value* of $h(x)$, i.e.,

$$H(X) := \langle h(X) \rangle = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) \ .$$

If the probability measure P is unclear from the context, we will include it in the notation as a subscript, i.e., we write $H(X)_P$.

Similarly, the *min-entropy*, denoted $H_{\min}$, is defined as the *minimum* entropy $H(E_x)$ of the events E_x, i.e.,

$$H_{\min}(X) := \min_{x \in \mathcal{X}} h(x) = -\log_2 \max_{x \in \mathcal{X}} P_X(x) \ .$$

A slightly different entropy measure is the *max-entropy*, denoted $H_{\max}$. Despite the similarity of its name to the above measure, the definition does not rely on the entropy of events, but rather on the cardinality of the support $\text{supp} P_X := \{x \in \mathcal{X} : P_X(x) > 0\}$ of P_X,

$$H_{\max}(X) := \log_2 \bigl|\text{supp} P_X\bigr|.$$

It is easy to verify that the entropies defined above are related by

$$H_{\min}(X) \leq H(X) \leq H_{\max}(X) \ , \tag{3.2}$$

with equality if the probability mass function P_X is flat. Furthermore, they have various properties in common. The following holds for H, $H_{\min}$, and $H_{\max}$; to keep the notation simple, however, we only write H.

1. H is invariant under permutations of the elements, i.e., $H(X) = H(\pi(X))$, for any permutation π.

2. H is nonnegative.[3]

3. H is upper bounded by the logarithm of the alphabet size, i.e., $H(X) \leq \log_2 |\mathcal{X}|$.

4. H equals zero if and only if exactly one of the entries of P_X equals one, i.e., if $|\text{supp} P_X| = 1$.

3.1.4 Conditional entropy

In information theory, one typically wants to quantify the uncertainty about some data X, given that one already has information Y. To capture such situations, we need to generalize the entropy measures introduced in Section 3.1.3.

Let X and Y be random variables with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, and define, analogously to (3.1),

$$h(x|y) := -\log_2 P_{X|Y=y}(x) \ , \tag{3.3}$$

for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then the *Shannon entropy of X conditioned on Y* is again defined as an expectation value,

$$H(X|Y) := \langle h(X|Y) \rangle = -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} P_{XY}(x,y) \log_2 P_{X|Y=y}(x) \ .$$

[3] Note that this will no longer be true for the conditional entropy of quantum states.

For the definition of the *min-entropy of X given Y*, the expectation value is replaced by a minimum, i.e.,

$$H_{\min}(X|Y) := \min_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} h(x|y) = -\log_2 \max_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} P_{X|Y=y}(x) \ .$$

Finally, the *max-entropy of X given Y* is defined by

$$H_{\max}(X|Y) := \max_{y\in\mathcal{Y}} \log_2 |\mathrm{supp} P_{X|Y=y}| \ .$$

The conditional entropies H, $H_{\min}$, and $H_{\max}$ satisfy the rules listed in Section 3.1.3. Furthermore, the entropies can only decrease when conditioning on an additional random variable Z, i.e.,

$$H(X|Y) \geq H(X|YZ) \ . \tag{3.4}$$

This relation is also known as *strong subadditivity* and we will prove it in the more general quantum case.

Finally, it is straightforward to verify that the Shannon entropy H satisfies the *chain rule*

$$H(X|YZ) = H(XY|Z) - H(Y|Z) \ .$$

In particular, if we omit the random variable Z, we get

$$H(X|Y) = H(XY) - H(Y)$$

that is, the uncertainty of X given Y can be seen as the uncertainty about the pair (X, Y) minus the uncertainty about Y. We note here that a slightly modified version of the chain rule also holds for $H_{\min}$ and $H_{\max}$, but we will not go further into this.

3.1.5 Mutual information

Let X and Y be two random variables. The *(Shannon) mutual information between X and Y*, denoted $I(X:Y)$ is defined as the amount by which the Shannon entropy on X decreases when one learns Y,

$$I(X:Y) := H(X) - H(X|Y) \ .$$

More generally, given an additional random variable Z, the *(Shannon) mutual information between X and Y conditioned on Z*, $I(X:Y|Z)$, is defined by

$$I(X:Y|Z) := H(X|Z) - H(X|YZ) \ .$$

It is easy to see that the mutual information is symmetric under exchange of X and Y, i.e.,

$$I(X:Y|Z) = I(Y:X|Z) \ .$$

Furthermore, because of the strong subadditivity (3.4), the mutual information cannot be negative, and $I(X:Y) = 0$ holds if and only if X and Y are mutually independent. More generally, $I(X:Y|Z) = 0$ if and only if $X \leftrightarrow Z \leftrightarrow Y$ is a Markov chain.

3.1.6 Smooth min- and max- entropies

The dependency of the min- and max-entropy of a random variable on the underlying probability mass functions is discontinuous. To see this, consider a random variable X with alphabet $\{1, \dots, 2^\ell\}$ and probability mass function P_X^ε given by

$$
\begin{aligned}
P_X^\varepsilon(1) &= 1 - \varepsilon \\
P_X^\varepsilon(x) &= \frac{\varepsilon}{2^\ell - 1} \quad \text{if } x > 1 \ ,
\end{aligned}
$$

where $\varepsilon \in [0, 1]$. It is easy to see that, for $\varepsilon = 0$,

$$H_{\max}(X)_{P_X^0} = 0$$

whereas, for any $\varepsilon > 0$,

$$H_{\max}(X)_{P_X^\varepsilon} = \ell \ .$$

Note also that the trace distance between the two distributions satisfies $\delta(P_X^0, P_X^\varepsilon) = \varepsilon$. That is, an arbitrarily small change in the distribution can change the entropy $H_{\max}(X)$ by an arbitrary amount. In contrast, a small change of the underlying probability mass function is often irrelevant in applications. This motivates the following definition of *smooth* min- and max-entropies, which extends the above definition.

Let X and Y be random variables with joint probability mass function P_{XY}, and let $\varepsilon \geq 0$. The *ε-smooth min-entropy of X conditioned on Y* is defined as

$$H_{\min}^\varepsilon(X|Y) := \max_{Q_{XY} \in \mathcal{B}^\varepsilon(P_{XY})} H_{\min}(X|Y)_{Q_{XY}}$$

where the maximum ranges over the ε-ball $\mathcal{B}^\varepsilon(P_{XY})$ of probability mass functions Q_{XY} satisfying $\delta(P_{XY}, Q_{XY}) \leq \varepsilon$. Similarly, the *$\varepsilon$-smooth max-entropy of X conditioned on Y* is defined as

$$H_{\max}^\varepsilon(X|Y) := \min_{Q_{XY} \in \mathcal{B}^\varepsilon(P_{XY})} H_{\max}(X|Y)_{Q_{XY}} \ .$$

Note that the original definitions of $H_{\min}$ and $H_{\max}$ can be seen as the special case where $\varepsilon = 0$.

3.1.7 Shannon entropy as a special case of min- and max-entropy

We have already seen that the Shannon entropy always lies between the min- and the max-entropy (see (3.2)). In the special case of n-tuples of *i.i.d.* random variables, the gap between $H_{\min}^\varepsilon$ and $H_{\max}^\varepsilon$ approaches zero with increasing n, which means that all entropies become identical. This is expressed by the following lemma.

Lemma 3.1.4. *For any $n \in \mathbb{N}$, let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sequence of i.i.d. pairs of random variables, i.e., $P_{X_1 Y_1 \cdots X_n Y_n} = P_{XY}^{\times n}$. Then*

$$
\begin{aligned}
H(X|Y)_{P_{XY}} &= \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} H_{\min}^\varepsilon(X_1 \cdots X_n | Y_1 \cdots Y_n) \\
H(X|Y)_{P_{XY}} &= \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} H_{\max}^\varepsilon(X_1 \cdots X_n | Y_1 \cdots Y_n) \ .
\end{aligned}
$$

Proof. The lemma is a consequence of the law of large numbers (see Section 2.3.7), applied to the random variables $Z_i := h(X_i|Y_i)$, for $h(x|y)$ defined by (3.3). More details are given in the exercises. □

3.2 An example application: channel coding

3.2.1 Definition of the problem

Consider the following scenario. A sender, traditionally called *Alice*, wants to send a message M to a receiver, *Bob*. They are connected by a communication channel $\mathbf{p}$ that takes inputs X from Alice and outputs Y on Bob's side (see Section 2.3.8). The channel might be noisy, which means that Y can differ from X. The challenge is to find an appropriate encoding scheme that allows Bob to retrieve the correct message M, except with a small error probability ε. As we shall see, ε can always be made arbitrarily small (at the cost of the amount of information that can be transmitted), but it is generally impossible to reach $\varepsilon = 0$, i.e., Bob cannot retrieve M with absolute certainty.

To describe the encoding and decoding process, we assume without loss of generality[4] that the message M is represented as an ℓ-bit string, i.e., M takes values from the set $\{0,1\}^\ell$. Alice then applies an *encoding function* $\mathrm{enc}_\ell : \{0,1\}^\ell \to \mathcal{X}$ that maps M to a channel input X. On the other end of the line, Bob applies a *decoding function* $\mathrm{dec}_\ell : \mathcal{Y} \to \{0,1\}^\ell$ to the channel output Y in order to retrieve M'.

$$M \xrightarrow[\mathrm{enc}_\ell]{} X \xrightarrow[\mathbf{p}]{} Y \xrightarrow[\mathrm{dec}_\ell]{} M' \; . \tag{3.5}$$

The transmission is successful if $M = M'$. More generally, for any fixed encoding and decoding procedures enc_ℓ and dec_ℓ, and for any message $m \in \{0,1\}^\ell$, we can define

$$p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(m) := P[\mathrm{dec}_\ell \circ \mathbf{p} \circ \mathrm{enc}_\ell(M) \neq M | M = m]$$

as the probability that the decoded message $M' := \mathrm{dec}_\ell \circ \mathbf{p} \circ \mathrm{enc}_\ell(M)$ generated by the process (3.5) does not coincide with M.[5]

In the following, we analyze the maximum number of message bits ℓ that can be transmitted in one use of the channel $\mathbf{p}$ if we tolerate a maximum error probability ε,

$$\ell^\varepsilon(\mathbf{p}) := \max\{\ell \in \mathbb{N} : \exists\, \mathrm{enc}_\ell, \mathrm{dec}_\ell : \max_m p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(m) \leq \varepsilon\} \; .$$

3.2.2 The general channel coding theorem

The *channel coding theorem* provides a lower bound on the quantity $\ell^\varepsilon(\mathbf{p})$. It is easy to see from the formula below that reducing the maximum tolerated error probability by a factor of 2 comes at the cost of reducing the number of bits that can be transmitted reliably by 1. It can also be shown that the bound is almost tight (up to terms $\log_2 \frac{1}{\varepsilon}$).

[4] Note that all our statements will be independent of the actual representation of M. The only quantity that matters is the alphabet size of M, i.e., the total number of possible values.

[5] Here we interpret a channel as a probabilistic function from the input to the output alphabets.

Theorem 3.2.1. *For any channel* $\mathbf{p}$ *and any* $\varepsilon \geq 0$*,*

$$\ell^{\varepsilon}(\mathbf{p}) \geq \max_{P_X}\big(H_{\min}(X) - H_{\max}(X|Y)\big) - \log_2 \frac{1}{\varepsilon} - 3 \ ,$$

where the entropies on the right hand side are evaluated for the random variables X *and* Y *jointly distributed according to* $P_{XY} = P_X\mathbf{p}$*.*[6]

The proof idea is illustrated in Fig. 3.1.

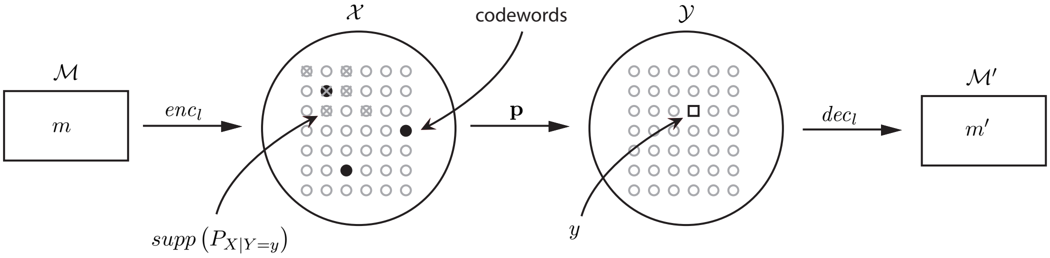


Figure 3.1: The figure illustrates the proof idea of the channel coding theorem. The range of the encoding function enc_ℓ is called *code* and their elements are the *codewords.*

Proof. The argument is based on a *randomized construction* of the encoding function. Let P_X be the distribution that maximizes the right hand side of the claim of the theorem and let ℓ be

$$\ell = \lfloor H_{\min}(X) - H_{\max}(X|Y) - \log_2 \frac{2}{\varepsilon} \rfloor. \tag{3.6}$$

In a first step, we consider an encoding function enc_ℓ *chosen at random* by assigning to each $m \in \{0,1\}^\ell$ a value $\mathrm{enc}_\ell(m) := X$ where X is chosen according to P_X. We then show that for a decoding function dec_ℓ that maps $y \in \mathcal{Y}$ to an arbitrary value $m' \in \{0,1\}^\ell$ that is *compatible* with y, i.e., $\mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}$, the error probability for a message M chosen uniformly at random satisfies

$$\big\langle p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(M)\big\rangle = P[\mathrm{dec}_\ell \circ \mathbf{p} \circ \mathrm{enc}_\ell(M) \neq M] \leq \frac{\varepsilon}{2} \ . \tag{3.7}$$

In a second step, we use this bound to show that there exist $\mathrm{enc}'_{\ell-1}$ and $\mathrm{dec}'_{\ell-1}$ such that

$$p_{\mathrm{err}}^{\mathrm{enc}'_{\ell-1},\mathrm{dec}'_{\ell-1}}(m) \leq \varepsilon \quad \forall m \in \{0,1\}^{\ell-1} \ . \tag{3.8}$$

[6] See also (2.7).

We then have

$$\begin{aligned}\ell^{\varepsilon}(\mathbf{p}) &\geq \ell - 1 \\ &= \lfloor H_{\min}(X) - H_{\max}(X|Y) - \log_2(2/\varepsilon) \rfloor - 1 \\ &\geq H_{\min}(X) - H_{\max}(X|Y) - \log_2(1/\varepsilon) - 3 .\end{aligned}$$

To prove (3.7), let enc_ℓ and M be chosen at random as described, let $Y := \mathbf{p} \circ \mathrm{enc}_\ell(M)$ be the channel output, and let $M' := \mathrm{dec}_\ell(Y)$ be the decoded message. We then consider any pair (m, y) such that $P_{MY}(m, y) > 0$. It is easy to see that, conditioned on the event that $(M, Y) = (m, y)$, the decoding function dec_ℓ described above can only fail, i.e., produce an $M' \neq M$, if there exists $m' \neq m$ such that $\mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}$. Hence, the probability that the decoding fails is bounded by

$$P[M \neq M' | M = m, Y = y] \leq P[\exists m' \neq m : \mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}] \ . \tag{3.9}$$

Furthermore, by the union bound, we have

$$P[\exists m' \neq m : \mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}] \leq \sum_{m' \neq m} P[\mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}] \ .$$

Because, by construction, $\mathrm{enc}_\ell(m')$ is a value chosen at random according to the distribution P_X, the probability in the sum on the right hand side of the inequality is given by

$$\begin{aligned}P[\mathrm{enc}_\ell(m') \in \mathrm{supp} P_{X|Y=y}] &= \sum_{x \in \mathrm{supp} P_{X|Y=y}} P_X(x) \\ &\leq |\mathrm{supp} P_{X|Y=y}| \max_x P_X(x) \\ &\leq 2^{-(H_{\min}(X) - H_{\max}(X|Y))} \ ,\end{aligned}$$

where the last inequality follows from the definitions of $H_{\min}$ and $H_{\max}$. Combining this with the above and observing that there are only $2^\ell - 1$ values $m' \neq m$, we find

$$P[M \neq M' | M = m, Y = y] \leq 2^{\ell - (H_{\min}(X) - H_{\max}(X|Y))} \leq \frac{\varepsilon}{2} \ .$$

Because this holds for any m and y, we have

$$P[M \neq M'] \leq \max_{m,y} P[M \neq M' | M = m, Y = y] \leq \frac{\varepsilon}{2} \ .$$

This immediately implies that (3.7) holds *on average* over all choices of enc_ℓ. But this also implies that there exists at least one specific choice for enc_ℓ such that (3.7) holds.

It remains to show inequality (3.8). For this, we divide the set of messages $\{0,1\}^\ell$ into two equally large sets $\underline{\mathcal{M}}$ and $\overline{\mathcal{M}}$ such that $p_{\mathrm{err}}^{\mathrm{enc}_\ell, \mathrm{dec}_\ell}(\underline{m}) \leq p_{\mathrm{err}}^{\mathrm{enc}_\ell, \mathrm{dec}_\ell}(\overline{m})$ for any $\underline{m} \in \underline{\mathcal{M}}$ and $\overline{m} \in \overline{\mathcal{M}}$. We then have

$$\max_{m \in \underline{\mathcal{M}}} p_{\mathrm{err}}^{\mathrm{enc}_\ell, \mathrm{dec}_\ell}(m) \leq \min_{m \in \overline{\mathcal{M}}} p_{\mathrm{err}}^{\mathrm{enc}_\ell, \mathrm{dec}_\ell}(m) \leq 2^{-(\ell - 1)} \sum_{m \in \overline{\mathcal{M}}} p_{\mathrm{err}}^{\mathrm{enc}_\ell, \mathrm{dec}_\ell}(m) \ .$$

Using (3.7), we conclude

$$\max_{m \in \underline{\mathcal{M}}} p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(m) \leq 2 \sum_{m \in \{0,1\}^\ell} 2^{-\ell} p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(m) = 2\big\langle p_{\mathrm{err}}^{\mathrm{enc}_\ell,\mathrm{dec}_\ell}(M)\big\rangle \leq \varepsilon \ .$$

Inequality (3.8) then follows by defining $\mathrm{enc}'_{\ell-1}$ as the encoding function enc_ℓ restricted to $\underline{\mathcal{M}}$, and adapting the decoding function accordingly.

□

3.2.3 Channel coding for i.i.d. channels

Realistic communication channels (e.g., an optical fiber) can usually be used repeatedly. Moreover, such channels often are accurately described by an i.i.d. noise model. In this case, the transmission of n subsequent signals over the physical channel corresponds to a single use of a channel of the form $\mathbf{p}^{\times n} = \mathbf{p} \times \cdots \mathbf{p}$. To determine the amount of information that can be transmitted from a sender to a receiver using the physical channel n times is thus given by Theorem 3.2.1 applied to $\mathbf{p}^{\times n}$.

In applications, the number n of channel uses is typically large. It is thus convenient to measure the capacity of a channel in terms of the asymptotic rate

$$\mathrm{rate}(\mathbf{p}) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \ell^{\varepsilon}(\mathbf{p}^{\times n}) \tag{3.10}$$

The computation of the rate will rely on the following corollary, which follows from Theorem 3.2.1 and the definition of smooth entropies.

Corollary 3.2.2. *For any channel* $\mathbf{p}$ *and any* $\varepsilon, \varepsilon', \varepsilon'' \geq 0$,

$$\ell^{\varepsilon+\varepsilon'+\varepsilon''}(\mathbf{p}) \geq \max_{P_X}\big(H_{\min}^{\varepsilon'}(X) - H_{\max}^{\varepsilon''}(X|Y)\big) - \log_2 \frac{1}{\varepsilon} - 3$$

where the entropies on the right hand side are evaluated for $P_{XY} := P_X \mathbf{p}$.

Combining this with Lemma 3.1.4, we get the following lower bound for the rate of a channel.

Theorem 3.2.3. *For any channel* $\mathbf{p}$

$$\mathrm{rate}(\mathbf{p}) \geq \max_{P_X}\big(H(X) - H(X|Y)\big) = \max_{P_X} I(X:Y) \ .$$

where the entropies on the right hand side are evaluated for $P_{XY} := P_X \mathbf{p}$.

3.2.4 The converse

We conclude our treatment of channel coding with a proof sketch which shows that the bound given in Theorem 3.2.3 is tight. The main ingredient to the proof is the *information processing inequality*

$$I(U:W) \leq I(U:V)$$

which holds for any random variables such that $U \leftrightarrow V \leftrightarrow W$ is a Markov chain. The inequality is proved by

$$I(U:W) \leq I(U:W) + I(U:V|W) = I(U:VW) = I(U:V) + I(U:W|V) = I(U:V) \ ,$$

where the first inequality holds because the mutual information cannot be negative and the last equality follows because $I(U:W|V) = 0$ (see end of Section 3.1.5). The remaining equalities are essentially rewritings of the chain rule (for the Shannon entropy).

Let now M, X, Y, and M' be defined as in (3.5). If the decoding is successful then $M = M'$ which implies

$$H(M) = I(M:M') \ . \tag{3.11}$$

Applying the information processing inequality first to the Markov chain $M \leftrightarrow Y \leftrightarrow M'$ and then to the Markov chain $M \leftrightarrow X \leftrightarrow Y$ gives

$$I(M:M') \leq I(M:Y) \leq I(X:Y) \ .$$

Combining this with (3.11) and assuming that the message M is uniformly distributed over the set $\{0,1\}^\ell$ of bitstrings of length ℓ gives

$$\ell = H(M) \leq \max_{P_X} I(X:Y) \ .$$

It is straightforward to verify that the statement still holds approximately if ℓ on the left hand side is replaced by ℓ^ε, for some small decoding error $\varepsilon > 0$. Taking the limits as in (3.10) finally gives

$$\mathrm{rate}(\mathbf{p}) \leq \max_{P_X} I(X:Y) \ .$$

4 Quantum States and Operations

The mathematical formalism used in quantum information theory to describe quantum mechanical systems is in many ways more general than that of typical introductory books on quantum mechanics. This is why we devote a whole chapter to it. The main concepts to be treated in the following are *density operators*, which represent the state of a system, as well as *positive-valued measures (POVMs)* and *completely positive maps (CPMs)*, which describe measurements and, more generally, the evolution of a system.

4.1 Preliminaries

4.1.1 Hilbert spaces and operators on them

An *inner product space* is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) equipped with an inner product $(\cdot, \cdot)$. A *Hilbert space* $\mathcal{H}$ is an inner product space such that the metric defined by the norm $\|\alpha\| \equiv \sqrt{(\alpha, \alpha)}$ is *complete*, i.e., every Cauchy sequence is convergent. We will often deal with finite-dimensional spaces, where the completeness condition always holds, i.e., inner product spaces are equivalent to Hilbert spaces.

We denote the set of *homomorphisms* (i.e., the linear maps) from a Hilbert space $\mathcal{H}$ to a Hilbert space $\mathcal{H}'$ by $\mathrm{Hom}(\mathcal{H}, \mathcal{H}')$. Furthermore, $\mathrm{End}(\mathcal{H})$ is the set of *endomorphism* (i.e., the homomorphisms from a space to itself) on $\mathcal{H}$, that is, $\mathrm{End}(\mathcal{H}) = \mathrm{Hom}(\mathcal{H}, \mathcal{H})$. The identity operator $\alpha \mapsto \alpha$ that maps any vector $\alpha \in \mathcal{H}$ to itself is denoted by id.

The *adjoint* of a homomorphism $S \in \mathrm{Hom}(\mathcal{H}, \mathcal{H}')$, denoted S^*, is the unique operator in $\mathrm{Hom}(\mathcal{H}', \mathcal{H})$ such that

$$(\alpha', S\alpha) = (S^* \alpha', \alpha) ,$$

for any $\alpha \in \mathcal{H}$ and $\alpha' \in \mathcal{H}'$. In particular, we have $(S^*)^* = S$. If S is represented as a matrix, then the adjoint operation can be thought of as the conjugate transpose.

In the following, we list some properties of endomorphisms $S \in \mathrm{End}(\mathcal{H})$.

- S is *normal* if $SS^* = S^* S$.
- S is *unitary* if $SS^* = S^* S = \mathrm{id}$. Unitary operators S are always normal.
- S is *Hermitian* if $S^* = S$. Hermitian operators are always normal.
- S is *positive* if $(\alpha, S\alpha) \geq 0$ for all $\alpha \in \mathcal{H}$. Positive operators are always Hermitian. We will sometimes write $S \geq 0$ to express that S is positive.
- S is a *projector* if $SS = S$. Projectors are always positive.

Given an orthonormal basis $\{e_i\}_i$ of $\mathcal{H}$, we also say that S is *diagonal with respect to* $\{e_i\}_i$ if the matrix $(S_{i,j})$ defined by the elements $S_{i,j} = (e_i, S e_j)$ is diagonal.

4.1.2 The bra-ket notation

In this script, we will make extensive use of a variant of Dirac's *bra-ket notation*, where vectors are interpreted as operators. More precisely, we identify any vector $\alpha \in \mathcal{H}$ with an endomorphism $|\alpha\rangle \in \mathrm{Hom}(\mathbb{C}, \mathcal{H})$, called *ket*, and defined as

$$|\alpha\rangle : \quad \gamma \mapsto \alpha\gamma$$

for any $\gamma \in \mathbb{C}$. The adjoint $|\alpha\rangle^*$ of this mapping is called *bra* and denoted by $\langle\alpha|$. It is easy to see that $\langle\alpha|$ is an element of the *dual space* $\mathcal{H}^* := \mathrm{Hom}(\mathcal{H}, \mathbb{C})$, namely the linear functional defined by

$$\langle\alpha| : \quad \beta \mapsto (\alpha, \beta)$$

for any $\beta \in \mathcal{H}$.

Using this notation, the concatenation $\langle\alpha||\beta\rangle$ of a bra $\langle\alpha| \in \mathrm{Hom}(\mathcal{H}, \mathbb{C})$ with a ket $|\beta\rangle \in \mathrm{Hom}(\mathbb{C}, \mathcal{H})$ results in an element of $\mathrm{Hom}(\mathbb{C}, \mathbb{C})$, which can be identified with $\mathbb{C}$. It follows immediately from the above definitions that, for any $\alpha, \beta \in \mathcal{H}$,

$$\langle\alpha||\beta\rangle \equiv (\alpha, \beta) \ .$$

We will thus in the following denote the scalar product by $\langle\alpha|\beta\rangle$.

Conversely, the concatenation $|\beta\rangle\langle\alpha|$ is an element of $\mathrm{End}(\mathcal{H})$ (or, more generally, of $\mathrm{Hom}(\mathcal{H}, \mathcal{H}')$ if $\alpha \in \mathcal{H}$ and $\beta \in \mathcal{H}'$ are defined on different spaces). In fact, any endomorphism $S \in \mathrm{End}(\mathcal{H})$ can be written as a linear combination of such concatenations, i.e.,

$$S = \sum_i |\beta_i\rangle\langle\alpha_i|$$

for some families of vectors $\{\alpha_i\}_i$ and $\{\beta_i\}_i$. For example, the identity $\mathrm{id} \in \mathrm{End}(\mathcal{H})$ can be written as

$$\mathrm{id} = \sum_i |e_i\rangle\langle e_i|$$

for any basis $\{e_i\}$ of $\mathcal{H}$.

4.1.3 Tensor products

Given two Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, the *tensor product* $\mathcal{H}_A \otimes \mathcal{H}_B$ is defined as the Hilbert space spanned by elements of the form $|\alpha\rangle \otimes |\beta\rangle$, where $\alpha \in \mathcal{H}_A$ and $\beta \in \mathcal{H}_B$, such that the following equivalences hold

- $(\alpha + \alpha') \otimes \beta = \alpha \otimes \beta + \alpha' \otimes \beta$
- $\alpha \otimes (\beta + \beta') = \alpha \otimes \beta + \alpha \otimes \beta'$
- $\mathbf{0} \otimes \beta = \alpha \otimes \mathbf{0} = \mathbf{0}$

for any $\alpha, \alpha' \in \mathcal{H}_A$ and $\beta, \beta' \in \mathcal{H}_B$, where $\mathbf{0}$ denotes the zero vector. Furthermore, the inner product of $\mathcal{H}_A \otimes \mathcal{H}_B$ is defined by the linear extension (and completion) of

$$\langle \alpha \otimes \beta | \alpha' \otimes \beta' \rangle = \langle \alpha | \alpha' \rangle \langle \beta | \beta' \rangle \ .$$

For two homomorphisms $S \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}'_A)$ and $T \in \mathrm{Hom}(\mathcal{H}_B, \mathcal{H}'_B)$, the tensor product $S \otimes T$ is defined as

$$(S \otimes T)(\alpha \otimes \beta) \equiv (S\alpha) \otimes (T\beta) \tag{4.1}$$

for any $\alpha \in \mathcal{H}_A$ and $\beta \in \mathcal{H}_B$. The space spanned by the products $S \otimes T$ can be canonically identified[1] with the tensor product of the spaces of the homomorphisms, i.e.,

$$\mathrm{Hom}(\mathcal{H}_A, \mathcal{H}'_A) \otimes \mathrm{Hom}(\mathcal{H}_B, \mathcal{H}'_B) \cong \mathrm{Hom}(\mathcal{H}_A \otimes \mathcal{H}_B, \mathcal{H}'_A \otimes \mathcal{H}'_B) \ . \tag{4.2}$$

This identification allows us to write, for instance,

$$|\alpha\rangle \otimes |\beta\rangle = |\alpha \otimes \beta\rangle \ ,$$

for any $\alpha \in \mathcal{H}_A$ and $\beta \in \mathcal{H}_B$.

4.1.4 Trace and partial trace

The *trace* of an endomorphism $S \in \mathrm{End}(\mathcal{H})$ over a Hilbert space $\mathcal{H}$ is defined by[2]

$$\mathrm{tr}(S) \equiv \sum_i \langle e_i | S | e_i \rangle$$

where $\{e_i\}_i$ is any orthonormal basis of $\mathcal{H}$. The trace is well defined because the above expression is independent of the choice of the basis, as one can easily verify.

The trace operation tr is obviously linear, i.e.,

$$\mathrm{tr}(uS + vT) = u\mathrm{tr}(S) + v\mathrm{tr}(T) \ ,$$

for any $S, T \in \mathrm{End}(\mathcal{H})$ and $u, v \in \mathbb{C}$. It also commutes with the operation of taking the adjoint,[3]

$$\mathrm{tr}(S^*) = \mathrm{tr}(S)^* \ .$$

Furthermore, the trace is cyclic, i.e.,

$$\mathrm{tr}(ST) = \mathrm{tr}(TS) \ .$$

[1] That is, the mapping defined by (4.1) is an isomorphism between these two vector spaces.

[2] More precisely, the trace is only defined for *trace class operators* over a separable Hilbert space. However, all endomorphisms on a finite-dimensional Hilbert space are trace class operators.

[3] The adjoint of a complex number $\gamma \in \mathbb{C}$ is simply its complex conjugate.

Also, it is easy to verify[4] that the trace $\mathrm{tr}(S)$ of a positive operator $S \geq 0$ is positive. More generally

$$(S \geq 0) \wedge (T \geq 0) \implies \mathrm{tr}(ST) \geq 0 \ . \tag{4.3}$$

The *partial trace*[5] tr_B is a mapping from the endomorphisms $\mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$ on a product space $\mathcal{H}_A \otimes \mathcal{H}_B$ onto the endomorphisms $\mathrm{End}(\mathcal{H}_A)$ on $\mathcal{H}_A$. It is defined by the linear extension of the mapping.[6]

$$\mathrm{tr}_B : \quad S \otimes T \mapsto \mathrm{tr}(T) S \ ,$$

for any $S \in \mathrm{End}(\mathcal{H}_A)$ and $T \in \mathrm{End}(\mathcal{H}_B)$.

Similarly to the trace operation, the partial trace tr_B is linear and commutes with the operation of taking the adjoint. Furthermore, it commutes with the left and right multiplication with an operator of the form $T_A \otimes \mathrm{id}_B$ where $T_A \in \mathrm{End}(\mathcal{H}_A)$.[7] That is, for any operator $S_{AB} \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$,

$$\mathrm{tr}_B\bigl(S_{AB}(T_A \otimes \mathrm{id}_B)\bigr) = \mathrm{tr}_B(S_{AB}) T_A \tag{4.4}$$

and

$$\mathrm{tr}_B\bigl((T_A \otimes \mathrm{id}_B) S_{AB}\bigr) = T_A \mathrm{tr}_B(S_{AB}) \ . \tag{4.5}$$

We will also make use of the property that the trace on a bipartite system can be decomposed into partial traces on the individual subsystems, i.e.,

$$\mathrm{tr}(S_{AB}) = \mathrm{tr}(\mathrm{tr}_B(S_{AB})) \tag{4.6}$$

or, more generally, for an operator $S_{ABC} \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C)$,

$$\mathrm{tr}_{AB}(S_{ABC}) = \mathrm{tr}_A(\mathrm{tr}_B(S_{ABC})) \ .$$

4.1.5 Decompositions of operators and vectors

Spectral decomposition. Let $S \in \mathrm{End}(\mathcal{H})$ be normal and let $\{e_i\}_i$ be an orthonormal basis of $\mathcal{H}$. Then there exists a unitary $U \in \mathrm{End}(\mathcal{H})$ and an operator $D \in \mathrm{End}(\mathcal{H})$ which is diagonal with respect to $\{e_i\}_i$ such that

$$S = U D U^* \ .$$

[4] The assertion can, for instance, be proved using the spectral decomposition of S and T (see below for a review of the spectral decomposition).

[5] Here and in the following, we will use subscripts to indicate the space on which an operator acts.

[6] Alternatively, the partial trace tr_B can be defined as a product mapping $\mathcal{I} \otimes \mathrm{tr}$ where $\mathcal{I}$ is the identity operation on $\mathrm{End}(\mathcal{H}_A)$ and tr is the trace mapping elements of $\mathrm{End}(\mathcal{H}_B)$ to $\mathrm{End}(\mathbb{C})$. Because the trace is a completely positive map (see definition below) the same is true for the partial trace.

[7] More generally, the partial trace commutes with any mapping that acts like the identity on $\mathrm{End}(\mathcal{H}_B)$.

The spectral decomposition implies that, for any normal $S \in \mathrm{End}(\mathcal{H})$, there exists a basis $\{e_i\}_i$ of $\mathcal{H}$ with respect to which S is diagonal. That is, S can be written as

$$S = \sum_i \alpha_i |e_i\rangle\langle e_i| \tag{4.7}$$

where α_i are the eigenvalues of S.

Expression (4.7) can be used to give a meaning to a complex function $f : \mathbb{C} \to \mathbb{C}$ applied to a normal operator S. We define $f(S)$ by

$$f(S) \equiv \sum_i f(\alpha_i)|e_i\rangle\langle e_i| \ .$$

Polar decomposition. Let $S \in \mathrm{End}(\mathcal{H})$. Then there exists a unitary $U \in \mathrm{End}(\mathcal{H})$ such that

$$S = \sqrt{SS^*}U$$

and

$$S = U\sqrt{S^*S} \ .$$

Singular decomposition. Let $S \in \mathrm{End}(\mathcal{H})$ and let $\{e_i\}_i$ be an orthonormal basis of $\mathcal{H}$. Then there exist unitaries $U, V \in \mathrm{End}(\mathcal{H})$ and an operator $D \in \mathrm{End}(\mathcal{H})$ which is diagonal with respect to $\{e_i\}_i$ such that

$$S = VDU \ .$$

In particular, for any $S \in \mathrm{Hom}(\mathcal{H}, \mathcal{H}')$, there exist bases $\{e_i\}_i$ of $\mathcal{H}$ and $\{e'_i\}_i$ of $\mathcal{H}'$ such that the matrix defined by the elements (e'_i, Se_j) is diagonal.

Schmidt decomposition. The Schmidt decomposition can be seen as a version of the singular decomposition for vectors. The statement is that any vector $\Psi \in \mathcal{H}_A \otimes \mathcal{H}_B$ can be written in the form

$$\Psi = \sum_i \gamma_i e_i \otimes e'_i$$

where $e_i \in \mathcal{H}_A$ and $e'_i \in \mathcal{H}_B$ are eigenvectors of the operators $\rho_A := \mathrm{tr}_B(|\Psi\rangle\langle\Psi|)$ and $\rho_B := \mathrm{tr}_A(|\Psi\rangle\langle\Psi|)$, respectively, and where γ_i^2 are the corresponding eigenvalues. In particular, the existence of the Schmidt decomposition implies that ρ_A and ρ_B have the same nonzero eigenvalues.

4.1.6 Operator norms and the Hilbert-Schmidt inner product

The *Hilbert-Schmidt inner product* between two operators $S, T \in \mathrm{End}(\mathcal{H})$ is defined by

$$(S, T) := \mathrm{tr}(S^* T) \ .$$

The induced norm $\|S\|_2 := \sqrt{(S, S)}$ is called *Hilbert-Schmidt norm.* If S is normal with spectral decomposition $S = \sum_i \alpha_i |e_i\rangle\langle e_i|$ then

$$\|S\|_2 = \sqrt{\sum_i |\alpha_i|^2} \ .$$

An important property of the Hilbert-Schmidt inner product (S, T) is that it is positive whenever S and T are positive.

Lemma 4.1.1. *Let $S, T \in \mathrm{End}(\mathcal{H})$. If $S \geq 0$ and $T \geq 0$ then*

$$\mathrm{tr}(ST) \geq 0 \ .$$

Proof. If S is positive we have $S = \sqrt{S}^2$ and $T = \sqrt{T}^2$. Hence, using the cyclicity of the trace, we have

$$\mathrm{tr}(ST) = \mathrm{tr}(V^* V)$$

where $V = \sqrt{S}\sqrt{T}$. Because the trace of a positive operator is positive, it suffices to show that $V^* V \geq 0$. This, however, follows from the fact that, for any $\phi \in \mathcal{H}$,

$$\langle \phi | V^* V | \phi \rangle = \|V\phi\|^2 \geq 0 \ .$$

□

The *trace norm* of S is defined by

$$\|S\|_1 := \mathrm{tr}|S|$$

where

$$|S| := \sqrt{S^* S} \ .$$

If S is normal with spectral decomposition $S = \sum_i \alpha_i |e_i\rangle\langle e_i|$ then

$$\|S\|_1 = \sum_i |\alpha_i| \ .$$

The following lemma provides a useful characterization of the trace norm.

Lemma 4.1.2. *For any $S \in \mathrm{End}(\mathcal{H})$,*

$$\|S\|_1 = \max_U |\mathrm{tr}(US)|$$

where U ranges over all unitaries on $\mathcal{H}$.

Proof. We need to show that, for any unitary U,

$$|\mathrm{tr}(US)| \leq \mathrm{tr}|S| \tag{4.8}$$

with equality for some appropriately chosen U.

Let $S = V|S|$ be the polar decomposition of S. Then, using the Cauchy-Schwarz inequality

$$|\mathrm{tr}(Q^* R)| \leq \|Q\|_2 \|R\|_2$$

with $Q := \sqrt{|S|}V^* U^*$ and $R := \sqrt{|S|}$ we find

$$\big|\mathrm{tr}(US)\big| = \big|\mathrm{tr}(UV|S|)\big| = \big|\mathrm{tr}(UV\sqrt{|S|}\sqrt{|S|})\big| \leq \sqrt{\mathrm{tr}(UV|S|V^*U^*)\mathrm{tr}(|S|)} = \mathrm{tr}(|S|)\ ,$$

which proves (4.8). Finally, it is easy to see that equality holds for $U := V^*$. □

4.1.7 The vector space of Hermitian operators

The set of Hermitian operators on a Hilbert space $\mathcal{H}$, in the following denoted $\mathrm{Herm}(\mathcal{H})$, forms a real vector space. Furthermore, equipped with the Hilbert Schmidt inner product defined in the previous section, $\mathrm{Herm}(\mathcal{H})$ is an inner product space.

If $\{e_i\}_i$ is an orthonormal basis of $\mathcal{H}$ then the set of operators $E_{i,j}$ defined by

$$E_{i,j} := \begin{cases} \frac{1}{2}|e_i\rangle\langle e_j| + \frac{1}{2}|e_j\rangle\langle e_i| & \text{if } i \leq j \\ \frac{i}{2}|e_i\rangle\langle e_j| - \frac{i}{2}|e_j\rangle\langle e_i| & \text{if } i > j \end{cases}$$

forms an orthonormal basis of $\mathrm{Herm}(\mathcal{H})$. We conclude from this that

$$\dim \mathrm{Herm}(\mathcal{H}) = (\dim \mathcal{H})^2\ . \tag{4.9}$$

For two Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, we have in analogy to (4.2)

$$\mathrm{Herm}(\mathcal{H}_A) \otimes \mathrm{Herm}(\mathcal{H}_B) \cong \mathrm{Herm}(\mathcal{H}_A \otimes \mathcal{H}_B)\ . \tag{4.10}$$

To see this, consider the canonical mapping from $\mathrm{Herm}(\mathcal{H}_A) \otimes \mathrm{Herm}(\mathcal{H}_B)$ to $\mathrm{Herm}(\mathcal{H}_A \otimes \mathcal{H}_B)$ defined by (4.1). It is easy to verify that this mapping is injective. Furthermore, because by (4.9) the dimension of both spaces equals $\dim(\mathcal{H}_A)^2 \dim(\mathcal{H}_B)^2$, it is a bijection, which proves (4.10).

4.2 Postulates of quantum mechanics

Despite more than one century of research, numerous questions related to the foundations of quantum mechanics are still unsolved (and highly disputed). For example, no fully satisfying explanation for the fact that quantum mechanics has its particular mathematical structure has been found so far. As a consequence, some of the aspects to be discussed in the following, e.g., the postulates of quantum mechanics, might appear to lack a clear motivation.

In this section, we pursue one of the standard approaches to quantum mechanics. It is based on a number of postulates about the states of physical systems as well as their evolution. (For more details, we refer to Section 2 of [10], where an equivalent approach is described.) The postulates are as follows:

1. States: The set of states of an isolated physical system is in one-to-one correspondence to the projective space of a Hilbert space $\mathcal{H}$. In particular, any physical state can be represented by a *normalized vector* $\phi \in \mathcal{H}$ which is unique up to a phase factor. In the following, we will call $\mathcal{H}$ the *state space* of the system.[8]

2. Composition: For two physical systems with state spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, the state space of the product system is isomorphic to $\mathcal{H}_A \otimes \mathcal{H}_B$. Furthermore, if the individual systems are in states $\phi \in \mathcal{H}_A$ and $\phi' \in \mathcal{H}_B$, then the joint state is
$$\Psi = \phi \otimes \phi' \in \mathcal{H}_A \otimes \mathcal{H}_B \ .$$

3. Evolutions: For any possible evolution of an isolated physical system with state space $\mathcal{H}$ and for any fixed time interval $[t_0, t_1]$ there exists a *unitary* U describing the mapping of states $\phi \in \mathcal{H}$ at time t_0 to states
$$\phi' = U\phi$$
at time t_1. The unitary U is unique up to a phase factor.

4. Measurements: For any measurement on a physical system with state space $\mathcal{H}$ there exists an *observable* O with the following properties. O is a Hermitian operator on $\mathcal{H}$ such that each eigenvalue x of O corresponds to a possible measurement outcome. If the system is in state $\phi \in \mathcal{H}$, then the probability of observing outcome x when applying the measurement is given by
$$P_X(x) = \mathrm{tr}(P_x |\phi\rangle\langle\phi|)$$
where P_x denotes the projector onto the eigenspace belonging to the eigenvalue x, i.e., $O = \sum_x x P_x$. Finally, the state ϕ'_x of the system after the measurement, conditioned on the event that the outcome is x, equals
$$\phi'_x := \sqrt{\frac{1}{P_X(x)}} P_x \phi \ .$$

4.3 Quantum states

In quantum information theory, one often considers situations where the state or the evolution of a system is only partially known. For example, we might be interested in

[8] In quantum mechanics, the elements $\phi \in \mathcal{H}$ are also called *wave functions*.

a scenario where a system might be in two possible states ϕ_0 or ϕ_1, chosen according to a certain probability distribution. Another simple example is a system consisting of two correlated parts A and B in a state

$$\Psi = \sqrt{\frac{1}{2}}\big(e_0 \otimes e_0 + e_1 \otimes e_1\big) \in \mathcal{H}_A \otimes \mathcal{H}_B \ , \tag{4.11}$$

where $\{e_0, e_1\}$ are orthonormal vectors in $\mathcal{H}_A = \mathcal{H}_B$. From the point of view of an observer that has no access to system B, the state of A does not correspond to a fixed vector $\phi \in \mathcal{H}_A$, but is rather described by a mixture of such states. In this section, we introduce the density operator formalism, which allows for a simple and convenient characterization of such situations.

4.3.1 Density operators — Definition and properties

The notion of *density operators* has been introduced independently by von Neumann and Landau in 1927. Since then, it has been widely used in quantum statistical mechanics and, more recently, in quantum information theory.

Definition 4.3.1. A *density operator* ρ on a Hilbert space $\mathcal{H}$ is a normalized positive operator on $\mathcal{H}$, i.e., $\rho \geq 0$ and $\mathrm{tr}(\rho) = 1$. The set of density operators on $\mathcal{H}$ is denoted by $\mathcal{S}(\mathcal{H})$. A density operator is said to be *pure* if it has the form $\rho = |\phi\rangle\langle\phi|$. If $\mathcal{H}$ is d-dimensional and ρ has the form $\rho = \frac{1}{d} \cdot \mathrm{id}$ then it is called *fully mixed.*

It follows from the spectral decomposition theorem that any density operator can be written in the form

$$\rho = \sum_x P_X(x)|e_x\rangle\langle e_x|$$

where P_X is the probability mass function defined by the eigenvalues $P_X(x)$ of ρ and $\{e_x\}_x$ are the corresponding eigenvectors. Given this representation, it is easy to see that a density operator is pure if and only if exactly one of the eigenvalues equals 1 whereas the others are 0. In particular, we have the following lemma.

Lemma 4.3.2. *A density operator ρ is pure if and only if* $\mathrm{tr}(\rho^2) = 1$.

4.3.2 Quantum-mechanical postulates in the language of density operators

In a first step, we adapt the postulates of Section 4.2 to the notion of density operators. At the same time, we generalize them to situations where the evolution and measurements only act on parts of a composite system.

1. States: The states of a physical system are represented as density operators on a state space $\mathcal{H}$. For an isolated system whose state, represented as a vector, is $\phi \in \mathcal{H}$, the corresponding density operator is defined by $\rho := |\phi\rangle\langle\phi|$.[9]

[9] Note that this density operator is pure.

2. Composition: The states of a composite system with state spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ are represented as density operators on $\mathcal{H}_A \otimes \mathcal{H}_B$. Furthermore, if the states of the individual subsystems are independent of each other and represented by density operators ρ_A and ρ_B, respectively, then the state of the joint system is $\rho_A \otimes \rho_B$.

3. Evolution: Any isolated evolution of a subsystem of a composite system over a fixed time interval $[t_0, t_1]$ corresponds to a unitary on the state space $\mathcal{H}$ of the subsystem. For a composite system with state space $\mathcal{H}_A \otimes \mathcal{H}_B$ and isolated evolutions on both subsystems described by U_A and U_B, respectively, any state ρ_{AB} at time t_0 is transformed into the state[10]

$$\rho'_{AB} = (U_A \otimes U_B)(\rho_{AB})(U_A^* \otimes U_B^*) \tag{4.12}$$

   at time t_1.[11]

4. Measurement: Any isolated measurement on a subsystem of a composite system is specified by a Hermitian operator, called *observable*. When applying a measurement $O_A = \sum_x x P_x$ on the first subsystem of a composite system $\mathcal{H}_A \otimes \mathcal{H}_B$ whose state is ρ_{AB}, the probability of observing outcome x is

$$P_X(x) = \mathrm{tr}(P_x \otimes \mathrm{id}_B \rho_{AB}) \tag{4.13}$$

   and the post-measurement state conditioned on this outcome is

$$\rho'_{AB,x} = \frac{1}{P_X(x)}(P_x \otimes \mathrm{id}_B)\rho_{AB}(P_x \otimes \mathrm{id}_B) \ . \tag{4.14}$$

It is straightforward to verify that these postulates are indeed compatible with those of Section 4.2. What is new is merely the fact that the evolution and measurements can be restricted to individual subsystems of a composite system. As we shall see, this extension is, however, very powerful because it allows us to examine parts of a subsystem without the need of keeping track of the state of the entire system.

4.3.3 Partial trace and purification

Let $\mathcal{H}_A \otimes \mathcal{H}_B$ be a composite quantum system which is initially in a state $\rho_{AB} = |\Psi\rangle\langle\Psi|$ for some $\Psi \in \mathcal{H}_A \otimes \mathcal{H}_B$. Consider now an experiment which is restricted to the first subsystem. More precisely, assume that subsystem A undergoes an isolated evolution, described by a unitary U_A, followed by an isolated measurement, described by an observable $O_A = \sum_x x P_x$.

According to the above postulates, the probability of observing an outcome x is then given by

$$P_X(x) = \mathrm{tr}\big((P_x \otimes \mathrm{id}_B)(U_A \otimes U_B)\rho_{AB}(U_A^* \otimes U_B^*)\big)$$

[10] In particular, if $\mathcal{H}_B = \mathbb{C}$ is trivial, this expression equals $\rho'_A = U_A \rho_A U_A^*$.

[11] By induction, this postulate can be readily generalized to composite systems with more than two parts.

where U_B is an arbitrary isolated evolution on $\mathcal{H}_B$. Using rules (4.6) and (4.4), this can be transformed into

$$P_X(x) = \mathrm{tr}\big(P_x U_A \mathrm{tr}_B(\rho_{AB}) U_A^\dagger\big) \ ,$$

which is independent of U_B. Observe now that this expression could be obtained equivalently by simply applying the above postulates to the *reduced state* $\rho_A := \mathrm{tr}_B(\rho_{AB})$. In other words, the reduced state already fully characterizes all observable properties of the subsystem $\mathcal{H}_A$.

This principle, which is sometimes called *locality*, plays a crucial role in many information-theoretic considerations. For example, it implies that it is impossible to influence system $\mathcal{H}_A$ by local actions on system $\mathcal{H}_B$. In particular, communication between the two subsystems is impossible as long as their evolution is determined by local operations $U_A \otimes U_B$.

In this context, it is important to note that the reduced state ρ_A of a pure joint state ρ_{AB} is not necessarily pure. For instance, if the joint system is in state $\rho_{AB} = |\Psi\rangle\langle\Psi|$ for Ψ defined by (4.11) then

$$\rho_A = \frac{1}{2}|e_0\rangle\langle e_0| + \frac{1}{2}|e_1\rangle\langle e_1| \ , \tag{4.15}$$

i.e., the density operator ρ_A is fully mixed. In the next section, we will give an interpretation of non-pure, or *mixed*, density operators.

Conversely, any mixed density operator can be seen as part of a pure state on a larger system. More precisely, given ρ_A on $\mathcal{H}_A$, there exists a pure density operator ρ_{AB} on a joint system $\mathcal{H}_A \otimes \mathcal{H}_B$ (where the dimension of $\mathcal{H}_B$ is at least as large as the rank of ρ_A) such that

$$\rho_A = \mathrm{tr}_B(\rho_{AB}) \tag{4.16}$$

A pure density operator ρ_{AB} for which (4.16) holds is called a *purification* of ρ_A.

4.3.4 Mixtures of states

Consider a quantum system $\mathcal{H}_A$ whose state depends on a classical value Z and let $\rho_A^z \in \mathcal{S}(\mathcal{H}_A)$ be the state of the system conditioned on the event $Z = z$. Furthermore, consider an observer who does not have access to Z, that is, from his point of view, Z can take different values distributed according to a probability mass function P_Z.

Assume now that the system $\mathcal{H}_A$ undergoes an evolution U_A followed by a measurement $O_A = \sum_x x P_x$ as above. Then, according to the postulates of quantum mechanics, the probability mass function of the measurement outcomes x conditioned on the event $Z = z$ is given by

$$P_{X|Z=z}(x) = \mathrm{tr}(P_x U_A \rho_A^z U_A^*) \ .$$

Hence, from the point of view of the observer who is unaware of the value Z, the probability mass function of X is given by

$$P_X(x) = \sum_z P_Z(z) P_{X|Z=z}(x) \ .$$

By linearity, this can be rewritten as

$$P_X(x) = \mathrm{tr}(P_x U_A \rho_A U_A^*) \ . \tag{4.17}$$

where

$$\rho_A := \sum_z P_Z(z) \rho_A^z \ .$$

Alternatively, expression (4.17) can be obtained by applying the postulates of Section 4.3.2 directly to the density operator ρ_A defined above. In other words, from the point of view of an observer not knowing Z, the situation is consistently characterized by ρ_A.

We thus arrive at a new interpretation of mixed density operators. For example, the density operator

$$\rho_A = \frac{1}{2}|e_0\rangle\langle e_0| + \frac{1}{2}|e_1\rangle\langle e_1| \tag{4.18}$$

defined by (4.15) corresponds to a situation where either state e_0 or e_1 is prepared, each with probability $\frac{1}{2}$. The *decomposition* according to (4.18) is, however, not unique. In fact, the same state could be written as

$$\rho_A = \frac{1}{2}|\tilde{e}_0\rangle\langle \tilde{e}_0| + \frac{1}{2}|\tilde{e}_1\rangle\langle \tilde{e}_1|$$

where $\tilde{e}_0 := \frac{1}{2}(e_0 + e_1)$ and $\tilde{e}_1 := \frac{1}{2}(e_0 - e_1)$. That is, the system could equivalently be interpreted as being prepared either in state $\tilde{e}_0$ or $\tilde{e}_1$, each with probability $\frac{1}{2}$.

It is important to note, however, that any predictions one can possibly make about observations restricted to system $\mathcal{H}_A$ are fully determined by the density operator ρ_A, and, hence do not depend on the choice of the interpretation. That is, whether we see the system $\mathcal{H}_A$ as a part of a larger system $\mathcal{H}_A \otimes \mathcal{H}_B$ which is in a pure state (as in Section 4.3.3) or as a mixture of pure states (as proposed in this section) is irrelevant as long as we are only interested in observable quantities derived from system $\mathcal{H}_A$.

4.3.5 Hybrid classical-quantum states

We will often encounter situations where parts of a system are quantum mechanical whereas others are classical. A typical example is the scenario described in Section 4.3.4, where the state of a quantum system $\mathcal{H}_A$ depends on the value of a classical random variable Z.

Since a classical system can be seen as a special type of a quantum system, such situations can be described consistently using the density operator formalism introduced above. More precisely, the idea is to represent the states of classical values Z by mutually orthogonal vectors on a Hilbert space. For example, the density operator describing the scenario of Section 4.3.4 would read

$$\rho_{AZ} = \sum_z P_Z(z) \rho_A^z \otimes |e_z\rangle\langle e_z| \ ,$$

where $\{e_z\}_z$ is a family of orthonormal vectors on $\mathcal{H}_Z$.

More generally, we use the following definition of *classicality.*

Definition 4.3.3. Let $\mathcal{H}_A$ and $\mathcal{H}_Z$ be Hilbert spaces and let $\{e_z\}_z$ be a fixed orthonormal basis of $\mathcal{H}_Z$. Then a density operator $\rho_{AZ} \in \mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_Z)$ is said to be *classical on* $\mathcal{H}_Z$ *(with respect to* $\{e_z\}_z$*)* if[12]

$$\rho_{AZ} \in \mathrm{Herm}(\mathcal{H}_A) \otimes \mathrm{span}\{|e_z\rangle\langle e_z|\}_z$$

4.3.6 Distance between states

Given two quantum states ρ and σ, we might ask how well we can distinguish them from each other. The answer to this question is given by the trace distance, which can be seen as a generalization of the corresponding distance measure for classical probability mass functions as defined in Section 2.3.6.

Definition 4.3.4. The *trace distance* between two density operators ρ and σ on a Hilbert space $\mathcal{H}$ is defined by

$$\delta(\rho, \sigma) := \frac{1}{2}\big\|\rho - \sigma\big\|_1 \ .$$

It is straightforward to verify that the trace distance is a metric on the space of density operators. Furthermore, it is unitarily invariant, i.e., $\delta(U\rho U^*, U\sigma U^*) = \delta(\rho, \sigma)$, for any unitary U.

The above definition of trace distance between density operators is consistent with the corresponding classical definition of Section 2.3.6. In particular, for two classical states $\rho = \sum_z P(z)|e_z\rangle\langle e_z|$ and $\sigma = \sum_z Q(z)|e_z\rangle\langle e_z|$ defined by probability mass functions P and Q, we have

$$\delta(\rho, \sigma) = \delta(P, Q) \ .$$

More generally, the following lemma implies that for any (not necessarily classical) ρ and σ there is always a measurement O that "conserves" the trace distance.

Lemma 4.3.5. *Let* $\rho, \sigma \in \mathcal{S}(\mathcal{H})$. *Then*

$$\delta(\rho, \sigma) = \max_O \delta(P, Q)$$

where the maximum ranges over all observables $O \in \mathrm{Herm}\mathcal{H}$ *and where* P *and* Q *are the probability mass functions of the outcomes when applying the measurement described by* O *to* ρ *and* σ*, respectively.*

[12] If the classical system $\mathcal{H}_Z$ itself has a tensor product structure (e.g., $\mathcal{H}_Z = \mathcal{H}_{Z'} \otimes \mathcal{H}_{Z''}$) we typically assume that the basis used for defining classical states has the same product structure (i.e., the basis vectors are of the form $e = e' \otimes e''$ with $e' \in \mathcal{H}_{Z'}$ and $e'' \in \mathcal{H}_{Z''}$).

Proof. Define $\Delta := \rho - \sigma$ and let $\Delta = \sum_i \alpha_i |e_i\rangle\langle e_i|$ be a spectral decomposition. Furthermore, let R and S be positive operators defined by

$$R = \sum_{i\colon \alpha_i \geq 0} \alpha_i |e_i\rangle\langle e_i|$$
$$S = -\sum_{i\colon \alpha_i < 0} \alpha_i |e_i\rangle\langle e_i| \ ,$$

that is,

$$\Delta = R - S \tag{4.19}$$
$$|\Delta| = R + S \ . \tag{4.20}$$

Finally, let $O = \sum_x x P_x$ be a spectral decomposition of O, where each P_x is a projector onto the eigenspace corresponding to the eigenvalue x. Then

$$\delta(P, Q) = \frac{1}{2}\sum_x \bigl|P(x) - Q(x)\bigr| = \frac{1}{2}\sum_x \bigl|\mathrm{tr}(P_x \rho) - \mathrm{tr}(P_x \sigma)\bigr| = \frac{1}{2}\sum_x \bigl|\mathrm{tr}(P_x \Delta)\bigr| \ . \tag{4.21}$$

Now, using (4.19) and (4.20),

$$\bigl|\mathrm{tr}(P_x \Delta)\bigr| = \bigl|\mathrm{tr}(P_x R) - \mathrm{tr}(P_x S)\bigr| \leq \bigl|\mathrm{tr}(P_x R)\bigr| + \bigl|\mathrm{tr}(P_x S)\bigr| = \mathrm{tr}(P_x |\Delta|) \ , \tag{4.22}$$

where the last equality holds because of (4.3). Inserting this into (4.21) and using $\sum_x P_x = \mathrm{id}$ gives

$$\delta(P, Q) \leq \frac{1}{2}\sum_x \mathrm{tr}\bigl(P_x |\Delta|\bigr) = \frac{1}{2}\mathrm{tr}\bigl(|\Delta|\bigr) = \frac{1}{2}\|\Delta\|_1 = \delta(\rho, \sigma) \ .$$

This proves that the maximum $\max_O \delta(P, Q)$ on the right hand side of the assertion of the lemma cannot be larger than $\delta(\rho, \sigma)$. To see that equality holds, it suffices to verify that the inequality in(4.22) becomes an equality if for any x the projector P_x either lies in the support of R or in the support of S. Such a choice of the projectors is always possible because R and S have mutually orthogonal support. □

An implication of Lemma 4.3.5 is that the trace distance between two states ρ and σ can be interpreted as the *maximum distinguishing probability*, i.e., the maximum probability by which a difference between ρ and σ can be detected (see Lemma 2.3.1). Another consequence of Lemma 4.3.5 is that the trace distance cannot increase under the partial trace, as stated by the following lemma.

Lemma 4.3.6. *Let ρ_{AB} and σ_{AB} be bipartite density operators and let $\rho_A := \mathrm{tr}_B(\rho_{AB})$ and $\sigma_A := \mathrm{tr}_B(\sigma_{AB})$ be the reduced states on the first subsystem. Then*

$$\delta(\rho_A, \sigma_A) \leq \delta(\rho_{AB}, \sigma_{AB}) \ .$$

Proof. Let P and Q be the probability mass functions of the outcomes when applying a measurement O_A to ρ_A and σ_A, respectively. Then, for an appropriately chosen O_A, we have according to Lemma 4.3.5

$$\delta(\rho_A, \sigma_A) = \delta(P, Q) \ . \tag{4.23}$$

Consider now the observable O_{AB} on the joint system defined by $O_{AB} := O_A \otimes \mathrm{id}_B$. It follows from property (4.4) of the partial trace that, when applying the measurement described by O_{AB} to the joint states ρ_{AB} and σ_{AB}, we get the same probability mass functions P and Q. Now, using again Lemma 4.3.5,

$$\delta(\rho_{AB}, \sigma_{AB}) \geq \delta(P, Q) \ . \tag{4.24}$$

The assertion follows by combining (4.23) and (4.24). □

The significance of the trace distance comes mainly from the fact that it is a bound on the probability that a difference between two states can be seen. However, in certain situations, it is more convenient to work with an alternative notion of distance, called *fidelity.*

Definition 4.3.7. The *fidelity* between two density operators ρ and σ on a Hilbert space $\mathcal{H}$ is defined by

$$F(\rho, \sigma) := \big\|\rho^{\frac{1}{2}}\sigma^{\frac{1}{2}}\big\|_1$$

where $\|S\|_1 := \mathrm{tr}\big(\sqrt{S^* S}\big)$.

To abbreviate notation, for two vectors $\phi, \psi \in \mathcal{H}$, we sometimes write $F(\phi, \psi)$ instead of $F(|\phi\rangle\langle\phi|, |\psi\rangle\langle\psi|)$, and, similarly, $\delta(\phi, \psi)$ instead of $\delta(|\phi\rangle\langle\phi|, |\psi\rangle\langle\psi|)$. Note that the fidelity is always between 0 and 1, and that $F(\rho, \rho) = 1$.

The fidelity is particularly easy to compute if one of the operators, say σ, is pure. In fact, if $\sigma = |\psi\rangle\langle\psi|$, we have

$$F(\rho, |\psi\rangle\langle\psi|) = \|\rho^{\frac{1}{2}}\sigma^{\frac{1}{2}}\|_1 = \mathrm{tr}\Big(\sqrt{\sigma^{\frac{1}{2}}\rho\sigma^{\frac{1}{2}}}\Big) = \mathrm{tr}\Big(\sqrt{|\psi\rangle\langle\psi|\rho|\psi\rangle\langle\psi|}\Big) = \sqrt{\langle\psi|\rho|\psi\rangle} \ .$$

In particular, if $\rho = |\phi\rangle\langle\phi|$, we find

$$F(\phi, \psi) = |\langle\phi|\psi\rangle| \ . \tag{4.25}$$

The fidelity between pure states thus simply corresponds to the (absolute value of the) scalar product between the states.

The following statement from Uhlmann generalizes this statement to arbitrary states.

Theorem 4.3.8 (Uhlmann)**.** *Let ρ_A and σ_A be density operators on a Hilbert space $\mathcal{H}_A$. Then*

$$F(\rho_A, \sigma_A) = \max_{\rho_{AB}, \sigma_{AB}} F(\rho_{AB}, \sigma_{AB}) \ .$$

where the maximum ranges over all purifications ρ_{AB} and σ_{AB} of ρ_A and σ_A, respectively.

Proof. Because any finite-dimensional Hilbert space can be embedded into any other Hilbert space with higher dimension, we can assume without loss of generality that $\mathcal{H}_A$ and $\mathcal{H}_B$ have equal dimension.

Let $\{e_i\}_i$ and $\{f_i\}_i$ be orthonormal bases of $\mathcal{H}_A$ and $\mathcal{H}_B$, respectively, and define

$$\Theta := \sum_i e_i \otimes f_i \ .$$

Furthermore, let $W \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B)$ be the transformation of the basis $\{e_i\}_i$ to the basis $\{f_i\}_i$, that is,

$$W : \, e_i \mapsto f_i \ .$$

Writing out the definition of Θ, it is easy to verify that, for any $S_B \in \mathrm{End}(\mathcal{H}_B)$,

$$(\mathrm{id}_A \otimes S_B)\Theta = (S'_A \otimes \mathrm{id}_B)\Theta \tag{4.26}$$

where $S'_A := W^{-1} S_B^T W$, and where S_B^T denotes the transpose of S_B with respect to the basis $\{f_i\}_i$.

Let now $\rho_{AB} = |\Psi\rangle\langle\Psi|$ and let

$$\Psi = \sum_i \alpha_i e'_i \otimes f'_i$$

be a Schmidt decomposition of Ψ. Because the coefficients α_i are the square roots of the eigenvalues of ρ_A, we have

$$\Psi = (\sqrt{\rho_A} \otimes \mathrm{id}_B)(U_A \otimes U_B)\Theta$$

where U_A is the transformation of $\{e_i\}_i$ to $\{e'_i\}_i$ and, likewise, U_B is the transformation of $\{f_i\}_i$ to $\{f'_i\}_i$. Using (4.26), this can be rewritten as

$$\Psi = (\sqrt{\rho_A} V \otimes \mathrm{id}_B)\Theta$$

for $V := U_A W^{-1} U_B^T W$ unitary. Similarly, for $\sigma_{AB} = |\Psi'\rangle\langle\Psi'|$, we have

$$\Psi' = (\sqrt{\sigma_A} V' \otimes \mathrm{id}_B)\Theta$$

for some appropriately chosen unitary V'. Thus, using (4.25), we find

$$F(\rho_{AB}, \sigma_{AB}) = |\langle\Psi|\Psi'\rangle| = \langle\Theta|V^* \sqrt{\rho_A}\sqrt{\sigma_A} V'|\Theta\rangle = \mathrm{tr}(V^* \sqrt{\rho_A}\sqrt{\sigma_A} V') \ ,$$

where the last equality is a consequence of the definition of Θ. Using the fact that any unitary V' can be obtained by an appropriate choice of the purification σ_{AB}, this can be rewritten as

$$F(\rho_{AB}, \sigma_{AB}) = \max_U \mathrm{tr}(U \sqrt{\rho_A}\sqrt{\sigma_A}) \ .$$

The assertion then follows because, by Lemma 4.1.2,

$$F(\rho_A, \sigma_A) = \|\sqrt{\rho_A}\sqrt{\sigma_A}\|_1 = \max_U \mathrm{tr}(U \sqrt{\rho_A}\sqrt{\sigma_A}) \ .$$

□

Uhlmann's theorem is very useful for deriving properties of the fidelity, as, e.g., the following lemma.

Lemma 4.3.9. *Let ρ_{AB} and σ_{AB} be bipartite states. Then*

$$F(\rho_{AB}, \sigma_{AB}) \leq F(\rho_A, \sigma_A) \ .$$

Proof. According to Uhlmann's theorem, there exist purifications ρ_{ABC} and σ_{ABC} of ρ_{AB} and σ_{AB} such that

$$F(\rho_{AB}, \sigma_{AB}) = F(\rho_{ABC}, \sigma_{ABC}) \ . \tag{4.27}$$

Trivially, ρ_{ABC} and σ_{ABC} are also purifications of ρ_A and σ_A, respectively. Hence, again by Uhlmann's theorem,

$$F(\rho_A, \sigma_A) \geq F(\rho_{ABC}, \sigma_{ABC}) \ . \tag{4.28}$$

Combining (4.27) and (4.28) concludes the proof. □

The trace distance and the fidelity are related to each other. In fact, for pure states, represented by normalized vectors ϕ and ψ, we have

$$\delta(\phi, \psi) = \sqrt{1 - F(\phi, \psi)^2} \ . \tag{4.29}$$

To see this, let $\phi^\perp$ be a normalized vector orthogonal to ϕ such that $\psi = \alpha\phi + \beta\phi^\perp$, for some $\alpha, \beta \in \mathbb{R}^+$ such that $\alpha^2 + \beta^2 = 1$. (Because the phases of both $\phi, \phi^\perp, \psi$ are irrelevant, the coefficients α and β can without loss of generality assumed to be real and positive.) The operators $|\phi\rangle\langle\phi|$ and $|\psi\rangle\langle\psi|$ can then be written as matrices with respect to the basis $\{\phi, \phi^\perp\}$,

$$|\phi\rangle\langle\phi| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$
$$|\psi\rangle\langle\psi| = \begin{pmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{pmatrix}$$

In particular, the trace distance takes the form

$$\delta(\phi, \psi) = \frac{1}{2}\big\||\phi\rangle\langle\phi| - |\psi\rangle\langle\psi|\big\|_1 = \frac{1}{2}\left\|\begin{pmatrix} 1 - |\alpha|^2 & -\alpha\beta^* \\ -\alpha^*\beta & -|\beta|^2 \end{pmatrix}\right\|_1 \ .$$

The eigenvalues of the matrix on the right hand side are $\alpha_0 = \beta$ and $\alpha_1 = -\beta$. We thus find

$$\delta(\phi, \psi) = \frac{1}{2}\big(|\alpha_0| + |\alpha_1|\big) = \beta \ .$$

Furthermore, by the definition of β, we have

$$\beta = \sqrt{1 - |\langle\phi|\psi\rangle|^2} \ .$$

The assertion (4.29) then follows from (4.25).

Equality (4.29) together with Uhlmann's theorem are sufficient to prove one direction of the following lemma.

Lemma 4.3.10. *Let ρ and σ be density operators. Then*

$$1 - F(\rho, \sigma) \leq \delta(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)^2} \ .$$

Proof. We only prove the second inequality. For a proof of the first, we refer to [10].

Consider two density operators ρ_A and σ_A and let ρ_{AB} and σ_{AB} be purifications such that

$$F(\rho_A, \sigma_A) = F(\rho_{AB}, \sigma_{AB})$$

as in Uhlmann's theorem. Combining this with equality (4.29) and Lemma 4.3.6, we find

$$\sqrt{1 - F(\rho_A, \sigma_A)^2} = \sqrt{1 - F(\rho_{AB}, \sigma_{AB})^2} = \delta(\rho_{AB}, \sigma_{AB}) \geq \delta(\rho_A, \sigma_A) \ .$$

□

4.4 Evolution and measurements

Let $\mathcal{H}_A \otimes \mathcal{H}_B$ be a composite system. We have seen in the previous sections that, as long as we are only interested in the observable quantities of subsystem $\mathcal{H}_A$, it is sufficient to consider the corresponding reduced state ρ_A. So far, however, we have restricted our attention to scenarios where the evolution of this subsystem is isolated.

In the following, we introduce tools that allow us to consistently describe the behavior of subsystems in the general case where there is interaction between $\mathcal{H}_A$ and $\mathcal{H}_B$. The basic mathematical objects to be introduced in this context are *completely positive maps (CPMs)* and *positive operator valued measures (POVMs)*, which are the topic of this section.

4.4.1 Completely Positive Maps (CPMs)

Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be the Hilbert spaces describing certain (not necessarily disjoint) parts of a physical system. The evolution of the system over a time interval $[t_0, t_1]$ induces a mapping $\mathcal{E}$ from the set of states $\mathcal{S}(\mathcal{H}_A)$ on subsystem $\mathcal{H}_A$ at time t_0 to the set of states $\mathcal{S}(\mathcal{H}_B)$ on subsystem $\mathcal{H}_B$ at time t_1. This and the following sections are devoted to the study of this mapping.

Obviously, not every function $\mathcal{E}$ from $\mathcal{S}(\mathcal{H}_A)$ to $\mathcal{S}(\mathcal{H}_B)$ corresponds to a physically possible evolution. In fact, based on the considerations in the previous sections, we have the following requirement. If ρ is a mixture of two states ρ_0 and ρ_1, then we expect that $\mathcal{E}(\rho)$ is the mixture of $\mathcal{E}(\rho_0)$ and $\mathcal{E}(\rho_1)$. In other words, a physical mapping $\mathcal{E}$ needs to conserve the convex structure of the set of density operators, that is,

$$\mathcal{E}\bigl(p\rho_0 + (1-p)\rho_1\bigr) = p\mathcal{E}(\rho_0) + (1-p)\mathcal{E}(\rho_1) \ , \tag{4.30}$$

for any $\rho_0, \rho_1 \in \mathcal{S}(\mathcal{H}_A)$ and any $p \in [0, 1]$.

As we shall see, any mapping from $\mathcal{S}(\mathcal{H}_A)$ to $\mathcal{S}(\mathcal{H}_B)$ that satisfies (4.30) corresponds to a physical process (and vice versa). In the following, we will thus have a closer look at these mappings.

For our considerations, it will be convenient to embed the mappings from $\mathcal{S}(\mathcal{H}_A)$ to $\mathcal{S}(\mathcal{H}_B)$ into the space of mappings from $\mathrm{End}(\mathcal{H}_A)$ to $\mathrm{End}(\mathcal{H}_B)$. The convexity requirement (4.30) then turns into the requirement that the mapping is linear. In addition, the requirement that density operators are mapped to density operators will correspond to two properties, called *complete positivity* and *trace preservation*.

The definition of complete positivity is based on the definition of positivity.

Definition 4.4.1. A linear map $\mathcal{E} \in \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$ is said to be *positive* if $\mathcal{E}(S) \geq 0$ for any $S \geq 0$.

An simple example of a positive map is the *identity map* on $\mathrm{End}(\mathcal{H}_A)$, in the following denoted $\mathcal{I}_A$. A more interesting example is $\mathcal{T}_A$ defined by

$$\mathcal{T}_A : S \mapsto S^T \ ,$$

where S^T denotes the transpose with respect to some fixed basis. To see that $\mathcal{T}_A$ is positive, note that $S \geq 0$ implies $\langle \phi | S | \phi \rangle \geq 0$ for any vector ϕ. Hence $\langle \phi | S^T | \phi \rangle = \overline{\langle \phi | S | \phi \rangle} \geq 0$, from which we conclude $S^T \geq 0$.

Remarkably, positivity of two maps $\mathcal{E}$ and $\mathcal{F}$ does not necessarily imply positivity of the tensor map $\mathcal{E} \otimes \mathcal{F}$ defined by

$$(\mathcal{E} \otimes \mathcal{F})(S \otimes T) := \mathcal{E}(S) \otimes \mathcal{F}(T) \ .$$

In fact, it is straightforward to verify that the map $\mathcal{I}_A \otimes \mathcal{T}_{A'}$ applied to the positive operator $\rho_{AA'} := |\Psi\rangle\langle\Psi|$, for Ψ defined by (4.11), results in a non-positive operator.

To guarantee that tensor products of mappings such as $\mathcal{E} \otimes \mathcal{F}$ are positive, a stronger requirement is needed, called *complete positivity*.

Definition 4.4.2. A linear map $\mathcal{E} \in \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$ is said to be *completely positive* if for any Hilbert space $\mathcal{H}_R$, the map $\mathcal{E} \otimes \mathcal{I}_R$ is positive.

Definition 4.4.3. A linear map $\mathcal{E} \in \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$ is said to be *trace preserving* if $\mathrm{tr}(\mathcal{E}(S)) = \mathrm{tr}(S)$ for any $S \in \mathrm{End}(\mathcal{H}_A)$.

We will use the abbreviation *CPM* to denote completely positive maps. Moreover, we denote by $\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$ the set of trace-preserving completely positive maps from $\mathrm{End}(\mathcal{H}_A)$ to $\mathrm{End}(\mathcal{H}_B)$.

4.4.2 The Choi-Jamiolkowski isomorphism

The Choi-Jamiolkowski isomorphism is a mapping that relates CPMs to density operators. Its importance results from the fact that it essentially reduces the study of CPMs to the study of density operators. In other words, it allows us to translate mathematical statements that hold for density operators to statements for CPMs (and vice versa).

Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be Hilbert spaces, let $\mathcal{H}_{A'}$ be isomorphic to $\mathcal{H}_A$, and define the normalized vector $\Psi = \Psi_{A'A} \in \mathcal{H}_{A'} \otimes \mathcal{H}_A$ by

$$\Psi = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} e_i \otimes e_i$$

where $\{e_i\}_{i=1,\ldots,d}$ is an orthonormal basis of $\mathcal{H}_A \cong \mathcal{H}_{A'}$ and $d = \dim(\mathcal{H}_A)$.

Definition 4.4.4. The *Choi-Jamiolkowski mapping (relative to the basis $\{e_i\}_i$)* is the linear function τ from $\mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$ to $\mathrm{End}(\mathcal{H}_{A'} \otimes \mathcal{H}_B)$ defined by

$$\tau : \mathcal{E} \mapsto (\mathcal{I}_{A'} \otimes \mathcal{E})(|\Psi\rangle\langle\Psi|) \ .$$

Lemma 4.4.5. *The Choi-Jamiolkowski mapping*

$$\tau : \mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B)) \longrightarrow \mathrm{End}(\mathcal{H}_{A'} \otimes \mathcal{H}_B)$$

is an isomorphism. Its inverse τ^{-1} maps any $\rho_{A'B}$ to

$$\tau^{-1}(\rho_{A'B}) : S_A \mapsto d \cdot \mathrm{tr}_{A'}\Big(\big(\mathcal{T}_{A\to A'}(S_A) \otimes \mathrm{id}_B\big)\rho_{A'B}\Big) \ ,$$

where $\mathcal{T}_{A\to A'} : \mathrm{End}(\mathcal{H}_A) \to \mathrm{End}(\mathcal{H}_{A'})$ is defined by

$$\mathcal{T}_{A\to A'}(S_A) := \sum_{i,j} |e_i\rangle_{A'}\langle e_j|_A S_A |e_i\rangle_A \langle e_j|_{A'} \ .$$

Proof. It suffices to verify that the mapping τ^{-1} defined in the lemma is indeed an inverse of τ. We first check that $\tau \circ \tau^{-1}$ is the identity on $\mathrm{End}(\mathcal{H}_{A'} \otimes \mathcal{H}_B)$. That is, we show that for any operator $\rho_{A'B} \in \mathrm{End}(\mathcal{H}_{A'} \otimes \mathcal{H}_B)$, the operator

$$\tau(\tau^{-1}(\rho_{A'B})) := d \cdot (\mathcal{I}_{A'} \otimes \mathrm{tr}_{A'})\Big(\big((\mathcal{I}_{A'} \otimes \mathcal{T}_{A\to A'})(|\Psi\rangle\langle\Psi|) \otimes \mathrm{id}_B\big)(\mathrm{id}_{A'} \otimes \rho_{A'B})\Big) \quad (4.31)$$

equals $\rho_{A'B}$ (where we have written $\mathcal{I}_{A'} \otimes \mathrm{tr}_{A'}$ instead of $\mathrm{tr}_{A'}$ to indicate that the trace only acts on the second subsystem $\mathcal{H}_{A'}$). Inserting the definition of Ψ, we find

$$\begin{aligned}\tau(\tau^{-1}(\rho_{A'B})) &= d \cdot (\mathcal{I}_{A'} \otimes \mathrm{tr}_{A'})\big(\sum_{i,j}(|e_i\rangle\langle e_j|_{A'} \otimes |e_j\rangle\langle e_i|_{A'} \otimes \mathrm{id}_B)(\mathrm{id}_{A'} \otimes \rho_{A'B})\big) \\ &= \sum_{i,j}(|e_i\rangle\langle e_i|_{A'} \otimes \mathrm{id}_B)\rho_{A'B}(|e_j\rangle\langle e_j|_{A'} \otimes \mathrm{id}_B) = \rho_{A'B} \ ,\end{aligned}$$

which proves the claim that $\tau \circ \tau^{-1}$ is the identity.

It remains to show that τ is injective. For this, let $S_A \in \mathrm{End}(\mathcal{H}_A)$ be arbitrary and note that

$$(\mathcal{T}_{A\to A'}(S_A) \otimes \mathrm{id}_A)\Psi = (\mathrm{id}_{A'} \otimes S_A)\Psi \ .$$

Together with the fact that $\mathrm{tr}_{A'}(|\Psi\rangle\langle\Psi|) = \frac{1}{d}\mathrm{id}_A$ this implies

$$\begin{aligned}\mathcal{E}(S_A) &= d \cdot \mathcal{E}\big(S_A \mathrm{tr}_{A'}(|\Psi\rangle\langle\Psi|)\big) \\ &= d \cdot \mathrm{tr}_{A'}\Big((\mathcal{I}_{A'} \otimes \mathcal{E})\big((\mathrm{id}_{A'} \otimes S_A)|\Psi\rangle\langle\Psi|\big)\Big) \\ &= d \cdot \mathrm{tr}_{A'}\Big((\mathcal{I}_{A'} \otimes \mathcal{E})\big((\mathcal{T}_{A\to A'}(S_A) \otimes \mathrm{id}_A)|\Psi\rangle\langle\Psi|\big)\Big) \\ &= d \cdot \mathrm{tr}_{A'}\Big((\mathcal{T}_{A\to A'}(S_A) \otimes \mathrm{id}_A)(\mathcal{I}_{A'} \otimes \mathcal{E})(|\Psi\rangle\langle\Psi|)\Big) \ .\end{aligned}$$

Assume now that $\tau(\mathcal{E}) = 0$. Then, by definition, $(\mathcal{I}_{A'} \otimes \mathcal{E})(|\Psi\rangle\langle\Psi|) = 0$. By virtue of the above equality, this implies $\mathcal{E}(S_A) = 0$ for any S_A and, hence, $\mathcal{E} = 0$. In other words, $\tau(\mathcal{E}) = 0$ implies $\mathcal{E} = 0$, i.e., τ is injective. □

In the following, we focus on trace-preserving CPMs. The set $\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$ obviously is a subset of $\mathrm{Hom}(\mathrm{End}(\mathcal{H}_A), \mathrm{End}(\mathcal{H}_B))$. Consequently, $\tau(\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B))$ is also a subset of $\mathrm{End}(\mathcal{H}_{A'} \otimes \mathcal{H}_B)$. It follows immediately from the complete positivity property that $\tau(\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B))$ only contains positive operators. Moreover, by the trace-preserving property, any $\rho_{A'B} \in \tau(\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B))$ satisfies

$$\mathrm{tr}_B(\rho_{A'B}) = \frac{1}{d}\mathrm{id}_{A'} \ . \tag{4.32}$$

In particular, $\rho_{A'B}$ is a density operator.

Conversely, the following lemma implies[13] that any density operator $\rho_{A'B}$ that satisfies (4.32) is the image of some trace-preserving CPM. We therefore have the following characterization of the image of $\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$ under the Choi-Jamiolkowski isomorphism,

$$\tau(\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)) = \{\rho_{A'B} \in \mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_B) : \mathrm{tr}_B(\rho_{A'B}) = \tfrac{1}{d}\mathrm{id}_{A'}\} \ .$$

Lemma 4.4.6. *Let $\Phi \in \mathcal{H}_{A'} \otimes \mathcal{H}_B$ such that $\mathrm{tr}_B(|\Phi\rangle\langle\Phi|) = \frac{1}{d}\mathrm{id}_{A'}$. Then the mapping $\mathcal{E} := \tau^{-1}(|\Phi\rangle\langle\Phi|)$ has the form*

$$\mathcal{E} : S_A \mapsto U S_A U^*$$

where $U \in \mathrm{End}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is an isometry, i.e., $U^*U = \mathrm{id}_A$.

Proof. Using the expression for $\mathcal{E} := \tau^{-1}(|\Phi\rangle\langle\Phi|)$ provided by Lemma 4.4.5, we find, for any $S_A \in \mathrm{End}(\mathcal{H}_A)$,

$$\begin{aligned}\mathcal{E}(S_A) &= d \cdot \mathrm{tr}_{A'}\big((\mathcal{T}_{A\to A'}(S_A) \otimes \mathrm{id}_B)|\Phi\rangle\langle\Phi|\big) \\ &= d \cdot \sum_{i,j} \langle e_i|S_A|e_j\rangle (\langle e_i| \otimes \mathrm{id}_B)|\Phi\rangle\langle\Phi|(|e_j\rangle \otimes \mathrm{id}_B) \\ &= \sum_{i,j} E_i S_A E_j^* \ ,\end{aligned}$$

where $E_i := \sqrt{d} \cdot (\langle e_i| \otimes \mathrm{id}_B)|\Phi\rangle\langle e_i|$. Defining $U := \sum_i E_i$, we conclude that $\mathcal{E}$ has the desired form, i.e., $\mathcal{E}(S_A) = U S_A U^*$.

To show that U is an isometry, let

$$\Phi = \frac{1}{\sqrt{d}} \sum_i e_i \otimes f_i$$

[13] See the argument in Section 4.4.3.

be a Schmidt decomposition of Φ. (Note that, because $\mathrm{tr}_B(|\Phi\rangle\langle\Phi|)$ is fully mixed, the basis $\{e_i\}$ can be chosen to coincide with the basis used for the definition of τ.) Then $(\langle e_i| \otimes \mathrm{id}_B)|\Phi\rangle = |f_i\rangle$ and, hence,

$$U^* U = d \sum_{i,j} |e_j\rangle\langle\Phi|(|e_j\rangle \otimes \mathrm{id}_B)(\langle e_i| \otimes \mathrm{id}_B)|\Phi\rangle\langle e_i| = \mathrm{id}_A \ .$$

□

4.4.3 Stinespring dilation

The following lemma will be of crucial importance for the interpretation of CPMs as physical maps.

Lemma 4.4.7 (Stinespring dilation). *Let $\mathcal{E} \in \mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$. Then there exists an isometry $U \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B \otimes \mathcal{H}_R)$, for some Hilbert space $\mathcal{H}$, such that*

$$\mathcal{E} : \, S_A \mapsto \mathrm{tr}_R(U S_A U^*) \ .$$

Proof. Let $\mathcal{E}_{A\to B} := \mathcal{E}$, define $\rho_{AB} := \tau(\mathcal{E})$, and let ρ_{ABR} be a purification of ρ_{AB}. We then define $\mathcal{E}' = \mathcal{E}'_{A\to(B,R)} := \tau^{-1}(\rho_{ABR})$. According to Lemma 4.4.6, because $\mathrm{tr}_{BR}(\rho_{ABR})$ is fully mixed, $\mathcal{E}'_{A\to(B,R)}$ has the form

$$\mathcal{E}'_{A\to(B,R)} : \, S_A \mapsto U S_A U^* \ .$$

The assertion then follows from the fact that the diagram below commutes, which can be readily verified from the definition of the Choi-Jamiolkowski isomorphism. (Note that the arrow on the top corresponds to the operation $\mathcal{E}' \mapsto \mathrm{tr}_R \circ \mathcal{E}'$.)

$$\begin{tikzcd}
\mathcal{E}_{A\to B} \arrow[d, "\tau"'] & \mathcal{E}'_{A\to(B,R)} \arrow[l, "\mathrm{tr}_R"'] \\
\rho_{A'B} \arrow[r, "\text{purif.}"'] & \rho_{A'BR} \arrow[u, "\tau^{-1}"']
\end{tikzcd}$$

□

We can use Lemma 4.4.7 to establish a connection between general trace-preserving CPMs and the evolution postulate of Section 4.3.2. Let $\mathcal{E} \in \mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_A)$ and let $U \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_A \otimes \mathcal{H}_R)$ be the corresponding Stinespring dilation, as defined by Lemma 4.4.7. Furthermore, let $\tilde{U} \in \mathrm{Hom}(\mathcal{H}_A \otimes \mathcal{H}_R, \mathcal{H}_A \otimes \mathcal{H}_R)$ be a unitary embedding of U in $\mathcal{H}_A \otimes \mathcal{H}_R$, i.e., $\tilde{U}$ is unitary and, for some fixed $w_0 \in \mathcal{H}_R$, satisfies

$$\tilde{U} : \, v \otimes w_0 \mapsto U v \ .$$

Using the fact that U is an isometry, it is easy to see that there always exists such a $\tilde{U}$.

By construction, the unitary $\tilde{U}$ satisfies

$$\mathcal{E}(S_A) = \mathrm{tr}_R\big(\tilde{U}(S_A \otimes |w_0\rangle\langle w_0|)\tilde{U}^*\big)$$

for any operator S_A on $\mathcal{H}_A$. Hence, the mapping $\mathcal{E}$ on $\mathcal{H}_A$ can be seen as a unitary on an extended system $\mathcal{H}_A \otimes \mathcal{H}_R$ (with $\mathcal{H}_R$ being initialized with a state w_0) followed by a partial trace over $\mathcal{H}_R$. In other words, any possible mapping from density operators to density operators that satisfies the convexity criterion (4.30) (this is exactly the set of trace-preserving CPMs) corresponds to a unitary evolution of a larger system.

4.4.4 Operator-sum representation

As we have seen in the previous section, CPMs can be represented as unitaries on a larger system. In the following, we consider an alternative and somewhat more economic[14] description of CPMs.

Lemma 4.4.8 (Operator-sum representation). *For any $\mathcal{E} \in \mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$ there exists a family $\{E_x\}_x$ of operators $E_x \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B)$ such that*

$$\mathcal{E} :\ S_A \mapsto \sum_x E_x S_A E_x^* \tag{4.33}$$

and $\sum_x E_x^* E_x = \mathrm{id}_A$.

Conversely, any mapping $\mathcal{E}$ of the form (4.33) *is contained in* $\mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_B)$.

Proof. By Lemma 4.4.7, there exists operators $U \in \mathrm{Hom}(\mathcal{H}_A, \mathcal{H}_B \otimes \mathcal{H}_R)$ such that

$$\mathcal{E}(S_A) = \mathrm{tr}_R(U S_A U^*) = \sum_x (\mathrm{id}_B \otimes \langle f_x|) U S_A U^* (\mathrm{id}_B \otimes |f_x\rangle) ,$$

where $\{f_x\}_x$ is an orthonormal basis of $\mathcal{H}_R$. Defining

$$E_x := (\mathrm{id}_B \otimes \langle f_x|) U ,$$

the direct assertion follows from the fact that

$$\sum_x E_x^* E_x = \sum_x U^* (\mathrm{id}_B \otimes |f_x\rangle)(\mathrm{id}_B \otimes \langle f_x|) U = U^* U = \mathrm{id} ,$$

which holds because U is an isometry.

The converse assertion can be easily verified as follows. The fact that any mapping of the form (4.33) is positive follows from the observation that $E_x S_A E_x^*$ is positive whenever S_A is positive. To show that the mapping is trace-preserving, we use

$$\mathrm{tr}(\mathcal{E}(S_A)) = \sum_x \mathrm{tr}(E_x S_A E_x^*) = \sum_x \mathrm{tr}(E_x^* E_X S_A) = \mathrm{tr}(\mathrm{id}_A S_A) .$$

□

[14] In the sense that there is less redundant information in the description of the CPM.

Note that the family $\{E_x\}_x$ is not uniquely determined by the CPM $\mathcal{E}$. This is easily seen by the following example. Let $\mathcal{E}$ be the trace-preserving CPM from $\mathrm{End}(\mathcal{H}_A)$ to $\mathrm{End}(\mathcal{H}_B)$ defined by

$$\mathcal{E}: \; S_A \mapsto \mathrm{tr}(S_A)|w\rangle\langle w|$$

for any operator $S_A \in \mathrm{End}(\mathcal{H}_A)$ and some fixed $w \in \mathcal{H}_B$. That is, $\mathcal{E}$ maps any density operator to the state $|w\rangle\langle w|$. It is easy to verify that this CPM can be written in the form (4.33) for

$$E_x := |w\rangle\langle e_x|$$

where $\{e_x\}_x$ in an arbitrary orthonormal basis of $\mathcal{H}_A$.

4.4.5 Measurements as CPMs

An elegant approach to describe measurements is to use the notion of classical states. Let ρ_{AB} be a density operator on $\mathcal{H}_A \otimes \mathcal{H}_B$ and let $O = \sum_x x P_x$ be an observable on $\mathcal{H}_A$. Then, according to the the measurement postulate of Section 4.3.2, the measurement process produces a classical value X distributed according to the probability distribution P_X specified by (4.13), and the post-measurement state $\rho'_{AB,x}$ conditioned on the outcome x is given by (4.14). This situation is described by a density operator

$$\rho'_{XAB} := \sum_x P_X(x)|e_x\rangle\langle e_x| \otimes \rho'_{AB,x} \; .$$

on $\mathcal{H}_X \otimes \mathcal{H}_A \otimes \mathcal{H}_B$ which is classical on $\mathcal{H}_X$ (with respect to some orthonormal basis $\{e_x\}_x$). Inserting the expressions for P_X and $\rho'_{AB,x}$, this operator can be rewritten as

$$\rho'_{XAB} = \sum_x |e_x\rangle\langle e_x| \otimes (P_x \otimes \mathrm{id}_B)\rho_{AB}(P_x \otimes \mathrm{id}_B) \; .$$

Note that the mapping $\mathcal{E}$ from ρ_{AB} to ρ'_{XAB} can be written in the operator-sum representation (4.33) with

$$E_x := |x\rangle \otimes P_x \otimes \mathrm{id}_B \; ,$$

where

$$\sum_x E_x^* E_x = \sum_x P_x \otimes \mathrm{id}_B = \mathrm{id}_{AB} \; .$$

It thus follows from Lemma 4.4.8 that the mapping

$$\mathcal{E}: \; \rho_{AB} \mapsto \rho'_{XAB}$$

is a trace-preserving CPM.

This is a remarkable statement. According to the Stinespring dilation theorem, it tells us that any measurement can be seen as a unitary on a larger system. In other words, a measurement is just a special type of evolution of the system.

4.4.6 Positive operator valued measures (POVMs)

When analyzing a physical system, one is often only interested in the probability distribution of the observables (but not in the post-measurement state). Consider a system that first undergoes an evolution characterized by a CPM and, after that, is measured. Because, as argued above, a measurement can be seen as a CPM, the concatenation of the evolution and the measurement is again a CPM $\mathcal{E} \in \mathrm{TPCPM}(\mathcal{H}_A, \mathcal{H}_X \otimes \mathcal{H}_B)$. If the measurement outcome X is represented by orthogonal vectors $\{e_x\}_x$ of $\mathcal{H}_X$, this CPM has the form

$$\mathcal{E} : S_A \mapsto \sum_x |e_x\rangle\langle e_x| \otimes E_x S_A E_x^* \ .$$

In particular, if we apply the CPM $\mathcal{E}$ to a density operator ρ_A, the distribution P_X of the measurement outcome X is given by

$$P_X(x) = \mathrm{tr}(E_x \rho_A E_x^*) = \mathrm{tr}(M_x \rho_A) \ ,$$

where $M_x := E_x^* E_x$.

From this we conclude that, as long as we are only interested in the probability distribution of X, it suffices to characterize the evolution and the measurement by the family of operators M_x. Note, however, that the operators M_x do not fully characterize the full evolution. In fact, distinct operators E_x can give raise to the same operator $M_x = E_x^* E_x$.

It is easy to see from Lemma 4.4.8 that the family $\{M_x\}_x$ of operators defined as above satisfies the following definition.

Definition 4.4.9. A *positive operator valued measure (POVM) (on $\mathcal{H}$)* is a family $\{M_x\}_x$ of positive operators $M_x \in \mathrm{Herm}(\mathcal{H})$ such that

$$\sum_x M_x = \mathrm{id}_{\mathcal{H}} \ .$$

Conversely, any POVM $\{M_x\}_x$ corresponds to a (not unique) physically possible evolution followed by a measurement. This can easily be seen by defining a CPM by the operator-sum representation with operators $E_x := \sqrt{M_x}$.

4.4.7 The diamond norm of CPMs

Let $\mathcal{E}$ and $\mathcal{F}$ be arbitrary CPMs from $\mathcal{S}(\mathcal{H})$ to $\mathcal{S}(\mathcal{H}')$. The defining demand on the definition of the wanted distance measure $d(..,..)$ between the CPMs $\mathcal{E}$ and $\mathcal{F}$ is that it is proportional to the maximal probability for distinguishing the maps $\mathcal{E}$ and $\mathcal{F}$ in an experiment. After our discussion of the trace distance between states in an earlier chapter it is natural to propose the distance measure

$$\tilde{d}(\mathcal{E}, \mathcal{F}) := \max_{\rho \in \mathcal{S}(\mathcal{H}^{(\mathrm{in})})} \|\mathcal{E}(\rho) - \mathcal{F}(\rho)\|_1$$

if one recalls the "maximal distinguishing probability property" of the trace distance. Up to a factor 1/2 this is the maximal probability to distinguish the CPMs $\mathcal{E}$ and $\mathcal{F}$ in an experiment which works with initial states in the Hilbert space $\mathcal{H}$. But this is *not* the best way to distinguish the CPMs $\mathcal{E}$ and $\mathcal{F}$ in an experiment!Note that in our naive definition above we have excluded the possibility to consider initial states in "larger" Hilbert spaces in the maximization-procedure. The probability to distinguish the CPMs $\mathcal{E}$ and $\mathcal{F}$ in an experiment may increase if we "enlarge" the input Hilbert space $\mathcal{H}$ by an additional tensor space factor;

$$\mathcal{H} \rightsquigarrow \mathcal{H} \otimes \mathcal{H}_E;$$

and apply the CPMs $\mathcal{E}$ and $\mathcal{F}$ as $\mathcal{E} \otimes \mathcal{I}_E$ and $\mathcal{F} \otimes \mathcal{I}_E$ to states in $\mathcal{S}(\mathcal{H} \otimes \mathcal{H}_E)$. These replacements lead to a simultaneous replacement of the output Hilbert space:

$$\mathcal{H}' \rightsquigarrow \mathcal{H}' \otimes \mathcal{H}_E.$$

Let us have a closer look at an explicit example to recognize that situations occur in which

$$\tilde{d}(\mathcal{E}, \mathcal{F}) < \tilde{d}(\mathcal{E} \otimes \mathcal{I}_E, \mathcal{F} \otimes \mathcal{I}_E)$$

for some Hilbert space $\mathcal{H}_E$. This shows why we discard the immediate use of $\tilde{d}(\mathcal{E}, \mathcal{F})$ but use a distance measure of the form $\tilde{d}(\mathcal{E} \otimes \mathcal{I}_E, \mathcal{F} \otimes \mathcal{I}_E)$ instead. We will still have to figure out the optimal choice for the Hilbert space $\mathcal{H}_E$ which will lead to the definition of the so called "diamond norm".

Example 4.4.10. *Let $\mathcal{H} \cong \mathcal{H}' \cong \mathcal{H}_E \cong \mathbb{C}^2$, define*

$$\begin{array}{llll} \mathcal{E}: & \mathcal{S}(\mathcal{C}^2) & \rightarrow & \mathcal{S}(\mathcal{C}^2) \\ & \rho & \mapsto & \mathcal{E}(\rho) = (1-p)\rho + \frac{p}{2}\mathbb{I}_{\mathbb{C}^2} \end{array}$$

and set $\mathcal{F} := \mathcal{I} := \mathcal{I}_{\mathbb{C}^2}$. We are trying to show that

$$\tilde{d}(\mathcal{E}, \mathcal{I}) < \tilde{d}(\mathcal{E} \otimes \mathcal{I}_E, \mathcal{I} \otimes \mathcal{I}_E).$$

We first compute the left hand sight explicitly and prove the inequality afterwards building on the explicit result derived for the left hand sight.

The left hand sight. *According to the proposed distance measure $\tilde{d}(..,..)$,*

$$\tilde{d}(\mathcal{E}, \mathcal{I}) = \max_{\rho \in \mathcal{S}(\mathcal{H})} \|\mathcal{E}(\rho) - \mathcal{I}(\rho)\|_1$$

To compute this expression we first prove two claims.

Claim 1: *The distance $\|\mathcal{E}(\rho) - \mathcal{F}(\rho)\|_1$ is maximal for pure states $\rho = |\psi\rangle\langle\psi|$, $\psi \in \mathcal{H}$.*
Proof: *The state ρ can be written in the form*

$$\rho = p\rho_1 + (1-p)\rho_2$$

(ρ_1 and ρ_2 have support on orthogonal subspaces) whenever the state ρ isn't pure. In this case we observe

$$\begin{array}{rcl} \|\mathcal{E}(\rho) - \mathcal{F}(\rho)\|_1 & \leq & p\|\mathcal{E}(\rho_1) - \mathcal{F}(\rho_1)\|_1 + (1-p)\|\mathcal{E}(\rho_2) - \mathcal{F}(\rho_2)\|_1 \\ & \leq & \max\{\|\mathcal{E}(\rho_1) - \mathcal{F}(\rho_1)\|_1, \|\mathcal{E}(\rho_2) - \mathcal{F}(\rho_2)\|_1\}, \end{array}$$

where we have used the linearity of CPMs and the triangle inequality in the first step. The application of this to smaller and smaller subsystems leads to pure states in the end. This proves the claim.

Claim 2: *The distance $\|\mathcal{E}(\rho) - \mathcal{I}(\rho)\|_1$ is invariant under unitary transformations of ρ, i.e.,*

$$\|\mathcal{E}(\rho) - \rho\|_1 = \|\mathcal{E}(U\rho U^*) - U\rho U^*\|_1.$$

Proof: *Because of the invariance of the trace norm under unitaries,*

$$\begin{aligned}\|\mathcal{E}(\rho) - \rho\|_1 &= \|U\mathcal{E}(\rho)U^* - U\rho U^*\|_1 \\ &= \|\mathcal{E}(U\rho U^*) - U\rho U^*\|_1,\end{aligned}$$

where we have used the explicit definition of the map $\mathcal{E}$ in the second step. This proves the claim.

Together, these two claims imply that we can use any pure state $\rho = |\psi\rangle\langle\psi|$ to maximize $\|\mathcal{E}(\rho) - \rho\|_1$. We chose $|\psi\rangle = |0\rangle$ where $\{|0\rangle, |1\rangle\}$ is the computational basis of $\mathbb{C}^2$. We get

$$\tilde{d}(\mathcal{E}, \mathcal{I}) = \left\| \begin{pmatrix} -\frac{p}{2} & 0 \\ 0 & \frac{p}{2} \end{pmatrix} \right\|_1 = p.$$

Proof of the inequality. *Now that we have computed $\tilde{d}(\mathcal{E}, \mathcal{I})$ we have a closer look at an experiment where the experimentalist implements the maps $\mathcal{E}$ and $\mathcal{I}$ as $\mathcal{E} \otimes \mathcal{I}_E = \mathcal{E} \otimes \mathcal{I}$ and $\mathcal{I} \otimes \mathcal{I}_E = \mathcal{I} \otimes \mathcal{I}$, respectively. We thus have to show that*

$$\tilde{d}(\mathcal{E}, \mathcal{I}) < \tilde{d}(\mathcal{E} \otimes \mathcal{I}_E, \mathcal{I} \otimes \mathcal{I}_E).$$

According to the definition of $\tilde{d}(..,..)$ it is sufficient to find a state $\rho \in \mathcal{S}(\mathbb{C}^2 \otimes \mathbb{C}^2)$ such that

$$\|\mathcal{E} \otimes \mathcal{I}(\rho) - \mathcal{I} \otimes \mathcal{I}(\rho)\|_1 \geq \tilde{d}(\mathcal{E}, \mathcal{I}) = p.$$

For simplicity, we assume $p = 1/2$. Our ansatz for ρ is the Bell state $|\beta_0\rangle\langle\beta_0|$.

Definition 4.4.11. The *Bell states* or *EPR pairs* are four specific two-qubit states $\beta_0, ..., \beta_3$ defined by

$$|\beta_\mu\rangle := \sum_{a,b \in \{0,1\}} \frac{1}{\sqrt{2}} (\sigma_\mu)_{ab} |a, b\rangle.$$

Hence,

$$|\beta_0\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle).$$

Inserting this state in $\|\mathcal{E} \otimes \mathcal{I}(\rho) - \mathcal{I} \otimes \mathcal{I}(\rho)\|_1$ gives approximately (for $p = 1/2$)

$$0.9789 > 1/2 = \tilde{d}(\mathcal{E}, \mathcal{I}),$$

where you may use Mathematica to diagonalize the resulting 4×4 matrix. This proves the inequality.

To summarize, we have found out that there exist situations in which

$$\tilde{d}(\mathcal{E},\mathcal{I}) < \tilde{d}(\mathcal{E}\otimes\mathcal{I}_E,\mathcal{I}\otimes\mathcal{I}_E).$$

This forces us to use

$$d(\mathcal{E},\mathcal{F}) := \max_{\rho\in\mathcal{S}(\mathcal{H}\otimes\mathcal{H}_E)} \|\mathcal{E}\otimes\mathcal{I}_E(\rho) - \mathcal{F}\otimes\mathcal{I}_E(\rho)\|_1$$

instead of our naive approach above. Next one asks how the distinguishing probability depends on the choice of the Hilbert space $\mathcal{H}_E$. To that purpose we are stating and proving two lemmas. In the final definition of distance between CPMs we will then use a Hilbert space $\mathcal{H}_E$ which maximizes the probability for distinguishing the CPMs $\mathcal{E}$ and $\mathcal{F}$.

Lemma 4.4.12. *Let ρ_{AB} be a pure state on a Hilbert space $\mathcal{H}_A\otimes\mathcal{H}_B$ and let $\rho'_{AB'}$ be an arbitrary state on a Hilbert space $\mathcal{H}_A\otimes\mathcal{H}_{B'}$, such that*

$$\mathrm{tr}_B\rho_{AB} = \mathrm{tr}_{B'}\rho'_{AB'}.$$

Then there exists a CPM $\mathcal{E}:\ \mathcal{S}(\mathcal{H}_B)\to\mathcal{S}(\mathcal{H}_{B'})$, such that

$$\rho'_{AB'} = \mathcal{I}_A\otimes\mathcal{E}(\rho_{AB}).$$

Proof. Assume that $\rho'_{AB'}$ is pure. Since ρ_{AB} is pure (and by the assumption made in the lemma) there exist states $\psi\in\mathcal{H}_A\otimes\mathcal{H}_B$ and $\psi'\in\mathcal{H}_A\otimes\mathcal{H}_{B'}$, such that $\rho_{AB}=|\psi\rangle\langle\psi|$ and $\rho'_{AB'}=|\psi'\rangle\langle\psi'|$. Let

$$|\psi\rangle = \sum_i \sqrt{\lambda_i}|v_i\rangle_A\otimes|w_i\rangle_B$$

and

$$|\psi'\rangle = \sum_i \sqrt{\lambda'_i}|v'_i\rangle_A\otimes|w'_i\rangle_{B'}$$

be the Schmidt decompositions of $|\psi\rangle$ and $|\psi'\rangle$. Without loss of generality we assume $|v_i\rangle_A = |v'_i\rangle_A$ because v_i and v'_i are both eigenvectors of the operator $\rho_A := \mathrm{tr}_B\rho_{AB} = \mathrm{tr}_{B'}\rho'_{AB'}$. Define the map

$$U := \sum_i |w'_i\rangle_{B'}\langle w_i|_B.$$

This map U is an isometry because $\{w_i\}_i$ and $\{w'_i\}_i$ are orthonormal systems in $\mathcal{H}_B$ and $\mathcal{H}_{B'}$, respectively. Consequently,

$$\psi' = (\mathrm{id}_A\otimes U)(\psi)$$

which proves the lemma for $\rho'_{AB'}$ being pure.
Now let's assume that $\rho'_{AB'}$ isn't pure and consider the purification $\rho'_{AB'R}$ of $\rho'_{AB'}$. Then (according to the statement proved so far) there exists a map

$$U:\ \mathcal{H}_B\to\mathcal{H}_{B'}\otimes\mathcal{H}_R,$$

such that

$$\rho'_{AB'R} = (\mathrm{id}_A \otimes U)\rho_{AB}(\mathrm{id}_A \otimes U^*).$$

Now we simply define $\mathcal{E} := \mathrm{tr}_R \circ \mathrm{ad}_{\mathrm{id}_A \otimes U}$ and thus

$$\rho'_{AB'} = \mathcal{I}_A \otimes \mathcal{E}(\rho_{AB})$$

which concludes the proof. □

Let us come back to the question about the best choice for the Hilbert space $\mathcal{H}_E$ appearing in the definition of the distance measure in the space of CPMs. Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two CPMs from $\mathcal{S}(\mathcal{H}_A)$ to $\mathcal{S}(\mathcal{H}'_A)$ and let ρ_A be a state in $\mathcal{S}(\mathcal{H}_A)$, $\rho_{AR} \in \mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_A)$ be the purification of ρ_A, ρ'_{AB} be a state in $\mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that $\rho_A = \mathrm{tr}_B \rho'_{AB}$. Because ρ_{AR} is pure there exists a state $\psi \in \mathcal{H}_A \otimes \mathcal{H}_A$, such that $\rho_{AR} = |\psi\rangle\langle\psi|$ and according to the Schmidt decomposition there exist $v_i \in \mathcal{H}_A$ and real numbers $\lambda_i \in \mathbb{R}$ such that

$$\psi = \sum_i \sqrt{\lambda_i} v_i \otimes v_i.$$

According to the lemma we just proved there exists a CPM $\mathcal{G} : \mathcal{S}(\mathcal{H}_A) \to \mathcal{S}(\mathcal{H}_B)$ such that

$$\rho'_{AB} = \mathcal{I}_A \otimes \mathcal{G}(\rho_{AR}).$$

The CPMs $\mathcal{E}_1$ and $\mathcal{E}_2$ act only on states in $\mathcal{S}(\mathcal{H}_A)$ and thus they act on the states ρ_{AR} and ρ'_{AB} as

$$\begin{aligned} \mathcal{E}_1 \otimes \mathcal{I}_B(\rho'_{AB}) &= (\mathcal{I}_A \otimes \mathcal{G}) \circ (\mathcal{E}_1 \otimes \mathcal{I}_A)(\rho_{AR}) \\ \mathcal{E}_2 \otimes \mathcal{I}_B(\rho'_{AB}) &= (\mathcal{I}_A \otimes \mathcal{G}) \circ (\mathcal{E}_2 \otimes \mathcal{I}_A)(\rho_{AR}). \end{aligned}$$

We have proved in an earlier chapter about quantum states and operations that trace preserving CPMs can never increase the distance between states. We thus get

$$\|\mathcal{E}_1 \otimes \mathcal{I}_B(\rho'_{AB}) - \mathcal{E}_2 \otimes \mathcal{I}_B(\rho'_{AB})\|_1 \leq \|\mathcal{E}_1 \otimes \mathcal{I}_A(\rho_{AR}) - \mathcal{E}_2 \otimes \mathcal{I}_A(\rho_{AR})\|_1.$$

This inequality holds for any choice of $\mathcal{H}_B$ and states in $\mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_B)$. We conclude that the right hand sight of our the inequality describes the best way to distinguish the CPMs $\mathcal{E}_1$ and $\mathcal{E}_2$ in an experiment. Consequently, this is the best choice for the distance measure between CPMs. This distance measure is induced by the following norm.

Definition 4.4.13 (Diamond norm for CPMs)**.** Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces and let

$$\mathcal{E} : \mathcal{S}(\mathcal{H}) \to \mathcal{S}(\mathcal{G})$$

be a CPM. Then the *diamond norm* $\|\mathcal{E}\|_\diamond$ of $\mathcal{E}$ is defined as

$$\|\mathcal{E}\|_\diamond := \|\mathcal{E} \otimes \mathcal{I}_{\mathcal{H}}\|_1,$$

where $\|\cdot\|_1$ denotes the so called *trace norm* for resources which is defined as

$$\|\Psi\|_1 := \max_{\rho \in \mathcal{S}(\mathcal{L}_1 \otimes \mathcal{L}_2)} \|\Psi(\rho)\|_1$$

where $\Psi : \mathcal{S}(\mathcal{L}_1) \to \mathcal{S}(\mathcal{L}_2)$ denotes an arbitrary CPM.

5 Basic protocols

5.1 Teleportation

Bennett, Brassard, Crépeau, Jozsa, Peres, Wootters, 1993.

> "An unknown quantum state $|\phi\rangle$ can be disassembled into, then later reconstructed from, purely classical information and purely nonclassical Einstein-Podolsky-Rosen (EPR) correlations. To do so the sender, Alice, and the receiver, Bob, must prearrange the sharing of an EPR-correlated pair of particles. Alice makes a joint measurement on her EPR particle and the unknown quantum system, and sends Bob the classical result of this measurement. Knowing this, Bob can convert the state of his EPR particle into an exact replica of the unknown state $|\phi\rangle$ which Alice destroyed."

With EPR correlations, Bennett *et al.* mean our familiar ebit $\frac{1}{\sqrt{2}}|00+11\rangle$. In more precise terms, we are interested in performing the following task:

Task: Alice wants to communicate the unknown state ρ of one qubit in system S to Bob. They share one Bell state. She can also send him two classical bits.

The protocol that achieves this, makes integral use of the *Bell measurement.* This is a measurement of two qubits and consists of projectors onto the four *Bell states*

$$|\psi^{00}\rangle = \frac{1}{\sqrt{2}}|00+11\rangle$$

$$|\psi^{01}\rangle = \frac{1}{\sqrt{2}}|00-11\rangle$$

$$|\psi^{10}\rangle = \frac{1}{\sqrt{2}}|01+10\rangle$$

$$|\psi^{11}\rangle = \frac{1}{\sqrt{2}}|01-10\rangle.$$

More compactly, we can write

$$|\psi^{ij}\rangle = \mathrm{id}\otimes\sigma^{ij}|\psi_{00}\rangle$$

where $\sigma^{ij} = \sigma_x^i \sigma_z^j$. For simplicity of the exposition, let $\rho = |\phi\rangle\langle\phi|$ be a pure state, $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$ (the more general case of mixed ρ follows then by linearity of the protocol). The global state before the protocol is therefore given by $|\phi\rangle_S \otimes |\psi^{00}\rangle_{AB}$. The protocol is as follows:

Protocol

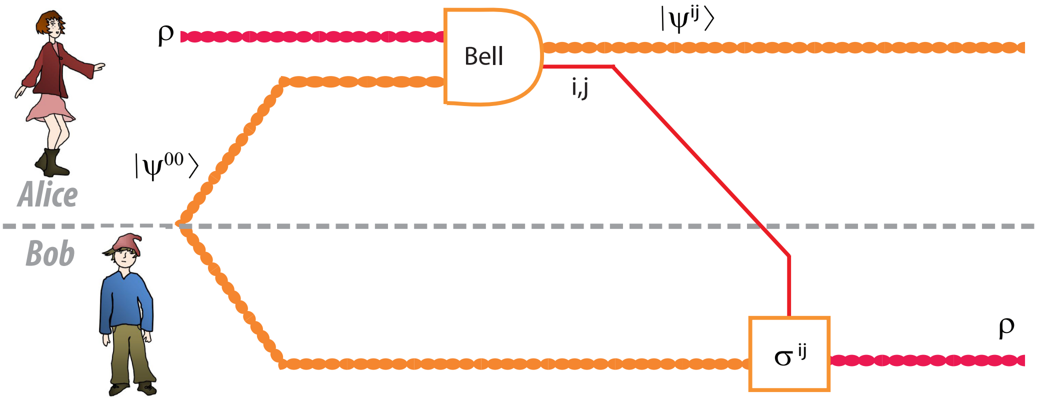
1. Alice measures S and A (her half of the entangled state) in the Bell basis.

Alice's outcome	Global projector	Resulting global state
$00: \ \|\psi^{00}\rangle_{SA}$	$\|\psi^{00}\rangle\langle\psi^{00}\|_{SA} \otimes \mathrm{id}_B$	$\|\psi^{00}\rangle_{SA} \otimes (\alpha\|0\rangle + \beta\|1\rangle)_B$
$01: \ \|\psi^{01}\rangle_{SA}$	$\|\psi^{01}\rangle\langle\psi^{01}\|_{SA} \otimes \mathrm{id}_B$	$\|\psi^{01}\rangle_{SA} \otimes (\alpha\|0\rangle - \beta\|1\rangle)_B$
$10: \ \|\psi^{10}\rangle_{SA}$	$\|\psi^{10}\rangle\langle\psi^{10}\|_{SA} \otimes \mathrm{id}_B$	$\|\psi^{10}\rangle_{SA} \otimes (\beta\|0\rangle + \alpha\|1\rangle)_B$
$11: \ \|\psi^{11}\rangle_{SA}$	$\|\psi^{11}\rangle\langle\psi^{11}\|_{SA} \otimes \mathrm{id}_B$	$\|\psi^{11}\rangle_{SA} \otimes (\beta\|0\rangle - \alpha\|1\rangle)_B$

2. Alice sends the classical bits that describe her outcome, i, j, to Bob.
3. Bob applies σ^{ij} on his qubit.

The resulting state is $|\phi\rangle$ as one easily verifies by direct computation or alternatively by the observation that

$$\sigma^{ij} \otimes \sigma^{ij}|\psi^{0}0\rangle = |\psi^{0}0\rangle.$$



Note that each outcome is equally probable and that entanglement between ρ and the rest of the universe is preserved.

Diagrammatically, we can summarise the teleportation as the following conversion of resources:

$$\begin{array}{c}\xrightarrow{2}\\ \overset{1}{\sim\!\sim}\end{array} \geq \overset{1}{\rightsquigarrow}$$

where the straight arrow represents the sending of a classical bit, the wiggly line an ebit and the wiggly arrow the sending of a qubit. The inequality sign means that there exists a protocol that can transform the *resources* of one ebit and two bits of classical communication into the resource of sending one qubit.

5.2 Superdense coding

Superdense coding answers the question of how many classical bits we can send with one use of a quantum channel if we are allowed to use preshared ebits.

Task Alice wants to send two classical bits, i and j, to Bob. They share one Bell state. She can also send him one qubit.

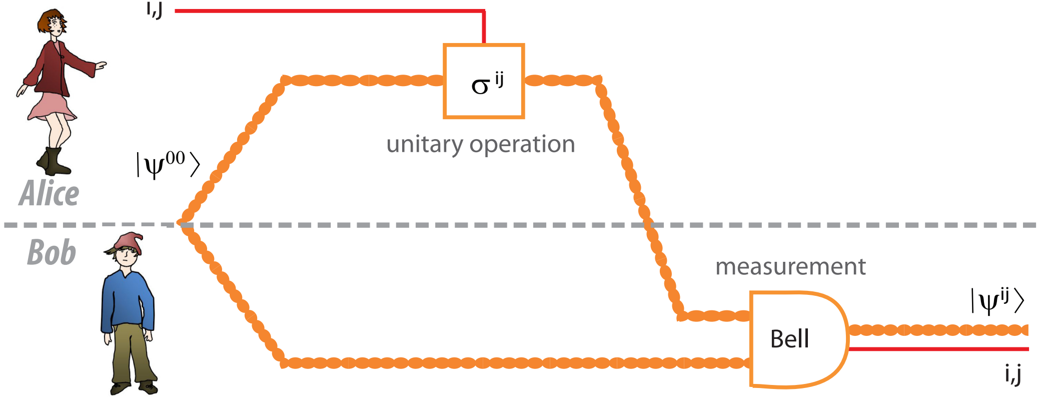
Protocol

1. Alice applies a local unitary operation, σ^{ij}, on her half of the entangled state.

i, j	Global operation	Resulting state
00	$\mathrm{id}_A \otimes \mathrm{id}_B \frac{\|00\rangle+\|11\rangle}{\sqrt{2}}$	$\frac{\|00\rangle+\|11\rangle}{\sqrt{2}} = \|\psi^{00}\rangle$
01	$\sigma^x_A \otimes \mathrm{id}_B \frac{\|00\rangle+\|11\rangle}{\sqrt{2}}$	$\frac{\|00\rangle-\|11\rangle}{\sqrt{2}} = \|\psi^{01}\rangle$
10	$\sigma^y_A \otimes \mathrm{id}_B \frac{\|00\rangle+\|11\rangle}{\sqrt{2}}$	$\frac{\|01\rangle+\|10\rangle}{\sqrt{2}} = \|\psi^{10}\rangle$
11	$\sigma^z_A \otimes \mathrm{id}_B \frac{\|00\rangle+\|11\rangle}{\sqrt{2}}$	$\frac{\|01\rangle-\|10\rangle}{\sqrt{2}} = \|\psi^{11}\rangle$

   Recall, that the states $|\psi^{ij}\rangle$ form a basis for two qubits: the Bell basis.

2. Alice sends her qubit to Bob.
3. Bob measures the two qubits in the Bell basis. Outcome of his measurement: i, j.



We can summarise the task of superdense coding in the following diagram:

$$\begin{array}{c}\overset{1}{\rightsquigarrow} \\ \overset{1}{\sim}\end{array} \geq \overset{2}{\rightarrow}$$

In order to show that this inequality is tight, i.e. that we cannot send more than two classical bits with one ebit and one use of a qubit channel, we will need some more technology - in particular the concept of quantum entropy.

5.3 Entanglement Conversion

With teleportation and superdense coding we have seen two tasks that can be solved nicely when we have access to ebits. In a realistic scenario, unfortunately, it is difficult to obtain or generate ebits exactly. It is therefore important to understand when and how we can *distill* ebits from other quantum correlations or more generally, how to convert one type of quantum correlation into another one. In this section, we will consider the simplest instance of this problem, namely the conversion of one bipartite pure state into another one. Before we state the main result, we need to do some preparatory work and introduce the concept of majorisation.

5.3.1 Majorisation

Given two d-dimensional real vectors x and y with entries in non-increasing order (i.e. $x_i \geq x_{i+1}$ and $y_i \geq y_{i+1}$) and of the same length $\sum_i x_i = \sum_i y_i$ we say that y *majorises* x, and write $x \prec y$ if

$$\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i$$

for all $k \in \{1, \dots, d\}$.

Lemma 5.3.1. *If y majorises x, then there exists a set of permutation matrices with associated probability $\{\pi_i, p_i\}$ such that*

$$x = \sum_i p_i \pi_i y.$$

Proof. We prove lemma inductively. Clearly the case $d = 1$ is true and we will therefore focus on the inductive step $d - 1 \mapsto d$.

$y \succ x$ implies that $x_1 \leq y_1$, which in turn implies that there exists j such that $y_j \leq x_1 \leq y_{j-1} \leq y_1$. Consequently, there is a $t \in [0,1]$ such that $x_1 = ty_1 + (1-t)y_j$. Let T be the transposition that interchanges places 1 and j and let $P = t\text{id} + (1-t)T$. Then $Py = (x_1, \underbrace{y_2, \dots, y_{j-1}, (1-t)y_1 + ty_j, \dots}_{\tilde{y}})$. It remains to show that $\tilde{y} \succ \tilde{x}$, where the latter is just x without x_1, since then the result follows by applying the inductive hypothesis to $\tilde{x}$ and $\tilde{y}$. This is shown as follows. For $k < j$:

$$\sum_{i=2}^{k} x_i \leq \sum_{i=2}^{k} x_1 \leq \sum_{i=2}^{k} y_{j-1} \leq \sum_{i=2}^{k} y_i.$$

For $k \geq j$:

$$\begin{aligned}
\sum_{i=1}^{k-1} \tilde{x}_i &= \sum_{i=2}^{k} x_i \\
&\leq \left(\sum_{j=2}^{k} y_i \right) + (y_1 - x_1) \\
&\leq \left(\sum_{i=2:i\neq j}^{k} y_i \right) + (y_1 - ty_j - (1-t)y_1 + y_j) \\
&= \left(\sum_{i=2:i\neq j}^{k} y_i \right) + (ty_1 + (1-t)y_j) \\
&= \sum_{i=1}^{k-1} \tilde{y}_i
\end{aligned}$$

□

Lemma 5.3.2. *Let A and B and $C = A + B$ be Hermitian operators with eigenvalues a, b and c ordered non-increasingly, then $c \prec a + b$*

Proof.

$$\begin{aligned}
\sum_{i=1}^{k} c_i &= \max_{V:|V|=k} \mathrm{tr} P_V (A + B) \\
&\leq \max_{V:|V|=k} \mathrm{tr} P_V A + \max_{W:|W|=k} \mathrm{tr} P_W B \\
&= \sum_{i=1}^{k} a_i + \sum_{i=1}^{k} b_i
\end{aligned}$$

where we used Ky Fan's principle which characterises the largest (and also the largest k) eigenvalues in a variational way. □

Corollary 5.3.3. *Let r and s be the eigenvalues (incl. multiplicities) of density matrices ρ and σ in non-increasing order. Then $s \succ r$ iff there exists a finite set of unitaries and associated probabilities $\{U_i, p_i\}$ such that*

$$\rho = \sum_i p_i U_i \sigma U_i^{-1}$$

Proof. If $s \succ r$, then according to Lemma 5.3.1 there exists a set of permutation matrices π_i (which are in particular unitary) and probabilities p_i such that $r = \sum_i p_i \pi_i s$. Inserting

$U\rho U^{-1} = \mathrm{diag}(r)$ and $V\sigma V^{-1} = \mathrm{diag}(s)$ for unitaries U and V arising from the spectral decomposition we find

$$U\rho U^{-1} = \sum_i p_i \pi_i V\sigma V^{-1}\pi_i^{-1}$$

which is equivalent to the claim for $U_i := U^{-1}\pi_i V$.
Conversely, Lemma 5.3.2 applied to $\rho = \sum_i p_i U_i \sigma U_i^{-1}$ implies

$$s = \mathrm{EV}(\sigma) = \sum_i \mathrm{EV}(p_i U_i \sigma U_i^{-1}) \succ \mathrm{EV}(\rho) = r$$

□

We now want to argue that any measurement on Bob's side of the state $|\psi\rangle$ can be replaced by a measurement on Alice's side and a unitary on Bob's side dependent on Alice's measurement outcome. Note that this is only possible since we know the state on which the measurement will be applied – without this knowledge this is impossible. In order to see how it works, we write $|\psi\rangle$ in its Schmidt decomposition

$$|\psi\rangle = \sum_i \psi_i |i\rangle_A |i\rangle_B$$

and express Bob's the Kraus operators of Bob's measurement B_k (i.e. $\sum_k B_k^\dagger B_k = \mathrm{id}$) in his Schmidt basis

$$B_k = \sum_{ij} b_{k,ji} |j\rangle\langle i|_B.$$

We now define measurement operators for Alice

$$A_k' = \sum_{ij} b_{k,ji} |j\rangle\langle i|_A$$

and note that

$$\mathrm{id} \otimes B_k |\psi\rangle = F A_k' \otimes \mathrm{id}|\psi\rangle$$

where F is the operator exchanging the two systems that we have encountered previously.[1] This shows in particular that the Schmidt coefficients of $\mathrm{id} \otimes B_k|\psi\rangle$ and $A_k' \otimes \mathrm{id}|\psi\rangle$ are identical. Therefore, there exist unitaries U_k and V_k such that

$$\mathrm{id} \otimes B_k |\psi\rangle = U_k \otimes V_k \cdot A_k' \otimes \mathrm{id}|\psi\rangle$$

which means that we can simulate the measurement on Bob's side on $|\psi\rangle$ by a measurement on Alice's side (with Kraus operators $A_k = U_k A_k'$) followed by a unitary V_k on Bob's side.

This way we can reduce an arbitrary LOCC protocol between Alice and Bob (applied to $|\psi\rangle$) by a measurement on Alice's side followed by a unitary on Bob's side conditioned on Alice's measurement outcome.

This preparation will allows us to prove the following result due to Nielsen.

[1] $F = \sum_{ij} |j\rangle\langle i| \otimes |i\rangle\langle j|$

Theorem 5.3.4. *$|\psi\rangle$ can be transformed into $|\phi\rangle$ by LOCC iff $r \prec s$, where r and s are the local eigenvalues of $|\psi\rangle$ and $|\phi\rangle$, respectively.*

Proof. Define $\rho_{AB} = |\psi\rangle\langle\psi|_{AB}$ and $\sigma_{AB} = |\phi\rangle\langle\phi|_{AB}$ with reduced states ρ_A and σ_A. By the above it suffices to consider protocols where Alice performs a measurement with Kraus operators A_k followed by a unitary V_k on Bob's side. Since the protocol must transform Alice's local state for each measurement outcome into the local part of the final state, we have

$$A_k \rho_A A_k^\dagger = p_k \sigma_A \tag{5.1}$$

for all k, where p_k is the probability to obtain outcome k. Let

$$A_k \sqrt{\rho_A} = |A_k \sqrt{\rho_A}| U_k = \sqrt{A_k \rho_A A_k} U_k$$

be the polar decomposition of the LHS. Multiplying this equation with its hermitian conjugate and using (5.1) we find

$$\sqrt{\rho_A} A_k^\dagger A_k \sqrt{\rho_A} = p_k U_k^\dagger \sigma_A U_k.$$

Summing over k yields

$$\rho_A = \sum_k p_k U_k^\dagger \sigma_A U_k \tag{5.2}$$

which by Corollary 5.3.3 implies that $r \prec s$.

In order to see the opposite direction, note that $r \prec s$ implies that there exist probabilities p_k and unitaries U_k such that (5.2) holds. We then define

$$A_k := \sqrt{p_k \sigma_A} U_k^\dagger \sqrt{\rho_A}^{-1}$$

where we assume for simplicity that ρ_A is invertible (the other case can be considered a limiting case). It is easy to verify that $\sum_k A_k^\dagger A_k = \mathrm{id}$. Clearly

$$A_k \rho_A A_k^\dagger = p_k \sigma_k$$

and therefore there exist unitaries V_k on Bob's side such that the final state is $|\phi\rangle$. □

6 Entropy of quantum states

In Chapter 3 we have discussed the definitions and properties of classical entropy measures and we have learned about their usefulness in the discussion of the channel coding theorem. After the introduction of the quantum mechanical basics in chapter 4 (and after the short insertion about the non-classicality in quantum theory) we are ready to introduce the notion of entropy in the quantum mechanical context. Textbooks usually start the discussion of quantum mechanical entropy with the definition of the so called von Neumann entropy and justify the explicit expression as being the most natural analog of the classical Shannon entropy for quantum systems. But this explanation is not completely satisfactory. Hence a lot of effort is made to replace the von Neumann entropy by the quantum version of the min-entropy which can be justified by its profound operational interpretation (recall for example the discussion of the channel coding theorem where we worked with the min-entropy and where the Shannon entropy only appears as a special case).

One can prove that the smooth min-entropy of a product state $\rho^{\otimes n}$ converges for large n to n-times the von Neumann entropy of the state ρ. The quantum mechanical min-entropy thus generalizes the von Neumann entropy in some sense. But since this work is still in progress we forgo this modern point of view and begin with the definition of the von Neumann entropy and only indicate at the end of the chapter these new developments.

6.1 Motivation and definitions

Let $\mathcal{H}_Z$ be a Hilbert space of dimension n which is spanned by the linearly independent family $\{|z\rangle\}_z$ and consider an arbitrary state ρ on $\mathcal{H}_Z$ which is classical with respect to $\{|z\rangle\}_z$. Hence,

$$\rho = \sum_z P_Z(z)|z\rangle\langle z|,$$

where $P_Z(z)$ is the probability distribution for measuring $|z\rangle$ in a measurement of ρ in the basis $\{|z\rangle\}_z$. Our central demand on the definition of the entropy measures of quantum states is that they generalize the classical entropies. More precisely, we demand that the evaluation of the quantum entropy on ρ yields the corresponding classical entropy of the distribution $P_Z(z)$. The following definitions meet these requirements as we will see below.

Definition 6.1.1. Let ρ be an arbitrary state on a Hilbert space $\mathcal{H}_A$. Then the *von Neumann entropy* H is the quantum mechanical generalization of the Shannon entropy. It is defined by

$$H(A)_\rho := -\mathrm{tr}(\rho \log \rho).$$

The *quantum mechanical min-entropy* $H_{\min}$ generalizes the classical min-entropy. It is defined by

$$H_{\min}(A)_\rho := -\log_2 \|\rho\|_\infty.$$

The *quantum mechanical max-entropy* $H_{\max}$ generalizes the classical max-entropy. It is defined by

$$H_{\max}(A)_\rho := \log_2 |\text{supp}(\rho)|,$$

where $\text{supp}(\rho)$ denotes the support of the operator ρ.

Now, we check if our requirement from above really is fulfilled. To that purpose we consider again the state

$$\rho_Z = \sum_z P_Z(z)|z\rangle\langle z|.$$

Since the map $\rho \to \rho \log \rho$ is defined through the eigenvalues of ρ,

$$H(Z)_\rho = -\text{tr}(\rho \log \rho) = -\sum_z P_Z(z) \log_2 P_Z(z),$$

which reproduces that the Shannon entropy as demanded. Recall that $\|\rho\|_\infty$ is the operator norm which equals the greatest eigenvalue of the operator ρ. Thus, the quantum mechanical min-entropy reproduces the classical min-entropy:

$$H_{\min}(Z)_\rho = -\log_2 \|\rho\|_\infty = -\log \max_{z \in \mathcal{Z}} P_Z(z).$$

To show that the classical max-entropy emerges as a special case from the quantum mechanical max-entropy we make the simple observation

$$H_{\max}(Z)_\rho = \log_2 |\text{supp}\rho| = \log_2 |\text{supp } P_Z|.$$

Notation. Let ρ_{AB} be a density operator on the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ and let ρ_A and ρ_B be defined as the partial traces

$$\rho_A := \text{tr}_B\ \rho_{AB}, \quad \rho_B := \text{tr}_A\ \rho_{AB}.$$

Then the entropies of the states $\rho_{AB} \in \mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_B)$, $\rho_A \in \mathcal{S}(\mathcal{H}_A)$ and $\rho_B \in \mathcal{S}(\mathcal{H}_B)$ are denoted by

$$H(AB)_\rho := H(AB)_{\rho_{AB}},\ H(A)_\rho := H(A)_{\rho_A},\ H(B)_{\rho_{AB}} := H(B)_{\rho_B}.$$

(other conventions are for instance $H(\rho)$ or $H(\rho_A)$, and there may be a few of those left in these notes)

6.2 Properties of the von Neumann entropy

In the present section we state and prove some basic properties of the von Neumann entropy.

Lemma 6.2.1. *Let ρ be an arbitrary state on $\mathcal{H}_A$. Then,*

$$H(A)_\rho \geq 0,$$

with equality iff ρ is pure.

Proof. Let $\{|j\rangle\}_j$ be a complete orthonormal system which diagonalizes ρ, i.e.,

$$\rho = \sum_j p_j |j\rangle\langle j|,$$

with $\sum_j p_j = 1$. Therefore,

$$H(A)_\rho = -\sum_j p_j \log p_j. \tag{6.1}$$

The function $-x \log x$ is positive on $[0, 1]$. Consequently, the RHS above is positive which shows that the entropy is non-negative. It is left to show that $H(A)_\rho = 0$ iff ρ is pure.

Assume $H(A)_\rho = 0$. Since the function $-x \log x$ is non-negative on $[0, 1]$ each term in the summation in (6.1) has to vanish separately. Thus, either $p_k = 0$ or $p_k = 1$ for all k. Because of the constraint $\sum_j p_j = 1$ exactly one coefficient p_m is equal to one whereas all the others vanish. We conclude that ρ describes the pure state $|m\rangle$.

Assume ρ is the pure state $|\phi\rangle$. Hence,

$$\rho = |\phi\rangle\langle\phi|$$

which yields $H(A)_\rho = 0$. □

Lemma 6.2.2. *The von Neumann entropy is invariant under similarity transformations, i.e.,*

$$H(A)_\rho = H(A)_{U\rho U^{-1}}$$

for $U \in GL(\mathcal{H}_A)$.

Proof. Let $f : \mathbb{R} \to \mathbb{R}$ be a function and let M be an operator on a Hilbert space $\mathcal{H}$. Recall that

$$f(M) := V^{-1} f(VMV^{-1})V,$$

where $V \in \mathrm{GL}(\mathcal{H})$ diagonalizes M. Now we show that

$$f(UMU^{-1}) = Uf(M)U^{-1}$$

for $U \in \mathrm{GL}(\mathcal{H})$ arbitrary. Let D denote the diagonal matrix similar to M. The operator VU^{-1} diagonalizes UMU^{-1}. According to the definition above,

$$f(UMU^{-1}) = UV^{-1} f(VU^{-1}UMU^{-1}UV^{-1})VU^{-1} = UV^{-1} f(VMV^{-1})VU^{-1}.$$

On the other hand

$$Uf(M)U^{-1} = UV^{-1} f(VMV^{-1})VU^{-1}.$$

This claims the assertion from above. Since the trace is unaffected by similarity transformations we conclude the proof by setting $M = \rho$ and $f(x) = x \log(x)$. □

Lemma 6.2.3. *Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be Hilbert spaces, let $|\psi\rangle$ be is a pure state on $\mathcal{H}_A \otimes \mathcal{H}_B$ and let $\rho_{AB} := |\psi\rangle\langle\psi|$. Then,*

$$H(A)_\rho = H(B)_\rho.$$

Proof. According to the Schmidt decomposition there exist orthonormal families $\{|i_A\rangle\}$ and $\{|i_B\rangle\}$ in $\mathcal{H}_A$ and $\mathcal{H}_B$, respectively, and positive real numbers $\{\lambda_i\}$ with the property $\sum_i \lambda_i^2 = 1$ such that

$$|\psi\rangle = \sum_i \lambda_i |i_A\rangle \otimes |i_B\rangle.$$

Hence, $\mathrm{tr}_B(\rho_{AB})$ and $\mathrm{tr}_A(\rho_{AB})$ have the same eigenvalues and thus, $H(A)_{\rho_{AB}} = H(B)_{\rho_{AB}}$. □

Lemma 6.2.4. *Let ρ_A and ρ_B be arbitrary states. Then,*

$$H(AB)_{\rho_A \otimes \rho_B} = H(A)_{\rho_A} + H(B)_{\rho_B}.$$

Proof. Let $\{p_i^A\}_i$ ($\{p_j^B\}_j$) and $\{|i_A\rangle\}_i$ ($\{|j_B\rangle\}_j$) be the eigenvalues and eigenvectors of the operators ρ_A (ρ_B). Hence,

$$\rho_A \otimes \rho_B = \sum_{ij} p_i^A p_j^B |i_A\rangle\langle i_A| \otimes |j_B\rangle\langle j_B|.$$

We deduce

$$\begin{aligned} H(AB)_{\rho_A \otimes \rho_B} &= -\sum_{ij} p_i^A p_j^B \log(p_i^A p_j^B) \\ &= H(A)_{\rho_A} + H(B)_{\rho_A}. \end{aligned}$$

□

Lemma 6.2.5. *Let ρ be a state on a Hilbert space $\mathcal{H}_A$ of the form*

$$\rho = p_1\rho_1 + ... + p_n\rho_n$$

with density operators $\{\rho_i\}_i$ having support on pairwise orthogonal subspaces of $\mathcal{H}$ and with $\sum_j p_j = 1$. Then,

$$H(A)_\rho = H_{class}(\{p_i\}_i) + \sum_j p_j H(A)_{\rho_j},$$

where $\{H_{class}(\{p_i\}_i)\}$ denotes the Shannon entropy of the probability distribution $\{p_i\}_i$.

Proof. Let $\{\lambda_j^{(i)}\}$ and $\{|j^{(i)}\rangle\}$ the eigenvalues and eigenvectors of the density operators $\{\rho_i\}$. Thus,

$$\rho = \sum_{i,j} p_i \lambda_j^{(i)} |j^{(i)}\rangle\langle j^{(i)}|$$

and consequently,

$$\begin{aligned}
H(A)_\rho &= -\sum_{i,j} p_i\lambda_j^{(i)}\log(p_i\lambda_j^{(i)}) \\
&= -\sum_i\left(\sum_j \lambda_j^{(i)}\right)p_i\log(p_i) - \sum_i p_i\sum_j \lambda_j^{(i)}\log(\lambda_j^{(i)}) \\
&= H_{\text{class}}(\{p_i\}) + \sum_i p_i H(A)_{\rho_i}.
\end{aligned}$$

□

A consequence of this lemma is that the entropy is concave. More precisely, let $\rho_1, ..., \rho_n$ be density operators on the same Hilbert space $\mathcal{H}_A$. Consider a mixture of those density operators according to a probability distribution $\{p_j\}_j$ on $\{1, ..., n\}$, $\rho = \sum_j p_j\rho_j$.

Then

$$H(A)_\rho \geq \sum_j p_j H(A)_{\rho_j}.$$

Proof. Let $\mathcal{H}_Z$ be an auxiliary Hilbert space of dimension n which is spanned by the linearly independent family $\{|i\rangle\}_i$ and let $\tilde{\rho}$ be the state

$$\tilde{\rho} := \sum_j p_j|j\rangle\langle j| \otimes \rho_A^{(j)}$$

on $\mathcal{H}_Z \otimes \mathcal{H}_A$ which is classical on $\mathcal{H}_Z$ with respect to $\{|i\rangle\}_i$. According to the strong subadditivity property

$$H(Z|A)_{\tilde{\rho}} \leq H(Z)_{\tilde{\rho}}$$

or equivalently,

$$H(ZA)_{\tilde{\rho}} \leq H(Z)_{\tilde{\rho}} + H(B)_{\tilde{\rho}}.$$

Using Lemma 6.2.6, we get

$$\begin{aligned}
H(ZA)_{\tilde{\rho}} &= H(\{p_j\}_j) + \sum_j p_j H(\rho_A^{(j)}) \\
H(Z)_{\tilde{\rho}} &= H(\{p_j\}_j) \\
H(B)_{\tilde{\rho}} &= H(p_1\rho_A^{(1)} + ... + \rho_A^{(n)}),
\end{aligned}$$

and thus,

$$p_1 H(\rho_A^{(1)}) + ... + p_n H(\rho_A^{(n)}) \leq H(p_1\rho_A^{(1)} + ... + \rho_A^{(n)})$$

□

Lemma 6.2.6. *Let $\mathcal{H}_A$ and $\mathcal{H}_Z$ be Hilbert spaces and let ρ_{AZ} be a state on $\mathcal{H}_A \otimes \mathcal{H}_Z$ which is classical on $\mathcal{H}_Z$ with respect to the basis $\{|z\rangle\}_z$ of $\mathcal{H}_Z$, i.e., ρ_{AZ} is of the form*

$$\rho_{AZ} = \sum_z P_Z(z)\rho_A^{(z)} \otimes |z\rangle\langle z|.$$

Then

$$H(AZ)_\rho = H_{class}(\{P_Z(z)\}_z) + \sum_z P_Z(z) H(A)_{\rho_A^{(z)}}.$$

Proof. Define

$$\tilde{\rho}_z := \rho_A^{(z)} \otimes |z\rangle\langle z|,$$

apply Lemma 6.2.5 with ρ_i replaced by $\tilde{\rho}_z$, use lemma 6.2.4 and apply Lemma 6.2.1. □

6.3 The conditional entropy and its properties

We have encountered the identity

$$H_{\text{class}}(X|Y) = H_{\text{class}}(XY) - H_{\text{class}}(Y)$$

for classical entropies in the chapter about classical information theory. We use exactly this identity to *define* conditional entropy in the context of quantum information theory.

Definition 6.3.1. Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be two Hilbert spaces and let ρ_{AB} be a state on $\mathcal{H}_A \otimes \mathcal{H}_B$. Then, the conditional entropy $H(A|B)_\rho$ is defined by

$$H(A|B)_{\rho_{AB}} := H(AB)_{\rho_{AB}} - H(B)_{\rho_{AB}}.$$

Recasting this defining equation leads immediately to the so called *chain rule*:

$$H(AB)_{\rho_{AB}} = H(A|B)_{\rho_{AB}} + H(B)_{\rho_{AB}}.$$

Lemma 6.3.2. *Let ρ_{AB} be a* pure *state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$. Then $H(A|B)_{\rho_{AB}} < 0$ iff ρ_{AB} is entangled, i.e. $H(AB)_{\rho_{AB}} \neq H(A)_{\rho_{AB}} + H(B)_{\rho_{AB}}$.*

Proof. Observe that

$$H(A|B)_{\rho_{AB}} = H(AB)_{\rho_{AB}} - H(B)_{\rho_{AB}}.$$

Recall from Lemma 6.2.1 that the entropy of a state is zero iff it is pure. The state $\text{tr}_A(\rho_{AB})$ is pure iff ρ_{AB} is not entangled. Thus. indeed $H(A|B)_{\rho_{AB}}$ is negative iff ρ_{AB} is entangled. □

Hence, the *the conditional entropy can be negative.*

Lemma 6.3.3. *Let $\mathcal{H}_A$, $\mathcal{H}_B$ and $\mathcal{H}_C$ be Hilbert spaces and let ρ_{ABC} be a state on $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. Then,*

$$H(A|B)_{\rho_{ABC}} = -H(A|C)_{\rho_{ABC}}.$$

Proof. We have seen in Lemma 6.2.3 that ρ_{ABC} pure implies that

$$H(AB)_\rho = H(C)_\rho,\ H(AC)_\rho = H(B)_\rho,\ H(BC)_\rho = H(A)_\rho.$$

Thus,

$$H(A|B)_\rho = H(AB)_\rho - H(B)_\rho = H(C)_\rho - H(AC)_\rho = -H(A|C)\rho.$$

□

Lemma 6.3.4. *Let $\mathcal{H}_A$ and $\mathcal{H}_Z$ be Hilbert spaces, let $\{|z\rangle\}_z$ be a complete orthonormal basis in $\mathcal{H}_Z$ and let ρ_{AZ} be classical on $\mathcal{H}_Z$ with respect to the basis $\{|z\rangle\}_z$, i.e.,*

$$\rho_{AZ} = \sum_z P_Z(z)\rho_A^{(z)} \otimes |z\rangle\langle z|.$$

Then the entropy conditioned on Z is

$$H(A|Z)_\rho = \sum_z P_Z(z)H(\rho_A^{(z)}).$$

Moreover,

$$H(A|Z)_\rho \geq 0.$$

Proof. Apply Lemma 6.2.6 to get

$$\begin{aligned} H(A|Z)_\rho &= H(AZ)_\rho - H(Z)_\rho \\ &= H_{\text{class}}(P_Z(z)) + \sum_z P_Z(z)H(\rho_A^{(z)}) - H_{\text{class}}(P_Z(z)) \\ &= \sum_z P_Z(z)H(\rho_A^{(z)}). \end{aligned}$$

In Lemma 6.2.1 we have seen that $H(\rho) \geq 0$ for all states ρ. Hence, $H(A|Z)_\rho \geq 0$. □

Now it's time to state one of the central identities in quantum information theory: the so called *strong subadditivity.*

Theorem 6.3.5. *Let ρ_{ABC} be a state on $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. Then,*

$$H(A|B)_{\rho_{ABC}} \geq H(A|BC)_{\rho_{ABC}}.$$

In textbooks you presently find complex proofs of this theorem based on the Araki-Lieb inequality (see for example [10]) . An alternative shorter proof can be found in [12].

Lemma 6.3.6. *Let ρ be an arbitrary state on a d-dimensional Hilbert space $\mathcal{H}$. Then,*

$$H(\rho) \leq \log_2 d,$$

with equality iff ρ is a completely mixed state, i.e., a state similar to $\frac{1}{d}id_{\mathcal{H}}$.

Proof. Let ρ be a state on $\mathcal{H}$ which maximizes the entropy and let $\{|j\rangle\}$ the diagonalizing basis, i.e.,

$$\rho = \sum_j p_j |j\rangle\langle j|.$$

The entropy does only depend on the state's eigenvalue, thus, in order to maximize the entropy, we are allowed to consider the entropy H as a function mapping ρ's eigenvalues $(p_1, ..., p_d) \in [0,1]^d$ to $\mathbb{R}$. Consequently, we have to maximize the function $H(p_1, ..., p_d)$

under the constraint $p_1 + ... + p_d = 1$. This is usually done using Lagrange multipliers. One gets $p_j = 1/d$ for all $j = 1, ..., d$ and therefore,

$$\rho = \frac{1}{d}\mathrm{id}_{\mathcal{H}}$$

(this is the completely mixed state). This description of the state uniquely characterizes the state independently of the choice of the basis the matrix above refers to since the identity $\mathrm{id}_{\mathcal{H}}$ is unaffected by similarity transformations. This proves that ρ is the only state that maximizes the entropy. The immediate observation that

$$S(\rho) = \log_2 d$$

concludes the proof. □

Lemma 6.3.7. *Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be two Hilbert spaces and let $d :=$ dim $\mathcal{H}_A$. Then,*

$$|H(A|B)_\rho| \leq \log_2(d).$$

Proof. Use Lemma 6.3.6 to get

$$H(A|B)_\rho \leq H(A)_\rho \leq \log_2(d)$$

and Lemma 6.3.3 to get

$$H(A|B)_{\rho_{AB}} = H(A|B)_{\rho_{ABC}} = -H(A|C)_{\rho_{ABC}} \geq -\log(d),$$

where ρ_{ABC} is a purification of ρ_{AB}. □

Lemma 6.3.8. *Let $\mathcal{H}_X$ and $\mathcal{H}_B$ be Hilbert spaces, $\{|x\rangle\}_z$ be a complete orthonormal basis in $\mathcal{H}_X$ and let ρ_{XB} be a state on $\mathcal{H}_X \otimes \mathcal{H}_B$ which is classical with respect to $\{|x\rangle\}_x$. Then,*

$$H(X|B)_\rho \geq 0$$

which means that the entropy of a classical system is non-negative.

Proof. Let $\mathcal{H}_{X'}$ be a Hilbert space isomorphic to $\mathcal{H}_X$ and let $\rho_{BXX'}$ be a state on $\mathcal{H}_B \otimes \mathcal{H}_X \otimes \mathcal{H}_{X'}$ defined by

$$\rho_{BXX'} := \sum_{x,j} P_X(x)\rho_B^{(x)} \otimes |x\rangle\langle x| \otimes |x\rangle\langle x|.$$

Hence,

$$H(X|B)_{\rho_{BXX'}} = H(BX)_{\rho_{BXX'}} - H(B)_{\rho_{BXX'}}$$

and

$$H(X|BX')_{\rho_{BXX'}} = H(BXX')_{\rho_{BXX'}} - H(BX')_{\rho_{BXX'}}.$$

According to the strong subadditivity

$$H(X|B)_{\rho_{BXX'}} \geq H(X|BX')_{\rho_{BXX'}}.$$

To prove the assertion we have to show that the RHS vanishes or equivalently that $H(BXX')_{\rho_{BXX'}}$ is equal to $H(BX')_{\rho_{BXX'}}$. Let $\rho_{BX'}$ denote the state which emerges from $\rho_{BXX'}$ after the application of $\mathrm{tr}_X(\cdot)$. Hence, $H(BX')_{\rho_{BXX'}} = H(BX')_{\rho_{BX'}}$. Further,

$$H(BX')_{\rho_{BX'}} = H(BX')_{\rho_{BX'} \otimes |0\rangle\langle 0|},$$

where $|0\rangle$ is a state in the basis $\{|x\rangle\}_z$ of the Hilbert space $\mathcal{H}_X$. Define the map

$$S: \; \mathcal{H}_X \otimes \mathcal{H}_{X'} \to \mathcal{H}_X \otimes \mathcal{H}_{X'}$$

by

$$\begin{aligned} S(|z0\rangle) &:= |zz\rangle \\ S(|zz\rangle) &:= |z0\rangle \\ S(|xy\rangle) &:= |xy\rangle, \text{ (otherwise)}. \end{aligned}$$

We observe,

$$[\mathcal{I}_B \otimes S]\rho_{BX'} \otimes |0\rangle\langle 0|[\mathcal{I}_B \otimes S]^{-1} = \rho_{BXX'}.$$

Obviously, $[\mathcal{I}_B \otimes S] \in \mathrm{GL}(\mathcal{H}_X \otimes \mathcal{H}_{X'})$ (the general linear group) and thus does not change the entropy:

$$H(BX')_{\rho_{BXX'}} = H(BX')_{\rho_{BX'} \otimes |0\rangle\langle 0|} = H(BXX')_{\rho_{BXX'}}.$$

□

Lemma 6.3.9. *Let $\mathcal{H}_A$, $\mathcal{H}_B$ and $\mathcal{H}_{B'}$ be Hilbert spaces, let ρ_{AB} be a state on $\mathcal{H}_A \otimes \mathcal{H}_B$, let*

$$\mathcal{E}: \; \mathcal{H}_B \to \mathcal{H}_{B'}$$

be a TPCPM($\mathcal{H}_B, \mathcal{H}_{B'}$) and let

$$\rho_{AB'} = [\mathcal{I}_A \otimes \mathcal{E}](\rho_{AB})$$

be a state on $\mathcal{H}_A \otimes \mathcal{H}_{B'}$. Then,

$$H(A|B)_{\rho_{AB}} \leq H(A|B')_{\rho_{AB'}}.$$

Proof. Let $|0\rangle$ be a state in an auxiliary Hilbert space $\mathcal{H}_R$. Then

$$\begin{aligned} H(A|B)_{\rho_{AB}} &= H(AB)_{\rho_{AB}} - H(B)_{\rho_{AB}} \\ &= H(ABR)_{\rho_{AB} \otimes |0\rangle\langle 0|} - H(BR)_{\rho_{AB} \otimes |0\rangle\langle 0|}. \end{aligned}$$

According to the Stinespring dilation the Hilbert space $\mathcal{H}_R$ can be chosen such that there exists a unitary U with the property

$$\mathrm{tr}_R \circ \mathrm{ad}_U(\xi \otimes |0\rangle\langle 0|) = \mathcal{E}(\xi),$$

where $\mathrm{ad}_U(\cdot) := U(\cdot)U^{-1}$ and $\xi \in \mathcal{S}(\mathcal{H}_B)$. Since the entropy is invariant under similarity transformations we can use this transformation U to get

$$\begin{aligned}
H(A|B)_{\rho_{AB}} &= H(AB'R)_{[\mathcal{I}_A\otimes\mathrm{ad}_U](\rho_{AB}\otimes|0\rangle\langle 0|)} - H(B'R)_{[\mathcal{I}_A\otimes\mathrm{ad}_U](\rho_{AB}\otimes|0\rangle\langle 0|)} \\
&= H(A|B'R)_{[\mathcal{I}_A\otimes\mathrm{ad}_U](\rho_{AB}\otimes|0\rangle\langle 0|)} \\
&\leq H(A|B')_{[\mathcal{I}_A\otimes\mathrm{tr}_R\circ\mathrm{ad}_U](\rho_{AB}\otimes|0\rangle\langle 0|)} \\
&= H(A|B')_{[\mathcal{I}_A\otimes\mathcal{E}](\rho_{AB})} \\
&= H(A|B')_{\rho_{AB'}},
\end{aligned}$$

where we have used the strong subadditivity and the Stinespring dilation. We get

$$H(A|B)_{\rho_{AB}} \leq H(A|B')_{\rho_{AB'}},$$

which concludes the proof.

□

6.4 The mutual information and its properties

Definition 6.4.1. Let ρ_{AB} a state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$. Then, the so called *mutual information* $I(A:B)$ is defined by

$$I(A:B) := H(A)_{\rho_{AB}} + H(B)_{\rho_{AB}} - H(AB)_{\rho_{AB}} = H(A)_{\rho_{AB}} - H(A|B)_{\rho_{AB}}$$

Let ρ_{ABC} a state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. Then, the so called *conditional mutual information* $I(A:B|C)$ is defined by

$$I(A:B|C) := H(A|C)_{\rho_{ABC}} - H(A|BC)_{\rho_{ABC}}$$

We observe that the definition of quantum mutual information and the definition of classical mutual information are formally identical. Next we prove a small number of properties of the mutual information.

Lemma 6.4.2. *Let ρ_{ABC} a state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. Then,*

$$I(A:B|C) \geq 0.$$

This Lemma is a direct corollary of the strong subadditivity property of conditional entropy.

Lemma 6.4.3. *Let $\mathcal{H}_A$, $\mathcal{H}_B$, $\mathcal{H}_{B'}$ be Hilbert spaces, let ρ_{AB} a state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ and let*

$$\mathcal{E}: \ \mathcal{H}_B \to \mathcal{H}_{B'}$$

be a TPCPM. Then,

$$I(A:B) \geq I(A:B').$$

This is an immediate consequence of Lemma 6.3.9.

Lemma 6.4.4. *Let $\mathcal{H}_A$, $\mathcal{H}_B$, $\mathcal{H}_C$ be Hilbert spaces and let ρ_{ABC} be a state on a Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. Then,*

$$I(A : BC) = I(A : B) + I(A : C|B).$$

To prove this statement we simply have to plug in the definition of mutual information and conditional mutual information.

Exercise (Bell state). Compute the mutual information $I(A : B)$ of a Bell state ρ_{AB}. You should get $H(A) = 1$, $H(A|B) = -1$ and thus $I(A : B) = 2$.

Exercise (Cat state). Let $\mathcal{H}_A$, $\mathcal{H}_B$, $\mathcal{H}_C$ and $\mathcal{H}_D$ be Hilbert spaces of quantum mechanical 2-level systems which are spanned by $\{|0\rangle_A, |1\rangle_A\}$, $\{|0\rangle_B, |1\rangle_B\}$, $\{|0\rangle_C, |1\rangle_C\}$ and $\{|0\rangle_D, |1\rangle_D\}$, respectively. Then, the so called *cat state* is defined the pure state

$$|\psi\rangle := \frac{1}{\sqrt{2}}(|0\rangle_A|0\rangle_B|0\rangle_C|0\rangle_D + |1\rangle_A|1\rangle_B|1\rangle_C|1\rangle_D).$$

Hence $\rho_{ABCD} = |\psi\rangle\langle\psi|$ is the corresponding density matrix. Compute the expressions $I(A : B)$, $I(A : B|C)$, $I(A : B|CD)$ and $I(A : BCD)$. During your calculations you should get

$$\begin{aligned} H(A)_\rho = H(B)_\rho = H(C)_\rho = H(D)_\rho &= 1,\\ H(AB)_\rho = H(AC)_\rho &= 1,\\ H(ABC)_\rho = H(D)_\rho &= 1,\\ H(ABCD)_\rho &= 0. \end{aligned}$$

6.5 Conditional Min-Entropy

In this section, we will introduce conditional min-entropy and discuss some of its properties and uses. The definition is a quantum generalisation of the classical conditional min-entropy which we have discussed some time ago, i.e. the maximum value of a conditional probability distribution is replaced by a maximum eigenvalue of a conditional operator – the the only change that we have to maximise over different versions of a conditional operator:

$$H_{\min}(A|B)_\rho = \max_{\sigma_B}(-\log \lambda_{\max}(\mathrm{id}_A \otimes \sigma_B^{-1/2} \rho_{AB} \mathrm{id}_A \otimes \sigma_B^{-1/2}))$$

where the maximisation is taken over density operators σ_B. σ_B^{-1} denotes the pseudo-inverse of σ_B, i.e. the operator

$$\sigma_B^{-1} = U\mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_\ell^{-1}, 0, \dots, 0)U^\dagger,$$

where $\sigma_B = U\mathrm{diag}(\lambda_1, \dots, \lambda_\ell, 0, \dots, 0)U^\dagger$ with $\lambda_1 \geq \cdots \geq \lambda_\ell > 0$ is the spectral decomposition of σ_B. There is an alternative way of writing the conditional min-entropy which often comes in handy when doing computations:

$$H_{\min}(A|B)_\rho = \max_{\sigma_B}(-\log \min\{\lambda : \lambda \mathrm{id}_A \otimes \sigma_B \geq \rho_{AB}\}).$$

The following lemma shows that conditional min-entropy characterises the maximum probability of guessing a value X correctly giving access to quantum information in a register B.

Lemma 6.5.1. *Let $\rho_{XB} = \sum_x |x\rangle\langle x| \otimes \rho_x$, then*

$$H_{\min}(X|B) = -\log p_{guess}(X|B)$$

where

$$p_{guess}(X|B) = \max_{\{E_x\} POVM} \sum_x \mathrm{tr}[\rho_x E_x]$$

is the maximum probability of guessing X correctly given access to B.

Proof. The proof uses semidefinite programming, an extension of linear programming. For a review see [14]. Defining $C = -\sum_x |x\rangle\langle x| \otimes \rho_x$, $\tilde{X} = \sum_x |x\rangle\langle x| \otimes E_x$, $A_{ij} = \mathrm{id} \otimes e_{ij}$ where e_{ij} denotes a matrix with a one in column i and row j and $b_{ij} := \delta_{ij}$, p_{guess} takes the classic form of a primal semidefinite programme:

$$-\min \mathrm{tr} C\tilde{X} : \tilde{X} \geq 0, \sum_{ij} \mathrm{tr} A_{ij}\tilde{X} = b_{ij}.$$

The dual programme is

$$\max \sum_{ij} b_{ij} y_{ij} : \sum_{ij} y_{ij} A_{ij} \leq C$$

Setting $y_{ij} := -\sigma_{ij}$ this SDP reads

$$\max -\mathrm{tr}\sigma : \mathrm{id} \otimes \sigma \geq \sum_x |x\rangle\langle x| \otimes \rho_x.$$

Both programmes are strictly feasible, since the points $X = \mathrm{id}$ and $\sigma = \mathrm{id}$ are feasible points, respectively. By semidefinite programming duality, the two programmes therefore have the same value. This proves the claim. □

Recall that by definition the conditional von Neumann entropy satisfies $H(A|B) = H(AB) - H(B)$. From the definition of the conditional min-entropy such an inequality is certainly non-obvious and indeed false when taken literally. For most purposes, a set of inequalities replaces this important equality (which is often known as a chain rule). To give you the flavor of such inequalities we will prove the most basic one:

Lemma 6.5.2.

$$H_{\min}(A|B) \geq H_{\min}(AB) - H_{\max}(B)$$

Proof.

$$\begin{aligned}
H_{\min}(A|B)_\rho &= \max_{\sigma_B}(-\log\min\{\lambda : \lambda \mathrm{id}_A \otimes \sigma_B \geq \rho_{AB}\}) && (6.2)\\
&\geq -\log\min\{\lambda : \lambda \mathrm{id}_A \otimes \frac{\rho_B^0}{|\mathrm{supp}\rho_B|} \geq \rho_{AB}\} && (6.3)\\
&= -\log\min\{\mu|\mathrm{supp}\rho_B| : \mu \mathrm{id}_A \otimes \mathrm{id}_B \geq \rho_{AB}\} && (6.4)\\
&= -\log\min\{\mu : \mu \mathrm{id}_A \otimes \mathrm{id}_B \geq \rho_{AB}\} - \log|\mathrm{supp}\rho_B| && (6.5)\\
&= H_{\min}(AB)_\rho - H_{\max}(B)_\rho && (6.6)
\end{aligned}$$

where ρ_B^0 denotes the projector onto the support of ρ_B. □

Strong subadditivity of von Neumann entropy is the inequality:

$$H(AB) + H(BC) \geq H(ABC) + H(B).$$

Using the definition of the conditional von Neumann entropy, this is equivalent to the inequality

$$H(A|B) \geq H(A|BC)$$

which is often interpreted as "conditioning reduces entropy". In this form, it has a direct analog for conditional min entropy:

Lemma 6.5.3.

$$H_{\min}(A|B) \geq H_{\min}(A|BC)$$

Proof. Since $\lambda\sigma_{BC} \geq \rho_{ABC}$ implies $\lambda\sigma_B \geq \rho_{AB}$ we find for the σ_{BC} that maximises the expression for $H_{\min}(A|BC)$

$$\begin{aligned} H_{\min}(A|BC)_\rho &= -\log\min\{\lambda : \lambda \mathrm{id}_A \otimes \sigma_{BC} \geq \rho_{ABC}\} && (6.7)\\ &\leq \max_{\sigma_B}(-\log\min\{\lambda : \lambda \mathrm{id}_A \otimes \sigma_B \geq \rho_{AB}\}) = H_{\min}(A|B)_\rho. && (6.8) \end{aligned}$$

□

In the exercises, you will show how these two lemmas also hold for the smooth min and max-entropy. Combined with the asymptotic equipartition property that we discussed in the part on classical information theory you will then prove strong subadditivity of von Neumann entropy. The very fundamental result by the mathematical physicist Beth Ruskai and Elliot Lieb was proven in 1973 and remains the only known inequality for the von Neumann entropy — there may be more, we just haven't discovered them yet!

7 Resources Inequalities

We have seen that ebits, classical communication and quantum communication can be seen as valuable resources with which we can achieve certain tasks. An important example was the teleportation protocol which shows one ebit and two bits of classical communication can simulate the transmission of one qubit. In the following we will develop a framework for the transformation resources and present a technique that allows to show the optimality of certain transformations.

7.1 Resources and Inequalities

We will consider a setup with two parties, Alice and Bob, who wish to convert one type of resource to another (one may also consider more than two parties, but this is a little outside the scope of this course). The resources we consider are:

- $\overset{n}{\rightsquigarrow}$ perfect quantum channel
  (Alice sends n qubits to Bob)
- $\overset{n}{\rightarrow}$ perfect classical channel
  (Alice sends n bits to Bob)
- $\overset{n}{\sim\!\sim}$ shared entanglement, or *ebits*
  (Alice and Bob share n Bell pairs)
- $\overset{n}{\text{—}}$ shared bits

A resource inequality is a relation $X \geq Y$ which is to be interpreted as "we can obtain Y using X". Formally, there exists a protocol to simulate resources Y using only resources X and local operations. The example to keep in mind is the teleportation protocol which achieves:

$$\begin{array}{c}\overset{2}{\rightarrow}\\ \overset{1}{\sim\!\sim}\end{array} \geq \overset{1}{\rightsquigarrow}$$

Sometimes, our resources are noisy and we do not require the resource conversion to be perfect. We can then still use resource inequalities to formulate our results but it becomes a little cumbersome as you can see in the case of Shannon's noiseless coding theorem for a channel $P_{Y|X}$:

$\xrightarrow{n} P_{Y|X} \geq_{\epsilon} \xrightarrow{n(\max_{P_X} I(X;Y)-\epsilon)}$, for all $\epsilon > 0$ and n large enough.

In the remainder we will only be concerned with an exact conversion of perfect resources with the main goal to show that the teleportation and superdense coding protocols are optimal.

7.2 Monotones

Given a class of quantum operations, a *monotone* M is a function from states into the real numbers that has the property that it does not increase under any operations from the class. Rather than making this definition too formal (e.g. by specifying exactly on which systems the operations act), we will consider a few characteristic examples.

Example 7.2.1. *For bipartite states, the quantum mutual information is a monotone for the class of local operations. More precisely, given a bipartite state ρ_{AB} and a local quantum operation (CPTP map), say on Bob's side, $\Lambda : \mathrm{End}(B) \mapsto \mathrm{End}(B')$*

$$I(A:B) \geq I(A:B').$$

Proof. Let $U_{B\to B'B''}$ be a Stinespring dilation of Λ. Since an isometry does not change the entropy, we have

$$I(A:B) = I(A:B'B'')$$

The RHS can be expanded as

$$I(A:B'B'') = I(A:B') + I(A:B''|B').$$

Strong subadditivity implies that the second term is nonnegative which leads us to the desired conclusion. □

A similar argument shows that

$$I(A:B|E) \geq I(A:B'|E).$$

where ρ_{ABE} is an arbitrary extension of ρ_{AB}, i.e. satisfies $\mathrm{tr}_E \rho_{ABE} = \rho_{AB}$.

Example 7.2.2 (Squashed entanglement)**.** *The squashed entanglement of a state ρ_{AB} is given by*

$$E_{sq}(A:B) := \frac{1}{2} \inf_E I(A:B|E)$$

where the minimisation extends over all extensions ρ_{ABE} of ρ_{AB}. Note that we do not impose a limit on the dimension of E!(That is why we do not know whether the minimum is achieved and write inf *rather than* min.*) Squashed entanglement is a monotone under local operations and classical communication. That squashed entanglement is monotone under local operations follows immediately from the previous example. We just only need to verify that it does not increase under classical communication.*

Proof. Alice will send classical system C to Bob (e.g. a bit string).

We want to compare $E_{sq}(AC : B)$ and $E_{sq}(A : BC)$. For any extension E, we have

$$\begin{aligned}
I(B : AC|E) &= H(B|E) - H(B|ACE) \\
&\geq H(B|EC) - H(B|AEC) \quad \text{(strong subadditivity)} \\
&= I(B : A|EC) \\
&= I(BC : A|EC) \quad EC =: E' \\
&\geq \min_{E'} I(BC : A|E')
\end{aligned}$$

This shows that $E_{sq}(AC : B) \geq E_{sq}(A : BC)$. By symmetry then $E_{sq}(AC : B) = E_{sq}(A : BC)$. □

7.3 Teleportation is Optimal

We will first show how to use monotones in order to prove that any protocol for teleportation of m qubits needs at least n ebits, regardless of how much classical communication the protocol uses. In our graphical notation this reads:

$$\begin{matrix}\overset{\infty}{\rightarrow} \\ \overset{n}{\sim\!\sim}\end{matrix} \geq \overset{m}{\rightsquigarrow} \quad \text{implies } n \geq m\ .$$

Note first that by sending m halves of ebits down the quantum channel on the RHS of

$$\begin{matrix}\overset{\infty}{\rightarrow} \\ \overset{n}{\sim\!\sim}\end{matrix} \geq \overset{m}{\rightsquigarrow}$$

we find

$$\begin{matrix}\overset{\infty}{\rightarrow} \\ \overset{n}{\sim\!\sim}\end{matrix} \geq \overset{m}{\sim\!\sim}$$

so we only need to show that we cannot increase the number of ebits by classical communication. This sounds easy, but in fact needs our monotone squashed entanglement. Since every possible extension ρ_{ABE} of a pure state ρ_{AB} (for instance the n ebits) is of the form $\rho_{ABE} = \rho_{AB} \otimes \rho_E$ we find

$$2E_{sq}(A : B)_{\overset{n}{\sim\!\sim}} = \inf_E I(A : B|E) = I(A : B) = 2n.$$

So we start the protocol with correlations of n bits measured in units of squashed entanglement. Then we perform a protocol which does not increase the squashed entanglement since it does only involve local operations and classical communication. The final state can therefore have at most n units of squashed entanglement. So, if the final state consists of m qubits as we require, then $m \leq n$, since we had otherwise increased the squashed entanglement. □

In fact, the statement also holds if one requires the transformation to only work approximately. The proof is then a little more technical and needs a result about the continuity of squashed entanglement.

One can also prove that one needs at least two bits of classical communication in order to teleport one qubit, regardless of how many ebits one has available. But we will leave this to the exercises.

7.4 Superdense Coding is Optimal

We want to prove that we need at least one qubit channel in order to send two classical bits, regardless of how many ebits we have available:

$$\overset{\overset{n}{\rightsquigarrow}}{\underset{\approx}{\infty}} \geq \overset{\overset{2m}{\rightarrow}}{\underset{\approx}{\infty}} \qquad \text{implies } m \leq n$$

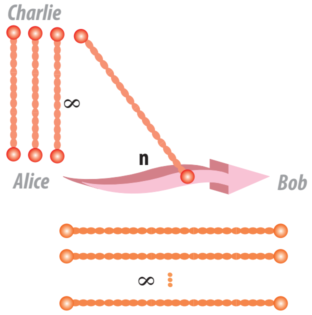
Note that concatenation of

$$\overset{\overset{n}{\rightsquigarrow}}{\underset{\approx}{\infty}} \geq \overset{\overset{2m}{\rightarrow}}{\underset{\approx}{\infty}}$$

with teleportation yields

$$\overset{\overset{n}{\rightsquigarrow}}{\underset{\approx}{\infty}} \geq \overset{\overset{m}{\rightsquigarrow}}{\underset{\approx}{\infty}} .$$

Now we have to prove that this implies $n \geq m$, i.e. entanglement does not help us to send more qubits. For this, we consider an additional player Charlie who holds system C and shares ebits with Alice. Let B_i be Bob's initial system, Q an n qubit system that Alice sends to Bob, Λ Bob's local operation and B_f Bob's final system. Clearly, if an n qubit channel could simulate an m qubit channel for $m > n$, then Alice could send m fresh halves of ebits that she shares with Charlie to Bob, thereby increasing the quantum mutual information between Charlie and Bob by $2m$.



We are now going to show that the amount of quantum mutual information that Bob and Charlie share cannot increase by more two times the number of qubits that he receives from Alice, i.e. by $2n$. For this we estimate Bob's final quantum mutual information with Charlie as

$$\begin{aligned} I(C:B_f) &\leq I(C:B_iQ) \\ &= I(C:B_i) + I(C:Q|B_i) \\ &\leq I(C:B_i) + 2n \end{aligned}$$

Therefore $m \leq n$. This concludes our proof that the superdense coding protocol is optimal.

Interestingly, for this argument, we did not use a monotone such as squashed entanglement from above. We merely used the property that the quantum mutual information cannot increase by too much under communication. Quantities that have the opposite behaviour (i.e. can increase sharply when only few qubits are communicated) are known as *lockable quantities* and have been in the focus of the attention in quantum information theory in recent years. So, we might also say that the quantum mutual information is *nonlockable.*

7.5 Entanglement

We have already encountered the word entanglement many times. Formally, we say that a quantum state ρ_{AB} is *separable* if it can be written as a convex combination of product states, i.e.

$$\rho_{AB} = \sum_k p_k \tau_k \otimes \sigma_k$$

where the p_k form a probability distribution and the ρ_k are states on A and the σ_k are states on B. A state is then called *entangled* if it is not *separable.*

Characteristic examples of separable states are

- $\rho_{AB} = |\phi\rangle\langle\phi|_A \otimes |\psi\rangle\langle\psi|_B$

- $\rho_{AB} = \mathrm{id}_{AB} = \frac{1}{4}\mathrm{id}_A \otimes \mathrm{id}_B$

- $\rho_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$

Characteristic examples of entangled states are

- In most situations (e.g. teleportation), ebits are the most useful entangled states. They are therefore also known as maximally entangled states (as well as all pure states of the form $U \otimes V \frac{1}{\sqrt{d}} \sum_i |ii\rangle_{AB}$ $|A| = |B| = d$.)

- Non-maximally entangled pure states of the form $\sum_i \alpha_i |ii\rangle$, where the α_i are not all of equal magnitude. In certain cases they can be converted (distilled) into maximally entangled states (of lower dimension) using Nielsen's majorisation criterion.

- The totally antisymmetric state $\rho_{AB} = \frac{1}{d(d-1)} \sum_{i<j} |ij - ji\rangle\langle ij - ji|_{AB}$ can be seen to be entangled, since every pure state supported on the antisymmetric subspace is entangled.

Since ebits are so useful, we can ask ourselves how many ebits we can extract per given copy of ρ_{AB}, as the number of copies approaches infinity. Formally, this number is known as the *distillable entanglement* of ρ_{AB}:

$$E_D(\rho_{AB}) = \lim_{\epsilon \mapsto 0} \lim_{n \mapsto \infty} \sup_{\Lambda \text{ LOCC}} \{\frac{m}{n} : \langle ebit|^{\otimes m} \Lambda(\rho_{AB}^{\otimes n})|ebit\rangle^{\otimes m} \geq 1 - \epsilon\}$$

This number is obviously very difficult to compute, but there is a whole theory of entanglement measures out there with the aim to provide upper bounds on distillable entanglement. A particularly easy upper bound is given by the squashed entanglement.

$$E_{sq}(\rho_{AB}) \geq E_D(\rho_{AB}).$$

The proof uses only the monotonicity of squashed entanglement under LOCC operations and the fact that the squashed entanglement of a state that is close to n ebits (in the purified distance) is close to n. In the exercise you will show that squashed entanglement of separable state is zero. This then immediately implies that one cannot extract any ebits from separable states!

8 Bell inequalities and non-locality

Many physicists have felt rather uncomfortable with the fact that quantum mechanics only makes statistical predictions. A famous example is Einstein, Podolsky, and Rosen, who argued that quantum mechanics is incomplete [5].

In an attempt to remedy this "problem", one could imagine that quantum mechanics is merely an effective statistical theory of a more fundamental deterministic theory. (Something in the spirit of statistical mechanics based on classical mechanics.) In such a theory, the randomness in the quantum measurements is merely a result of our ignorance about some hidden degrees of freedoms or parameters. Keeping Einstein in mind, a condition we would like to impose is that there should be no action at a distance. In other words, this hypothetical theory should respect special relativity and not allow superluminal communication. This class of theories are often referred to as local hidden variable theories [2] (where "local" stands for the absence of action at distance).

As we will see, the measurement statistics resulting from any local hidden variable theory has to satisfy certain types of inequalities, often referred to as Bell inequalities [2, 3] (after John Stewart Bell, who first came up with the idea). A consequence of these inequalities is that if one finds a setting (within a theory, or in a real experiment) where these inequalities are violated, then one can conclude that a hidden variable theory is not a good model in this case. For example, we will later construct measurements within quantum mechanics that violate these inequalities, which shows that we cannot replace quantum mechanics by a local hidden variable theory. Furthermore, many experiments have shown violations of Bell inequalities (up to some loopholes).

However, before we turn to the actual Bell inequalities, we will first approach the whole idea from a rather different angle, in the form of a game. The reason is to show that even in a very classical setting, in a game that we all easily could play, the access to a relatively simple quantum device can improve our chances to win this game beyond what we can do if we are restricted to classical means. This highlights that quantum correlations in some sense go beyond classical correlations.

8.1 A game

We here consider a very simple (but maybe slightly odd-looking) game, where we have a referee and two players, Alice and Bob. The game is cooperative, so Alice and Bob either both win or both loose.

The rules are as follows (see figure 8.1):

- The referee asks Alice a question $a \in \{0, 1\}$, and Alice gives an answer $x \in \{1, -1\}$.

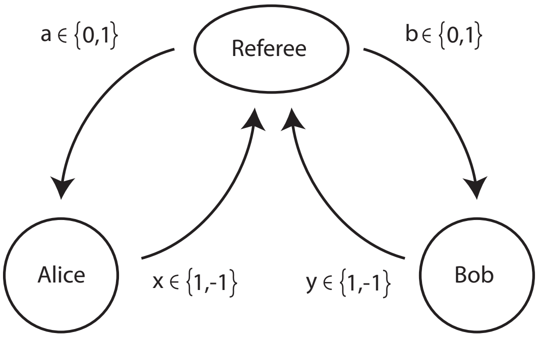


Figure 8.1: The referee asks asks Alice question "a", and Bob question "b", which each can be one of two questions (0 or 1). Their answers we denote by x and y, which each can take one of two values 1 and -1. Alice and Bob cannot communicate with each other during the game, and depending of the combination of questions and answers, they either both win or both lose.

- The referee asks Bob a question $b \in \{0, 1\}$, and Bob gives an answer $y \in \{1, -1\}$.
- Before the game starts, Alice and Bob can meet to decide on some strategy. However, during the game they cannot communicate with each other. Moreover, Alice does not know which question Bob receives, nor his answer, and vice versa. (Think of Alice and Bob having to go into separate rooms when the game starts, without their mobile phones.)
- The referee randomly asks the four possible combinations of questions (a, b) with equal probability 1/4.
- In the cases $(a, b) \in \{(0, 0), (0, 1), (1, 0)\}$ Alice and Bob win if the product of their answers is $x \cdot y = 1$, while in the case $(a, b) = (1, 1)$ they win if their answers are such that $x \cdot y = -1$.

The game can be summarized by the following table

$P_{A,B}(a, b)$	a	b	$x \cdot y$
1/4	0	0	1
1/4	0	1	1
1/4	1	0	1
1/4	1	1	-1

We let $\mathcal{W}$ denote the set of quadruples of questions and answers (a, b, x, y) for which Alice and Bob win, i.e.,

$$\begin{aligned}\mathcal{W} = \Big\{&(0,0,1,1),(0,0,-1,-1),\\ &(0,1,1,1),(0,1,-1,-1),\\ &(1,0,1,1),(1,0,-1,-1),\\ &(1,1,1,-1),(1,1,-1,1)\Big\}.\end{aligned} \tag{8.1}$$

8.1.1 Optimal deterministic strategies

The question is, what is the best strategy that Alice and Bob can use in order to maximize their chance of winning?

We shall first consider deterministic strategies. In a deterministic strategy, Alice's answer is a function of the question she receives, i.e., $x = f(a)$. Similarly for Bob, $y = g(b)$. (Since Alice and Bob cannot communicate, it follows that x cannot depend on b, and y cannot depend on a.) Assume for the moment that we could find functions f and g such that Alice and Bob always win the game.

According to the rules this means that

$$\left.\begin{array}{r} f(0)g(0) = 1 \\ f(0)g(1) = 1 \\ f(1)g(0) = 1 \\ f(1)g(1) = -1 \end{array}\right\} \Rightarrow \left\{\begin{array}{r} f(0) = g(0) \\ f(0) = g(1) \\ f(1) = g(0) \\ f(1) = -g(1) \end{array}\right. \tag{8.2}$$

A quick inspection yields $g(1) = -g(1)$, which is a contradiction since $g(1) \in \{1, -1\}$. We can thus conclude that there exists no deterministic strategy that lets Alice and Bob win in all the four cases.

Could we find a deterministic strategy that wins in three cases? Indeed we can, namely if Alice and Bob both answer "1" irrespective of the question they receive. This choice makes them win in three out of four cases. We can thus conclude that the optimal winning probability for a deterministic strategy is $P_{\text{win}} = 3/4$. (Since all four combinations of questions occur with equal probability we cannot do better than this for any strategy that wins in three out of four cases.)

8.1.2 Optimal (classically local) probabilistic strategies

One could imagine that Alice and Bob could do better with some probabilistic strategy. This means that their answers are not functions of the questions, but rather determined randomly according to some probability distribution.

Since Alice and Bob are allowed to meet before the game starts, they can jointly create a random number z, with distribution P_Z. Later they can use this when they pick which functions to use when they determine their outputs, i.e., $x = f_z(a)$ and $y = g_z(b)$. The statistics resulting from this strategy can be described by a conditional probability distribution. We let $P_{X,Y|a,b}(x, y)$ denote the probability that their answers are x, y conditioned on the questions being a, b. In the present case, the conditional probability distribution reads

$$P_{X,Y|a,b}(x,y) = \sum_z P_Z(z)\delta_{x,f_z(a)}\delta_{y,g_z(b)}, \tag{8.3}$$

where the delta function $\delta_{x,f(a)}$ means that the outcome x is equal to $f_z(a)$ with probability 1.

Nothing says that Alice and Bob have to determine their outputs deterministically, given z and their questions. Their answers x and y could be determined randomly according to some conditional distributions $P_{X|a,z}(x)$ and $P_{Y|b,z}(y)$, respectively. In total, the statistics of their answers would thus be characterized by the conditional distribution

$$P_{X,Y|a,b}(x,y) = \sum_z P_Z(z)P_{X|a,z}(x)P_{Y|b,z}(y). \tag{8.4}$$

Note that the random variable Z does not depend on a or b, since Z is determined before Alice and Bob receive the questions a and b. We refer to conditional distributions like in (8.4) as *classically local.*

At first sight it might look like if the set of classically local distributions, as defined by equation (8.4), is larger than the set spanned by equation (8.3). However, this is not the case. The reason is that if a and x only can take a finite number of values, then every conditional distribution can be written $P_{X|a}(x) = \sum_f P(f)\delta_{x,f(x)}$, where the sum spans over the set of all possible mappings f. From this we can conclude that the extreme points of the set of classically local distributions are of the form $\delta_{x,f(a)}\delta_{y,g(b)}$, for different choices of f and g. Hence, both equation (8.3) and (8.4) span the set of classically local distributions (sometimes also referred to as the "Bell polytope"). In other words, all probabilistic strategies can be regarded as convex combinations of local deterministic strategies.

Returning to our game, we see that given a conditional probability distribution $P_{X,Y|a,b}$ (not necessarily classically local) the probability that Alice and Bob win is given by

$$\begin{aligned} P_{\text{win}} &= \sum_{(a,b,x,y)\in\mathcal{W}} P_{X,Y,A,B}(x,y,a,b) \\ &= \sum_{(a,b,x,y)\in\mathcal{W}} P_{X,Y|a,b}(x,y)P_{A,B}(a,b) \\ &= \frac{1}{4} \sum_{(a,b,x,y)\in\mathcal{W}} P_{X,Y|a,b}(x,y). \end{aligned} \tag{8.5}$$

The winning probability, P_{win}, is clearly linear in the conditional probability distribution $P_{X,Y|a,b}$. We can thus conclude that the maximum of P_{win} over the convex set of classically local distributions is attained at one of the extreme points. Since the extreme points are the local deterministic strategies, it follows that $3/4$ is the maximum winning probability among all strategies leading to classically local distributions.

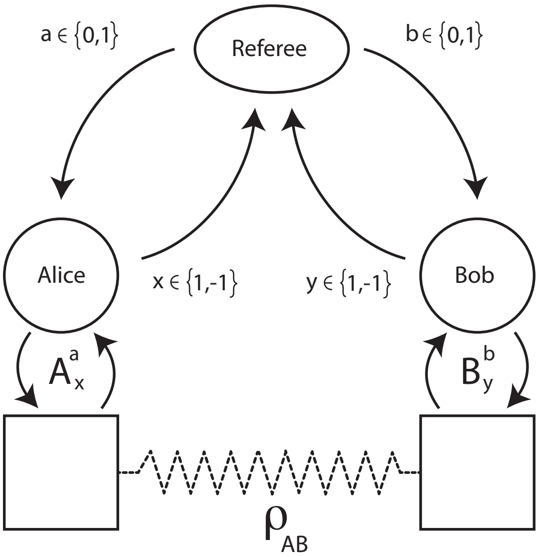


Figure 8.2: Alice and Bob can improve their chances of winning the game by using a quantum device. Before the game starts they create a pair of particles in the maximally entangled state $\rho_{AB} = |\psi\rangle\langle\psi|$, where $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$. When the game starts, Alice and Bob each bring one of these particles with them. Given the question a, Alice measures the POVM $\{A^a\}_x$ on her particle, and returns the measurement outcome x as her answer. Similarly, Bob measures the POVM $\{B^b_y\}_y$ and returns the answer y.

8.1.3 A strategy based on a quantum device

In the previous section we tacitly assumed that Alice and Bob are limited to classical operations, in the sense that they can only generate and share classical randomness and correlations. However, one could imagine Alice and Bob to establish a pair of entangled particles before the game starts, and each bring with them one particle in this pair (see figure 8.2). We will see that Alice and Bob can increase their chance of winning beyond the value 3/4 with a clever choice of measurements.

Suppose that Alice and Bob share a quantum state ρ_{AB}. For each question a, Alice measures a POVM $\{A^a_x\}_x$, and outputs the measurement outcome x as the answer. Analogously, Bob measures the POVM $\{B^b_y\}_y$ if he gets question b, and lets the measurement outcome be his answer. The resulting conditional probability distribution of this procedure is

$$P_{X,Y|a,b}(x,y) = \mathrm{tr}(A^a_x \otimes B^b_y \rho_{AB}), \tag{8.6}$$

where the tensor product $A^a_x \otimes B^b_y$ corresponds to the fact that Alice and Bob measure their POVMs on separate locations and systems.

We shall now consider a specific choice of state ρ_{AB} and families of POVMs that makes Alice and Bob win our game with a higher probability than in the classical case. We let ρ_{AB} be the maximally entangled state

$$\rho_{AB} = |\psi\rangle\langle\psi|, \quad |\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle).$$

Define the following family of normalized states

$$\begin{aligned} |\phi_1(\theta)\rangle =& \cos(\theta)|0\rangle + \sin(\theta)|1\rangle, \\ |\phi_{-1}(\theta)\rangle =& -\sin(\theta)|0\rangle + \cos(\theta)|1\rangle. \end{aligned} \tag{8.7}$$

Via these states we define Alice and Bob's POVMs.

Alice have the POVM $\{A_1^0, A_{-1}^0\}$ that corresponds to the question $a = 0$, and the POVM $\{A_1^1, A_{-1}^1\}$ corresponding to question $a = 1$. We let

$$\begin{aligned} A_1^0 =& |\phi_1(0)\rangle\langle\phi_1(0)|, & A_{-1}^0 =& |\phi_{-1}(0)\rangle\langle\phi_{-1}(0)|, \\ A_1^1 =& |\phi_1(\pi/4)\rangle\langle\phi_1(\pi/4)|, & A_{-1}^1 =& |\phi_{-1}(\pi/4)\rangle\langle\phi_{-1}(\pi/4)|. \end{aligned}$$

Analogously, Bob has the two POVMs $\{B_1^0, B_{-1}^0\}$, $\{B_1^1, B_{-1}^1\}$ defined as

$$\begin{aligned} B_1^0 =& |\phi_1(\pi/8)\rangle\langle\phi_1(\pi/8)|, & B_{-1}^0 =& |\phi_{-1}(\pi/8)\rangle\langle\phi_{-1}(\pi/8)|, \\ B_1^1 =& |\phi_1(-\pi/8)\rangle\langle\phi_1(-\pi/8)|, & B_{-1}^1 =& |\phi_{-1}(-\pi/8)\rangle\langle\phi_{-1}(-\pi/8)|. \end{aligned}$$

By combining Eqs. (8.5) with (8.6) and inserting the above POVM elements, we obtain

$$\begin{aligned} P_{\text{win}} =& \frac{1}{4} \sum_{(a,b,x,y)\in\mathcal{W}} \text{tr}(A_x^a \otimes B_y^b \rho_{AB}) \\ =& \frac{1}{8} \sum_{(0,0,x,y)\in\mathcal{W}} \Big| \langle\phi_x(0)|0\rangle\langle\phi_y(\pi/8)|0\rangle + \langle\phi_x(0)|1\rangle\langle\phi_y(\pi/8)|1\rangle \Big|^2 \\ &+ \frac{1}{8} \sum_{(0,1,x,y)\in\mathcal{W}} \Big| \langle\phi_x(0)|0\rangle\langle\phi_y(-\pi/8)|0\rangle + \langle\phi_x(0)|1\rangle\langle\phi_y(-\pi/8)|1\rangle \Big|^2 \\ &+ \frac{1}{8} \sum_{(1,0,x,y)\in\mathcal{W}} \Big| \langle\phi_x(\phi/4)|0\rangle\langle\phi_y(\pi/8)|0\rangle + \langle\phi_x(\pi/4)|1\rangle\langle\phi_y(\phi/8)|1\rangle \Big|^2 \\ &+ \frac{1}{8} \sum_{(1,1,x,y)\in\mathcal{W}} \Big| \langle\phi_x(\pi/4)|0\rangle\langle\phi_y(-\pi/8)|0\rangle + \langle\phi_x(\pi/4)|1\rangle\langle\phi_y(-\pi/8)|1\rangle \Big|^2 \\ =& \frac{1}{2}(1 + \frac{1}{\sqrt{2}}) \\ \approx& 0.85 \end{aligned} \tag{8.8}$$

Hence, the winning probability is larger than 3/4.

A consequence of this observation is that the set of conditional probability distributions that we can reach via quantum states and local measurements, as in equation (8.6), is strictly larger than the set of conditional probability distributions as in equation (8.4). In the following section we shall reformulate this observation in more "traditional" terms.

8.2 The CHSH inequality

After this prelude, we now return to our main purpose, namely to introduce Bell inequalities. As mentioned in the beginning of this chapter, the idea is to find an inequality that is satisfied by all local hidden variable theories. The setting is more or less identical to our game in the previous section, but the roles and meanings of the various objects have changed.

In a deterministic theory, e.g., classical mechanics, the outcomes of the measurements are uniquely determined by the choice of measurement and the actual state of the system. In a hidden variable theory this "actual state" corresponds to a parameter z. We imagine that Alice and Bob are space-like separated, such that no signal have time to propagate between Alice and Bob during the time it takes them to perform their measurements, and thus we do not allow Alice's outcome x to depend on Bob's choice of measurement b, or Bob's outcome y to depend on Alice's choice of measurement a. (This is the assumption of locality.) Hence, in a deterministic local hidden variable theory, Alice's measurement outcome is uniquely determined by her choice of measurement a and the hidden variable z, i.e., $x = f_z(a)$. Similarly, Bob's measurement outcome is uniquely determined by his choice of measurement and the hidden variable, $y = g_z(b)$.

The randomness in the measurement outcomes comes from the fact that we do not know what value the hidden variable z has. The distribution of z is described by P_Z. The statistics of the whole experiment is hence captured by a conditional distribution as in equation (8.3), and is thus classically local.

A Bell inequality is an inequality that is satisfied by any measurement statistics that can be described by a classically local distribution, and thus has to be satisfied by any local hidden variable theory. There exist several different types of Bell inequalities, but we shall here consider a specific version called the CHSH inequality [8] (after Clauser, Horne, Shimony, and Holt).

Lemma 8.2.1. *The function*

$$F(x,y) := |x+y| + |x-y| \tag{8.9}$$

is such that

$$\max_{(x,y)\in[-1,1]\times[-1,1]} F(x,y) = 2. \tag{8.10}$$

This follows from the fact that F is convex. Furthermore, on the set $|x| = 1, |y| = 1$ it takes the value $F(x,y) = 2$. Hence, $F(x,y) \leq 2$ for $|x| \leq 1$, $|y| \leq 1$.

Theorem 8.2.2 (CHSH inequality)**.** *Let X and Y be random variables that take values in the interval $[-1,1]$, with a classically local conditional probability distribution $P_{X,Y|a,b}$,* *where* $a, b \in \{0,1\}$. *Then*

$$\begin{aligned}&\Big|\big\langle XY|(a,b)=(0,0)\big\rangle + \big\langle XY|(a,b)=(0,1)\big\rangle\\ &+\big\langle XY|(a,b)=(1,0)\big\rangle - \big\langle XY|(a,b)=(1,1)\big\rangle\Big| \leq 2.\end{aligned} \tag{8.11}$$

The quantity $\big\langle XY|(a,b)=(0,0)\big\rangle$ is the expectation value of the product of the two measurement outcomes X and Y, given that we know that we have made measurements $a=0$ and $b=0$. The total expression in equation (8.11) is thus a linear combination of such expectation values, for the four possible combinations of experiments (a,b).

Proof. We first define

$$\langle X|a,z\rangle := \sum_x x P_{X|a,z}(x), \quad \langle Y|b,z\rangle := \sum_y y P_{Y|b,z}(y). \tag{8.12}$$

Note that since X and Y take values in the interval $[-1,1]$ it follows that $|\langle X|a,z\rangle| \leq 1$ and $|\langle Y|b,z\rangle| \leq 1$. We combine this observation with the triangle inequality, and Lemma 8.2.1, to obtain

$$\begin{aligned}&\Big|\big\langle XY|(a,b)=(0,0)\big\rangle + \big\langle XY|(a,b)=(0,1)\big\rangle\\ &+\big\langle XY|(a,b)=(1,0)\big\rangle - \big\langle XY|(a,b)=(1,1)\big\rangle\Big|\\ \leq&\sum_z P_Z(z)\Big|\big\langle X|a=0,z\big\rangle\big\langle Y|b=0,z\big\rangle\\ &+\big\langle X|a=0,z\big\rangle\big\langle Y|b=1,z\big\rangle\\ &+\big\langle X|a=1,z\big\rangle\big\langle Y|b=0,z\big\rangle\\ &-\big\langle X|a=1,z\big\rangle\big\langle Y|b=1,z\big\rangle\Big|\\ \leq&\sum_z P_Z(z)\Big|\Big[\big\langle X|a=0,z\big\rangle+\big\langle X|a=1,z\big\rangle\Big]\big\langle Y|b=0,z\big\rangle\\ &+\Big[\big\langle X|a=0,z\big\rangle-\big\langle X|a=1,z\big\rangle\Big]\big\langle Y|b=1,z\big\rangle\Big|\\ \leq&\sum_z P_Z(z) F\Big(\big\langle X|a=0,z\big\rangle, \big\langle X|a=1,z\big\rangle\Big)\\ \leq&2.\end{aligned} \tag{8.13}$$

□

8.3 Violation of the CHSH inequality

Here we demonstrate by an explicit example (this is the example of section 8.1.3 in disguise) that the CHSH inequality can be violated within quantum mechanics.

Consider the four observables

$$\mathcal{A}^0 := \sigma_z, \quad \mathcal{A}^1 := \sigma_x, \tag{8.14}$$

$$\mathcal{B}^0 := \frac{1}{\sqrt{2}}(\sigma_z + \sigma_x), \quad \mathcal{B}^1 := \frac{1}{\sqrt{2}}(\sigma_z - \sigma_x). \tag{8.15}$$

The observables $\mathcal{A}^0$ and $\mathcal{A}^1$ correspond to Alice's two choices of measurements, and similarly $\mathcal{B}^0$ and $\mathcal{B}^1$ are the observables of Bob's two measurements.

If we let X denote the measurement outcome of Alice's measurement, and Y is Bob's measurement outcome, then $\langle XY|a,b\rangle = \mathrm{tr}(A^a \otimes B^b \rho_{AB})$. Note that the eigenvalues of $\mathcal{A}^0$, $\mathcal{A}^1$, $\mathcal{B}^0$, and $\mathcal{B}^1$, are all ± 1. Hence, the measurement outcomes X, Y take values in the interval $[-1,1]$ for all values of a and b. For this setup we find

$$\begin{aligned}
&\big\langle XY|(a,b)=(0,0)\big\rangle + \big\langle XY|(a,b)=(0,1)\big\rangle \\
&+ \big\langle XY|(a,b)=(1,0)\big\rangle - \big\langle XY|(a,b)=(1,1)\big\rangle \\
=&\langle\psi|\mathcal{A}^0\otimes\mathcal{B}^0|\psi\rangle + \langle\psi|\mathcal{A}^0\otimes\mathcal{B}^1|\psi\rangle + \langle\psi|\mathcal{A}^1\otimes\mathcal{B}^0|\psi\rangle - \langle\psi|\mathcal{A}^1\otimes\mathcal{B}^1|\psi\rangle \\
=&\sqrt{2}\langle\psi|\sigma_z\otimes\sigma_z|\psi\rangle + \sqrt{2}\langle\psi|\sigma_x\otimes\sigma_x|\psi\rangle \\
=&2\sqrt{2}.
\end{aligned} \tag{8.16}$$

Hence, the CHSH inequality can be violated, which means that we can never find a local hidden variable theory that "simulates" quantum mechanics.

As a side remark we note that the measurements of the observables $\mathcal{A}^0, \mathcal{A}^1, \mathcal{B}^0, \mathcal{B}^1$ gives precisely the POVMs in section 8.1.3 (as the pairs of eigenvectors of respective observable)

$$\begin{array}{lll}
\mathcal{A}^0: & |\phi_1(0)\rangle, & |\phi_{-1}(0)\rangle, \\
\mathcal{A}^1: & |\phi_1(\pi/4)\rangle, & |\phi_{-1}(\pi/4)\rangle, \\
\mathcal{B}^0: & |\phi_1(\pi/8)\rangle, & |\phi_{-1}(\pi/8)\rangle, \\
\mathcal{B}^1: & |\phi_1(-\pi/8)\rangle, & |\phi_{-1}(-\pi/8)\rangle.
\end{array}$$

Assuming quantum mechanics to be the correct description of the world, the above result excludes local hidden variable theories. But what if one does not trust quantum mechanics? The Bell inequalities provide means to exclude local hidden variable theories even if you do not want to subscribe to quantum mechanics. It is essentially enough to gather measurement statistics (together with some weak theoretical assumptions) that yields a violation of the Bell inequalities. A number of experiments have been performed that rather clearly suggest violations of Bell inequalities. However, due to various limitations in the implementations (often referred to as loopholes) we can strictly speaking not be entirely sure that Mother Nature is not conspiring to fool us into believing that we see Bell violations in our experiments, while we actually do not [6, 7]. (However, most physicists find the results convincing enough.)

Here we should also take the opportunity to point out that one can construct hidden variable theories that do give the same predictions as quantum mechanics ("pilot-wave theory" or "Bohmian mechanics", see [1]) but the price is that we have to give up locality, i.e., these theories contain superluminal communication in some sense.

8.4 Cirel'son's inequality

So far we have seen that every local hidden variable theory (or more precisely every classically local distribution) gives rise to the bound 2 in the CHSH inequality. However, the previous example shows that we can reach up to $2\sqrt{2}$ within quantum mechanics. The question is if we can reach even higher values. By a quantum analogue of the CHSH inequality we can show that local quantum measurements cannot yield anything beyond the value $2\sqrt{2}$.

Lemma 8.4.1. *The function* $H(x) := \sqrt{2+2x} + \sqrt{2-2x}$ *is such that* $\max_{x\in[-1,1]} H(x) = 2\sqrt{2}$.

Theorem 8.4.2. *(Cirel'son's inequality [4]) Let* $\mathcal{A}^0$, $\mathcal{A}^1$, $\mathcal{B}^0$, $\mathcal{B}^1$ *be observables with spectrum in* $[-1,1]$. *Let* $\rho_{AB} \in \mathcal{S}_=(\mathcal{H}_A \otimes \mathcal{H}_B)$. *Then*

$$\begin{aligned}\Big|&\operatorname{tr}(\mathcal{A}^0 \otimes \mathcal{B}^0 \rho_{AB}) + \operatorname{tr}(\mathcal{A}^0 \otimes \mathcal{B}^1 \rho_{AB}) \\ &+ \operatorname{tr}(\mathcal{A}^1 \otimes \mathcal{B}^0 \rho_{AB}) - \operatorname{tr}(\mathcal{A}^1 \otimes \mathcal{B}^1 \rho_{AB})\Big| \leq 2\sqrt{2}.\end{aligned} \tag{8.17}$$

Proof. Let $\rho_{AB} = \sum_k \lambda_k |\psi_k\rangle\langle\psi_k|$ be an eigenvalue decomposition (or any other convex decomposition into pure states). The following is essentially just an application of the triangle inequality and the Cauchy-Schwarz inequality.

$$\begin{aligned}
&\Big|\operatorname{tr}(\mathcal{A}^0 \otimes \mathcal{B}^0 \rho_{AB}) + \operatorname{tr}(\mathcal{A}^0 \otimes \mathcal{B}^1 \rho_{AB}) \\
&\quad + \operatorname{tr}(\mathcal{A}^1 \otimes \mathcal{B}^0 \rho_{AB}) - \operatorname{tr}(\mathcal{A}^1 \otimes \mathcal{B}^1 \rho_{AB})\Big| \\
&\leq \sum_k \lambda_k \Big|\langle\psi_k|\mathcal{A}^0 \otimes \mathcal{B}^0|\psi_k\rangle + \langle\psi_k|\mathcal{A}^0 \otimes \mathcal{B}^1|\psi_k\rangle \\
&\quad + \langle\psi_k|\mathcal{A}^1 \otimes \mathcal{B}^0|\psi_k\rangle - \langle\psi_k|\mathcal{A}^1 \otimes \mathcal{B}^1|\psi_k\rangle\Big| \\
&\leq \sum_k \lambda_k \Big|\Big(\langle\psi_k|\mathcal{A}^0 \otimes \hat{1} + \langle\psi_k|\mathcal{A}^1 \otimes \hat{1}\Big)\hat{1} \otimes \mathcal{B}^0|\psi_k\rangle\Big| \\
&\quad + \sum_k \lambda_k \Big|\Big(\langle\psi_k|\mathcal{A}^0 \otimes \hat{1} - \langle\psi_k|\mathcal{A}^1 \otimes \hat{1}\Big)\hat{1} \otimes \mathcal{B}^1|\psi_k\rangle\Big| \\
&\leq \sum_k \lambda_k \Big\|\mathcal{A}^0 \otimes \hat{1}|\psi_k\rangle + \mathcal{A}^1 \otimes \hat{1}|\psi_k\rangle\Big\| \, \Big\|\hat{1} \otimes \mathcal{B}^0|\psi_k\rangle\Big\| \\
&\quad + \sum_k \lambda_k \Big\|\mathcal{A}^0 \otimes \hat{1}|\psi_k\rangle - \mathcal{A}^1 \otimes \hat{1}|\psi_k\rangle\Big\| \, \Big\|\hat{1} \otimes \mathcal{B}^1|\psi_k\rangle\Big\| \\
&\leq \sum_k \lambda_k \Big[\big\||\alpha_k\rangle + |\gamma_k\rangle\big\| + \big\||\alpha_k\rangle - |\gamma_k\rangle\big\|\Big],
\end{aligned} \tag{8.18}$$

where

$$|\alpha_k\rangle := \mathcal{A}^0 \otimes \hat{1}|\psi_k\rangle, \quad |\gamma_k\rangle := \mathcal{A}^1 \otimes \hat{1}|\psi_k\rangle. \tag{8.19}$$

In the last inequality above, we made use of the assumption that $\mathcal{B}^b$ has its spectrum in $[-1,1]$. Due to the analogous assumption on $\mathcal{A}^a$, it follows that $\|\alpha_k\| \leq 1$ and $\|\gamma_k\| \leq 1$, and $x_k := \mathrm{Re}\langle\alpha_k|\gamma_k\rangle \in [-1,1]$. If we combine these observations with Lemma 8.4.1 we find

$$\begin{aligned}
&\sum_k \lambda_k \Big[\big\| |\alpha_k\rangle + |\gamma_k\rangle \big\| + \big\| |\alpha_k\rangle - |\gamma_k\rangle \big\| \Big] \\
\leq &\sum_k \lambda_k \Big[\sqrt{2 + 2\mathrm{Re}\langle\alpha_k|\gamma_k\rangle} + \sqrt{2 - 2\mathrm{Re}\langle\alpha_k|\gamma_k\rangle} \Big] \\
= &\sum_k \lambda_k \big[\sqrt{2 + 2x_k} + \sqrt{2 - 2x_k} \big] \\
= &\sum_k \lambda_k H(x_k) \\
\leq &2\sqrt{2}.
\end{aligned} \tag{8.20}$$

□

8.5 Beyond quantum

A recurrent theme throughout this chapter has been that of two subsystems that are prevented from communicating while we perform experiments on them. We have furthermore seen that all our considerations can be phrased in terms of different classes of conditional probability distributions. Here we take this line of thoughts to its ultimate conclusion, and ask what kind of conditional probability distributions we obtain if the only restriction we impose is that they should not allow Alice and Bob to communicate. First of all, what do we mean by this? As an example, suppose that we would have the conditional probability distribution $P_{X,Y|a,b}(x,y) = \delta_{x,b}\delta_{y,a}$. This is a deterministic distribution where Alice's output x is always equal to Bob's input b, and where Bob's output y is always equal to Alice's input a. In other words, if we had a device (or a pair of devices) that would behave according to this conditional distribution, it would allow Alice and Bob to communicate perfectly.

To exclude all communication, Alice should not be able to detect Bob's inputs, and vice versa. Moreover, they should not even have an increased chance at *guessing* correctly what inputs the other is choosing. To impose this, we must make sure that if Bob does not know Alice's measurement outcomes x, then his conditional probability distribution $P_{Y|a,b}(y) := \sum_x P_{X,Y|a,b}(x,y)$ does not depend on Alice's choices a, and similarly that Alice's marginal distribution does not depend on Bob's choices. We formalize this with the following definition.

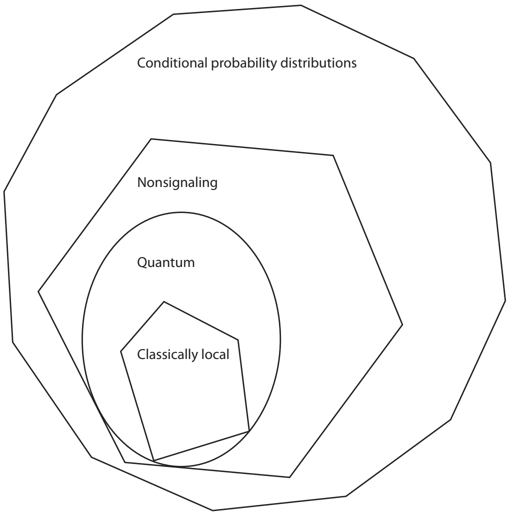


Figure 8.3: The bipartite conditional probability distributions with a fixed and finite number of possible inputs and outputs x, y, a, b form a convex polytope. In this polytope, the non-signaling, quantum, and classically local distributions form a sequence of nested subsets. The classically local and the non-signaling sets are both convex polytopes, while the quantum set is convex but not a polytope.

Definition 8.5.1. A conditional probability distribution is called *non-signaling* if

$$\begin{aligned}\sum_y P_{X,Y|a,b}(x,y) &= \sum_y P_{X,Y|a,b'}(x,y), \quad \forall a, b, b', \quad \forall x,\\ \sum_x P_{X,Y|a,b}(x,y) &= \sum_x P_{X,Y|a',b}(x,y), \quad \forall a, a', b, \quad \forall y.\end{aligned} \tag{8.21}$$

We can now check that all conditional distributions arising from bipartite quantum measurements, as in equation (8.6), are non-signaling. By letting $\rho_A = \mathrm{tr}_B \rho_{AB}$, and use the fact that $\{B_y^b\}_y$ is a POVM, we find

$$\begin{aligned}\sum_y P_{X,Y|a,b}(x,y) =& \sum_y \mathrm{tr}(A_x^a \otimes B_y^b \rho_{AB})\\ =& \mathrm{tr}(A_x^a \otimes \hat{1}_B \rho_{AB})\\ =& \mathrm{tr}(A_x^a \rho_A).\end{aligned} \tag{8.22}$$

Hence, Alice cannot notice any of Bob's choices. By instead summing over x, we find the analogous statement for Bob.

We can conclude that bipartite quantum measurements always give rise to non-signaling distributions. However, there do exist non-signaling distributions that cannot be reached by local quantum measurements. One way to prove this is to show that there exists a non-signaling distribution that violates Cirel'son's inequality in Theorem (8.4.2). The following non-signaling distribution gives such an example.

$$\begin{aligned}
&P_{X,Y|(a,b)=(0,0)}(1,1) = \frac{1}{2}, \quad P_{X,Y|(a,b)=(0,0)}(-1,-1) = \frac{1}{2},\\
&P_{X,Y|(a,b)=(0,1)}(1,1) = \frac{1}{2}, \quad P_{X,Y|(a,b)=(0,1)}(-1,-1) = \frac{1}{2},\\
&P_{X,Y|(a,b)=(1,0)}(1,1) = \frac{1}{2}, \quad P_{X,Y|(a,b)=(1,0)}(-1,-1) = \frac{1}{2},\\
&P_{X,Y|(a,b)=(1,1)}(1,-1) = \frac{1}{2}, \quad P_{X,Y|(a,b)=(1,1)}(-1,1) = \frac{1}{2}.
\end{aligned} \tag{8.23}$$

The first row of the above equations tells us that if the measurement settings are $a = 0$ and $b = 0$, then Alice and Bob always get the same output, where the two possibilities $x = y = 1$ and $x = y = -1$ occur with equal probability. The same is true for the settings $(0,1)$ and $(1,0)$, while if they happen to choose the settings $a = 1$ and $b = 1$, then Alice and Bob are guaranteed to get opposite outputs $x = -y$. Note that if Alice does not know Bob's outputs, then Alice only sees that she gets the outputs 1 or -1 with equal probability, irrespective of Bob's input b (as well as irrespective of her own input a). The analogous conclusion holds for Bob, and thus (8.23) describes a non-signaling distribution.

For this choice of conditional distribution we furthermore have

$$\begin{aligned}
&\Big\langle XY|(a,b) = (0,0)\Big\rangle + \Big\langle XY|(a,b) = (0,1)\Big\rangle\\
&\quad + \Big\langle XY|(a,b) = (1,0)\Big\rangle - \Big\langle XY|(a,b) = (1,1)\Big\rangle\\
&=4.
\end{aligned} \tag{8.24}$$

Hence, this goes beyond Cirel'son's bound of $2\sqrt{2}$, and we can conclude that this conditional distribution cannot be generated by local quantum measurements.

Note that if we could build a pair of devices that implement the conditional probability distribution in equation (8.23), then Alice and Bob would be able win the game that we introduced in Section 8.1 each time they play it. Unfortunately, it looks like our universe does not allow such machines.

If such correlations do not exist in our world, why should we bother about them? One reason is that these more general types of theories may help us to understand why quantum mechanics looks the way it does. For example, one could ask why quantum mechanics does not allow these more 'violent' types of correlations. After all, they are allowed by relativity theory, so why should nature not allow them? (Well, as far as we know they are not allowed in nature, but who knows.) As of today, no one really has a good answer, and questions like this are presently a rather active research topic. (For an introduction, see [11].)

Bibliography

[1] D. Z. Albert. Bohm's alternative to quantum mechanics. *Scientific American*, May:58, 1994.

[2] J. Bell. *Introduction to the Hidden-Variable Question.* Academic Press, 1971.

[3] J. S. Bell. On the Einstein Podolsky Rosen paradox. *Physics*, 1:195–200, 1964.

[4] B.S. Cirel'son. Quantum generalizations of bell's inequality. *Letters in Mathematical Physics*, 4:93–100, 1980.

[5] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47(10), 1935.

[6] N. Gisin. Bell inequalities: many questions, a few answers. `http://arxiv.org/abs/quant-ph/0702021`.

[7] A. Lamas-Linares J. Skaar V. Scarani V. Makarov I. Gerhardt, Q. Liu and C. Kurtsiefer. Experimentally faking the violation of bell's inequalities. *Physical Review Letters*, 107:170404, 2011.

[8] A. Shimony J. F. Clauser, M.A. Horne and R. A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23:880–884, 1969.

[9] D. H. Mellor. *Probability: A Philosophical Introduction.* Routledge, 2005.

[10] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[11] S. Popescu. Why isn't nature more non-local? *Nature Physics*, 2:507–508, 2006.

[12] Renato Renner. Security of quantum key distribution. *quant-ph/0512258*, December 2005.

[13] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[14] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM review*, 38:49, 1996.

[15] Monty Hall problem. `http://en.wikipedia.org/wiki/Monty_Hall_problem`.

[16] S. Wolf. Einführung in die Quanteninformatik. `http://qi.ethz.ch/edu/qiHS08/`. (Regular course in fall semester).